Information Science and Statistics

Ingo Steinwart • Andreas Christmann

Support Vector Machines

Springer

Information Science and Statistics

Series Editors:

M. Jordan
J. Kleinberg
B. Schölkopf

Information Science and Statistics

For other titles published in this series, go to
www.springer.com/series/3816

Ingo Steinwart • Andreas Christmann

Support Vector Machines

Springer

Ingo Steinwart
Information Sciences Group (CCS-3)
Los Alamos National Laboratory
Los Alamos, NM 87545
USA

Andreas Christmann
University of Bayreuth
Department of Mathematics
Chair of Stochastics
95440 Bayreuth
Germany

Series Editors

Michael Jordan
Division of Computer Science
and Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
USA

Jon Kleinberg
Department of Computer Science
Cornell University
Ithaca, NY 14853
USA

Bernhard Schölkopf
Max Planck Institute
for Biological Cybernetics
Spemmannstrasse 38
72076 Tübingen
Germany

Previously published material included in the book:

Portions of Section 4.4 are based on: I. Steinwart, D. Hush, and C. Scovel (2006), An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inf. Theory*, **52**, 4635–4643

Portions of Section 9.4 are based on:

A. Christmann and I. Steinwart (2008), Consistency of kernel based quantile regression. *Appl. Stoch. Models Bus. Ind.*, **24**, 171–183.

ISBN: 978-0-387-77241-7 e-ISBN: 978-0-387-77242-4
DOI: 10.1007/978-0-387-77242-4

Library of Congress Control Number: 2008932159

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Für Wiebke, Johanna, Wenke und Anke. (I.S.)
Für Anne, Hanna, Thomas und meine Eltern. (A.C.)

Preface

Every mathematical discipline goes through three periods of development: the naive, the formal, and the critical.
David Hilbert

The goal of this book is to explain the principles that made support vector machines (SVMs) a successful modeling and prediction tool for a variety of applications. We try to achieve this by presenting the basic ideas of SVMs together with the latest developments and current research questions in a unified style. In a nutshell, we identify at least three reasons for the success of SVMs: their ability to learn well with only a very small number of free parameters, their robustness against several types of model violations and outliers, and last but not least their computational efficiency compared with several other methods.

Although there are several roots and precursors of SVMs, these methods gained particular momentum during the last 15 years since Vapnik (1995, 1998) published his well-known textbooks on statistical learning theory with a special emphasis on support vector machines. Since then, the field of machine learning has witnessed intense activity in the study of SVMs, which has spread more and more to other disciplines such as statistics and mathematics. Thus it seems fair to say that several communities are currently working on support vector machines and on related kernel-based methods. Although there are many interactions between these communities, we think that there is still room for additional fruitful interaction and would be glad if this textbook were found helpful in stimulating further research. Many of the results presented in this book have previously been scattered in the journal literature or are still under review. As a consequence, these results have been accessible only to a relatively small number of specialists, sometimes probably only to people from one community but not the others. In view of the importance of SVMs for statistical machine learning, we hope that the unified presentation given here will make these results more accessible to researchers and users from different

communities (e.g.; from the fields of statistics, mathematics, computer science, bioinformatics, data and text mining, and engineering).

As in most monographs, the selection of topics treated in this textbook is biased for several reasons. We have of course focused on those that particularly interest us and those that we have been working on during the last decade. We also decided to concentrate on some important and selected topics, so for these topics we can offer not only the results but also the proofs. This is in contrast to some other textbooks on SVMs or statistical machine learning in general, but we try to follow the path described by Devroye *et al.* (1996) and Györfi *et al.* (2002). Moreover, some topics, such as the robustness properties of SVMs, a detailed treatment of loss functions and reproducing kernel Hilbert spaces, recent advances in the statistical analysis of SVMs, and the relationship between good learning properties and good robustness properties such as a bounded influence function and a bounded maxbias, are not covered by other currently available books on SVMs. On the other hand, the area of statistical machine learning is nowadays so rich and progressing so rapidly that covering all aspects in detail in a single book hardly seems possible. The consequence is of course that several important and interesting topics of SVMs and related methods are not covered in this monograph. This includes, for example, SVMs for anomaly detection, kernel principal component analysis, kernel-based independence measures, structured estimation, recent progress in computational algorithms, boosting, Bayesian approaches, and the analysis of time series or text data. A reader interested in these topics will get useful information in the books by Vapnik (1995, 1998), Cristianini and Shawe-Taylor (2000), Hastie *et al.* (2001), Schölkopf and Smola (2002), Shawe-Taylor and Cristianini (2004), and Bishop (2006), among others. Moreover, many of the most recent developments can be found in journals such as *Journal of Machine Learning Research*, *Machine Learning*, and *Annals of Statistics*, or in the proceedings to conferences such as *NIPS* or *COLT*.

The process of writing this book took about four years. Springer asked one of us (A.C.), after a talk on robustness properties of SVMs, whether he was willing to write a book on this topic. After a few weeks to convince Ingo to write a joint textbook, our plan was to write a very condensed book of about 200–250 pages within one-and-a-half years. However, this soon turned out to be unrealistic because we aimed to include some of our own research results that were partially written simultaneously to the book. Moreover, we totally underestimated the richness of the available literature on SVMs and the field's speed of progress.

Acknowledgments

First of all, we would like to thank our wives, Anne and Wiebke, for their permanent support and understanding during the entire process of writing

the book. Without their patience it would have been completely impossible for us to write this book.

My (I.S.) special thanks go to Dieter Eilts, who played so sensationally at the European Championship in 1996 that the semi-final was the classic England vs. Germany. This in turn caused Bob Williamson, who stayed during the time in London, to flee from the crowded city into a secondhand bookstore, where he found a book on entropy numbers by my Ph.D. adviser Prof. Bernd Carl. Bob, who was working closely with Alex Smola and Bernhard Schölkopf, contacted Bernd Carl, and after some time Alex and Bernhard were visiting us in Jena. Inspired by their visit, I decided to work on SVMs after finishing my Ph.D. thesis, and it seems fair to say that without Dieter Eilts (and, of course, the other persons mentioned) this book would have never been written. I also thank Prof. Andreas Defant and Prof. Klaus Floret for teaching me to look behind the formulas of mathematics, and Prof. Bernd Carl and Prof. Erich Novak, who greatly supported me when I changed my focus of research to SVMs.

My (A.C.) special thanks go to Prof. Dr. Ursula Gather from the University of Dortmund for bringing robust and non-parametric statistics, rates of convergence, and other topics of statistics to my attention during my studies and later on when I was a member of the Collaborative Research Center 475 on "Reduction of Complexity in Multivariate Data Structures." Her experience and her support were always very valuable to me.

We would like to thank the following friends and colleagues who have, over the years, collaborated with us on various aspects of the work described here, provided numerous helpful discussions, and brought several references to our attention: F. Alexander, M. Anghel, P. Bartlett, J. Beirlant, O. Bousquet, C. Croux, I. Daubechies, P. L. Davies, M. Debruyne, H. Dette, U. Einmahl, L. Fernholz, L. Rosasco, P. Filzmoser, U. Gather, M. Hallin, X. He, M. Hein, A. Hinrichs, J. Howse, M. Hubert, D. Hush, T. Joachims, M. Kohl, V. Koltchinskii, N. List, G. Lugosi, S. Mendelson, S. Morgenthaler, K. Mosler, S. Mukherjee, H. Oja, D. Paindaveine, S. Portnoy, H. Rieder, E. Ronchetti, P. J. Rousseeuw, P. Ruckdeschel, S. Rüping, M. Salibian-Barrera, B. Schölkopf, C. Scovel, H. U. Simon, S. Smale, A. Smola, J. Suykens, V. Temlyakov, J. Teugels, J. Theiler, A. Tsybakov, D. Tyler, S. Van Aelst, A. Van Messem, U. von Luxburg, C. Weihs, B. Williamson, T. Zhang, and D. X. Zhou.

We deeply thank F. Alexander, M. Anghel, U. Einmahl, U. Gather, M. Hein, A. Hinrichs, M. Kohl, P. Ruckdeschel, C. Scovel, J. Theiler, and A. Van Messem, who critically read parts of the book and made many valuable suggestions.

We thank Springer and, in particular, Executive Editor John Kimmel and Frank Glanz for their support and constant advice and the editors of the Springer Series on Information Science and Statistics, especially Prof. Bernhard Schölkopf, for their support. We would also like to thank many unknown

reviewers who read one or more chapters of earlier versions of the manuscript for constructive criticism that led to considerable improvements.

Finally, we thank the Los Alamos LDRD program for basic research for grants and support for the first author. We thank the Deutsche Forschungsgesellschaft (DFG); the Collaborative Research Center SFB 475 and the interdisciplinary research program DoMuS, both at the University of Dortmund in Germany; and the Fonds voor Wetenschappelijk Onderzoek (FWO) in Belgium for grants and additional support for the second author.

Ingo Steinwart
Los Alamos National Laboratory
Group CCS-3
Los Alamos, NM, USA

Andreas Christmann
University of Bayreuth
Department of Mathematics
Bayreuth, GERMANY

Spring 2008

Reading Guide

This book contains both the foundations and advanced material on support vector machines, and as such it can serve several purposes.

First, it can serve as a textbook on SVMs for a one-semester course for graduate students by explaining the key ingredients and principles of support vector machines. For example, the following chapters or parts of them can be used in this respect: the introduction in Chapter 1 written in a tutorial style, Chapter 2 (on loss functions), Chapter 4 (on kernels and reproducing kernel Hilbert spaces), and Chapter 6 (on the statistical analysis of SVMs) are prerequisites for the understanding of SVMs and hence they should be included. This core material can be complemented for example by Chapter 5 (on infinite sample versions of SVMs), or Chapters 8 and 9 (on classification and regression) to present more concrete applications. Finally, Chapter 10 (on robustness properties) and Chapter 11 (on computational aspects) broaden the knowledge of SVMs.

Second, an advanced course for graduate students can cover the remaining parts of the book, such as surrogate loss functions, additional concentration inequalities, the parts of the chapters on classification or regression not treated in the first course, additional results on robustness properties of SVMs, and the chapter explaining how SVMs fit in as a tool in a whole data mining strategy. The second course can thus be based for example on Chapters 3, 7, 8 or 9, and 12.

Last but not least, the somewhat more advanced topics may also be interesting to researchers joining, or already working in, the field of statistical machine learning theory. The chapters and sections containing such more advanced material are indicated by an asterisk (*) in the title.

Besides the introduction, all chapters contain various exercises with levels of difficulty (indicated by a scale of one to four stars ($\star$)) ranging from those of a more repetitive nature to a serious challenge.

Moreover, to keep the book as self-contained as possible, we also added an extensive appendix that collects necessary notions and results from several disciplines, such as topology, probability theory and statistics, functional analysis, and convex analysis.

A website for this book is located at

`http://www.staff.uni-bayreuth.de/~btms01/svm.html`

Contents

1

Introduction

Overview. *The goal of this introduction is to give a gentle and informal overview of what this book is about. In particular, we will discuss key concepts and questions on statistical learning. Furthermore, the underlying ideas of support vector machines are presented, and important questions for understanding their learning mechanisms will be raised.*

Usage. *This introduction serves as a tutorial that presents key concepts and questions of this book. By connecting these to corresponding parts of the book, it is furthermore a guidepost for reading this book.*

It is by no means a simple task to give a precise definition of learning and, furthermore, the notion of learning is used in many different topics. Rather than attempting to give a definition of learning on our own, we thus only mention two possible versions. Following the Encyclopædia Britannica (online version), *learning is the*

> *"process of acquiring modifications in existing knowledge, skills, habits, or tendencies through experience, practice, or exercise."*

Simon (1983, p. 28), who was awarded the Nobel Prize in Economic Sciences in 1978, defined learning in the following way:

> *"Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time."*

1.1 Statistical Learning

Statistical learning is a particular mathematical formulation of the general concept of learning. Before we present this formulation, let us first describe and motivate it in a rather informal way. To this end, let us assume that we want to relate a certain type of *input* value or measurement(s) to an *output* or response. Knowing whether such a dependence structure between input and output values exists, and if so which functional relationship describes it, might be of interest in real-life applications such as the following:

(a) make a diagnosis based on some clinical measurements;

(b) assign the ASCII code to digitalized images of handwritten characters;
(c) predict whether a client will pay back a loan to a bank;
(d) assess the price of a house based on certain characteristics;
(e) estimate the costs of claims of insurees based on insurance data.

Another characteristic of the applications we have in mind is that there is a need to *automatically* assign the response. For example, this may be the case because the structure of the measurements is too complex to be reasonably understood by human experts, or the amount of measurements is too vast to be manually processed in a timely fashion. The examples above illustrate, however, that we often do not have a reasonable description of a functional relationship between the measurements and the desired response that can easily be formalized and implemented on a computer. One way to resolve this problem is the one taken by statistical learning theory or machine learning. In this approach, it is assumed that we have already gathered a finite set of input values together with corresponding output values. For example, in the scenarios above, we could have collected: *(a)* data on some patients who had already developed the disease, *(b)* handwritten characters with manually assigned ASCII code, *(c)* data on clients who got a loan from the bank within the last ten years, *(d)* the prices of recently sold houses, or *(e)* insurance data of previous years. In the machine learning approach, the limited number of input values with known output values are then used to "learn" the assumed but unknown functional relationship between the input values and the output values by an algorithm, which in turn makes it possible to predict the output value for *future* input values. This is a crucial point. In many applications, the collected data set consisting of input values and output values can be thought of as a finite sample taken from all possible input and output values. However, the goal is not to find a suitable description of the dependency between input and output values of the collected data set (because we already know the output values) but to find a prediction rule for output values that works well for new, so far unseen input values.

In order to formalize this approach, we first assume that all input values x are contained in a known set X that describes their format and their range. In the examples above, this could be *(a)* the set of possible values of clinically measured parameters, *(b)* the set of all possible sixteen by sixteen digitalized black and white images, *(c)* the set of all possible information on the clients, *(d)* the set of possible configurations and locations of houses in a town, or *(e)* the set of all possible collected personal information of insurees. In addition, we assume that we have a known set Y that describes the format and the range of possible responses. For example, this could simply be the set $\{\text{"negative"},\text{"positive"}\}$ or $\{-1,+1\}$ when we want to diagnose a disease, while in the other examples Y could be the set of ASCII codes related to numerals or letters, a price range, or a range that contains all possible claim amounts.

As already mentioned informally, we assume in the machine learning approach that we have collected a sequence $D := ((x_1, y_1), \ldots, (x_n, y_n))$ of input/output pairs that are used to "learn" a function $f : X \to Y$ such that $f(x)$ is a good approximation of the possible response y to an arbitrary x. Obviously, in order to find such a function, it is necessary that the already collected data D have something in common with the new and unseen data. In the framework of statistical learning theory, this is guaranteed by assuming that both past and future pairs (x, y) are *independently* generated by the same, but of course *unknown*, probability distribution P on $X \times Y$. In other words, a pair (x, y) is generated in two steps. First, the input value x is generated according to the marginal distribution P_X. Second, the output value y is generated according to the conditional probability $\mathrm{P}(\,\cdot\,|x)$ on Y given the value of x. Note that by letting x be generated by an *unknown* distribution P_X, we basically assume that we have no control over how the input values have been and will be observed. Furthermore, assuming that the output value y to a given x is stochastically generated by $\mathrm{P}(\,\cdot\,|x)$ accommodates the fact that in general the information contained in x may not be sufficient to determine a single response in a deterministic manner. In particular, this assumption includes the two extreme cases where either all input values determine an (almost surely) unique output value or the input values are completely irrelevant for the output value. Finally, assuming that the conditional probability $\mathrm{P}(\,\cdot\,|x)$ is *unknown* contributes to the fact that we assume that we do not have a reasonable description of the relationship between the input and output values. Note that this is a fundamental difference from *parametric models*, in which the relationship between the inputs x and the outputs y is assumed to follow some unknown function $f \in \mathcal{F}$ from a *known, finite-dimensional* set of functions $\mathcal{F}$.

So far, we have only described the nature of the data with which we are dealing. Our next goal is to describe what we actually mean by "learning." To this end, we assume that we have means to assess the quality of an estimated response $f(x)$ when the true input and output pair is (x, y). To simplify things, we assume throughout this book that the set of possible responses Y is a subset of $\mathbb{R}$ and that all estimated responses are real-valued. Moreover, we assume that our quality measure is a non-negative real number $L(x, y, f(x))$ that is smaller when the estimated response $f(x)$ is better. In other words, we have a function $L : X \times Y \times \mathbb{R} \to [0, \infty)$, which in the following is called a *loss function*, where the term "loss" indicates that we are interested in small values of L. Of course, in order to assess the quality of a learned function f, it does not suffice to know the value $L(x, y, f(x))$ for a particular choice of (x, y), but in fact we need to quantify how small the *function* $(x, y) \mapsto L(x, y, f(x))$ is. Clearly, there are many different ways to do this, but in statistical learning theory one usually considers the *expected* loss of f, that is, the quantity

$$\mathcal{R}_{L,\mathrm{P}}(f) := \int_{X \times Y} L\big(x, y, f(x)\big)\, d\mathrm{P}(x, y) = \int_X \int_Y L\big(x, y, f(x)\big)\, d\mathrm{P}(y|x) d\mathrm{P}_X\,,$$

which in the following is called the *risk* of f.[1] Let us consider a few examples to motivate this choice.

For the first example, we recall the scenario *(b)*, where the goal was to assign ASCII codes to digitalized images of handwritten characters. A straightforward loss function for this problem is the function $L(x, y, f(x))$, which equals one whenever the prediction $f(x)$ does not equal the true ASCII code y and is otherwise equal to zero. Now assume that we have already learned a function f and our goal is now to apply f to new bitmaps $x_{n+1}, \ldots, x_m$ that correspond to the ASCII codes $y_{n+1}, \ldots, y_m$. Then intuitively a function is better the fewer errors it makes on $x_{n+1}, \ldots, x_m$, and obviously this is equivalent to saying that the average future empirical loss

$$\frac{1}{m-n} \sum_{i=n+1}^{m} L\big(x_i, y_i, f(x_i)\big) \tag{1.1}$$

should be as small as possible. Now recall that we always assume that $(x_{n+1}, y_{n+1}), \ldots, (x_m, y_m)$ are independently generated by the same distribution P, and consequently the law of large numbers (see Theorem A.4.8) shows that, for $m \to \infty$, the average empirical loss in (1.1) converges in probability to $\mathcal{R}_{L,\mathrm{P}}(f)$. In other words, the risk is indeed a very reasonable quality measure for the function f.

In the example above the risk directly assesses how well a function f performs a certain task, namely classification. In contrast to this, our second example considers a case where the main learning objective is to estimate a function f^* and the risk used is only a relatively arbitrary tool to describe this objective. To be more concrete, let us recall the scenario *(d)*, where we wished to assess the prices of houses. According to our general assumption, these prices have the unknown distributions $\mathrm{P}(\,\cdot\,|x)$, where x may describe the configuration and location of a certain house. Now, depending on the particular application, it could be reasonable to estimate the center of the conditional probability distribution such as the conditional mean or the conditional median of Y given x. Let us write $f^*(x)$ for either of them. Intuitively, a good estimator $f(x)$ of the quantity $f^*(x)$ should be close to it, and hence we could consider loss functions of the form

$$L_p(x, y, f(x)) := |f^*(x) - f(x)|^p,$$

where $p > 0$ is some fixed number.[2] The average loss (i.e., the risk) of an estimator f then becomes

[1] Throughout the introduction, we ignore technicalities such as measurability, integrability, etc., to simplify the presentation. In the subsequent chapters, however, we will seriously address these issues.

[2] The experienced reader probably noticed that for these learning goals one often uses the least squares loss or the absolute distance loss, respectively. However, these loss functions are strictly speaking only *tools* to determine f^* and do not define the goal itself. While later in the book we will also use these loss functions

$$\mathcal{R}_{L_p,\mathrm{P}}(f) = \int_X \left| f^*(x) - f(x) \right|^p d\mathrm{P}_X(x) \,,$$

which obviously is a reasonable quality measure with a clear intuitive meaning. Note, however, that, unlike in the first example, we are not able to actually compute the loss functions L_p since we do not know f^*. Moreover, there seems to be no natural choice for p either, though at least for the problem of estimating the conditional mean we will see in Example 2.6 that $p = 2$ is in some sense a canonical choice. Similarly, we will see in Example 3.67 that, under some relatively mild assumptions on the conditional distributions $\mathrm{P}(\,\cdot\,|x)$, the choice $p = 1$ is suitable for the problem of estimating the conditional median.

Let us now return to the general description of the learning problem. To this end, recall that a function is considered to be better the smaller its risk $\mathcal{R}_{L,\mathrm{P}}(f)$ is. Hence it is natural to consider the smallest possible risk,

$$\mathcal{R}^*_{L,\mathrm{P}} := \inf_{f:X\to\mathbb{R}} \mathcal{R}_{L,\mathrm{P}}(f) \,,$$

where the infimum is taken over the set of *all* possible functions. Note that considering all possible functions is in general necessary since we do not make assumptions on the distribution P. Nevertheless, for particular loss functions, we can actually consider smaller sets of functions without changing the quantity $\mathcal{R}^*_{L,\mathrm{P}}$. For example, in the scenario where we wish to assign ASCII codes to digitalized images, it clearly makes no difference whether we consider all functions or just all functions that take values in the ASCII codes of interest. Moreover, we will see in Section 5.5 that in many cases it suffices to consider sets of functions that are in some sense dense.

So far, we have described that we wish to find a function $f : X \to \mathbb{R}$ that (approximately) minimizes the risk $\mathcal{R}_{L,\mathrm{P}}$. If the distribution P is known, this is often a relatively easy task, as we will see in Section 3.1.[3] In our setup, however, the distribution P is unknown, and hence it is in general impossible to find such an (approximate) minimizer without additional information. In the framework of statistical learning theory, this information comes in the form of the already collected finite data set $D := ((x_1, y_1), \ldots, (x_n, y_n))$, where all n data points (x_i, y_i) are assumed to be generated independently from the same distribution P. Based on this data set, we then want to build a function $f_D : X \to \mathbb{R}$ whose risk $\mathcal{R}_{L,\mathrm{P}}(f_D)$ is close to the minimal risk $\mathcal{R}^*_{L,\mathrm{P}}$. Since the process of building such a function should be done in a systematic manner, we restrict our considerations throughout this book to *learning methods*, that

as tools, it is conceptionally important to distinguish between the *learning goal* and the *tools to achieve this goal*. A systematic treatment of this difference will be given in Chapter 3.

[3] In this case there is, of course, no need to learn since we already know the distribution P that describes the desired functional relationship between input and output values. Nonetheless, we will see that we gain substantial insight into the learning process by considering the case of known P.

is, to deterministic methods that assign to *every* finite sequence D a *unique* function f_D. Now one way to formalize what is meant by saying that a learning method is able to learn is the notion of *universal consistency*. This notion which we will discuss in detail in Section 6.1, requires that, for *all* distributions P on $X \times Y$, the functions f_D produced by the learning method satisfy

$$\mathcal{R}_{L,\mathrm{P}}(f_D) \to \mathcal{R}^*_{L,\mathrm{P}}\,, \qquad n \to \infty, \tag{1.2}$$

in probability. In other words, we wish to have a learning method that in the long run finds functions with near optimal response performance *without* knowing any specifics of P. Although this generality with respect to P may seem to be a very strong goal, it has been known since a seminal paper by Stone (1977) that for certain loss functions there do exist such learning methods. Moreover, in Section 6.4 and Chapter 9, we will see that the learning methods we will deal with, namely *support vector machines* (SVMs), are often universally consistent.

One clear disadvantage of the notion of universal consistency is that no speed of convergence is quantified in (1.2). In particular, we cannot *a priori* exclude the possibility that universally consistent methods are only able to find functions f_D with near optimal response performances for extremely large values of n. Unfortunately, it turns out by the so-called *no-free-lunch theorem* shown by Devroye (1982) that in general this issue cannot be resolved. To be more precise, we will see in Section 6.1 that for every learning method and every *a priori* fixed speed of convergence there exists a distribution P for which the learning method cannot achieve (1.2) with the prescribed speed. Having said that, it is, however, well-known that for many learning methods it is possible to derive uniform convergence rates, and sometimes also asymptotic distributions, under certain *additional* assumptions on P. For the earlier mentioned and more restrictive case of parametric models and for local asymptotic optimality results, we refer to Lehmann and Casella (1998) and LeCam (1986), respectively. Moreover, if P is an element of a neighborhood of a parametric model and a robust statistical method is used, we refer to Huber (1964, 1967, 1981), Hampel *et al.* (1986), and Rieder (1994). On the other hand, convergence rates for regression with smooth but otherwise unspecified target functions are discussed in great detail by Györfi *et al.* (2002). While one can show that convergence rates for regression can also be obtained by certain SVMs, we will mainly focus on convergence rates for classification. In particular, we will present a class of mild assumptions on P in Chapter 8 that, while realistic in many cases, still allow us to derive reasonable learning rates. Finally, one should always keep in mind that the existence of convergence rates only provides theoretical assurance up to a certain degree since in practice we can almost never rigorously prove that the required assumptions are met.

1.2 Support Vector Machines: An Overview

Let us now describe the basic ideas of support vector machines. In order to fix ideas, we focus here in the introduction on only one particular type of learning problem, namely binary classification. For convenience, let us in this and the following section assume that the loss function depends on x only via $f(x)$ such that we can simply write $L(y, f(x))$ instead of $L(x, y, f(x))$. As in example *(a)* mentioned earlier, where the goal was to make a diagnosis, the goal in binary classification is to estimate a response that only has two states. Consequently, we define the set of possible response values by $Y := \{-1, +1\}$. Moreover, the *classification loss function* L_{class} commonly used in binary classification only penalizes misclassifications (i.e., $L_{\text{class}}(y, f(x))$ equals 1 if $\operatorname{sign} f(x) \neq y$ and equals 0 otherwise). Finally, we assume that all possible input values are contained in some set, say $X \subset \mathbb{R}^d$.

Let us now recall that the *learning goal* was to find a function f^* that (approximately) achieves the smallest possible risk,

$$\mathcal{R}^*_{L,\mathrm{P}} = \inf_{f:X\to\mathbb{R}} \mathcal{R}_{L,\mathrm{P}}(f)\,, \tag{1.3}$$

where $L := L_{\text{class}}$. Since the distribution P generating the input/output pairs is unknown, the risk $\mathcal{R}_{L,\mathrm{P}}(f)$ is unknown and consequently we cannot directly find f^*. To resolve this problem, it is tempting to replace the risk $\mathcal{R}_{L,\mathrm{P}}(f)$ in (1.3) by its empirical counterpart

$$\mathcal{R}_{L,\mathrm{D}}(f) := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))\,,$$

where $D := ((x_1, y_1), \ldots, (x_n, y_n))$ is the finite sequence of already gathered samples.[4] Unfortunately, however, even though the law of large numbers shows that $\mathcal{R}_{L,\mathrm{D}}(f)$ is an approximation of $\mathcal{R}_{L,\mathrm{P}}(f)$ for each *single* f, solving

$$\inf_{f:X\to\mathbb{R}} \mathcal{R}_{L,\mathrm{D}}(f) \tag{1.4}$$

does not in general lead to an approximate minimizer of $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$. To see this, consider the function that classifies all x_i in D correctly but equals 0 everywhere else. Then this function is clearly a solution of (1.4), but since this function only memorizes D, it is in general a very poor approximation of (1.3). This example is an extreme form of a phenomenon called *overfitting*, in which the learning method produces a function that models too closely the output values in D and, as a result, has a poor performance on future data.

One common way to avoid overfitting is to choose a small set $\mathcal{F}$ of functions $f : X \to \mathbb{R}$ that is assumed to contain a reasonably good approximation of the solution of (1.3). Then, instead of minimizing $\mathcal{R}_{L,\mathrm{D}}(\,\cdot\,)$ over all functions, one minimizes only over $\mathcal{F}$; i.e., one solves

[4] The corresponding empirical distribution is denoted by $\mathrm{D} := \frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i,y_i)}$.

$$\inf_{f \in \mathcal{F}} \mathcal{R}_{L,\mathrm{D}}(f)\,. \tag{1.5}$$

This approach, which is called *empirical risk minimization* (ERM), often tends to produce approximate solutions of the infinite-sample counterpart of (1.5),

$$\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} := \inf_{f \in \mathcal{F}} \mathcal{R}_{L,\mathrm{P}}(f)\,. \tag{1.6}$$

In particular, we will see in Section 6.3 that this is true if $\mathcal{F}$ is finite or if $\mathcal{F}$ can at least be approximated by a finite set of functions. Unfortunately, however, this approach has two serious issues. The first one is that in the problems we are interested in our knowledge of P is in general not rich enough to identify a set $\mathcal{F}$ such that a solution of (1.6) is a reasonably good approximation of the solution of (1.3). In other words, we usually cannot guarantee that the *model error* or *approximation error*

$$\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} - \mathcal{R}^*_{L,\mathrm{P}} \tag{1.7}$$

is sufficiently small. The second issue is that solving (1.5) may be computationally infeasible. For example, the 0/1 loss function used in binary classification is non-convex, and as a consequence solving (1.5) is often NP-hard, as Höffgen *et al.* (1995) showed.

To resolve the first issue, one usually increases the size of the set $\mathcal{F}$ with the sample size n so that the approximation error (1.7) decreases with the sample size. In this approach, it is crucial to ensure that solving (1.5) for larger sets $\mathcal{F}$ still leads to approximate solutions of (1.6), which is usually achieved by controlling the growth of $\mathcal{F}$ with the sample size. The second issue is often resolved by replacing the risk $\mathcal{R}_{L,\mathrm{D}}(\,\cdot\,)$ in (1.5) by a suitable surrogate that is computationally more attractive. Various proposed learning methods follow these two basic ideas in one form or another, and a complete account of these methods would fill another textbook. Consequently, we will focus in the following on how support vector machines implement these two basic strategies.

Let us first explain the idea of how SVMs make the optimization problem computationally feasible. Here the first step is to replace the 0/1 classification loss by a *convex* surrogate. The most common choice in this regard is the hinge loss, which is defined by

$$L_{\mathrm{hinge}}(y,t) := \max\{0, 1 - yt\}, \qquad y \in \{-1,+1\},\, t \in \mathbb{R};$$

see also Figure 3.1. It is easy to see that the corresponding empirical risk $\mathcal{R}_{L_{\mathrm{hinge}},\mathrm{D}}(f)$ is convex in f, and consequently, if we further assume that $\mathcal{F}$ is a convex set, we end up with the convex optimization problem

$$\inf_{f \in \mathcal{F}} \mathcal{R}_{L_{\mathrm{hinge}},\mathrm{D}}(f)\,, \tag{1.8}$$

which defines our *learning method*. Before we present another step SVMs take to make the optimization problem more attractive, let us briefly note that

using the surrogate loss function L_{hinge} instead of the non-convex classification loss raises a new issue. To explain this, let us assume that (1.8) is well-behaved in the sense that solving (1.8) leads to a function f_D whose risk $\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f_D)$ is close to

$$\mathcal{R}^*_{L_{\text{hinge}},\mathrm{P},\mathcal{F}} := \inf_{f\in\mathcal{F}} \mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f)\,.$$

Moreover, we assume that the *approximation error* $\mathcal{R}^*_{L_{\text{hinge}},\mathrm{P},\mathcal{F}} - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}}$ with respect to the hinge loss is small. In other words, we assume that we are in a situation where we can hope to find a function $f_D \in \mathcal{F}$ such that

$$\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f_D) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}}$$

is small. However, we are actually interested in learning with respect to the classification risk; i.e., we want $\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_D) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}}$ to be small. Obviously, one way to ensure this is to establish inequalities between the two differences. Fortunately, for the hinge and classification losses, it will turn out in Section 2.3 that relatively elementary considerations yield

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \;\le\; \mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}}$$

for all functions $f : X \to \mathbb{R}$. Therefore, the idea of using the hinge loss is justified as long as we can guarantee that our learning method learns well in terms of the hinge loss. Unfortunately, however, for several other learning problems, such as estimating the conditional median, it turns out to be substantially more difficult to establish inequalities that relate the risk used in the *learning method* to the risk defining the *learning goal*. Since this is a rather general issue, we will develop in Chapter 3 a set of tools that make it possible to derive such inequalities in a systematic way.

Let us now return to support vector machines. So far, we have seen that for binary classification they replace the non-convex 0/1-classification loss by a convex surrogate loss such as the hinge loss. Now, the second step of SVMs toward computational feasibility is to consider very specific sets of functions, namely *reproducing kernel Hilbert spaces*[5] H. These spaces will be introduced and investigated in detail in Chapter 4, and hence we skip a formal definition here in the introduction and only mention that for now we may simply think of them as Hilbert spaces that consist of functions $f : X \to \mathbb{R}$. We will see in Chapter 4 that every RKHS possesses a unique function $k : X \times X \to \mathbb{R}$, called its *kernel*, that can be used to describe all functions contained in H. Moreover, the value $k(x, x')$ can often be interpreted as a measure of dissimilarity between the input values x and x'. Let us fix such an RKHS H and denote its norm by $\|\cdot\|_H$. For a fixed real number $\lambda > 0$, support vector machines then find a minimizer of

$$\inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)\,, \tag{1.9}$$

[5] We will often use the abbreviation RKHS for such Hilbert spaces.

where the *regularization term* $\lambda\|f\|_H^2$ is used to penalize functions f with a large RKHS norm. One motivation of this regularization term is to reduce the danger of overfitting; rather complex functions $f \in H$, which model too closely the output values in the training data set D, tend to have large H-norms. The regularization term penalizes such functions more than "simple" functions.

If L is a convex loss function such as the hinge loss, the objective function in (1.9) becomes convex in f. Using this, we will see in Section 5.1 and Chapter 11 that (1.9) has in basically all situations of interest a *unique* and *exact* minimizer, which in the following we denote by $f_{\mathrm{D},\lambda}$. Moreover, we will see in Section 5.1 that this minimizer is of the form

$$f_{\mathrm{D},\lambda} = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)\,, \tag{1.10}$$

where $k : X \times X \to \mathbb{R}$ is the kernel that belongs to H and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ are suitable coefficients. In other words, the minimizer $f_{\mathrm{D},\lambda}$ is a weighted average of (at most) n functions $k(x_i, \cdot)$, where the weights α_i are data-dependent. A remarkable consequence of the representation given in (1.10) is the fact that $f_{\mathrm{D},\lambda}$ is contained in a known *finite* dimensional space, namely the linear span of $k(x_i, \cdot)$, $1 \le i \le n$, even if the space H itself is substantially larger. This observation makes it possible to consider even *infinite* dimensional spaces H such as the one belonging to the popular *Gaussian radial basis function* (RBF) *kernel* (see Section 4.4 for a detailed account) defined by

$$k_\gamma(x, x') := \exp\left(-\gamma^{-2}\|x - x'\|_2^2\right)\,, \quad x, x' \in \mathbb{R}^d,$$

where $\gamma > 0$ is a fixed parameter called the *width*. Moreover, for particular loss functions such as the hinge loss, we will see below, and in much more detail in Chapter 11, that the solution $f_{\mathrm{D},\lambda}$ of (1.9) can be found by solving a *convex quadratic* optimization problem with linear constraints. For the hinge loss, we will further develop and analyze efficient algorithms to compute $f_{\mathrm{D},\lambda}$ in Section 11.2.

Although computational feasibility is an important feature of every learning algorithm, one of the most important features is definitely its ability to find decision functions having near optimal risks. Let us now motivate why SVMs can find such functions in many situations. To this end, we again restrict our considerations here in the introduction to SVMs using the hinge loss L. For this loss function, we easily check that $\mathcal{R}_{L,\mathrm{D}}(0) = 1$ for every sample set D and hence obtain

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 \;\le\; \lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) \;\le\; \mathcal{R}_{L,\mathrm{D}}(0) = 1\,,$$

where in the second estimate we used that, by definition, $f_{\mathrm{D},\lambda}$ is a solution of (1.9). Consequently, $f_{\mathrm{D},\lambda}$ is also a solution of the optimization problem

$$\min_{\|f\|_H \le \lambda^{-1/2}} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)\,.$$

Now smaller values of λ lead to larger sets $\{f \in H : \|f\|_H \leq \lambda^{-1/2}\}$, the minimum is computed over, and in addition smaller values of λ also reduce the influence of the regularization term $\lambda\|\cdot\|_H^2$. Consequently, the SVM optimization problem can be interpreted as an approximation of the ERM approach with increasing sets of functions. So far this interpretation is, however, little more than a vague analogy, but we will see in Section 6.4 that the techniques used to analyze ERM can actually be extended to a basic statistical analysis of SVMs. In particular, we will see there that with probability not smaller than $1 - e^{-\tau}$ we have

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) \;\leq\; \inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f) + \varepsilon(n,\lambda,\tau)\,, \qquad (1.11)$$

where $\tau > 0$ is arbitrary and the value $\varepsilon(n,\lambda,\tau)$, which we will derive explicitly, converges to 0 for $n \to \infty$. Consequently, we see that besides this statistical analysis we also need to understand how the right-hand side of (1.11) behaves as a function of λ. This will be one of the major topics of Chapter 5, where we will in particular show that

$$\lim_{\lambda\to 0}\big(\inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f)\big) = \inf_{f\in H} \mathcal{R}_{L,\mathrm{P}}(f)\,. \qquad (1.12)$$

Starting from this observation, we will further investigate in Section 5.5 under which conditions H can be used to approximate the minimal risk in the sense of $\inf_{f\in H} \mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}^*_{L,\mathrm{P}}$. In particular, we will see with some results from Section 4.4 that in almost all situations the RKHSs of the Gaussian RBF kernels satisfy this equality. Combining this with (1.11), we will then show in Section 6.4 that SVMs using such a kernel can be made universally consistent by using suitable null sequences (λ_n) of regularization parameters that depend only on the sample size n but *not* on the distribution P. In addition, similar consistency results for more general cases with *unbounded* Y are given in Section 9.2 for regression and Section 9.3 for quantile regression. Interestingly, the analysis of Section 6.4 further shows that we obtain learning rates for such sequences (λ_n) whenever we know the speed of convergence in (1.12). Unfortunately, however, the best possible rates this analysis yields can only be achieved by sequences (λ_n) that use knowledge about the speed of convergence in (1.12).[6] Since the latter is unknown in almost all applications, we thus need strategies that *adaptively* (i.e., without knowledge of P) find nearly optimal values for λ. In Section 6.5, we will analyze a very simple version of such a strategy that despite its simplicity resembles many ideas of commonly used, more complex strategies. In practice, not only the quantity λ but also some other so-called *hyperparameters* have an impact on the quality of $f_{\mathrm{D},\lambda}$. For example, if we use a Gaussian RBF kernel, we have to specify the value of the width γ. We will deal with this issue in Section 8.2 from a theoretical point of view and in Section 11.3 from a practical point of view.

[6] This phenomenon remains true for the more advanced analysis in Chapter 7.

The discussion above showed that, in order to understand when SVMs learn with a favorable learning rate, one needs to know when the RKHS H approximates the risk easily in the sense of the convergence given in (1.12). By the no-free-lunch theorem mentioned earlier and (1.11), the latter requires assumptions on the distribution P and the space H. In Section 8.2, we will present such an assumption for binary classification that *(a)* is weak enough to be likely met in practice and *(b)* still allows a reasonable mathematical treatment. Roughly speaking (see Figures 8.1 and 8.2 for some illustrations), this type of assumption describes how the data are typically concentrated in the vicinity of the decision boundary.

The ability to learn from a finite number of data points with a reasonable algorithmic complexity is in general not enough for a learning method to be successful from both a theoretical and an applied point of view. Indeed, already Hadamard (1902) thought that a well-posed mathematical problem should have the property that there *exists a unique solution that additionally depends continuously on the data*. In our case, this means that a few outlying data points that are far away from the pattern set by the majority of the data should influence $f_{\mathrm{D},\lambda}$ and its associated risk $\mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda})$ only in a continuous and bounded manner. More generally, a good learning method should give stable results for (almost all) distributions Q lying in a small neighborhood of the unknown distribution P. Therefore, we will describe in Chapter 10 some modern concepts of robust statistics such as the influence function, sensitivity curve, and maxbias. By applying these general concepts to SVMs, it will be shown that SVMs have—besides other good properties—the advantage of being *robust* if the loss function L and the kernel k are suitably chosen. In particular, weak conditions on L and k are derived that guarantee good robustness of SVM methods for large classes of probability distributions. These results will be derived not only for the classification case but also for quantile regression and regression for the mean, where for the latter two cases we assume an unbounded range Y of possible outputs. For the latter, we will need an explicit control over the growth of the loss function, and in Section 2.2 we therefore introduce a general notion, namely so-called *Nemitski loss functions*, that describes this growth. Although this notion was originally tailored to regression with unbounded Y, it will turn out that it is also a very fruitful concept for many other learning problems. Finally, we would like to point out that combining the robustness results from Chapter 10 with those from Chapter 9 on SVMs for regression shows that learning properties and robustness properties of SVMs are connected to each other. Roughly speaking, it will turn out that the SVMs with better robustness properties are able to learn over larger classes of distributions than those SVMs with worse robustness properties. Chapter 10 therefore complements recent *stability* results obtained by Poggio *et al.* (2004) and Mukherjee *et al.* (2006), who study the impact of one data point on SVMs under the boundedness assumption of the space of input and output values.

Besides robustness, another important property of learning methods is their ability to find decision functions that can be evaluated in a computationally efficient manner. For support vector machines, the time needed to evaluate $f_{\mathrm{D},\lambda}(x)$ is obviously proportional to the number of nonzero coefficients α_i in the representation (1.10). For SVMs used for binary classification, we will thus investigate the typical number of such coefficients in Sections 8.4 and 8.5. Here and also in Chapter 11, on computational aspects of SVMs, it will turn out that yet another time the choice of the loss function in (1.9) plays a crucial role.

Finally, it is important to mention that support vector machines are often used as *one* tool in data mining projects. Therefore, we will briefly describe in Chapter 12 a general and typical data mining strategy. In particular, we will show how SVMs can be a successful part of a data mining project and mention a few alternative statistical modeling tools often used for data mining purposes, such as generalized linear models and trees. Last but not least, a brief comparison of the advantages and disadvantages of such tools with respect to SVMs is given.

1.3 History of SVMs and Geometrical Interpretation

Considering regularized empirical (least squares) risks over reproducing kernel Hilbert spaces is a relatively old idea (see, e.g., Poggio and Girosi, 1990; Wahba, 1990; and the references therein). Although this view on support vector machines will be adopted throughout the rest of this book, it is nonetheless interesting to take a sidestep and have a look of how a geometric idea led to the first algorithms named "support vector machines." To this end, we again consider a binary classification problem with $Y = \{-1, +1\}$. For this learning problem, the original SVM approach by Boser *et al.* (1992) was derived from the *generalized portrait algorithm* invented earlier by Vapnik and Lerner (1963). Therefore, we begin by describing the latter algorithm. To this end, let us assume that our input space X is a subset of the Euclidean space $\mathbb{R}^d$. Moreover, we assume that we have a training set $D = ((x_1, y_1), \ldots, (x_n, y_n))$ for which there exists an element $w \in \mathbb{R}^d$ with $\|w\|_2 = 1$ and a real number $b \in \mathbb{R}$ such that

$$\begin{aligned} \langle w, x_i \rangle + b &> 0, && \text{for all } i \text{ with } y_i = +1, \\ \langle w, x_i \rangle + b &< 0, && \text{for all } i \text{ with } y_i = -1. \end{aligned}$$

In other words, the affine linear hyperplane described by (w, b) perfectly separates the training set D into the two groups $\{(x_i, y_i) \in D : y_i = +1\}$ and $\{(x_i, y_i) \in D : y_i = -1\}$. Now, the generalized portrait algorithm constructs a perfectly separating hyperplane, described by (w_D, b_D) with $\|w_D\|_2 = 1$, that has maximal *margin* (i.e., maximal distance to the points in D). Its resulting decision function is then defined by

$$f_D(x) \;:=\; \operatorname{sign}\big(\langle w_D, x\rangle + b_D\big) \tag{1.13}$$

for all $x \in \mathbb{R}^d$. In other words, the decision function f_D assigns negative labels to one affine half-space defined by the hyperplane (w_D, b_D) and positive labels to the other affine half-space. Now note that these half-spaces do not change if we consider $(\kappa w_D, \kappa b_D)$ for some $\kappa > 0$. Instead of looking for an element $w_D \in \mathbb{R}^d$ with $\|w_D\|_2 = 1$ that maximizes the margin, we can thus also fix a lower bound on the margin and look for a vector $w^* \in \mathbb{R}^d$ that respects this lower bound and has minimal norm. In other words, we can seek a solution $(w_D^*, b_D^*) \in \mathbb{R}^d \times \mathbb{R}$ of the optimization problem

$$\begin{array}{lll} \text{minimize} & \langle w, w\rangle & \text{over } w \in \mathbb{R}^d,\, b \in \mathbb{R} \\ \text{subject to} & y_i(\langle w, x_i\rangle + b) \geq 1 & i = 1, \ldots, n\,. \end{array} \tag{1.14}$$

Simple linear algebra shows that $w_D = w_D^*/\|w_D^*\|_2$ and $b_D = b_D^*/\|w_D^*\|_2$, and hence solving the optimization problem (1.14) indeed yields the affine hyperplane constructed by the generalized portrait algorithm.

Although geometrically compelling, the ansatz of the generalized portrait algorithm obviously has two shortcomings:

i) A *linear* form of the decision function may not be suitable for the classification task at hand. In particular, we may be confronted with situations in which the training set D cannot be linearly separated at all and hence (w_D, b_D) does not exist.

ii) In the presence of noise, it can happen that we need to misclassify some training points in order to avoid overfitting. In particular, if the dimension d is greater than or equal to the sample size n, overfitting can be a serious issue.

To resolve the first issue, the SVM initially proposed by Boser *et al.* (1992) maps the input data $(x_1, \ldots, x_n)$ into a (possibly infinite-dimensional) Hilbert space H_0, the so-called *feature space*, by a typically non-linear map $\Phi : X \to H_0$ called the *feature map*. Then the generalized portrait algorithm is applied to the mapped data set $((\Phi(x_1), y_1), \ldots, (\Phi(x_n), y_n))$; i.e., it is applied in H_0 instead of in X. In other words, we replace x and x_i in (1.13) and (1.14) by $\Phi(x)$ and $\Phi(x_i)$, respectively, and the vector w in (1.14) is chosen from the Hilbert space H_0. The corresponding learning method was initially called *maximal margin classifier*, and later also *hard margin SVM*.

We will see in Section 4.6 that for certain feature maps Φ the first issue of the generalized portrait algorithm is successfully addressed. In particular, we will show that there exist feature maps for which *every* training set without contradicting examples (i.e., without samples (x_i, y_i) and (x_j, y_j) satisfying $x_i = x_j$ and $y_i \neq y_j$) can be perfectly separated by a hyperplane in the feature space. The price for this high flexibility however is, that the separating hyperplane now lies in a high or even infinite-dimensional space, and hence the second issue of generating overfitted decision functions becomes even more serious.

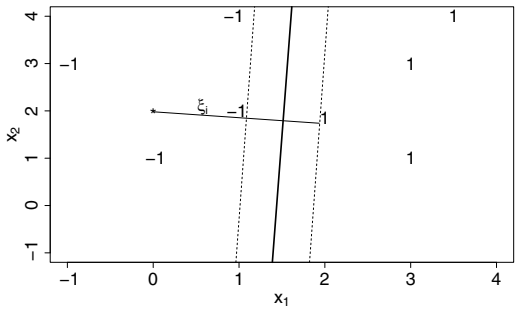


Fig. 1.1. Geometric interpretation of the soft margin SVM in a two-dimensional feature space.

This second issue was first addressed by the *soft margin support vector machine* of Cortes and Vapnik (1995). To explain their approach, recall that in the optimization problem (1.14) the constraints $y_i(\langle w, x_i\rangle + b) \geq 1$ forced the hyperplanes to make no errors on the training data set D. The approach of the soft margin SVM is thus to relax these constraints by requiring only that (w, b) satisfy $y_i\big(\langle w, x_i\rangle + b\big) \geq 1 - \xi_i$ for some so-called *slack variables* $\xi_i \geq 0$. However, if these slack variables are too large, the relaxed constraints would be trivially satisfied, and hence one has to add safeguards against such behavior. One way to do so is to add the slack variables to the objective function in (1.14).[7] Combining these modifications with the feature map idea leads to the quadratic optimization problem

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{n}\xi_i & \text{for } w \in H_0,\, b \in \mathbb{R},\, \xi \in \mathbb{R}^n \\ \text{subject to} & y_i(\langle w, \Phi(x_i)\rangle + b) \geq 1 - \xi_i, & i = 1, \ldots, n \\ & \xi_i \geq 0, & i = 1, \ldots, n, \end{array} \tag{1.15}$$

where $C > 0$ is a free (but fixed) parameter that is used to balance the first term of the objective function with the second. Note that, due to the special form of the supplemented term $C\sum_{i=1}^{n}\xi_i$, the objective function is still convex, or to be more precise, quadratic, while the constraints are all linear.

Although this optimization problem looks at first glance more complicated than that of the generalized portrait algorithm, it still enjoys a nice geometrical interpretation for *linear kernels* $k(x, x') := \langle x, x'\rangle$ where $x, x' \in \mathbb{R}^d$ (see Lemma 4.7). Let us illustrate this in the case where $X = H_0 = \mathbb{R}^2$ and

[7] Their original motivation for this step was a little more involved, but at this point we decided to slightly sacrifice historical accuracy for the sake of clarity.

$\Phi : X \to H_0$ is the identity; see Figure 1.1. To this end, we fix a vector $w = (w_1, w_2) \in \mathbb{R}^2$, where without loss of generality we assume $w_1 < 0$ and $w_2 > 0$. Moreover, we fix a sample (x, y) from D whose index denoting the sample number we omit here for notational reasons. Instead we denote the coordinates of x by indexes; i.e., we write $x = (x_1, x_2)$. Let us first consider the case $y = 1$. Since $\langle w, \Phi(x)\rangle = w_1 x_1 + w_2 x_2$ it is then easy to see that the linear constraint

$$y(\langle w, \Phi(x)\rangle + b) \geq 1 - \xi \tag{1.16}$$

requires a strictly positive slack variable (i.e., $\xi > 0$) if and only if

$$x_2 < \frac{1-b}{w_2} - \frac{w_1}{w_2}\, x_1. \tag{1.17}$$

In an analogous manner, we obtain in the case $y = -1$ that (1.16) requires a strictly positive slack variable if and only if

$$x_2 > \frac{-1-b}{w_2} - \frac{w_1}{w_2}\, x_1.$$

A comparison of these inequalities shows that both lines have equal slopes but different intercept terms. The latter terms define a tube of width $2/w_2$ in the x_2-direction around the affine hyperplane given by (w, b), which in our simple case is described by

$$x_2 = -\frac{b}{w_2} - \frac{w_1}{w_2}\, x_1.$$

Let us compare the decisions made by this hyperplane with the behavior of the slack variables. To this end, we again restrict our considerations to the case $y = 1$. Moreover, we assume that $x = (x_1, x_2)$ is correctly classified, i.e.,

$$x_2 > -\frac{b}{w_2} - \frac{w_1}{w_2}\, x_1,$$

and that x is contained inside the tube around the separating hyperplane. Then (1.17) is satisfied and hence we have a strictly positive slack variable that is penalized in the objective function of (1.15). In other words, the objective function in (1.15) penalizes margin errors (i.e., points inside the tube or lying in the wrong affine hyperplane) and not only classification errors (i.e., points lying in the wrong affine half-space).

Let us now relate the optimization problem (1.15) to the previous SVM formulation given by the optimization problem (1.9). To this end, we observe that the first set of linear constraints can be rewritten as $\xi_i \geq 1 - y_i(\langle w, \Phi(x_i)\rangle + b)$. Combining this constraint with the second set of constraints, namely $\xi_i \geq 0$, we see that the slack variables must satisfy

$$\xi_i \geq \max\bigl\{0, 1 - y_i(\langle w, \Phi(x_i)\rangle + b)\bigr\} = L\bigl(y_i, \langle w, \Phi(x_i)\rangle + b\bigr),$$

where L is the hinge loss introduced earlier. Obviously, the objective function in (1.15) becomes minimal in ξ_i if this inequality is actually an equality. For a

given $(w, b) \in H_0 \times \mathbb{R}$, let us now consider the function $f_{(w,b)} : X \to \mathbb{R}$ defined by $f_{(w,b)}(x) := \langle w, \Phi(x_i) \rangle + b$. Multiplying the objective function in (1.15) by $2\lambda := \frac{1}{nC}$ we can thus rewrite (1.15) in the form

$$\min_{(w,b)\in H_0\times\mathbb{R}} \lambda\langle w, w\rangle + \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f_{(w,b)}(x_i)\big). \tag{1.18}$$

Compared with the optimization problem (1.9), that is,

$$\inf_{f\in H} \lambda\|f\|_H^2 + \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i)\big),$$

there are obviously two major differences. The first one is that in the geometrical approach we consider a *general* Hilbert space H_0 and define a function $f_{(w,b)}$ in terms of an affine hyperplane specified by (w, b) in this space, while in (1.9) we start with an RKHS H and directly consider the functions contained in H. Remarkably, however, both approaches are equivalent if we fix b. More precisely, we will see in Section 4.2 that the functions $\langle w, \Phi(\,\cdot\,)\rangle$, $w \in H_0$, form an RKHS H whose norm can be computed by

$$\|f\|_H = \inf\big\{\|w\|_{H_0} : w \in H_0 \text{ with } f = \langle w, \Phi(\,\cdot\,)\rangle\big\}.$$

Consequently, (1.18) is equivalent to the optimization problem

$$\inf_{(f,b)\in H\times\mathbb{R}} \lambda\|f\|_H^2 + \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i) + b\big);$$

i.e., modulo the so-called *offset term* or *intercept term* b, the geometric approach is indeed equivalent to the RKHS approach (1.9). The offset term, however, makes a real difference and, in general, the decision functions produced by both approaches are different. This, of course, raises the question which optimization problem one should prefer. For very simple feature maps such as the identity map $\mathrm{id} : \mathbb{R}^d \to \mathbb{R}^d$, the offset term has obviously a clear advantage since it addresses translated data. Moreover, the version with the offset term is implemented in many standard software packages for SVMs. On the other hand, for more flexible feature maps such as those of Gaussian RBF kernels, which belong to the most important kernels in practice, the offset term has neither a known theoretical nor an empirical advantage. In addition, the theoretical analysis is often substantially complicated by the offset term. For the theoretical chapters of this book, we thus decided to exclusively consider the approach without an offset, while Chapter 11, which deals with computational aspects of SVMs, considers the approaches with and without an offset term. However, we sometimes mention theoretical results covering the offset term in the sections "Further Reading and Advanced Topics", such as in Section 10.7 for robustness properties of SVMs.

The optimization problem (1.15) has the drawback that it has to be solved in an often high- or even infinite-dimensional Hilbert space H_0. We will see in Chapter 11 that in practice one thus uses the Lagrange approach to compute the corresponding dual program. For the hinge loss function, for example, this dual program is given by

$$\begin{aligned} &\text{maximize } \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle \quad \text{over } \alpha \in [0, C]^n \\ &\text{subject to } \sum_{i=1}^{n} y_i \alpha_i = 0; \end{aligned}$$

see Example 11.3. Moreover, we will see that if $(\alpha_1^*, \ldots, \alpha_n^*)$ denotes a solution of this optimization problem, the solution (w_D^*, b_D^*) of (1.15) can be computed by

$$w_D^* = \sum_{i=1}^{n} y_i \alpha_i^* \Phi(x_i)$$

and

$$b_D^* = y_j - \sum_{i=1}^{n} y_i \alpha_i^* \langle \Phi(x_i), \Phi(x_j) \rangle,$$

where j is an index with $0 < \alpha_j^* < c$. Note that w_D^* only depends on the samples x_i whose weights satisfy $\alpha_i^* \neq 0$. Geometrically, this means that the affine hyperplane described by (w_D^*, b_D^*) is only "supported" by these $\Phi(x_i)$, and hence the corresponding data points (x_i, y_i) are called *support vectors*. As mentioned above, the decision function of the soft margin SVM is given by the constructed affine hyperplane,

$$f_{w_D^*, b_D^*}(x) = \langle w_D^*, \Phi(x) \rangle + b_D^* = \sum_{i=1}^{n} y_i \alpha_i^* \langle \Phi(x_i), \Phi(x) \rangle + b_D^*, \quad x \in X.$$

Now note that in both the dual optimization problem and the evaluation of the resulting decision function only inner products of Φ with itself occur. Thus, instead of computing the feature map directly, it suffices to know the function $\langle \Phi(\cdot), \Phi(\cdot) \rangle : X \times X \to \mathbb{R}$. Interestingly, there do exist cases in which this function can be computed *without* knowing the feature map Φ itself. The Gaussian RBF kernels are examples of such a case, but there are many more. In Chapter 4, we will thus systematically investigate kernels; i.e., functions $k : X \times X \to \mathbb{R}$ for which there exists a feature map satisfying

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle, \qquad x, x' \in X.$$

Obviously, using kernels directly instead of first computing feature maps works for all statistical methods and algorithms in which inner products of the feature map but not the feature map itself are needed. By using kernels, we can

thus build a non-linear algorithm from a linear one without changing the core design of the algorithm. This observation, known as the *"kernel-trick,"* was to the best of our knowledge first explicitly stated by Schölkopf *et al.* (1998); however, it had already been used earlier by Aizerman *et al.* (1964) and Boser *et al.* (1992). Since then, various algorithms have been "kernelized," such as principal component analysis or Fisher's discriminant analysis. We refer to Schölkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004) for detailed accounts. A second advantage of the kernel trick is that the input space X is no longer required to be a subset of $\mathbb{R}^d$ since all computations are done in the feature space. Interestingly, there exist various kernels that are defined on non-vectorial data such as text or DNA sequences. Therefore, the kernel trick indeed extends the applicability of methods that can be kernelized. We refer to Schölkopf and Smola (2002), Joachims (2002), Schölkopf *et al.* (2004), and Shawe-Taylor and Cristianini (2004) for various examples of kernel approaches for non-vectorial data.

1.4 Alternatives to SVMs

There exists a vast body of literature on both classification and regression procedures, and hence describing all these methods even briefly would fill yet another book. Nonetheless, it is always good to have alternatives, and hence we would like to briefly mention at least some of the most classical approaches. However, a comparison of a few of these methods is postponed to Section 12.2 on data mining, when our knowledge of SVMs will be more complete.

Besides the *least squares method*, which goes back to Gauss, Legendre, and Adrain, *linear discriminant analysis* is probably one of the oldest methods for pattern recognition. This procedure, which was developed by Sir R. A. Fisher in 1936, is strongly linked to multivariate normal distributions and uses affine hyperplanes as decision functions. A generalization of this procedure is *quadratic discriminant analysis*, which allows quadratic decision functions. Both methods are still used by many practitioners often with good success.

In 1956, one of the first iterative algorithms for learning a linear classification rule was proposed by Rosenblatt (1956, 1962) with the *perceptron*.

Another classical method is the *k-nearest-neighbor rule* which was introduced in 1951; see Fix and Hodges (1951, 1952). It has attracted many followers and is still used by many researchers. In addition, it was the first method for which universal consistency was established; see Stone (1977). The idea of k-nearest-neighbor methods for classificaton is to construct a decision function pointwise for each x by first determining the k points of D that are closest to x and then making the prediction for $y = 1$ if and only if the average of the k corresponding y-values is positive.

The goal in *cluster analysis* is to recognize clusters in unlabeled data. We refer to the books by Hartigan (1975) and Kaufman and Rousseeuw (2005) for an introduction to cluster analysis techniques.

Parametric *logistic regression* was proposed by Sir D. R. Cox to model binomial distributed outputs; see Cox and Snell (1989). This method is based on linear decision functions but does not make specific assumptions on the distribution of the inputs. Parametric logistic regression is a special case of *generalized linear models* in which the outputs are assumed to have a distribution from an exponential family, see McCullagh and Nelder (1989). Hastie and Tibshirani (1990) proposed a semi-parametric generalization called *generalized additive models* where the inputs may influence the outputs in an additive but not necessarily linear manner. The *lasso* (Tibshirani, 1996) is a method for regularizing a least squares regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients.

Other classical methods for classification and regression are *trees*, which were proposed by Breiman *et al.* (1984). The idea of trees is to partition the input space recursively into disjoint subsets such that points belonging to the same subset behave more homogeneously than points from different subsets. Trees often produce not only accurate results but are also able to uncover the predictive structure of the problem.

Neural networks in the context of machine learning are non-linear statistical data modeling tools that can be used to model complex relationships between inputs and outputs or to find patterns in data sets. In a neural network model, simple nodes (or "neurons") are connected together to form a network of nodes. The strength of the connections in the network depends on the data and may be time-dependent to allow for adaptivity. The motivation for neural networks, which were very popular in the 1990s, goes back to McCullogh and Pitts (1943) and Rosenblatt (1962). We refer also to Bishop (1996), Anthony and Bartlett (1999), and Vidyasagar (2002).

There also exist various other *kernel-based methods*. For *wavelets*, we refer to Daubechies (1991), and for *splines* to Wahba (1990). Recent developments for kernel-based methods in the context of SVMs are descibed by Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002), and Shawe-Taylor and Cristianini (2004).

Boosting algorithms are based on an adaptive aggregation to construct from a set of weak learners a strong learner; see Schapire (1990), Freund (1995), and Freund and Schapire (1997).

Finally, the books by Hastie *et al.* (2001), Duda *et al.* (2001), and Bishop (2006) give a broad overview of various techniques used in statistical machine learning, whereas both Devroye *et al.* (1996) and Györfi *et al.* (2002) treat several classification and regression methods in a mathematically more rigorous way.

2

Loss Functions and Their Risks

Overview. *We saw in the introduction that the learning problems we consider in this book can be described by loss functions and their associated risks. In this chapter, we present some common examples of such learning problems and introduce a general notion for losses and their risks. Furthermore, we discuss some elementary yet fundamental properties of these concepts.*

Prerequisites. *Basic knowledge of measure and integration theory provided in Section A.3.*

Usage. *Sections 2.1 and 2.2 are essential for the rest of this book, and Sections 2.3 and 2.4 are used whenever we deal with classification and regression problems, respectively.*

Every learning problem requires that we specify our learning goal, i.e., what we ideally would like to achieve. We saw in the introduction that the specification of the learning problems treated in this book needs a *loss* $L(x, y, f(x))$ that describes the cost of the discrepancy between the prediction $f(x)$ and the observation y at the point x. To the loss L we then associate a *risk* that is defined by the average future loss of f. This chapter introduces these concepts and presents important examples of learning goals described by losses. In addition, basic yet useful properties of risks are derived from properties of the corresponding losses.

2.1 Loss Functions: Definition and Examples

In this section, we will first introduce loss functions and their associated risks. We will then present some basic examples of loss functions that describe the most important learning scenarios we are dealing with in this book.

In order to avoid notational overload, we assume throughout this chapter that subsets of $\mathbb{R}^d$ are equipped with their Borel σ-algebra and that products of measurable spaces are equipped with the corresponding product σ-algebra.

Let us now recall from the introduction that we wish to find a function $f : X \to \mathbb{R}$ such that for $(x, y) \in X \times Y$ the value $f(x)$ is a good prediction of y at x. The following definition will help us to define what we mean by "good".

Definition 2.1. *Let $(X, \mathcal{A})$ be a measurable space and $Y \subset \mathbb{R}$ be a closed subset. Then a function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called a* ***loss function****, or simply a* ***loss****, if it is measurable.*

In the following, we will interpret $L(x, y, f(x))$ as the *cost*, or *loss*, of predicting y by $f(x)$ if x is observed, i.e., the smaller the value $L(x, y, f(x))$ is, the better $f(x)$ predicts y in the sense of L. From this it becomes clear that constant loss functions, such as $L := 0$, are rather meaningless for our purposes, since they do not distinguish between good and bad predictions.

Let us now recall from the introduction that our major goal is to have a small *average* loss for future unseen observations (x, y). This leads to the following definition.

Definition 2.2. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a probability measure on $X \times Y$. Then, for a measurable function $f : X \to \mathbb{R}$, the* ***L-risk*** *is defined by*

$$\mathcal{R}_{L,\mathrm{P}}(f) := \int_{X \times Y} L\big(x, y, f(x)\big)\, d\mathrm{P}(x, y) = \int_X \int_Y L\big(x, y, f(x)\big)\, d\mathrm{P}(y|x)\, d\mathrm{P}_X(x).$$

Note that the function $(x, y) \mapsto L(x, y, f(x))$ is measurable by our assumptions, and since it is also non-negative, the above integral over $X \times Y$ always exists, although it is not necessarily finite. In addition, our label space $Y \subset \mathbb{R}$ is closed, and hence Lemma A.3.16 ensures the existence of the *regular* conditional probability $\mathrm{P}(\cdot\,|x)$, used in the inner integral.

For a given sequence $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$, we write $\mathrm{D} := \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}$, where $\delta_{(x_i, y_i)}$ denotes the Dirac measure at (x_i, y_i). In other words, D is the empirical measure associated to D. The risk of a function $f : X \to \mathbb{R}$ with respect to this measure is called the **empirical L-risk**

$$\mathcal{R}_{L,\mathrm{D}}(f) = \frac{1}{n} \sum_{i=1}^n L\big(x_i, y_i, f(x_i)\big)\,. \tag{2.1}$$

Let us now assume for a moment that D is a sequence of i.i.d. observations generated by P and f satisfies $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$. Recalling the law of large numbers, we then see that the empirical risk $\mathcal{R}_{L,\mathrm{D}}(f)$ is close to $\mathcal{R}_{L,\mathrm{P}}(f)$ with high probability. In this sense, the L-risk of f can be seen as an approximation to the average loss on the observations D (and vice versa).

Now recall that $L(x, y, f(x))$ was interpreted as a cost that we wish to keep small, and hence it is natural to look for functions f whose risks are as small as possible. Since the smallest possible risk plays an important role throughout this book, we now formally introduce it.

Definition 2.3. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a probability measure on $X \times Y$. Then the minimal L-risk*

$$\mathcal{R}^*_{L,\mathrm{P}} := \inf\big\{\mathcal{R}_{L,\mathrm{P}}(f) \,\big|\, f : X \to \mathbb{R} \text{ measurable}\big\}$$

is called the ***Bayes risk*** *with respect to* P *and* L*. In addition, a measurable* $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ *with* $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ *is called a* ***Bayes decision function****.*

Usually the first step in solving a practical learning problem is finding a loss function that best describes the often only informally specified learning goal. In general, the choice of a suitable loss function strongly depends on the specific application, and hence only a few general statements are possible in this regard. However, there are a few basic learning scenarios that often fit the learning problem at hand, and hence we will formally introduce these scenarios and their corresponding loss functions now.

Example 2.4 (Standard binary classification). Let $Y := \{-1, 1\}$ and P be an unknown data-generating distribution on $X \times Y$. Then the informal goal in (binary) classification is to predict the label y of a pair (x, y) drawn from P if only x is observed. The most common loss function describing this learning goal is the **classification loss**[1] $L_{\text{class}} : Y \times \mathbb{R} \to [0, \infty)$, which is defined by

$$L_{\text{class}}(y, t) := \mathbf{1}_{(-\infty,0]}\big(y \operatorname{sign} t\big)\,, \qquad y \in Y,\, t \in \mathbb{R}\,, \tag{2.2}$$

where we use the convention $\operatorname{sign} 0 := 1$. Note that L_{class} only penalizes predictions t whose signs disagree with that of y, so it indeed reflects our informal learning goal. Now, for a measurable function $f : X \to \mathbb{R}$, an elementary calculation shows

$$\begin{aligned}\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) &= \int_X \eta(x)\mathbf{1}_{(-\infty,0)}(f(x)) + (1 - \eta(x))\mathbf{1}_{[0,\infty)}(f(x))\, d\mathrm{P}_X(x) \\ &= \mathrm{P}\big(\{(x, y) \in X \times Y : \operatorname{sign} f(x) \neq y\}\big)\,,\end{aligned}$$

where $\eta(x) := \mathrm{P}(y = 1|x)$, $x \in X$. From this we conclude that f is a Bayes decision function if and only if $(2\eta(x) - 1) \operatorname{sign} f(x) \geq 0$ for P_X-almost all $x \in X$. In addition, this consideration yields

$$\mathcal{R}^*_{L_{\text{class}},\mathrm{P}} = \int_X \min\{\eta, 1 - \eta\}\, d\mathrm{P}_X\,. \qquad \lhd$$

The loss function L_{class} equally weights both types of errors, namely $y = 1$ while $f(x) < 0$, and $y = -1$ while $f(x) \geq 0$. This particularly makes sense in situations in which one wishes to *categorize* objects such as hand-written characters or images. In many practical situations, however, both error types should be weighted differently. For example, if one wants to detect computer network intrusions, then depending on the available resources for investigating alarms and the sensitivity of the network, the two types of errors, namely false alarms and undetected intrusions, are likely to have different actual costs.

[1] Formally, L_{class} is not a loss function; however, we can canonically identify it with the loss function $(x, y, t) \mapsto L_{\text{class}}(y, t)$, and hence we usually do not distinguish between L_{class} and its associated loss function. Since this kind of identification also occurs in the following examples, we will later formalize it in Definition 2.7.

Since this example is rather typical for classification problems in which the goal is to *detect* certain objects or events, we now present a weighted version of the classification scenario above.

Example 2.5 (Weighted binary classification). Let $Y := \{-1, 1\}$ and $\alpha \in (0, 1)$. Then the **α-weighted classification loss** $L_{\alpha\text{-class}} : Y \times \mathbb{R} \to [0, \infty)$ is defined by

$$L_{\alpha\text{-class}}(y, t) := \begin{cases} 1 - \alpha & \text{if } y = 1 \text{ and } t < 0 \\ \alpha & \text{if } y = -1 \text{ and } t \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

for all $y \in Y$, $t \in \mathbb{R}$. Obviously we have $2L_{1/2\text{-class}} = L_{\text{class}}$, i.e., the standard binary classification scenario is a special case of the general weighted classification scenario. Now, given a probability measure P on $X \times Y$ and a measurable $f : X \to \mathbb{R}$, the $L_{\alpha\text{-class}}$-risk can be computed by

$$\mathcal{R}_{L_{\alpha\text{-class}},\mathrm{P}}(f) = (1 - \alpha) \int_{f<0} \eta \, d\mathrm{P}_X + \alpha \int_{f \geq 0} (1 - \eta) \, d\mathrm{P}_X \,,$$

where again $\eta(x) := \mathrm{P}(y = 1|x)$, $x \in X$. From this we easily conclude that f is a Bayes decision function if and only if $(\eta(x) - \alpha) \operatorname{sign} f(x) \geq 0$ for P_X-almost all $x \in X$. Finally, the Bayes $L_{\alpha\text{-class}}$-risk is

$$\mathcal{R}^*_{L_{\alpha\text{-class}},\mathrm{P}} = \int_X \min\bigl\{(1 - \alpha)\eta, \alpha(1 - \eta)\bigr\} \, d\mathrm{P}_X \,. \qquad \triangleleft$$

In the two examples above the goal was to predict labels y from the set $\{-1, 1\}$. In the next example, we wish to predict general real-valued labels.

Example 2.6 (Least squares regression). The informal goal in regression is to predict the label $y \in Y := \mathbb{R}$ of a pair (x, y) drawn from an unknown probability measure P on $X \times Y$ if only x is observed. The most common way to formalize this goal is based on the **least squares loss** $L_{\mathrm{LS}} : Y \times \mathbb{R} \to [0, \infty)$ defined by

$$L_{\mathrm{LS}}(y, t) := (y - t)^2 \,, \qquad y \in Y, \, t \in \mathbb{R} \,. \tag{2.4}$$

In other words, the least squares loss penalizes the discrepancy between y and t *quadratically*. Obviously, for a measurable function $f : X \to \mathbb{R}$, the L_{LS}-risk is

$$\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}}(f) = \int_X \int_Y \bigl(y - f(x)\bigr)^2 \, d\mathrm{P}(y|x) \, d\mathrm{P}_X(x) \,.$$

By minimizing the inner integral with respect to $f(x)$, we then see that f is a Bayes decision function if and only if $f(x)$ almost surely equals the expected Y-value in x, i.e., if and only if

$$f(x) = \mathbb{E}_{\mathrm{P}}(Y|x) := \int_Y y \, d\mathrm{P}(y|x) \tag{2.5}$$

for P_X-almost all $x \in X$. Moreover, plugging $x \mapsto \mathbb{E}_{\mathrm{P}}(Y|x)$ into $\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}}(\,\cdot\,)$ shows that the Bayes L_{LS}-risk is the average conditional Y-variance, i.e.,

$$\mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}} = \int_X \mathbb{E}_{\mathrm{P}}(Y^2|x) - \big(\mathbb{E}_{\mathrm{P}}(Y|x)\big)^2 \, d\mathrm{P}_X(x)\,.$$

Finally, an elementary calculation shows that the *excess* L_{LS}-risk of $f : X \to \mathbb{R}$ is

$$\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}} = \int_X \big(\mathbb{E}_{\mathrm{P}}(Y|x) - f(x)\big)^2 d\mathrm{P}_X(x)\,,$$

i.e., if $\mathcal{R}_{L_{\mathrm{LS}},\mathrm{P}}(f)$ is close to $\mathcal{R}^*_{L_{\mathrm{LS}},\mathrm{P}}$, then f is close to the Bayes decision function in the sense of the $\|\,\cdot\,\|_{L_2(\mathrm{P}_X)}$. ◁

Using the least squares loss to make the informal regression goal precise seems to be rather arbitrary since, for example, for $p > 0$, the loss function

$$(y,t) \mapsto |y-t|^p\,, \qquad\qquad y \in \mathbb{R},\, t \in \mathbb{R}\,,$$

reflects the informal regression goal just as well. Nevertheless, the least squares loss is often chosen since it *"simplifies the mathematical treatment (and)... leads naturally to estimates which can be computed rapidly"*, as Györfi *et al.* (2002) write on p. 2. For SVMs, however, we will see later that none of these properties is exclusive for the least squares loss, and therefore we do not have to stick to the least squares loss for, e.g., computational reasons. On the other hand, the least squares loss is (essentially) the only loss whose Bayes decision functions have the form (2.5) for *all* distributions P with finite Bayes risk (see Proposition 3.44 for details), and hence the least squares loss is often the first choice when we wish to estimate the conditional expectations $\mathbb{E}_{\mathrm{P}}(Y|x)$, $x \in X$. Unfortunately, however, we will see in Chapter 10 that SVMs based on the least squares loss are rather sensitive to large deviations in y, and hence other losses may be preferred in some situations. We will discuss these questions in more detail in Sections 3.7 and 3.9 and Chapter 9.

A common feature of the loss functions above is that they are all *independent* of the input value x. Since this will also be the case for many other loss functions considered later, we introduce the following notion.

Definition 2.7. *A function $L : Y \times \mathbb{R} \to [0,\infty)$ is called a* ***supervised loss function****, or simply a* ***supervised loss****, if it is measurable.*

Note that a supervised loss L can be canonically identified with the loss function $\bar{L} : (x,y,t) \mapsto L(y,t)$. As in the examples above, we will thus write $\mathcal{R}_{L,\mathrm{P}}(f) := \mathcal{R}_{\bar{L},\mathrm{P}}(f)$ and $\mathcal{R}^*_{L,\mathrm{P}} := \mathcal{R}^*_{\bar{L},\mathrm{P}}$ in order to avoid notational overload.

Formally, we can also consider losses that are independent of y, i.e., we can introduce the following notion.

Definition 2.8. *A function $L : X \times \mathbb{R} \to [0,\infty)$ is called an* ***unsupervised loss function****, or simply an* ***unsupervised loss****, if it is measurable.*

Obviously, an unsupervised loss L can be canonically identified with the loss function $\bar{L} : (x, y, t) \mapsto L(x, t)$. As for supervised losses, we thus write

$$\mathcal{R}_{L,\mathrm{P}}(f) := \mathcal{R}_{\bar{L},\mathrm{P}}(f) = \int_X L\big(x, f(x)\big)\, d\mathrm{P}_X(x)$$

and $\mathcal{R}^*_{L,\mathrm{P}} := \mathcal{R}^*_{\bar{L},\mathrm{P}}$. Note that, in contrast to the risks for supervised losses, the risks for unsupervised losses are *independent of the supervisor* $\mathrm{P}(\,\cdot\,|x)$ that generates the labels. This explains the term "unsupervised loss". Since unsupervised losses do not depend on labeling information, these loss functions often occur in learning scenarios that lack labels in the available sample data. The two most important scenarios of this type are introduced in the following examples.

Example 2.9 (Density level detection). Let us suppose that we have some samples $D := (x_1, \ldots, x_n) \in X^n$ drawn in an i.i.d. fashion from an *unknown* distribution Q on X. Moreover, assume that our informal learning goal is to find the region where Q has relatively high concentration.

One way to formalize this learning goal is to assume that Q is absolutely continuous with respect to some known *reference measure* μ. Let $g : X \to [0, \infty)$ be the corresponding *unknown* density with respect to μ, i.e., $\mathrm{Q} = g\mu$. Then Q is highly concentrated in exactly the region where g is "large", i.e., our informal learning goal is to find the **density level sets** $\{g > \rho\}$ or $\{g \geq \rho\}$ for some fixed threshold $\rho > 0$. In order to find a formal specification of this learning goal, let us consider the unsupervised **density level detection (DLD) loss** $L_{\mathrm{DLD}} : X \times \mathbb{R} \to [0, \infty)$, which is defined by

$$L_{\mathrm{DLD}}(x, t) := \mathbf{1}_{(-\infty,0)}\big((g(x) - \rho)\operatorname{sign} t\big)\,, \qquad x \in X,\, t \in \mathbb{R}\,. \tag{2.6}$$

Note that for $f : X \to \mathbb{R}$ the loss $L_{\mathrm{DLD}}(x, f(x))$ penalizes the prediction $f(x)$ at x if either $f(x) \geq 0$ and $g(x) < \rho$, or $f(x) < 0$ and $g(x) > \rho$, whereas it ignores $f(x)$ if $g(x) = \rho$. In this sense, $\{f \geq 0\}$ is the prediction of f for our desired level set. In order to further formalize our informal learning goal, recall that the risks of unsupervised losses only depend on the marginal distributions P_X. In the density level detection scenario, we are mainly interested in the case $\mathrm{P}_X = \mu$, and thus we usually use the notation

$$\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f) := \mathcal{R}_{L_{\mathrm{DLD}},\mathrm{P}}(f) = \int_X L_{\mathrm{DLD}}\big(x, f(x)\big)\, d\mu(x)\,.$$

From (2.6) it is then easy to conclude that a measurable $f : X \to \mathbb{R}$ is a Bayes decision function with respect to μ if and only if $\{g > \rho\} \subset \{f \geq 0\} \subset \{g \geq \rho\}$ holds true up to μ-zero sets. Consequently, we always have $\mathcal{R}^*_{L_{\mathrm{DLD}},\mu} = 0$ and, in addition, if $\mu(\{g = \rho\}) = 0$, we find

$$\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f) = \mu\big(\{g \geq \rho\} \triangle \{f \geq 0\}\big)$$

for all measurable $f : X \to \mathbb{R}$, where $\triangle$ denotes the **symmetric difference** $A \triangle B := A \setminus B \cup B \setminus A$. In this sense, $\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f)$ measures how well $\{f \geq 0\}$ detects the level set $\{g \geq \rho\}$.

Finally, observe that, unlike for the supervised loss functions of the previous examples, we *cannot compute* $L_{\mathrm{DLD}}(x,t)$ since g is unknown to us. Consequently, we cannot use an ERM scheme based on L_{DLD} simply because we cannot compute the empirical risk $\mathcal{R}_{L_{\mathrm{DLD}},\mathrm{D}}(f)$ for any $f : X \to \mathbb{R}$. Moreover, note that for the same reason we cannot estimate the quality of a found approximation $\{f \geq 0\}$ by $\mathcal{R}_{L_{\mathrm{DLD}},\mathrm{D}}(f)$ either. Because of these disadvantages of L_{DLD}, we will investigate more accessible supervised *surrogate* losses for L_{DLD} in Section 3.8. ◁

The density level detection scenario is often used if one wants to identify *anomalous future samples* $x \in X$ on the basis of unlabeled training data $D := (x_1, \ldots, x_n) \in X^n$. To this end, it is assumed that anomalous samples are somewhat atypical in the sense that they are not clustered. In other words, they occur in regions with low concentration, and consequently they are described by a level set $\{g \geq \rho\}$ for some suitably specified ρ.

In some sense, the density level detection scenario is an unsupervised counterpart of binary classification, and in fact we will establish a precise connection between these two in Section 3.8. The following, last example describes in a similar way an unsupervised counterpart of the regression scenario.

Example 2.10 (Density estimation). Let μ be a known probability measure on X and $g : X \to [0,\infty)$ be an *unknown* density with respect to μ. Let us further assume that our goal is to estimate the density g. Then one possible way to specify this goal is based on the unsupervised loss $L_q : X \times \mathbb{R} \to [0,\infty)$, $q > 0$, defined by

$$L_q(x,t) := \bigl|g(x) - t\bigr|^q, \qquad x \in X,\, t \in \mathbb{R}. \tag{2.7}$$

As for the DLD problem, we are usually interested in distributions P with $\mathrm{P}_X = \mu$, and for such we have

$$\mathcal{R}_{L_q,\mathrm{P}}(f) = \int_X \bigl|g(x) - f(x)\bigr|^q \, d\mu(x)$$

for all measurable $f : X \to \mathbb{R}$. From this we find $\mathcal{R}^*_{L_q,\mathrm{P}} = 0$ and, in addition, it is not hard to see that every Bayes decision function equals g modulo some μ-zero set. ◁

The presented examples of unsupervised learning scenarios suggest that the absence of labels is characteristic for situations where unsupervised losses occur. However, we will see in Chapter 3 that unsupervised losses are also a very powerful tool for investigating certain questions related to supervised learning scenarios.

2.2 Basic Properties of Loss Functions and Their Risks

In this section, we introduce some additional features of loss functions such as convexity, continuity, and differentiability and relate these features to analogous features of the associated risks. Since the results of this section will be used throughout this book, we recommend that even the experienced reader becomes familiar with the material of this section.

Our first lemma shows that under some circumstances risk functionals are measurable.

Lemma 2.11 (Measurability of risks). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and $\mathcal{F} \subset \mathcal{L}_0(X)$ be a subset that is equipped with a complete and separable metric d and its corresponding Borel σ-algebra. Assume that the metric d* ***dominates the pointwise convergence****, i.e.,*

$$\lim_{n\to\infty} d(f_n, f) = 0 \qquad \Longrightarrow \qquad \lim_{n\to\infty} f_n(x) = f(x)\,, \qquad x \in X, \tag{2.8}$$

for all $f, f_n \in \mathcal{F}$. Then the evaluation map

$$\begin{aligned} \mathcal{F} \times X &\to \mathbb{R} \\ (f, x) &\mapsto f(x) \end{aligned}$$

is measurable, and consequently the map $(x, y, f) \mapsto L(x, y, f(x))$ defined on $X \times Y \times \mathcal{F}$ is also measurable. Finally, given a distribution P on $X \times Y$, the risk functional $\mathcal{R}_{L,\mathrm{P}} : \mathcal{F} \to [0, \infty]$ is measurable.

Proof. Since d dominates the pointwise convergence, we see that, for fixed $x \in X$, the $\mathbb{R}$-valued map $f \mapsto f(x)$ defined on $\mathcal{F}$ is continuous with respect to d. Furthermore, $\mathcal{F} \subset \mathcal{L}_0(X)$ implies that, for fixed $f \in \mathcal{F}$, the $\mathbb{R}$-valued map $x \mapsto f(x)$ defined on X is measurable. By Lemma A.3.17, we then obtain the first assertion. Since this implies that the map $(x, y, f) \mapsto (x, y, f(x))$ is measurable, we obtain the second assertion. The third assertion now follows from the measurability statement in Tonelli's Theorem A.3.10. □

Obviously, the metric defined by the supremum norm $\|\cdot\|_\infty$ dominates the pointwise convergence for every $\mathcal{F} \subset \mathcal{L}_\infty(X)$. Moreover, we will see in Section 4.2 that the metric of reproducing kernel Hilbert spaces also dominates the pointwise convergence.

Let us now consider some additional properties of loss functions and their risks. We begin with convexity.

Definition 2.12. *A loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is called* ***(strictly) convex*** *if $L(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ is (strictly) convex for all $x \in X$ and $y \in Y$.*

If L is a supervised or unsupervised loss function, then we call L (strictly) convex if its canonically associated loss function $\bar{L}$ is (strictly) convex. In the

following, we will analogously assign other properties to L via its identification with $\bar{L}$.

The next simple lemma, whose proof is left as an exercise, shows that convexity of the loss implies convexity of its risks.

Lemma 2.13 (Convexity of risks). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a (strictly) convex loss and P be a distribution on $X \times Y$. Then $\mathcal{R}_{L,\mathrm{P}} : \mathcal{L}_0(X) \to [0,\infty]$ is (strictly) convex.*

Besides convexity we also need some notions of continuity for loss functions. We begin with a qualitative definition.

Definition 2.14. *A loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is called **continuous** if $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is continuous for all $x \in X$, $y \in Y$.*

If we have a continuous loss function $L : X \times Y \times \mathbb{R} \to [0,\infty)$ and a sequence (f_n) of measurable functions $f_n : X \to \mathbb{R}$ that converges pointwise to a function $f : X \to \mathbb{R}$, then we obviously have $L(x,y,f_n(x)) \to L(x,y,f(x))$ for all $(x,y) \in X \times Y$. However, it is well-known from integration theory that such a convergence does *not* imply a convergence of the corresponding integrals, i.e., in general we cannot conclude $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}_{L,\mathrm{P}}(f)$. However, the following, weaker result always holds.

Lemma 2.15 (Lower semi-continuity of risks). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous loss, P be a distribution on $X \times Y$, and $(f_n) \subset \mathcal{L}_0(\mathrm{P}_X)$ be a sequence that converges to an $f \in \mathcal{L}_0(\mathrm{P}_X)$ in probability with respect to the marginal distribution P_X. Then we have*

$$\mathcal{R}_{L,\mathrm{P}}(f) \le \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)\,.$$

Proof. Since (f_n) converges in probability P_X, there exists a subsequence (f_{n_k}) of (f_n) with

$$\lim_{k\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_{n_k}) = \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)$$

and $f_{n_k}(x) \to f(x)$ for P_X-almost all $x \in X$. By the continuity of L, we then have $L(x,y,f_{n_k}(x)) \to L(x,y,f(x))$ for P-almost all $(x,y) \in X \times Y$, and hence Fatou's lemma (see Theorem A.3.4) gives

$$\begin{aligned}
\mathcal{R}_{L,\mathrm{P}}(f) &= \int_{X\times Y} \lim_{k\to\infty} L\bigl(x,y,f_{n_k}(x)\bigr)\, d\mathrm{P}(x,y) \\
&\le \liminf_{k\to\infty} \int_{X\times Y} L\bigl(x,y,f_{n_k}(x)\bigr)\, d\mathrm{P}(x,y) \\
&= \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)\,.
\end{aligned}$$

□

If we have an integrable majorant of the sequence $L(\cdot,\cdot,f_n(\cdot))$ in the proof above, Lebesgue's Theorem A.3.6 obviously gives $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}_{L,\mathrm{P}}(f)$. The following definition describes losses for which we have such a majorant.

Definition 2.16. *We call a loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ a* ***Nemitski loss*** *if there exist a measurable function $b : X \times Y \to [0,\infty)$ and an increasing function $h : [0,\infty) \to [0,\infty)$ such that*

$$L(x,y,t) \le b(x,y) + h(|t|)\,, \qquad (x,y,t) \in X \times Y \times \mathbb{R}. \tag{2.9}$$

Furthermore, we say that L is a ***Nemitski loss of order*** *$p \in (0,\infty)$ if there exists a constant $c > 0$ such that*

$$L(x,y,t) \le b(x,y) + c\,|t|^p\,, \qquad (x,y,t) \in X \times Y \times \mathbb{R}.$$

Finally, if P is a distribution on $X \times Y$ with $b \in \mathcal{L}_1(\mathrm{P})$, we say that L is a ***P-integrable Nemitski loss.***

Note that P-integrable Nemitski losses L satisfy $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for all $f \in L_\infty(\mathrm{P}_X)$, and consequently we also have $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ and $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. In addition, we should keep in mind that the notion of Nemitski losses will become of particular interest when dealing with *unbounded* Y, which is typical for the regression problems treated in Chapters 9 and 10.

Let us now investigate the continuity of risks based on Nemitski losses.

Lemma 2.17 (Continuity of risks). *Let P be a distribution on $X \times Y$ and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous, P-integrable Nemitski loss. Then the following statements hold:*

i) Let $f_n : X \to \mathbb{R}$, $n \ge 1$, be bounded, measurable functions for which there exists a constant $B > 0$ with $\|f_n\|_\infty \le B$ for all $n \ge 1$. If the sequence (f_n) converges P_X-almost surely to a function $f : X \to \mathbb{R}$, then we have

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) = \mathcal{R}_{L,\mathrm{P}}(f)\,.$$

ii) The map $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to [0,\infty)$ is well-defined and continuous.

iii) If L is of order $p \in [1,\infty)$, then $\mathcal{R}_{L,\mathrm{P}} : L_p(\mathrm{P}_X) \to [0,\infty)$ is well-defined and continuous.

Proof. i). Obviously, f is a bounded and measurable function with $\|f\|_\infty \le B$. Furthermore, the continuity of L shows

$$\lim_{n\to\infty} \bigl|L(x,y,f_n(x)) - L(x,y,f(x))\bigr| = 0$$

for P-almost all $(x,y) \in X \times Y$. In addition, we have

$$\begin{aligned}\bigl|L(x,y,f_n(x)) - L(x,y,f(x))\bigr| &\le 2b(x,y) + h(|f_n(x)|) + h(|f(x)|)\\ &\le 2b(x,y) + 2h(B)\end{aligned}$$

for all $(x,y) \in X \times Y$ and all $n \ge 1$. Since the function $2b(\,\cdot\,,\,\cdot\,) + 2h(B)$ is P-integrable, Lebesgue's theorem together with

$$\left|\mathcal{R}_{L,\mathrm{P}}(f_n) - \mathcal{R}_{L,\mathrm{P}}(f)\right| \leq \int_{X\times Y} \left|L(x,y,f_n(x)) - L(x,y,f(x))\right| d\mathrm{P}(x,y)$$

gives the assertion.

ii). Condition (2.9) together with $b \in \mathcal{L}_1(\mathrm{P})$ obviously ensures $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for all $f \in L_\infty(\mathrm{P}_X)$, i.e., $\mathcal{R}_{L,\mathrm{P}}$ actually maps $L_\infty(\mathrm{P}_X)$ into $[0,\infty)$. Moreover, the continuity is a direct consequence of *i).*

iii). Since L is a P-integrable Nemitski loss of order p, we obviously have $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for all $f \in L_p(\mathrm{P}_X)$. Now let $(f_n) \subset L_p(\mathrm{P}_X)$ be a convergent sequence with limit $f \in L_p(\mathrm{P}_X)$. Since convergence in $L_p(\mathrm{P}_X)$ implies convergence in probability P_X, Lemma 2.15 then yields

$$\mathcal{R}_{L,\mathrm{P}}(f) \leq \liminf_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n)\,.$$

Moreover, $\tilde{L}(x,y,t) := b(x,y) + c\,|t|^p - L(x,y,t)$ defines a continuous loss function, and hence Lemma 2.15 also gives

$$\begin{aligned}
\|b\|_{L_1(\mathrm{P})} + c\|f\|_p^p - \mathcal{R}_{L,\mathrm{P}}(f) &= \mathcal{R}_{\tilde{L},\mathrm{P}}(f) \\
&\leq \liminf_{n\to\infty} \mathcal{R}_{\tilde{L},\mathrm{P}}(f_n) \\
&= \liminf_{n\to\infty} \big(\|b\|_{L_1(\mathrm{P})} + c\|f_n\|_p^p - \mathcal{R}_{L,\mathrm{P}}(f_n)\big)\,.
\end{aligned}$$

Using that $\|\cdot\|_p^p$ is continuous on $L_p(\mathrm{P}_X)$, we thus obtain

$$\limsup_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) \leq \mathcal{R}_{L,\mathrm{P}}(f)\,. \qquad \square$$

Let us now turn to a quantitative notion of continuity for loss functions.

Definition 2.18. *A loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is called **locally Lipschitz continuous** if for all $a \geq 0$ there exists a constant $c_a \geq 0$ such that*

$$\sup_{\substack{x\in X\\ y\in Y}} \left|L(x,y,t) - L(x,y,t')\right| \;\leq\; c_a\,|t-t'|\,, \qquad t,t' \in [-a,a]\,. \tag{2.10}$$

*Moreover, for $a \geq 0$, the smallest such constant c_a is denoted by $|L|_{a,1}$. Finally, if we have $|L|_1 := \sup_{a\geq 0} |L|_{a,1} < \infty$, we call L **Lipschitz continuous**.*

Note that if Y is *finite* and $L : Y \times \mathbb{R} \to [0,\infty)$ is a *supervised* convex loss, then L is locally Lipschitz continuous since every convex function is locally Lipschitz continuous by Lemma A.6.5. Furthermore, a locally Lipschitz continuous loss L is a Nemitski loss since the definition of $|L|_{|t|,1}$ yields

$$L(x,y,t) \leq L(x,y,0) + |L|_{|t|,1}|t|\,, \qquad (x,y) \in X \times Y,\, t \in \mathbb{R}. \tag{2.11}$$

In particular, a locally Lipschitz continuous loss L is a P-integrable Nemitski loss if and only if $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$. Finally, if L is Lipschitz continuous, then L is a Nemitski loss of order $p = 1$.

The following lemma, whose proof is left as an exercise, relates the (local) Lipschitz continuity of L to the (local) Lipschitz continuity of its risk.

Lemma 2.19 (Lipschitz continuity of risks). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a locally Lipschitz continuous loss and P be a distribution on $X \times Y$. Then for all $B \geq 0$ and all $f, g \in L_\infty(\mathrm{P}_X)$ with $\|f\|_\infty \leq B$ and $\|g\|_\infty \leq B$, we have*

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| \;\leq\; |L|_{B,1} \cdot \|f - g\|_{L_1(\mathrm{P}_X)} .$$

Our next goal is to consider the differentiability of risks. To this end, we first have to introduce differentiable loss functions in the following definition.

Definition 2.20. *A loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is called **differentiable** if $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is differentiable for all $x \in X$, $y \in Y$. In this case, $L'(x,y,t)$ denotes the derivative of $L(x,y,\cdot)$ at $t \in \mathbb{R}$.*

In general, we cannot expect that the risk of a differentiable loss function is differentiable. However, for certain integrable Nemitski losses, we can actually establish the differentiability of the associated risk.

Lemma 2.21 (Differentiability of risks). *Let P be a distribution on $X \times Y$ and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a differentiable loss such that both L and $|L'| : X \times Y \times \mathbb{R} \to [0,\infty)$ are P-integrable Nemitski losses. Then the risk functional $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to [0,\infty)$ is Fréchet differentiable and its derivative at $f \in L_\infty(\mathrm{P}_X)$ is the bounded linear operator $\mathcal{R}'_{L,\mathrm{P}}(f) : L_\infty(\mathrm{P}_X) \to \mathbb{R}$ given by*

$$\mathcal{R}'_{L,\mathrm{P}}(f)g = \int_{X\times Y} g(x) L'\bigl(x,y,f(x)\bigr)\, d\mathrm{P}(x,y)\,, \qquad g \in L_\infty(\mathrm{P}_X).$$

Proof. We first observe that the derivative $L' : X \times Y \times \mathbb{R} \to \mathbb{R}$ is measurable since

$$L'(x,y,t) = \lim_{n\to\infty} \frac{L(x,y,t+1/n) - L(x,y,t)}{1/n}\,, \qquad (x,y,t) \in X \times Y \times \mathbb{R}.$$

Now let $f \in L_\infty(\mathrm{P}_X)$ and $(f_n) \subset L_\infty(\mathrm{P}_X)$ be a sequence with $f_n \neq 0$, $n \geq 1$, and $\lim_{n\to\infty} \|f_n\|_\infty = 0$. Without loss of generality, we additionally assume for later use that $\|f_n\|_\infty \leq 1$ for all $n \geq 1$. For $(x,y) \in X \times Y$ and $n \geq 1$, we now define

$$G_n(x,y) := \left| \frac{L\bigl(x,y,f(x)+f_n(x)\bigr) - L\bigl(x,y,f(x)\bigr)}{f_n(x)} - L'\bigl(x,y,f(x)\bigr) \right|$$

if $f_n(x) \neq 0$, and $G_n(x,y) := 0$ otherwise. Now an easy estimation gives

$$\begin{aligned}
&\left| \frac{\mathcal{R}_{L,\mathrm{P}}(f+f_n) - \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}'_{L,\mathrm{P}}(f) f_n}{\|f_n\|_\infty} \right| \\
&\leq \int_{X\times Y} \left| \frac{L\bigl(x,y,f(x)+f_n(x)\bigr) - L\bigl(x,y,f(x)\bigr) - f_n(x) L'\bigl(x,y,f(x)\bigr)}{\|f_n\|_\infty} \right| d\mathrm{P}(x,y) \\
&\leq \int_{X\times Y} G_n(x,y)\, d\mathrm{P}(x,y) \qquad (2.12)
\end{aligned}$$

for all $n \geq 1$. Furthermore, for $(x, y) \in X \times Y$, the definitions of G_n and $L'(x, y, \cdot)$ obviously yield

$$\lim_{n\to\infty} G_n(x, y) = 0. \tag{2.13}$$

Moreover, for $(x, y) \in X \times Y$ and $n \geq 1$ with $f_n(x) \neq 0$, the mean value theorem shows that there exists a $g_n(x, y)$ with $|g_n(x, y)| \in [0, |f_n(x)|]$ and

$$\frac{L\big(x, y, f(x) + f_n(x)\big) - L\big(x, y, f(x)\big)}{f_n(x)} = L'\big(x, y, f(x) + g_n(x, y)\big)\,.$$

Since $|L'|$ is a P-integrable Nemitski loss, there also exist a $b : X \times Y \to [0, \infty)$ with $b \in L_1(\mathrm{P})$ and an increasing function $h : [0, \infty) \to [0, \infty)$ with

$$|L'(x, y, t)| \leq b(x, y) + h(t)\,, \qquad (x, y, t) \in X \times Y \times \mathbb{R}.$$

This together with $\|f_n\|_\infty \leq 1$, $n \geq 1$, implies

$$\begin{aligned}\left|\frac{L\big(x, y, f(x) + f_n(x)\big) - L\big(x, y, f(x)\big)}{f_n(x)}\right| &\leq b(x, y) + h\big(|f(x) + g_n(x, y)|\big)\\ &\leq b(x, y) + h\big(\|f\|_\infty + 1\big)\end{aligned}$$

for all $(x, y) \in X \times Y$ and $n \geq 1$ with $f_n(x) \neq 0$. Since these estimates show that

$$G_n(x, y) \leq 2b(x, y) + 2h\big(\|f\|_\infty + 1\big)$$

for all $(x, y) \in X \times Y$ and $n \geq 1$, we then obtain the assertion by (2.12), (2.13), and Lebesgue's Theorem A.3.6. □

Our last goal in this section is to investigate loss functions that in some sense can be restricted to domains of the form $X \times Y \times [-M, M]$.

Definition 2.22. *We say that a loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ can be **clipped** at $M > 0$ if, for all $(x, y, t) \in X \times Y \times \mathbb{R}$, we have*

$$L(x, y, \widehat{t}\,) \;\leq\; L(x, y, t)\,,$$

*where $\widehat{t}$ denotes the **clipped** value of t at $\pm M$, that is*

$$\widehat{t} := \begin{cases} -M & \text{if } t < -M \\ t & \text{if } t \in [-M, M] \\ M & \text{if } t > M\,. \end{cases} \tag{2.14}$$

Moreover, we say that L can be clipped if it can be clipped at some $M > 0$.

For most losses, it is elementary to check whether they can be clipped, but for convex losses this work can be further simplified by the following elementary criterion.

Lemma 2.23 (Clipped convex losses). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss and $M > 0$. Then the following statements are equivalent:*

i) L can be clipped at M.
ii) For all $(x,y) \in X \times Y$, the function $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ has at least one global minimizer in $[-M,M]$.

Proof. For $(x,y) \in X \times Y$, we denote the set of minimizers of $L(x,y,\cdot)$ by $\mathcal{M}_{x,y} := \{t^* \in \mathbb{R} : L(x,y,t^*) = \inf_{t\in\mathbb{R}} L(x,y,t)\}$. For later use, note that the convexity of L implies that $\mathcal{M}_{x,y}$ is a closed interval by Lemma A.6.2.

i) $\Rightarrow$ *ii)*. Assume that there exists a pair $(x,y) \in X \times Y$ such that $\mathcal{M}_{x,y} \cap [-M,M] = \emptyset$. In the case $\mathcal{M}_{x,y} = \emptyset$, the convexity of L shows that $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is strictly monotone and hence L cannot be clipped at any real number. Therefore we may assume without loss of generality that $t := \inf \mathcal{M}_{x,y}$ satisfies $M < t < \infty$. However, in this case we have

$$L(x,y,\widehat{t}\,) = L(x,y,M) > L(x,y,t)\,,$$

i.e., L cannot be clipped at M.

ii) $\Rightarrow$ *i)*. Our assumption *ii)* guarantees $\mathcal{M}_{x,y} \cap [-M,M] \neq \emptyset$, and hence we have $\inf \mathcal{M}_{x,y} \leq M$ and $\sup \mathcal{M}_{x,y} \geq -M$. Moreover, the convexity of L shows that $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is increasing on $[\sup \mathcal{M}_{x,y}, \infty)$ and decreasing on $(-\infty, \inf \mathcal{M}_{x,y}]$, and hence L can be clipped at M. □

The criterion above will be of particular interest in Section 7.4, where we investigate the statistical properties of SVMs that use clippable losses. Therefore, it will be important to remember that, for the loss functions introduced in the following sections, condition *ii)* is usually elementary to check.

2.3 Margin-Based Losses for Classification Problems

In Examples 2.4 and 2.5, we considered the (weighted) binary classification scenario, which is described by the supervised loss functions L_{class} and $L_{\alpha\text{-class}}$, respectively. Now observe that both loss functions are not *convex*, which may lead to computational problems if, for example, one tries to minimize an empirical risk $\mathcal{R}_{L_{\text{class}},D}(\cdot)$ over some set $\mathcal{F}$ of functions $f : X \to \mathbb{R}$. This is the reason why many machine learning algorithms consider the empirical risk $\mathcal{R}_{L,D}(\cdot)$ of a *surrogate* loss function $L : Y \times \mathbb{R} \to [0,\infty)$ instead. In this section, we will introduce some commonly used surrogate losses and establish a few basic properties of these losses. Finally, we show why the hinge loss used in SVMs for classification is a good surrogate.

Throughout this section, we assume $Y := \{-1,1\}$. Let us begin with the following basic definition, which introduces a type of loss function often used in classification algorithms.

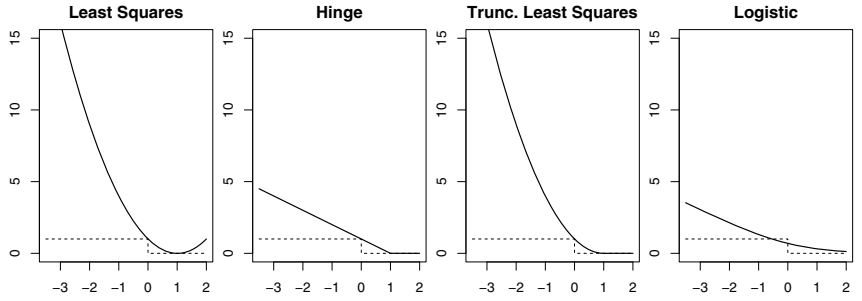


Fig. 2.1. The shape of the representing function φ for some margin-based loss functions considered in the text.

Definition 2.24. *A supervised loss $L : Y \times \mathbb{R} \to [0, \infty)$ is called* ***margin-based*** *if there exists a* ***representing function*** *$\varphi : \mathbb{R} \to [0, \infty)$ such that*

$$L(y, t) = \varphi(yt)\,, \qquad y \in Y,\, t \in \mathbb{R}.$$

The following lemma relates some simple properties of margin-based losses to analogous properties of their representing functions.

Lemma 2.25 (Properties of margin-based losses). *Let L be a margin-based loss represented by φ. Then the following statements are true:*

i) L is (strictly) convex if and only if φ is (strictly) convex.
ii) L is continuous if and only if φ is.
iii) L is (locally) Lipschitz continuous if and only if φ is.
iv) If L is convex, then it is locally Lipschitz continuous.
v) L is a P-integrable Nemitski loss for all measurable spaces X and all distributions P on $X \times Y$.

Proof. Recalling the definitions of Section 2.2, the first three assertions are trivial and *iv*) follows from Lemma A.6.5. Finally, *v)* follows from

$$L(y, t) \le \max\bigl\{\varphi(-t), \varphi(t)\bigr\}\,, \qquad y \in Y,\, t \in \mathbb{R}. \qquad \square$$

Note that the classification loss L_{class} is *not* margin-based, while many commonly used surrogates for L_{class} are margin-based. We are in particular interested in the following examples (see also Figure 2.1 for some illustrations).

Example 2.26. The **least squares loss** L_{LS} is margin-based since it satisfies

$$L_{\text{LS}}(y, t) = (y - t)^2 = (1 - yt)^2\,, \qquad y = \pm 1,\, t \in \mathbb{R}.$$

In addition, L_{LS} is obviously strictly convex, and for $a > 0$ its local Lipschitz constant is $|L_{\text{LS}}|_{a,1} = 2a + 2$ by Lemma A.6.8. Finally, L_{LS} can be clipped at $M = 1$. $\triangleleft$

Example 2.27. The **hinge loss** $L_{\text{hinge}} : Y \times \mathbb{R} \to [0, \infty)$ is defined by

$$L_{\text{hinge}}(y, t) := \max\{0, 1 - yt\}\,, \qquad y = \pm 1,\ t \in \mathbb{R}.$$

It is clearly a margin-based loss that linearly penalizes every prediction t with $yt \leq 1$. In addition, it is obviously convex and Lipschitz continuous with $|L_{\text{hinge}}|_1 = 1$. Finally, L_{hinge} can be clipped at $M = 1$. ◁

Example 2.28. The **truncated least squares loss** or **squared hinge loss** is defined by

$$L_{\text{trunc-ls}}(y, t) := \big(\max\{0, 1 - yt\}\big)^2\,, \qquad y = \pm 1,\ t \in \mathbb{R}.$$

It is obviously a margin-based loss that quadratically penalizes every prediction t with $yt \leq 1$. In addition, it is convex, and its local Lipschitz constants are $|L_{\text{trunc-ls}}|_{a,1} = 2a + 2$, $a > 0$. Finally, L_{LS} can be clipped at $M = 1$. ◁

Example 2.29. The **logistic loss for classification** $L_{\text{c-logist}}$ is defined by

$$L_{\text{c-logist}}(y, t) := \ln\big(1 + \exp(-yt)\big)\,, \qquad y = \pm 1,\ t \in \mathbb{R}.$$

It is obviously a margin-based loss function whose shape is close to that of the hinge loss. However, unlike the hinge loss, the logistic loss is infinitely many times differentiable. In addition, it is strictly convex and Lipschitz continuous with $|L_{\text{c-logist}}|_1 = 1$. Finally, $L_{\text{c-logist}}$ cannot be clipped. ◁

Let us finally investigate in which sense the hinge loss used in the soft margin SVM is a reasonable surrogate for the classification loss. To this end, we need the following elementary lemma.

Lemma 2.30. *For all $\eta \in [0, 1]$ and all $t \in [-1, 1]$, we have*

$$|2\eta - 1|\mathbf{1}_{(-\infty,0]}\big((2\eta - 1)\operatorname{sign} t\big) \;\leq\; |2\eta - 1| \cdot \big|t - \operatorname{sign}(2\eta - 1)\big|\,. \tag{2.15}$$

Proof. For $\eta = 1/2$, there is nothing to prove. In order to prove the other cases, let us first recall our convention $\operatorname{sign} 0 := 1$. For $\eta \in [0, 1/2)$ and $t \in [-1, 0)$, we now have $(2\eta - 1)\operatorname{sign} t > 0$, and hence the left-hand side of (2.15) equals zero. From this we immediately obtain the assertion. Moreover, for $t \in [0, 1]$, we have $(2\eta - 1)\operatorname{sign} t < 0$, which in turn yields

$$|2\eta - 1|\mathbf{1}_{(-\infty,0]}\big((2\eta - 1)\operatorname{sign} t\big) \leq |2\eta - 1| \cdot (t + 1) = |2\eta - 1| \cdot \big|t - \operatorname{sign}(2\eta - 1)\big|\,.$$

In other words, we have shown the assertion for $\eta < 1/2$. The case $\eta > 1/2$ can be shown completely analogously and is left as an additional exercise for the reader. □

With the help of the lemma above we can now investigate the relationship between the L_{hinge}-risk and the classification risk.

Theorem 2.31 (Zhang's inequality). *Given a distribution* P *on* $X \times Y$*, we write* $\eta(x) := \mathrm{P}(y = 1|x)$*,* $x \in X$*. Moreover, let* $f^*_{L_{\text{class}},\mathrm{P}}$ *be the Bayes classification function given by* $f^*_{L_{\text{class}},\mathrm{P}}(x) := \operatorname{sign}(2\eta(x) - 1)$*,* $x \in X$*. Then, for all measurable* $f : X \to [-1, 1]$*, we have*

$$\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} = \int_X |f(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \cdot |2\eta(x) - 1| \, d\mathrm{P}_X(x) \,.$$

Moreover, for every measurable $f : X \to \mathbb{R}$*, we have*

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \;\le\; \mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \,.$$

Proof. For $f : X \to [-1, 1]$, the definition of the hinge loss yields

$$\begin{aligned}\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) &= \int_X (1 - f(x))\,\eta(x) + (1 + f(x))(1 - \eta(x))\, d\mathrm{P}_X(x) \\ &= \int_X 1 + f(x)(1 - 2\eta(x))\, d\mathrm{P}_X(x) \,,\end{aligned}$$

which in turn implies $\mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} = \mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f^*_{L_{\text{class}},\mathrm{P}})$ since the hinge loss can be clipped at $M = 1$ by Lemma 2.23. From this we conclude that

$$\begin{aligned}\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} &= \int_X f(x)(1 - 2\eta(x)) + |2\eta(x) - 1| \, d\mathrm{P}_X(x) \\ &= \int_X |f(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \cdot |2\eta(x) - 1| \, d\mathrm{P}_X(x) \,,\end{aligned}$$

i.e., we have shown the first assertion. To prove the second assertion, we first use that L_{hinge} can be clipped at $M = 1$ to obtain

$$\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(\widehat{f}\,) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \;\le\; \mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}}$$

for the clipped version $\widehat{f}$ of a function $f : X \to \mathbb{R}$. Moreover, this clipped version also satisfies

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} = \mathcal{R}_{L_{\text{class}},\mathrm{P}}(\widehat{f}\,) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \,,$$

and consequently it suffices to show the second assertion for $f : X \to [-1, 1]$. Now recall Example 2.4, where we saw $\mathcal{R}^*_{L_{\text{class}},\mathrm{P}} = \mathcal{R}_{L_{\text{class}},\mathrm{P}}(f^*_{L_{\text{class}},\mathrm{P}})$ and

$$\begin{aligned}&\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \\ &= \int_X \eta\, \mathbf{1}_{(-\infty,0)}(f) + (1 - \eta)\, \mathbf{1}_{[0,\infty)}(f) - \min\{\eta, 1 - \eta\}\, d\mathrm{P}_X \\ &= \int_X |2\eta(x) - 1|\, \mathbf{1}_{(-\infty,0]}\big((2\eta(x) - 1) \operatorname{sign} f(x)\big)\, d\mathrm{P}_X(x) \,.\end{aligned}$$

Lemma 2.30 and the first assertion then yield the second assertion. □

Recall that the goal in binary classification was to find a function f whose excess classification risk $\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}}$ is small. By Theorem 2.31, we now see that we achieve this goal whenever $\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}}$ is small. In this sense, the hinge loss is a reasonable surrogate for the classification loss. Finally, note that we will show in Section 3.4 that the other margin-based losses introduced in this section are also reasonable surrogates.

Finally, observe that *all* calculations in the preceding proof are *solely* in terms of $\eta(x) = \mathrm{P}(y = 1|x)$ and $f(x)$. This observation will be the key trick for analyzing general surrogate losses in Chapter 3.

2.4 Distance-Based Losses for Regression Problems

In regression, the problem is to predict a real-valued output y given an input x. The discrepancy between the prediction $f(x)$ and the observation y is often measured by the least squares loss we introduced in Example 2.6. However, we also mentioned there that this is by no means the only reasonable loss. In this section, we therefore introduce some other loss functions for the regression problem. In addition, we establish some basic properties of these losses and their associated risks.

Let us begin with the following basic definitions.

Definition 2.32. *We say that a supervised loss $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is:*

i) ***distance-based*** *if there exists a* ***representing function*** *$\psi : \mathbb{R} \to [0, \infty)$ satisfying $\psi(0) = 0$ and*

$$L(y, t) = \psi(y - t)\,, \qquad y \in Y,\, t \in \mathbb{R};$$

ii) ***symmetric*** *if L is distance-based and its representing function ψ satisfies*

$$\psi(r) = \psi(-r)\,, \qquad r \in \mathbb{R}.$$

Obviously, the least squares loss as well as the family of losses mentioned after Example 2.6 are symmetric loss functions. Further examples of this type of loss will be presented later in this section. Let us first, however, establish some basic properties of distance-based losses and their associated risks. We begin with the following lemma, which relates properties of L with properties of ψ. Its proof is left as an exercise.

Lemma 2.33 (Properties of distance-based losses). *Let L be a distance-based loss with representing function $\psi : \mathbb{R} \to [0, \infty)$. Then we have:*

i) L is (strictly) convex if and only if ψ is (strictly) convex.
ii) L is continuous if and only if ψ is continuous.
iii) L is Lipschitz continuous if and only if ψ is Lipschitz continuous.

Note that the *local* Lipschitz continuity of ψ does *not* imply the local Lipschitz continuity of the corresponding distance-based loss function as, for example, the least squares loss shows.

Our next goal is to investigate under which conditions on the distribution P a distance-based loss function is a P-integrable Nemitski loss. This analysis will be conducted in two steps: *a)* the analysis of the integrals of the form

$$\mathcal{C}_{L,\mathrm{Q}}(t) := \int_{\mathbb{R}} L(y,t)\, d\mathrm{Q}(y)\,, \tag{2.16}$$

which occur for $\mathrm{Q} := \mathrm{P}(Y|x)$ as inner integrals in the definition of the L-risk, and *b)* a subsequent analysis of the averaging with respect to P_X. For the first step, we need the following definition, which will be used to describe the tail behavior of the conditional distributions $\mathrm{P}(Y|x)$.

Definition 2.34. *For a distribution* Q *on* $\mathbb{R}$*, the* p***-th moment****,* $p \in (0,\infty)$*, is defined by*

$$|\mathrm{Q}|_p := \left(\int_{\mathbb{R}} |y|^p \, d\mathrm{Q}(y)\right)^{1/p}.$$

Moreover, its ∞***-moment*** *is defined by* $|\mathrm{Q}|_\infty := \sup\bigl|\operatorname{supp} Q\bigr|$.

Note that in general the p-th moment of a distribution Q on $\mathbb{R}$ is not finite. In particular, we have $|\mathrm{Q}|_\infty < \infty$ if and only if Q has a bounded support. Moreover, for $0 < p \le q \le \infty$, we always have $|\mathrm{Q}|_p \le |\mathrm{Q}|_q$.

Besides controlling the tail behavior of the conditional distributions we also need to describe the growth behavior of the loss function considered. This is done in the following definition.

Definition 2.35. *Let* $p \in (0,\infty)$ *and* $L : \mathbb{R} \times \mathbb{R} \to [0,\infty)$ *be a distance-based loss with representing function* ψ*. We say that* L *is of:*

i) ***upper growth*** p *if there is a constant* $c > 0$ *such that*

$$\psi(r) \le c\left(|r|^p + 1\right), \qquad r \in \mathbb{R};$$

ii) ***lower growth*** p *if there is a constant* $c > 0$ *such that*

$$\psi(r) \ge c\left(|r|^p - 1\right), \qquad r \in \mathbb{R};$$

iii) ***growth type*** p *if* L *is of both upper and lower growth type* p*.*

Our next goal is to relate the tail behavior of the conditional distributions with the growth behavior of L and the integrals (2.16). To this end, recall that convex functions are locally Lipschitz continuous (see Lemma A.6.5), and hence, for *convex* distance-based loss functions L, the representing ψ is locally Lipschitz continuous on every interval $[-r,r]$. Consequently,

$$r \mapsto |\psi_{|[-r,r]}|_1\,, \qquad r \ge 0, \tag{2.17}$$

defines an increasing, non-negative function. The following lemma establishes some basic properties of this function and relates them to the growth type of distance-based loss functions.

Lemma 2.36 (Growth type and moments). *Let L be a distance-based loss with representing function ψ and Q be a distribution on $\mathbb{R}$. For $p \in (0,\infty)$, we then have:*

i) If ψ is convex and $\lim_{|r|\to\infty} \psi(r) = \infty$, then L is of lower growth type 1.
ii) If ψ is Lipschitz continuous, then L is of upper growth type 1.
iii) If ψ is convex, then for all $r > 0$ we have

$$|\psi_{|[-r,r]}|_1 \le \frac{2}{r} \, \|\psi_{|[-2r,2r]}\|_\infty \le 4|\psi_{|[-2r,2r]}|_1 \, .$$

iv) If L is convex and of upper growth type 1, then it is Lipschitz continuous.
v) If L is of upper growth type p, then there exists a constant $c_{L,p} > 0$ independent of Q such that

$$\mathcal{C}_{L,\mathrm{Q}}(t) \;\le\; c_{L,p}\bigl(|\mathrm{Q}|_p^p + |t|^p + 1\bigr) \, , \qquad t \in \mathbb{R}. \tag{2.18}$$

Moreover, L is a Nemitski loss of order p.
vi) If L is of lower growth type p, then there exists a constant $c_{L,p} > 0$ independent of Q such that

$$|\mathrm{Q}|_p^p \;\le\; c_{L,p}\bigl(\mathcal{C}_{L,\mathrm{Q}}(t) + |t|^p + 1\bigr) \, , \qquad t \in \mathbb{R}, \tag{2.19}$$

and

$$|t|^p \;\le\; c_{L,p}\bigl(\mathcal{C}_{L,\mathrm{Q}}(t) + |\mathrm{Q}|_p^p + 1\bigr) \, , \qquad t \in \mathbb{R}. \tag{2.20}$$

vii) If L is of growth type p, then we have $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$ if and only if $|\mathrm{Q}|_p < \infty$.

Proof. iii). Follows immediately from Lemma A.6.5.

iv). Follows from the left inequality of *iii)* and Lemma 2.33.

ii). Follows from $|\psi(s)| = |\psi(s) - \psi(0)| \le |\psi|_1 \, |s|$ for all $s \in \mathbb{R}$.

i). The assumption $\lim_{|r|\to\infty} \psi(r) = \infty$ implies that $|\psi_{|[-r,r]}|_1 > 0$ for all sufficiently large $r > 0$. Moreover, it shows that ψ is decreasing on $(-\infty, 0]$ and increasing on $[0,\infty)$. Consequently, we have $\|\psi_{|[-r,0]}\|_\infty = \psi(r)$ for $r \le 0$, and $\|\psi_{|[0,r]}\|_\infty = \psi(r)$ for $r \ge 0$. Now, the assertion follows from applying the first part of Lemma A.6.5 to the convex functions $\mathbf{1}_{(-\infty,0]}\psi$ and $\mathbf{1}_{[0,\infty)}\psi$.

v). Writing $c_p := \max\{1, 2^{p-1}\}$, the second assertion follows from

$$L(y,t) = \psi(y-t) \le c\,\bigl(c_p|y|^p + c_p|t|^p + 1\bigr) \, , \qquad y,t \in \mathbb{R}. \tag{2.21}$$

Using this inequality, we then immediately obtain

$$\mathcal{C}_{L,\mathrm{Q}}(t) = \int_{\mathbb{R}} \psi(y-t)\, d\mathrm{Q}(y) \le c\, c_p\bigl(|\mathrm{Q}|_p^p + |t|^p\bigr) + c \, .$$

vi). We fix a $t \in \mathbb{R}$ and write $c_p := \max\{1, 2^{p-1}\}$. Since without loss of generality we may assume $\mathcal{C}_{L,\mathrm{Q}}(t) < \infty$, we can estimate

$$\begin{aligned}|\mathrm{Q}|_p^p &= \int_{\mathbb{R}} |y|^p \, d\mathrm{Q}(y) \le c_p \int_{\mathbb{R}} |y-t|^p + |t|^p \, d\mathrm{Q}(y) \\ &= \frac{c_p}{c} \int_{\mathbb{R}} c\left(|y-t|^p - 1\right) d\mathrm{Q}(y) + c_p + c_p|t|^p \\ &\le \frac{c_p}{c}\, \mathcal{C}_{L,\mathrm{Q}}(t) + c_p + c_p|t|^p \,.\end{aligned}$$

Now we easily find (2.19). Moreover, (2.20) can be shown analogously.

vii). The assertion immediately follows from *v)* and *vi).* □

So far we have analyzed the interplay between the growth behavior of L and the tail behavior of the conditional distributions $\mathrm{P}(\cdot\,|x)$. Our next step is to investigate the effect of the integration with respect to P_X. To this end, we need the following definition.

Definition 2.37. *For a distribution* P *on* $X \times \mathbb{R}$*, the* ***average*** p***-th moment****,* $p \in (0,\infty)$*, is defined by*

$$|\mathrm{P}|_p := \left(\int_X \int_{\mathbb{R}} |y|^p \, d\mathrm{P}(x,y)\right)^{1/p} = \left(\int_X \left|\mathrm{P}(\cdot\,|x)\right|_p^p d\mathrm{P}_X(x)\right)^{1/p}.$$

Moreover, its ***average*** 0***-moment*** *is defined by* $|\mathrm{P}|_0 := 1$ *and its* ***average*** ∞***-moment*** *is defined by* $|\mathrm{P}|_\infty := \operatorname{ess-sup}_{x\in X} \left|\mathrm{P}(\cdot\,|x)\right|_\infty$.

Again, the p-th moment of a distribution P on $X \times \mathbb{R}$ is not necessarily finite. In particular, it is easy to see that $|\mathrm{P}|_\infty < \infty$ if and only if there is an $M > 0$ such that $\operatorname{supp} \mathrm{P}(\cdot\,|x) \subset [-M, M]$ for P_X-almost all $x \in X$. Finally, for $0 < p \le q \le \infty$ we again have $|\mathrm{P}|_p \le |\mathrm{P}|_q$.

Let us now investigate how average moments and risks interplay.

Lemma 2.38 (Average moments and risks). *Let* L *be a distance-based loss and* P *be a distribution on* $X \times Y$*. For* $p > 0$*, we then have:*

i) If L *is of upper growth type* p*, there exists a constant* $c_{L,p} > 0$ *independent of* P *such that, for all measurable* $f : X \to \mathbb{R}$*, we have*

$$\mathcal{R}_{L,\mathrm{P}}(f) \;\le\; c_{L,p}\big(|\mathrm{P}|_p^p + \|f\|_{L_p(\mathrm{P}_X)}^p + 1\big)\,. \tag{2.22}$$

Moreover, if $|\mathrm{P}|_p < \infty$*, then* L *is a* P*-integrable Nemitski loss of order* p*, and* $\mathcal{R}_{L,\mathrm{P}}(\cdot) : L_p(\mathrm{P}_X) \to [0,\infty)$ *is well-defined and continuous.*

ii) If L *is convex and of upper growth type* p *with* $p \ge 1$*, then for all* $q \in [p-1,\infty]$ *with* $q > 0$ *there exists a constant* $c_{L,p,q} > 0$ *independent of* P *such that, for all measurable* $f : X \to \mathbb{R}$ *and* $g : X \to \mathbb{R}$*, we have*

$$\begin{aligned}&\left|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\right| \\ &\le c_{L,p,q}\Big(|\mathrm{P}|_q^{p-1} + \|f\|_{L_q(\mathrm{P}_X)}^{p-1} + \|g\|_{L_q(\mathrm{P}_X)}^{p-1} + 1\Big)\|f-g\|_{L_{\frac{q}{q-p+1}}(\mathrm{P}_X)}.\end{aligned} \tag{2.23}$$

iii) If L is of lower growth type p, there exists a constant $c_{L,p} > 0$ independent of P such that, for all measurable $f : X \to \mathbb{R}$, we have

$$|\mathrm{P}|_p^p \;\le\; c_{L,p}\big(\mathcal{R}_{L,\mathrm{P}}(f) + \|f\|_{L_p(\mathrm{P}_X)}^p + 1\big) \tag{2.24}$$

and

$$\|f\|_{L_p(\mathrm{P}_X)}^p \;\le\; c_{L,p}\big(\mathcal{R}_{L,\mathrm{P}}(f) + |\mathrm{P}|_p^p + 1\big)\,. \tag{2.25}$$

Proof. i). Inequality (2.22) follows from integrating (2.18). The second assertion follows from Inequality (2.21) and the last assertion is a consequence of Lemma 2.17.

ii). We define $r(x,y) := |y| + |f(x)| + |g(x)| + 1$, $(x,y) \in X \times Y$. By Lemma 2.36, we then obtain

$$\begin{aligned}
\big|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\big| &\le \int_{X\times Y} \big|\psi\big(y - f(x)\big) - \psi\big(y - g(x)\big)\big| d\mathrm{P}(x,y) \\
&\le \int_{X\times Y} \big|\psi_{|[-r(x,y),r(x,y)]}\big|_1 \, \big|f(x) - g(x)\big| d\mathrm{P}(x,y) \\
&\le 2\int_{X\times Y} \frac{\|\psi_{|[-2r(x,y),2r(x,y)]}\|_\infty}{r(x,y)} \, \big|f(x) - g(x)\big| d\mathrm{P}(x,y) \\
&\le c\int_{X\times Y} \frac{|2r(x,y)|^p + 1}{2r(x,y)} \, \big|f(x) - g(x)\big| d\mathrm{P}(x,y)
\end{aligned}$$

for a suitable constant $c > 0$ only depending on L. Using $\frac{t^p+1}{t} \le 2t^{p-1}$ for all $t \ge 1$ and Hölder's inequality, we then conclude

$$\begin{aligned}
\big|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\big| &\le 2^p c \int_{X\times Y} |r(x,y)|^{p-1} \, \big|f(x) - g(x)\big| d\mathrm{P}(x,y) \\
&\le 2^p c \left(\int_{X\times Y} |r|^{(p-1)s} d\mathrm{P}\right)^{1/s} \left(\int_{X\times Y} \big|f - g\big|^{s'} d\mathrm{P}\right)^{1/s'},
\end{aligned}$$

where $s := \frac{q}{p-1}$ and $\frac{1}{s'} := 1 - \frac{1}{s} = 1 - \frac{p-1}{q} = \frac{q-p+1}{q}$. Using the definition of r, we further find

$$\begin{aligned}
\left(\int_{X\times Y} |r|^{(p-1)s} d\mathrm{P}\right)^{1/s} &= \left(\int_{X\times Y} \big(|y| + |f(x)| + |g(x)| + 1\big)^q d\mathrm{P}(x,y)\right)^{(p-1)/q} \\
&\le c_q\Big(|\mathrm{P}|_q + \|f\|_{L_q(\mathrm{P}_X)} + \|g\|_{L_q(\mathrm{P}_X)} + 1\Big)^{p-1}
\end{aligned}$$

for a suitable constant $c_q > 0$. By combining the estimates, we then obtain the assertion.

iii). The inequalities (2.24) and (2.25) follow from integrating (2.19) and (2.20), respectively. □

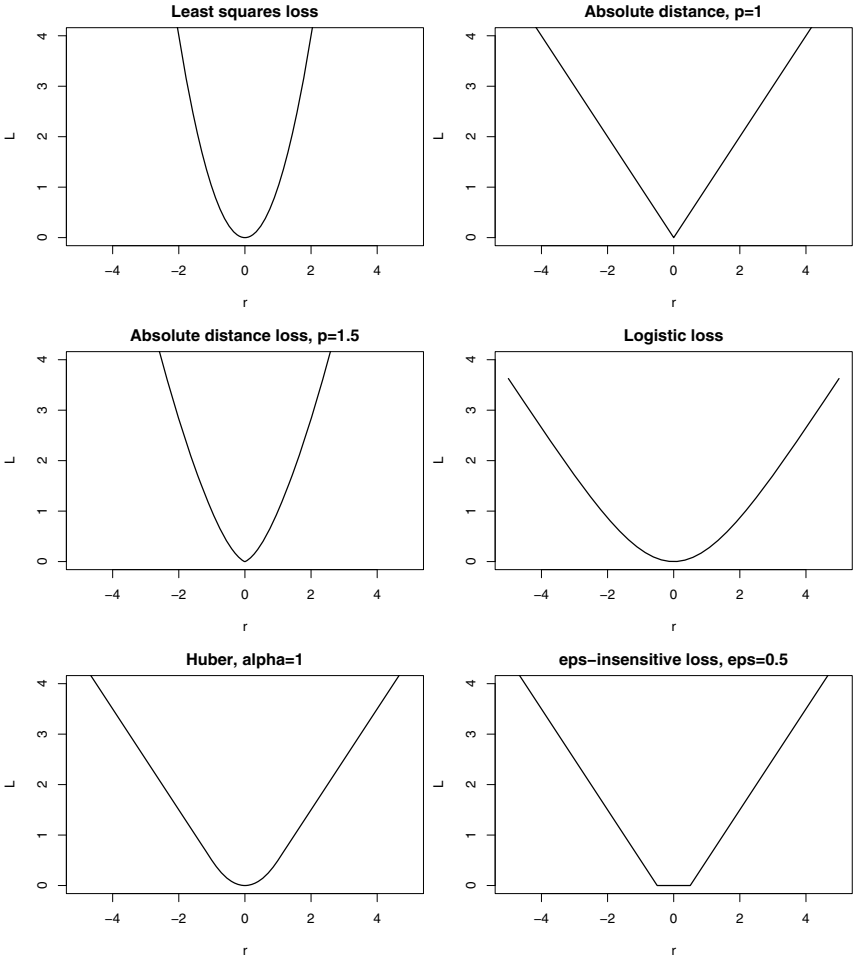


Fig. 2.2. The shape of the representing function ψ for some distance-based loss functions considered in the text.

If L is a distance-based loss function of growth type p and P is a distribution on $X \times \mathbb{R}$ with $|\mathrm{P}|_p = \infty$, the preceding lemma shows $\mathcal{R}_{L,\mathrm{P}}(f) = \infty$ for all $f \in L_p(\mathrm{P}_X)$. This suggests that we may even have $\mathcal{R}^*_{L,\mathrm{P}} = \infty$. However, in general, this is *not* the case, as Exercise 2.6 shows.

Let us finally consider some examples of distance-based loss functions (see also Figure 2.2 for some illustrations) together with some of their basic properties. We will see later in Section 3.7 that the first three losses can be used to estimate the conditional mean whenever $\mathrm{P}(\,\cdot\,|x)$ is symmetric.

Example 2.39. For $p > 0$, the **p-th power absolute distance loss** $L_{p\text{-dist}}$ is the distance-based loss function represented by

$$\psi(r) := |r|^p\,, \qquad r \in \mathbb{R}.$$

Note that for $p = 2$ this definition recovers the least squares loss. Moreover, for $p = 1$, we call $L_{p\text{-dist}}$ simply the **absolute distance loss**. It is not hard to see that $L_{p\text{-dist}}$ is of growth type p and that $L_{p\text{-dist}}$ is convex if and only if $p \geq 1$. Furthermore, $L_{p\text{-dist}}$ is strictly convex if and only if $p > 1$, and it is Lipschitz continuous if and only if $p = 1$. ◁

Example 2.40. The distance-based **logistic loss for regression** $L_{\text{r-logist}}$ is represented by

$$\psi(r) := -\ln \frac{4e^r}{(1+e^r)^2}\,, \qquad r \in \mathbb{R}.$$

Some simple calculations show that $L_{\text{r-logist}}$ is strictly convex and Lipschitz continuous, and consequently $L_{\text{r-logist}}$ is of growth type 1. ◁

Example 2.41. For $\alpha > 0$, **Huber's loss** $L_{\alpha\text{-Huber}}$ is the distance-based loss represented by

$$\psi(r) := \begin{cases} \frac{r^2}{2} & \text{if } |r| \leq \alpha \\ \alpha|r| - \frac{\alpha^2}{2} & \text{otherwise.} \end{cases}$$

Note that, for small r, Huber's loss has the shape of the least squares loss, whereas for large r it has the shape of the absolute distance loss. Consequently, $L_{\alpha\text{-Huber}}$ is convex but not strictly convex. Furthermore, it is Lipschitz continuous, and thus $L_{\alpha\text{-Huber}}$ is of growth type 1. Finally, note that the derivative of ψ equals the clipping operation (2.14) for $M = \alpha$. ◁

Example 2.42. For $\epsilon > 0$, the distance-based **ϵ-insensitive loss** $L_{\epsilon\text{-insens}}$ is represented by

$$\psi(r) := \max\bigl\{0, |r| - \epsilon\bigr\}\,, \qquad r \in \mathbb{R}.$$

The ϵ-insensitive loss ignores deviances smaller than ϵ, whereas it linearly penalizes larger deviances. It is easy to see that $L_{\epsilon\text{-insens}}$ is Lipschitz continuous and convex but not strictly convex. Therefore it is of growth type 1. We will see in Section 9.5 that this loss function can be used to estimate the conditional median, i.e., the median of $\mathrm{P}(\,\cdot\,|x)$, $x \in X$, whenever these conditional distributions are symmetric and have, for example, a Lebesgue density that is bounded away from zero on the support of $\mathrm{P}(\,\cdot\,|x)$. ◁

Example 2.43. For $\tau \in (0,1)$, the distance-based **pinball loss** $L_{\tau\text{-pin}}$ is represented by

$$\psi(r) = \begin{cases} -(1-\tau)r, & \text{if } r < 0 \\ \tau r, & \text{if } r \geq 0. \end{cases} \tag{2.26}$$

Obviously, this loss function is convex and Lipschitz continuous, but for $\tau \neq 1/2$ it is *not* symmetric. We will see in Sections 3.9 and 9.3 that this loss function can be used to estimate conditional τ-quantiles defined by

$$f^*_{\tau,\mathrm{P}}(x) := \left\{ t^* \in \mathbb{R} : \mathrm{P}\big((-\infty, t^*] \,|x\big) \geq \tau \text{ and } \mathrm{P}\big([t^*, \infty) \,|x\big) \geq 1 - \tau \right\}. \qquad \lhd$$

2.5 Further Reading and Advanced Topics

Loss functions and their associated risks have a long history in mathematical statistics and machine learning. For example, the least squares loss for regression was already used by Legendre, Gauss, and Adrain in the early 19th century (see, e.g., Harter, 1983; Stigler, 1981; and the references therein), and the classification loss function dates back to the beginning of machine learning.

In the statistical literature, density level detection has been studied by Hartigan (1987), Müller and Sawitzki (1991), Polonik (1995), Sawitzki (1996), and Tsybakov (1997), among others. Most of these authors focus on the so-called *excess mass approach*. Steinwart *et al.* (2005) showed that this approach is equivalent to an empirical risk minimization approach using a particular classification problem, and based on this observation the authors derived an SVM for the density level detection problem (see also Sections 3.8 and 8.6). Moreover, the risk based on the density level detection loss defined in (2.6) was proposed by Polonik (1995) and later also used by, e.g., Tsybakov (1997) and Ben-David and Lindenbaum (1997). Various applications of the DLD problem, such as cluster analysis, testing for multimodality, and spectral analysis, are described by Hartigan (1975), Müller and Sawitzki (1991), and Polonik (1995). Finally, using DLD for anomaly detection is a widely known technique; see Davies and Gather (1993) and Ripley (1996), for example.

It is well-known that empirical risk minimization for the classification loss typically leads to combinatorial optimization problems that in many cases are NP-hard to solve (see, e.g., Höffgen *et al.*, 1995). Using a margin-based loss as a surrogate for the classification loss is a well-known trick in machine learning to make the training process algorithmically more tractable (see, e.g., the motivation for the hinge loss by Cortes and Vapnik, 1995). In particular, for SVMs, the first surrogates for the classification loss were the hinge loss and its squared variant, the truncated least squares loss. Later, other loss functions, such as the least squares loss and the logistic loss, were introduced into the support vector machine literature by Suykens and Vandewalle (1999), see also Poggio and Girosi (1990), Wahba (1990), and Girosi *et al.* (1995) for earlier work in this direction, and Wahba (1999), respectively. Other margin-based loss functions used in the literature include the exponential loss $\varphi(t) := \exp(-t)$, $t \in \mathbb{R}$, used in the AdaBoost algorithm (see Freund and Schapire, 1996; Breiman, 1999b) and the loss $\varphi(t) := (1-t)^5$

used in the ARC-X4 procedure of Breiman (1998). Some further margin-based losses used in boosting algorithms are listed by Mason *et al.* (2000). Finally, Zhang's inequality was shown by Zhang (2004b).

The importance of Nemitski losses for conditional distributions with unbounded support was first discovered by De Vito *et al.* (2004), and the growth type of distance-based losses was introduced by Christmann and Steinwart (2007).

Huber's loss was proposed by Huber (1964) in the context of robust statistics, and the logistic loss function was already used in the Princeton study by Andrews *et al.* (1972). Moreover, the pinball loss was utilized by Koenker and Bassett (1978) in the context of quantile regression. Last but not least, for a comparison between the absolute distance loss and the least squares loss regarding computational speed for certain algorithms, we refer to Portnoy and Koenker (1997).

2.6 Summary

In this chapter, we introduced loss functions and their associated risks. We saw in Section 2.1 that loss functions can be used to formalize many learning goals, including classification, regression, and density level detection problems. We then investigated simple yet important properties of loss functions. Among them, the notion of integrable Nemitski losses will be a central tool in the following chapters.

Since the classification loss typically leads to computationally hard optimization problems, we presented margin-based surrogates in Section 2.3. For one of these surrogates, namely the hinge loss, we explicitly showed in Zhang's inequality how its excess risk relates to the excess classification risk. In Chapter 3, we will see that a similar relation holds for the other margin-based losses we presented.

Finally, we investigated distance-based loss functions for regression problems in Section 2.4. There, we first showed how the growth behavior of the loss function L together with the average conditional tail behavior of the distribution P determines whether L is a P-integrable Nemitski loss. These considerations will play a crucial role in Chapter 9, where we investigate the learning capabilities of SVMs in regression problems. At the end of Section 2.4, we presented some examples of distance-based losses, including the least squares loss, the pinball loss, the logistic loss, Huber's loss, and the ϵ-insensitive loss. In Chapter 3, we will investigate their relationships to each other.

2.7 Exercises

2.1. Convex and Lipschitz continuous risks ($\star$)
Prove Lemma 2.13 and Lemma 2.19.

2.2. Properties of some margin-based losses ($\star$)
Verify the assertions made in the examples of Section 2.3. Moreover, investigate the properties of the *exponential loss* represented by $\varphi(t) := \exp(-t)$, $t \in \mathbb{R}$, and the *sigmoid loss* represented by $\varphi(t) := 1 - \tanh(t)$, $t \in \mathbb{R}$.

2.3. A surrogate inequality for the logistic loss ($\star\star\star\star$)
Try to find an inequality between the excess classification risk and the excess $L_{\text{c-logist}}$-risk. Compare your findings with the inequality we will obtain in Section 3.4.

2.4. Properties of some distance-based losses ($\star$)
Verify the assertions made in the examples of Section 2.4.

2.5. Clippable convex distance-based losses ($\star\star$)
Let L be a distance-based loss function whose representing function ψ satisfies $\lim_{r\to\pm\infty} \psi(r) = \infty$. Show that L can be clipped at some $M > 0$ if and only if Y is bounded.

2.6. Infinite Bayes risk for regression ($\star\star$)
Let L be a distance-based loss of growth type p and $X := Y := \mathbb{R}$. Find a distribution P on $X \times Y$ such that $|\mathrm{P}|_p = \infty$ and $\mathcal{R}_{L,\mathrm{P}}(f) = \infty$ for all $f \in L_p(\mathrm{P}_X)$ but $\mathcal{R}^*_{L,\mathrm{P}} < \infty$.

Hint: Use a measurable function $g : X \to \mathbb{R}$ with $g \notin L_p(\mathrm{P}_X)$.

3

Surrogate Loss Functions (*)

Overview. *In many cases, the loss describing a learning problem is not suitable when designing a learning algorithm. A common approach to resolve this issue is to use a surrogate loss in the algorithm design. For example, we saw in the introduction that SVMs use the convex hinge loss instead of the discontinuous classification loss. The goal of this chapter is to systematically develop a theory that makes it possible to identify suitable surrogate losses for general learning problems.*

Prerequisites. *Besides Chapter 2 and Section A.3.3, only basic mathematical knowledge is required.*

Usage. *Sections 3.1 – 3.3 and 3.6 provide the theoretical framework required for Sections 3.4, 3.5, and 3.7 – 3.9, which deal with surrogate losses for common learning scenarios. These examples are important but not essential for classification, regression, and robustness, discussed in Chapters 8, 9, and 10, respectively. On the other hand, most of the material in this chapter is of general interest for machine learning and hence relatively independent of the rest of this book.*

In Chapter 2, we introduced some important learning scenarios and their corresponding loss functions. One way to design learning algorithms for these learning scenarios is to use a straightforward empirical risk minimization (ERM) ansatz based on the corresponding loss function. However, this approach may often be flawed, as the following examples illustrate:

- ERM optimization problems based on the classification loss are usually combinatorial problems, and even solving these problems approximately is often NP-hard.
- The least squares loss is known to be rather sensitive to outliers, and hence for certain data sets a (regularized) ERM approach based on this loss may fail, as we will see in Chapter 10.
- For some unsupervised learning scenarios, including the DLD scenario, we do not know the associated loss function since it depends on the unknown density.

These examples demonstrate that in many cases the loss function describing the learning problem is not suitable for a (regularized) ERM ansatz. Now recall that in the SVM approach discussed in the introduction one of the main ideas was to use the hinge loss function as a *surrogate* for the classification loss, and

consequently it is tempting to try surrogate losses in other learning scenarios, too. However, it is not hard to imagine that, given a *target loss*, not every loss function is a good surrogate, and hence we need some guidance for choosing a suitable surrogate loss.

Therefore, let us now describe what properties we do expect from good surrogate losses. To this end let, L_{tar} be a **target loss** that describes our learning goal and L_{sur} be a **surrogate loss**. Furthermore, assume that we have a learning method $\mathcal{A}$, e.g., a regularized L_{sur}-ERM approach, that asymptotically learns the surrogate learning problem defined by L_{sur}, i.e.,

$$\lim_{|D|\to\infty} \mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f_D) = \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} \tag{3.1}$$

holds in probability, where f_D is the decision function the method $\mathcal{A}$ produces for the training set D of length $|D|$. However, since our learning goal is defined by L_{tar}, we are actually interested in L_{tar}-consistency of $\mathcal{A}$, i.e., in

$$\lim_{|D|\to\infty} \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_D) = \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}\,. \tag{3.2}$$

Obviously, we obtain the latter if the convergence in (3.1) implies the convergence in (3.2). This leads to the first question we will address in this chapter.

Question 3.1. *Given a target loss L_{tar}, which surrogate losses L_{sur} ensure the implication*

$$\lim_{n\to\infty} \mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f_n) = \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} \qquad \Longrightarrow \qquad \lim_{n\to\infty} \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_n) = \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \tag{3.3}$$

for all sequences (f_n) of measurable functions $f_n : X \to \mathbb{R}$?

Question 3.1 is of purely asymptotic nature, i.e., it does consider any convergence rate in (3.1) or (3.2). Consequently, the surrogate losses that we find by answering Question 3.1 are a reasonable choice when dealing with consistency but may be less suitable when we wish to establish convergence rates for (3.2). This leads to the second question we will address.

Question 3.2. *Given a target loss L_{tar}, which surrogate losses L_{sur} allow us to deduce convergence rates for the right-hand side of (3.3) from convergence rates on the left-hand side of (3.3)?*

In particular, does there exist an increasing function $\Upsilon : [0,\infty) \to [0,\infty)$ that is continuous at 0 with $\Upsilon(0) = 0$ such that, for all measurable $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \;\le\; \Upsilon\bigl(\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}\bigr) \;?$$

Recall that we have already seen an example of such an inequality in Section 2.3, namely Zhang's inequality, which relates the excess classification risk to the excess hinge risk. In this chapter, we will systematically generalize the ideas used in the proof of that inequality to develop a general theory on

surrogate losses. The main results in this direction, including answers to the questions above, can be found in Sections 3.2 and 3.3. Furthermore, these general results will be applied to standard learning scenarios such as classification, regression, and density level detection in Sections 3.4, 3.5, 3.7, and 3.8.

3.1 Inner Risks and the Calibration Function

In order to address Questions 3.1 and 3.2, we need some tools and notions that will be introduced in this section. To this end let, us first recall that, given a loss function L and a distribution P on $X \times Y$, the L-risk of a measurable function $f : X \to \mathbb{R}$ is given by

$$\mathcal{R}_{L,\mathrm{P}}(f) = \int_X \int_Y L\big(x, y, f(x)\big)\, d\mathrm{P}(y|x)\, d\mathrm{P}_X(x). \tag{3.4}$$

Now, motivated by the calculations made in the proof of Zhang's inequality, the basic idea of our approach is to treat the inner and outer integrals *separately*. Besides some technical advantages, it will turn out that this approach has the important benefit of making our analysis rather independent of the specific distribution P, which, from the machine learning point of view, is unknown to us.

Let us begin with some fundamental definitions that will be used throughout this chapter.

Definition 3.3. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and* Q *be a distribution on Y. We define the* ***inner L-risks*** *of* Q *by*

$$\mathcal{C}_{L,\mathrm{Q},x}(t) := \int_Y L(x, y, t)\, d\mathrm{Q}(y)\,, \qquad x \in X,\, t \in \mathbb{R}\,.$$

Furthermore, the ***minimal inner L-risks*** *are denoted by*

$$\mathcal{C}^*_{L,\mathrm{Q},x} := \inf_{t \in \mathbb{R}} \mathcal{C}_{L,\mathrm{Q},x}(t)\,, \qquad x \in X.$$

Finally, if L is a supervised loss, we usually drop the subscript x in these notations, and for unsupervised losses we analogously omit the subscript Q*.*

Note that by (3.4) and the definition of the inner risks, we immediately obtain

$$\mathcal{R}_{L,\mathrm{P}}(f) = \int_X \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big)\, d\mathrm{P}_X(x)\,. \tag{3.5}$$

Our first goal is to establish the same relation between the minimal inner risks $\mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$, $x \in X$, and the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$. To this end, we have to recall the notion of a complete measurable space given after Lemma A.3.3.

Lemma 3.4 (Computation of Bayes risks). *Let X be a complete measurable space, $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, and P be a distribution on $X \times Y$. Then $x \mapsto \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$ is measurable and we have*

$$\mathcal{R}^*_{L,\mathrm{P}} = \int_X \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} \, d\mathrm{P}_X(x) \,. \tag{3.6}$$

Proof. Let us define $\varphi : X \times \mathbb{R} \to [0, \infty]$ by

$$\varphi(x, t) := \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(t) \,, \qquad\qquad x \in X,\, t \in \mathbb{R}.$$

Then φ is measurable by the measurability statement in Tonelli's theorem, and hence the first assertion follows from *iii)* of Lemma A.3.18 using $F(x) := \mathbb{R}$, $x \in X$. Consequently, the integral on the right-hand side of (3.6) exists, and it is easy to see that it satisfies

$$\mathcal{R}^*_{L,\mathrm{P}} = \inf_{f \in \mathcal{L}_0(X)} \int_X \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(f(x)) \, d\mathrm{P}_X(x) \geq \int_X \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} \, d\mathrm{P}_X(x) \,.$$

On the other hand, given $n \geq 1$, the second part of *iii)* in Lemma A.3.18 yields a measurable function $f_n : X \to \mathbb{R}$ with

$$\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(f_n(x)) \;\leq\; \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} + \frac{1}{n} \,, \qquad\qquad x \in X, \tag{3.7}$$

and hence we obtain

$$\mathcal{R}^*_{L,\mathrm{P}} \;\leq\; \mathcal{R}_{L,\mathrm{P}}(f_n) \;\leq\; \int_X \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} \, d\mathrm{P}_X(x) + \frac{1}{n} \,.$$

Letting $n \to \infty$ then yields the assertion. □

Lemma 3.4 shows that the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$ can be achieved by minimizing the inner risks $\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(\,\cdot\,)$, $x \in X$, which in general will be easier than a direct minimization of $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$. Now assume that $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Then the **excess risk** $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}$ is defined and can be computed by

$$\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} = \int_X \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(f(x)) - \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} \, d\mathrm{P}_X(x)$$

for all measurable $f : X \to \mathbb{R}$. Consequently, we can split the analysis of the excess risk into:

i) the analysis of the **inner excess risks** $\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(\,\cdot\,) - \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$, $x \in X$;
ii) the investigation of the integration with respect to P_X.

The benefit of this approach is that the analysis in *i)* only depends on P via the conditional distributions $\mathrm{P}(\,\cdot\,|x)$, and hence we can consider the excess inner risks $\mathcal{C}_{L,\mathrm{Q},x}(\,\cdot\,) - \mathcal{C}^*_{L,\mathrm{Q},x}$ for suitable classes of distributions Q on Y as a *template* for $\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(\,\cdot\,) - \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$. This leads to the following definition.

Definition 3.5. *Let $\mathcal{Q}$ be a set of distributions on Y. We say that a distribution P on $X \times Y$ is of* ***type*** *$\mathcal{Q}$ if $\mathrm{P}(\,\cdot\,|x) \in \mathcal{Q}$ for P_X-almost all $x \in X$.*

In view of Questions 3.1 and 3.2, we are mainly interested in functions $f : X \to \mathbb{R}$ that almost minimize the risk under consideration. Following the idea of splitting the analysis into the steps *i)* and *ii)*, we therefore write

$$\mathcal{M}_{L,\mathrm{Q},x}(\varepsilon) := \left\{ t \in \mathbb{R} : \mathcal{C}_{L,\mathrm{Q},x}(t) < \mathcal{C}^*_{L,\mathrm{Q},x} + \varepsilon \right\}, \qquad \varepsilon \in [0,\infty],$$

for the sets containing the **ε-approximate minimizers** of $\mathcal{C}_{L,\mathrm{Q},x}(\,\cdot\,)$. Moreover, the set of **exact minimizers** is denoted by

$$\mathcal{M}_{L,\mathrm{Q},x}(0^+) := \bigcap_{\varepsilon>0} \mathcal{M}_{L,\mathrm{Q},x}(\varepsilon)\,.$$

Again, for supervised and unsupervised losses, we usually omit the subscripts x and Q in the preceding definitions, respectively.

Before we investigate properties of the concepts above let us first illustrate these definitions with some examples. We begin with some margin-based losses introduced in Section 2.3. To this end, observe that any distribution Q on $Y := \{-1, 1\}$ can be uniquely described by an $\eta \in [0, 1]$ using the identification $\eta = \mathrm{Q}(\{1\})$. For a supervised loss $L : Y \times \mathbb{R} \to [0, \infty)$, we thus use the notation

$$\begin{aligned} \mathcal{C}_{L,\eta}(t) &:= \mathcal{C}_{L,\mathrm{Q}}(t)\,, \qquad t \in \mathbb{R}\,, \\ \mathcal{C}^*_{L,\eta} &:= \mathcal{C}^*_{L,\mathrm{Q}}\,, \end{aligned} \tag{3.8}$$

as well as $\mathcal{M}_{L,\eta}(0^+) := \mathcal{M}_{L,\mathrm{Q}}(0^+)$ and $\mathcal{M}_{L,\eta}(\varepsilon) := \mathcal{M}_{L,\mathrm{Q}}(\varepsilon)$ for $\varepsilon \in [0, \infty]$.

Example 3.6. Let L be the **least squares loss** defined in Example 2.26. For $t \in \mathbb{R}$ and $\eta \in [0, 1]$, a simple calculation then shows

$$\mathcal{C}_{L,\eta}(t) = \eta(1-t)^2 + (1-\eta)(1+t)^2 = 1 + 2t + t^2 - 4\eta t\,,$$

and hence elementary calculus gives $\mathcal{M}_{L,\eta}(0^+) = \{2\eta - 1\}$, $\mathcal{C}^*_{L,\eta} = 4\eta(1-\eta)$, and $\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = (t - 2\eta + 1)^2$ for all $t \in \mathbb{R}$ and $\eta \in [0, 1]$. ◁

Example 3.7. Recall that in Example 2.27 we defined the **hinge loss** by $L(y, t) := \max\{0, 1 - yt\}$, $y = \pm 1$, $t \in \mathbb{R}$. Now, for $\eta \in [0, 1]$ and $t \in \mathbb{R}$, a simple calculation shows

$$\mathcal{C}_{L,\eta}(t) = \begin{cases} \eta(1-t) & \text{if } t \le -1 \\ 1 + t(1-2\eta) & \text{if } t \in [-1, 1] \\ (1-\eta)(1+t) & \text{if } t \ge 1. \end{cases}$$

For $\eta \in [1/2, 1]$, we thus have

$$\mathcal{M}_{L,\eta}(0^+) = \begin{cases} [-1, 1] & \text{if } \eta = \frac{1}{2} \\ \{1\} & \text{if } \frac{1}{2} < \eta < 1 \\ [1, \infty) & \text{if } \eta = 1\,, \end{cases}$$

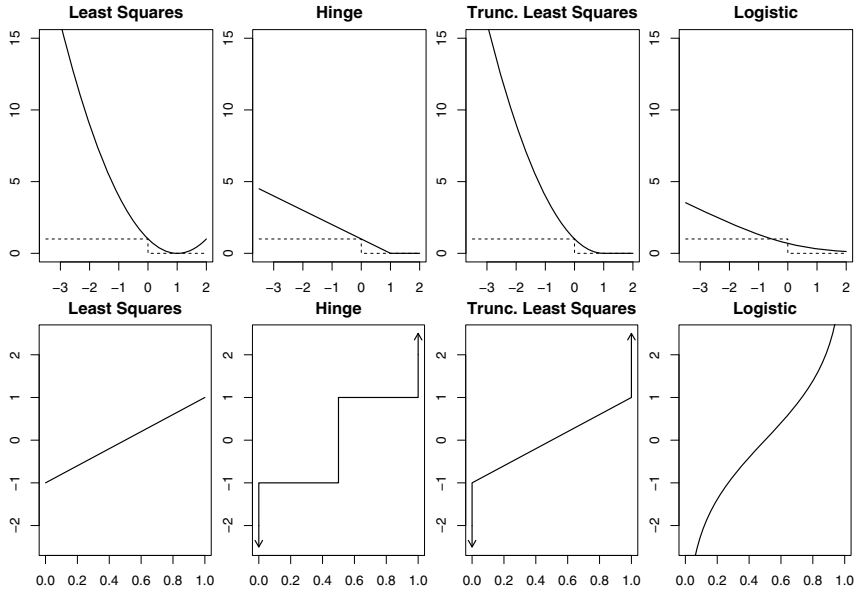


Fig. 3.1. The representing functions φ for some important margin-based loss functions L (top row) and their minimizing sets $\mathcal{M}_{L,\eta}(0^+)$, $\eta \in [0,1]$ (bottom row). For some losses and values of η, these sets are not singletons. This situation is displayed by vertical lines. Moreover, the arrows at the ends of some of these vertical lines indicate that the corresponding set is unbounded in the direction of the arrow.

$\mathcal{C}^*_{L,\eta} = 2(1-\eta)$, and

$$\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \begin{cases} 3\eta - 2 - \eta t & \text{if } t \le -1 \\ (1-t)(2\eta - 1) & \text{if } t \in [-1,1] \\ (t-1)(1-\eta) & \text{if } t \ge 1. \end{cases}$$

In addition, similar formulas hold for $\eta \in [0, 1/2]$ by symmetry. ◁

Both margin-based loss functions discussed above will serve us as surrogates for the classification loss. Therefore, let us now consider the inner risks and the set of minimizers for the standard classification loss itself.

Example 3.8. Recall that the standard **binary classification loss** is defined by $L(y,t) := \mathbf{1}_{(-\infty,0]}\big(y \operatorname{sign} t\big)$, $y \in Y$, $t \in \mathbb{R}$. For this loss, the inner risk is given by

$$\mathcal{C}_{L,\eta}(t) = \eta \mathbf{1}_{(-\infty,0)}(t) + (1-\eta)\mathbf{1}_{[0,\infty)}(t)$$

for all $\eta \in [0,1]$ and $t \in \mathbb{R}$. From this we easily conclude $\mathcal{C}^*_{L,\eta} = \min\{\eta, 1-\eta\}$, which in turn yields

$$\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = |2\eta - 1| \cdot \mathbf{1}_{(-\infty,0]}\big((2\eta - 1)\operatorname{sign} t\big) \tag{3.9}$$

for all $\eta \in [0,1]$ and $t \in \mathbb{R}$. Considering the cases $\varepsilon > |2\eta - 1|$ and $\varepsilon \leq |2\eta - 1|$ separately, we thus find

$$\mathcal{M}_{L,\eta}(\varepsilon) = \begin{cases} \mathbb{R} & \text{if } \varepsilon > |2\eta - 1| \\ \{t \in \mathbb{R} : (2\eta - 1)\operatorname{sign} t > 0\} & \text{if } 0 < \varepsilon \leq |2\eta - 1|\,. \end{cases} \qquad \triangleleft$$

Let us finally determine the inner risks and their minimizers for a more elaborate example.

Proposition 3.9 (Quantiles and the pinball loss). *For $\tau \in (0,1)$, let L be the τ-pinball loss defined in Example 2.43. Moreover, let Q be a distribution on $\mathbb{R}$ with $|\mathrm{Q}|_1 < \infty$ and let t^* be a τ-quantile of Q, i.e., we have*

$$\mathrm{Q}\big((-\infty, t^*]\big) \geq \tau \qquad \text{and} \qquad \mathrm{Q}\big([t^*, \infty)\big) \geq 1 - \tau\,.$$

Then there exist real numbers $q_+, q_- \geq 0$ such that $q_+ + q_- = \mathrm{Q}(\{t^*\})$ and

$$\mathcal{C}_{L,\mathrm{Q}}(t^* + t) - \mathcal{C}^*_{L,\mathrm{Q}} = tq_+ + \int_0^t \mathrm{Q}\big((t^*, t^* + s)\big)\, ds\,, \tag{3.10}$$

$$\mathcal{C}_{L,\mathrm{Q}}(t^* - t) - \mathcal{C}^*_{L,\mathrm{Q}} = tq_- + \int_0^t \mathrm{Q}\big((t^* - s, t^*)\big)\, ds\,, \tag{3.11}$$

for all $t \geq 0$. Moreover, we have

$$\mathcal{M}_{L,\mathrm{Q}}(0^+) = \{t^*\} \cup \big\{t > t^* \colon q_+ + \mathrm{Q}((t^*, t)) = 0\big\} \cup \big\{t < t^* \colon q_- + \mathrm{Q}((-t, t^*)) = 0\big\}.$$

Proof. Recall from Example 2.43 that distance-based τ-pinball loss is represented by

$$\psi(r) = \begin{cases} (\tau - 1)r, & \text{if } r < 0 \\ \tau r, & \text{if } r \geq 0. \end{cases}$$

Now let us consider the distribution $\mathrm{Q}^{(t^*)}$ defined by $\mathrm{Q}^{(t^*)}(A) := \mathrm{Q}(t^* + A)$ for all measurable $A \subset \mathbb{R}$. Then it is not hard to see that 0 is a τ-quantile of $\mathrm{Q}^{(t^*)}$. Moreover, we obviously have $\mathcal{C}_{L,\mathrm{Q}}(t^* + t) = \mathcal{C}_{L,\mathrm{Q}^{(t^*)}}(t)$, and hence we may assume without loss of generality that $t^* = 0$. Then our assumptions together with $\mathrm{Q}((-\infty, 0]) + \mathrm{Q}([0, \infty)) = 1 + \mathrm{Q}(\{0\})$ yield $\tau \leq \mathrm{Q}((-\infty, 0]) \leq \tau + \mathrm{Q}(\{0\})$, i.e., there exists a q_+ satisfying $0 \leq q_+ \leq \mathrm{Q}(\{0\})$ and

$$\mathrm{Q}((-\infty, 0]) = \tau + q_+\,. \tag{3.12}$$

Let us now prove the first expression for the inner risks of L. To this end, we first observe that for $t \geq 0$ we have

$$\int_{y<t} (y - t)\, d\mathrm{Q}(y) = \int_{y<0} y\, d\mathrm{Q}(y) - t\mathrm{Q}((-\infty, t)) + \int_{0 \leq y < t} y\, d\mathrm{Q}(y)$$

and

$$\int_{y\geq t}(y-t)\,d\mathrm{Q}(y) = \int_{y\geq 0} y\,d\mathrm{Q}(y) - t\mathrm{Q}([t,\infty)) - \int_{0\leq y<t} y\,d\mathrm{Q}(y)\,.$$

Consequently, we obtain

$$\begin{aligned}\mathcal{C}_{L,\mathrm{Q}}(t) &= (\tau-1)\int_{y<t}(y-t)\,d\mathrm{Q}(y) + \tau\int_{y\geq t}(y-t)\,d\mathrm{Q}(y)\\ &= \mathcal{C}_{L,\mathrm{Q}}(0) - \tau t + t\mathrm{Q}((-\infty,0)) + t\mathrm{Q}([0,t)) - \int_{0\leq y<t} y\,d\mathrm{Q}(y)\,.\end{aligned}$$

Moreover, using Lemma A.3.11, we find

$$\begin{aligned}t\,\mathrm{Q}([0,t)) - \int_{0\leq y<t} y\,d\mathrm{Q}(y) &= \int_0^t \mathrm{Q}([0,t))\,ds - \int_0^t \mathrm{Q}([s,t))\,ds\\ &= t\,\mathrm{Q}(\{0\}) + \int_0^t \mathrm{Q}((0,s))\,ds\,,\end{aligned}$$

and since (3.12) implies $\mathrm{Q}((-\infty,0))+\mathrm{Q}(\{0\}) = \mathrm{Q}((-\infty,0]) = \tau+q_+$, we thus obtain

$$\mathcal{C}_{L,\mathrm{Q}}(t) = \mathcal{C}_{L,\mathrm{Q}}(0) + tq_+ + \int_0^t \mathrm{Q}\big((0,s)\big)\,ds\,.$$

Moreover, applying this equation to the pinball loss with parameter $1-\tau$ and the distribution $\bar{\mathrm{Q}}$ defined by $\bar{\mathrm{Q}}(A) := \mathrm{Q}(-A)$, $A\subset\mathbb{R}$ measurable, gives a real number $0\leq q_-\leq \mathrm{Q}(\{0\})$ such that $\mathrm{Q}([0,\infty)) = 1-\tau+q_-$ and

$$\mathcal{C}_{L,\mathrm{Q}}(-t) = \mathcal{C}_{L,\mathrm{Q}}(0) + tq_- + \int_0^t \mathrm{Q}\big((-s,0)\big)\,ds$$

for all $t\geq 0$. Consequently, $t^* = 0$ is a minimizer of $\mathcal{C}_{L,\mathrm{Q}}(\,\cdot\,)$ and hence we find both (3.10) and (3.11). Moreover, combining $\mathrm{Q}([0,\infty)) = 1-\tau+q_-$ with (3.12), we find $q_+ + q_- = \mathrm{Q}(\{0\})$. Finally, the formula for the set of exact minimizers is an obvious consequence of (3.10) and (3.11). □

Let us now return to our general theory. We begin with the following lemma, which collects some useful properties of the sets $\mathcal{M}_{L,\mathrm{Q},x}(\,\cdot\,)$. Its proof is left as an exercise.

Lemma 3.10 (Properties of minimizers). *Let $L : X\times Y\times\mathbb{R}\to[0,\infty)$ be a loss and Q be a distribution on Y. For $x\in X$ and $t\in\mathbb{R}$, we then have:*

i) $\mathcal{M}_{L,\mathrm{Q},x}(0) = \emptyset$.
ii) $\mathcal{M}_{L,\mathrm{Q},x}(\varepsilon)\neq\emptyset$ *for some* $\varepsilon\in(0,\infty]$ *if and only if* $\mathcal{C}^*_{L,\mathrm{Q},x} < \infty$.
iii) $\mathcal{M}_{L,\mathrm{Q},x}(\varepsilon_1)\subset\mathcal{M}_{L,\mathrm{Q},x}(\varepsilon_2)$ *for all* $0\leq\varepsilon_1\leq\varepsilon_2\leq\infty$.
iv) $t\in\mathcal{M}_{L,\mathrm{Q},x}(0^+)$ *if and only if* $\mathcal{C}_{L,\mathrm{Q},x}(t) = \mathcal{C}^*_{L,\mathrm{Q},x}$ *and* $\mathcal{C}^*_{L,\mathrm{Q},x} < \infty$.
v) $t\in\mathcal{M}_{L,\mathrm{Q},x}(\infty)$ *if and only if* $\mathcal{C}_{L,\mathrm{Q},x}(t) < \infty$.

Our goal in the following two lemmas is to show that we can use the sets $\mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(\,\cdot\,)$ to construct (approximate) L-risk minimizers. Note that the main difficulty in these lemmas is to ensure the measurability of the (approximate) minimizers.

Lemma 3.11 (Existence of approximate minimizers). *Let X be a complete measurable space, $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, P be a distribution on $X \times Y$, and $\varepsilon \in (0,\infty]$. Then the following statements are equivalent:*

i) $\mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} < \infty$ *for* P_X*-almost all* $x \in X$.
ii) There exists a measurable $f : X \to \mathbb{R}$ *such that* $f(x) \in \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(\varepsilon)$ *for* P_X*-almost all* $x \in X$.

Proof. *ii)* $\Rightarrow$ *i)*. This immediately follows from *ii)* of Lemma 3.10.

i) $\Rightarrow$ *ii)*. Let $n \geq 1$ with $1/n < \varepsilon$. As in the proof of Lemma 3.4, we then obtain a measurable function $f_n : X \to \mathbb{R}$ satisfying (3.7) for all $x \in X$. Since $\mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} < \infty$ for P_X-almost all $x \in X$, we thus find the assertion. □

While the preceding lemma characterizes the situations where *uniform* ε-approximate minimizers exist, the following lemma characterizes L-risks that have an *exact* minimizer, i.e., a Bayes decision function.

Lemma 3.12 (Existence of exact minimizers). *Let X be a complete measurable space, $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, and P be a distribution on $X \times Y$ satisfying $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Then the following are equivalent:*

i) $\mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(0^+) \neq \emptyset$ *for* P_X*-almost all* $x \in X$.
ii) There exists a measurable $f^* : X \to \mathbb{R}$ *such that* $\mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}^*_{L,\mathrm{P}}$.

Moreover, if one of the conditions is satisfied, every Bayes decision function $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ *satisfies* $f^*_{L,\mathrm{P}}(x) \in \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(0^+)$ *for* P_X*-almost all* $x \in X$.

Proof. *i)* $\Rightarrow$ *ii)*. Let φ and F be defined as in the proof of Lemma 3.4. Using the last part of *iii)* in Lemma A.3.18, we then find a measurable $f^* : X \to \mathbb{R}$ with $f^*(x) \in \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(0^+)$ for P_X-almost all $x \in X$. Obviously, part *iv)* of Lemma 3.10 and Lemma 3.4 then show $\mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}^*_{L,\mathrm{P}}$.

ii) $\Rightarrow$ *i)*. Let $f^*_{L,\mathrm{P}}$ be a Bayes decision function, i.e., it satisfies $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$. Since $\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(f^*_{L,\mathrm{P}}) \geq \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$ for all $x \in X$, Lemma 3.4 together with $\mathcal{R}^*_{L,\mathrm{P}} < \infty$ then yields $\mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(f^*_{L,\mathrm{P}}) = \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$ for P_X-almost all $x \in X$. Moreover, $\mathcal{R}^*_{L,\mathrm{P}} < \infty$ implies $\mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x} < \infty$ for P_X-almost all $x \in X$, and hence we find $f^*_{L,\mathrm{P}}(x) \in \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x),x}(0^+)$ for P_X-almost all $x \in X$ by part *iv)* of Lemma 3.10. □

Let us now assume for a moment that we have two loss functions $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ such that

$$\emptyset \neq \mathcal{M}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}(0^+) \subset \mathcal{M}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(0^+)\,, \qquad x \in X. \tag{3.13}$$

Then Lemmas 3.4 and 3.12 show that every exact minimizer of $\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(\,\cdot\,)$ is also an exact minimizer of $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(\,\cdot\,)$, i.e., we have the implication

$$\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) = \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) = \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}\,. \tag{3.14}$$

However, exact minimizers do not necessarily exist, as one can see by combining Lemma 3.11 with Lemma 3.12, and even if they do exist, it is rather unlikely that we will find them by a learning procedure. On the other hand, we have already indicated in Chapter 1 that many learning procedures are able to find approximate minimizers, and therefore we need an *approximate* version of (3.14) to answer Question 3.1. Now, the key idea for establishing such a modification of (3.14) is to consider approximate versions of (3.13). To this end, we begin with the following fundamental definition.

Definition 3.13. *Let $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be loss functions, Q be a distribution on Y, and $x \in X$. Then we define the* ***calibration function*** *$\delta_{\max}(\,\cdot\,,\mathrm{Q},x) : [0,\infty] \to [0,\infty]$ of $(L_{\mathrm{tar}}, L_{\mathrm{sur}})$ by*

$$\delta_{\max}(\varepsilon,\mathrm{Q},x) := \begin{cases} \inf\limits_{\substack{t\in\mathbb{R}\\ t\notin\mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)}} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} & \text{if } \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} < \infty \\ \infty & \text{if } \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} = \infty \end{cases}$$

for all $\varepsilon \in [0,\infty]$. Moreover, we write $\delta_{\max,L_{\mathrm{tar}},L_{\mathrm{sur}}}(\varepsilon,\mathrm{Q},x) := \delta_{\max}(\varepsilon,\mathrm{Q},x)$ whenever it is necessary to explicitly mention the target and surrogate losses. Finally, if both losses are supervised, we usually omit the argument x.

The following lemma collects some simple though extremely important properties of the calibration function.

Lemma 3.14 (Properties of the calibration function). *Let $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be losses and Q be a distribution on Y. For all $x \in X$ and $\varepsilon \in [0,\infty]$, we then have:*

i) $\mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta_{\max}(\varepsilon,\mathrm{Q},x)) \subset \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$.
ii) $\mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta) \not\subset \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$ whenever $\delta > \delta_{\max}(\varepsilon,\mathrm{Q},x)$.

Consequently, the calibration function can be computed by

$$\delta_{\max}(\varepsilon,\mathrm{Q},x) = \max\bigl\{\delta \in [0,\infty] : \mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta) \subset \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)\bigr\}\,. \tag{3.15}$$

Finally, if both $\mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{Q},x} < \infty$ and $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} < \infty$, then for all $t \in \mathbb{R}$ we have

$$\delta_{\max}\bigl(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{Q},x},\mathrm{Q},x\bigr) \;\le\; \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x}\,. \tag{3.16}$$

Inequality (3.16) will be the key ingredient when we compare the excess L_{tar}-risk with the excess L_{sur}-risk since it allows us to compare the inner integrals of these risks. Furthermore, one can show by *ii)* that the calibration function is the optimal way to compare these inner integrals. We refer to Exercise 3.3 for details.

Proof. Let us first assume $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} = \infty$. Then we have $\delta_{\max}(\varepsilon, \mathrm{Q}, x) = \infty$ and hence *ii)* is trivially satisfied. Moreover, we have $\mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta_{\max}(\varepsilon, \mathrm{Q}, x)) = \emptyset$ by *ii)* of Lemma 3.10, and hence we obtain *i)*. Let us now assume $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} < \infty$. Then, for $t \in \mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta_{\max}(\varepsilon, \mathrm{Q}, x))$, we have

$$\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} < \delta_{\max}(\varepsilon, \mathrm{Q}, x) = \inf_{\substack{t' \in \mathbb{R} \\ t' \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)}} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t') - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x},$$

which shows $t \in \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$. For the proof of the second assertion, let us fix a δ with $\delta > \delta_{\max}(\varepsilon, \mathrm{Q}, x)$. By definition, this means

$$\inf_{\substack{t \in \mathbb{R} \\ t \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)}} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} = \delta_{\max}(\varepsilon, \mathrm{Q}, x) < \delta,$$

and hence there exists a $t \in \mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta)$ with $t \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$. This shows part *ii)*. Moreover, (3.15) is a direct consequence of *i)* and *ii)*.

Let us finally prove Inequality (3.16). To this end, we fix a $t \in \mathbb{R}$ and write $\varepsilon := \mathcal{C}_{L_{\mathrm{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{Q},x}$. Then have $t \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$, which implies $t \notin \mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(\delta_{\max}(\varepsilon, \mathrm{Q}, x))$ by *i)*. The latter means

$$\begin{aligned}\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) &\geq \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} + \delta_{\max}(\varepsilon, \mathrm{Q}, x) \\ &= \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} + \delta_{\max}\left(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{Q},x}, \mathrm{Q}, x\right). \end{aligned}$$

□

Due to algorithmic reasons, we are often interested in *convex* surrogate losses. For such surrogates, the calibration function can be easily computed.

Lemma 3.15 (Calibration function for convex surrogates). *Let* Q *be a distribution on* Y*,* $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0, \infty)$ *be a loss, and* $x \in X$*,* $\varepsilon > 0$ *such that* $\mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$ *is an interval. Moreover, let* $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0, \infty)$ *be a convex loss such that* $\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) < \infty$ *for all* $t \in \mathbb{R}$*. If* $\mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(0^+) \subset \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(0^+)$*, then we have*

$$\delta_{\max}(\varepsilon, \mathrm{Q}, x) = \min\left\{\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t_\varepsilon^-), \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t_\varepsilon^+)\right\} - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x}, \tag{3.17}$$

where we used the definitions $t_\varepsilon^- := \inf \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$*,* $t_\varepsilon^+ := \sup \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$*, and* $\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(\pm\infty) := \infty$*.*

Proof. Obviously, $\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(\,\cdot\,) : \mathbb{R} \to [0, \infty)$ is convex, and thus it is also continuous by Lemma A.6.2. Since $\mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$ is an interval, we then obtain

$$\delta_{\max}(\varepsilon, \mathrm{Q}, x) = \min\Big\{\inf_{t \leq t_\varepsilon^-} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t), \inf_{t \geq t_\varepsilon^+} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t)\Big\} - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x}.$$

Moreover, for $t < t_\varepsilon^-$, we have $t \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$ and hence $t \notin \mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(0^+)$. From this we conclude that $t \notin \mathcal{M}_{L_{\mathrm{sur}},\mathrm{Q},x}(0^+)$, i.e., $\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) > \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x}$. Consequently, the convexity of $t \mapsto \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t)$ shows that this map is strictly decreasing on $(-\infty, t_\varepsilon^-]$, and hence we obtain $\inf\{\mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) : t \leq t_\varepsilon^-\} = \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t_\varepsilon^-)$. For $t \geq t_\varepsilon^+$, we can argue analogously. □

Let us close this section with an example that illustrates how to compute the calibration function for specific loss functions.

Example 3.16. Let L be either the **least squares loss** L_{LS} or the **hinge loss** L_{hinge}. We write L_{class} for the binary classification loss and identify distributions Q on $\{-1,1\}$ by $\eta := \mathrm{Q}(\{1\})$. Then Lemma 3.15 together with Example 3.8 yields $\delta_{\max,L_{\mathrm{class}},L}(\varepsilon,\eta) = \infty$ if $\varepsilon > |2\eta-1|$. Moreover, for $0 < \varepsilon \leq |2\eta-1|$, we find

$$\delta_{\max,L_{\mathrm{class}},L}(\varepsilon,\eta) = \mathcal{C}_{L,\eta}(0) - \mathcal{C}^*_{L,\eta} = \begin{cases} (2\eta-1)^2 & \text{if } L = L_{\mathrm{LS}} \\ |2\eta-1| & \text{if } L = L_{\mathrm{hinge}} \end{cases}$$

by applying Examples 3.6 and 3.7, respectively. In particular note that in both cases we have $\delta_{\max,L_{\mathrm{class}},L}(\varepsilon,\eta) > 0$ for all $\eta \in [0,1]$ and all $\varepsilon > 0$. ◁

3.2 Asymptotic Theory of Surrogate Losses

In this section, we investigate the asymptotic relationship between excess risks in the sense of Question 3.1. The main result in this direction is the following theorem.

Theorem 3.17 (Asymptotic calibration of risks). *Let X be a complete measurable space, $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be losses, and P be a distribution on $X \times Y$ such that $\mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} < \infty$. Then*

$$x \mapsto \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)$$

is measurable for all $\varepsilon \in [0,\infty]$. In addition, consider the following statements:

i) For all $\varepsilon \in (0,\infty]$, we have $\mathrm{P}_X\big(\{x \in X : \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = 0\}\big) = 0$.
ii) For all $\varepsilon \in (0,\infty]$, there exists a $\delta > 0$ such that, for all measurable functions $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} + \delta \quad \Longrightarrow \quad \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} + \varepsilon\,. \tag{3.18}$$

Then we have ii) $\Rightarrow$ i). Furthermore, i) $\Rightarrow$ ii) holds if there exists a P_X-integrable function $b : X \to [0,\infty)$ such that, for all $x \in X$, $t \in \mathbb{R}$, we have

$$\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(t) \;\leq\; \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} + b(x)\,. \tag{3.19}$$

Proof. To show the measurability of $\delta_{\max}(\,\cdot\,, \mathrm{P}(\,\cdot\,|x), x)$, we may assume without loss of generality that we have $\mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} < \infty$ and $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} < \infty$ for all $x \in X$. We equip $[0,\infty]$ with the Borel σ-algebra and write $A := [\varepsilon,\infty]$. Furthermore, let $h : X \times \mathbb{R} \to [0,\infty]$ be defined by

$$h(x,t) := \mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\,, \qquad (x,t) \in X \times \mathbb{R}\,.$$

Then h is measurable and, for the set-valued function $F : X \to 2^{\mathbb{R}}$ defined by $F(x) := \{t \in \mathbb{R} : h(x,t) \in A\}$, $x \in X$, we have $\mathbb{R}\backslash\mathcal{M}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(\varepsilon) = F(x)$ for all $x \in X$. Furthermore, $\varphi : X \times \mathbb{R} \to [0,\infty]$ defined by

$$\varphi(x,t) := \mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\,, \qquad (x,t) \in X \times \mathbb{R}\,,$$

is measurable. Now, for all $x \in X$, our construction yields

$$\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = \inf_{t \in F(x)} \varphi(x,t)\,,$$

and hence $x \mapsto \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)$ is measurable by Lemma A.3.18.

ii) $\Rightarrow$ *i)*. Assume that *i)* is not true. Then there is an $\varepsilon \in (0,\infty]$ such that

$$B := \bigl\{x \in X : \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = 0 \text{ and } \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} < \infty\bigr\}$$

satisfies $\mathrm{P}_X(B) > 0$. Note that for $x \in B$ we have $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} < \infty$ by the very definition of the calibration function. In addition, for $x \in B$, we have $\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = 0$ and hence there exists a $t \in \mathbb{R}\backslash\mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)$. Using the notation of the first part of the proof, this t satisfies $h(x,t) \geq \varepsilon$ and hence we have $F(x) \neq \emptyset$. This shows $B \subset \operatorname{Dom} F$. By Lemma A.3.18, there thus exist measurable functions $f_n^{(1)} : X \to \mathbb{R}$ such that

$$\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\bigl(f_n^{(1)}(x)\bigr) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} \;\leq\; \frac{1}{n}$$

and

$$\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\bigl(f_n^{(1)}(x)\bigr) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} \;\geq\; \varepsilon$$

for all $x \in B$ and $n \geq 1$. Furthermore, by Lemma 3.11, we find measurable functions $f_n^{(2)} : X \to \mathbb{R}$, $n \geq 1$, with

$$\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\bigl(f_n^{(2)}(x)\bigr) \;<\; \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} + \frac{1}{n}$$

for P_X-almost all $x \in X$. We define $f_n : X \to \mathbb{R}$ by

$$f_n(x) := \begin{cases} f_n^{(1)}(x) & \text{if } x \in B \\ f_n^{(2)}(x) & \text{otherwise.} \end{cases}$$

Then f_n is measurable and our construction yields both

$$\begin{aligned} \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} &\geq \int_B \Bigl(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\bigl(f_n(x)\bigr) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\Bigr) d\mathrm{P}_X(x) \\ &\geq \varepsilon\, \mathrm{P}_X(B) \end{aligned}$$

and $\lim_{n\to\infty} \mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f_n) = \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}$. From this we conclude that *ii)* is not true.

i) $\Rightarrow$ *ii)*. Let us assume that *ii)* is not true. Then there exists an $\varepsilon_0 \in (0,\infty]$ such that for all $n \geq 1$ there exists a measurable function $f_n : X \to \mathbb{R}$ with $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \geq \varepsilon_0$ and

$$\frac{1}{n} \geq \mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} = \int_X \left| \mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f_n(x)\big) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} \right| d\mathrm{P}_X(x).$$

Hence there exists a sub-sequence (f_{n_i}) satisfying

$$\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f_{n_i}(x)\big) \to \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}$$

for P_X-almost all $x \in X$. Let us fix an $x \in X$ at which the convergence takes place and that additionally satisfies $\mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} < \infty$, $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} < \infty$, and $\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) > 0$ for *all* $\varepsilon > 0$. For later use, note that the probability for such an element x is 1 since $\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)$ is monotonically increasing in ε. Now, for $\varepsilon > 0$, there exists an i_0 such that for all $i \geq i_0$ we have

$$\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f_{n_i}(x)\big) < \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} + \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)\,.$$

By part *i*) of Lemma 3.14, this yields $\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(f_{n_i}(x)) < \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x} + \varepsilon$, i.e., we have shown

$$\lim_{i\to\infty} \mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big(f_{n_i}(x)\big) \;=\; \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\,. \tag{3.20}$$

Since the probability of the considered x was 1, the limit relation (3.20) holds for P_X-almost all $x \in X$, and hence we obtain $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_{n_i}) \to \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}$ by Lebesgue's convergence theorem and (3.19). However, this contradicts the fact that $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \geq \varepsilon_0$ holds for all $n \geq 1$. □

Theorem 3.17 shows that an almost surely strictly positive calibration function is *necessary* for a positive answer to Question 3.1, i.e., for having an implication of the form

$$\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f_n) \to \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f_n) \to \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \tag{3.21}$$

for *all* sequences (f_n) of measurable functions. Moreover, Theorem 3.17 also shows that an almost surely strictly positive calibration function is *sufficient* for (3.21) if the *additional* assumption (3.19) holds. In this regard, note that in general this additional assumption is *not* superfluous. For details, we refer to Exercise 3.11.

Let us now recall that from a machine learning point of view we are not interested in a single distribution since we do not know the data-generating distribution P. However, we may know that P is a distribution of a certain type $\mathcal{Q}$, and consequently the following definition is of great importance in practical situations.

Definition 3.18. *Let $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be two losses and $\mathcal{Q}$ be a set of distributions on Y. We say that L_{sur} is L_{tar}-**calibrated** with respect to $\mathcal{Q}$ if, for all $\varepsilon \in (0,\infty]$, $\mathrm{Q} \in \mathcal{Q}$, and $x \in X$, we have*

$$\delta_{\max}\left(\varepsilon, \mathrm{Q}, x\right) > 0\,.$$

Note that, using (3.15), we easily verify that L_{sur} is L_{tar}-calibrated with respect to $\mathcal{Q}$ if and only if for all $\varepsilon \in (0,\infty]$, $\mathrm{Q} \in \mathcal{Q}$, and $x \in X$ there is a $\delta \in (0,\infty]$ with

$$\mathcal{M}_{L_{\text{sur}},\mathrm{Q},x}(\delta) \subset \mathcal{M}_{L_{\text{tar}},\mathrm{Q},x}(\varepsilon)\,. \tag{3.22}$$

Now assume that our only information on the data-generating distribution P is that it is of some type $\mathcal{Q}$. Then Theorem 3.17 shows that we can only hope for a positive answer to Question 3.1 if our surrogate loss L_{sur} is L_{tar}-calibrated with respect to $\mathcal{Q}$. In this sense, calibration of L_{sur} is a first test on whether L_{sur} is a reasonable surrogate. The following corollary, whose proof is left as an exercise, shows that for some target losses this test is also sufficient.

Corollary 3.19. *Let X be a complete measurable space, $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be two losses, and $\mathcal{Q}$ be a set of distributions on Y. If L_{tar} is bounded, i.e., there is $B > 0$ with $L(x,y,t) \le B$ for all $(x,y,t) \in X \times Y \times \mathbb{R}$, then the following statements are equivalent:*

i) L_{sur} is L_{tar}-calibrated with respect to $\mathcal{Q}$.
ii) For all $\varepsilon \in (0,\infty]$ and all distributions P of type $\mathcal{Q}$ with $\mathcal{R}^*_{L_{\text{sur}},\mathrm{P}} < \infty$, there exists a $\delta \in (0,\infty]$ such that, for all measurable $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}} + \delta \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} + \varepsilon\,.$$

Recall that both the classification loss and the density level detection loss are bounded losses, and consequently the preceding corollary applies to these target losses. Moreover, for the classification loss being the target loss and the least squares or the hinge loss being the surrogate loss, we have already shown in Example 3.16 that the corresponding calibration function is strictly positive. Consequently, Corollary 3.19 shows that both loss functions are reasonable surrogates in an asymptotic sense. However, we have already seen in Zhang's inequality, see Theorem 2.31, that there is even a strong *quantitative* relationship between the excess classification risk and the excess hinge risk. Such stronger relationships are studied in the next section in more detail.

3.3 Inequalities between Excess Risks

If one wants to find a good surrogate loss L_{sur} for a given target loss L_{tar}, then implication (3.18) is in some sense a minimal requirement. However, we have already indicated in Question 3.2 that in many cases one actually needs quantified versions of (3.18), e.g., in terms of *inequalities* between the excess L_{tar}-risk and the excess L_{sur}-risk. Considering Theorem 3.17, such inequalities are readily available if, for all $\varepsilon > 0$, we *know* a $\delta(\varepsilon) > 0$ such that implication (3.18) holds for all measurable $f : X \to \mathbb{R}$. Indeed, for f with $\varepsilon := \mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} > 0$, we have $\mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}} \ge \delta(\varepsilon)$, or in other words

$$\delta\bigl(\mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}}\bigr) \;\le\; \mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}}\,. \tag{3.23}$$

In addition, if we define $\delta(0) := 0$, then this inequality actually holds for *all* measurable $f : X \to \mathbb{R}$. Unfortunately, however, the proof of Theorem 3.17 does not provide a constructive way to find a value for $\delta(\varepsilon)$, and hence we have so far no method to establish inequalities of the form (3.23). This problem is resolved in the following theorems for which we first introduce the Fenchel-Legendre bi-conjugate of a function.

Definition 3.20. *Let $I \subset \mathbb{R}$ be an interval and $g : I \to [0,\infty]$ be a function. Then the* ***Fenchel-Legendre bi-conjugate*** *$g^{**} : I \to [0,\infty]$ of g is the largest convex function $h : I \to [0,\infty]$ satisfying $h \leq g$. Moreover, we write $g^{**}(\infty) := \lim_{t\to\infty} g^{**}(t)$ if $I = [0,\infty)$.*

Note that if $g : [0,B] \to [0,\infty)$ is a strictly positive and increasing function on $(0,B]$ with $g(0) = 0$, then Lemma A.6.20 shows that its bi-conjugate g^{**} is also strictly positive on $(0,B]$. Furthermore, a similar result holds for continuous functions (see Lemma A.6.21). However, these results are in general false if one considers functions on $I := [0,\infty)$, as, e.g., the square root $\sqrt{\cdot} : [0,\infty) \to [0,\infty)$ shows.

Besides the Fenchel-Legendre bi-conjugate, we also need some additional notations and definitions. To this end, let X be a complete measurable space, $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss function, and P be a distribution on $X \times Y$ such that $\mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} < \infty$. For a measurable function $f : X \to \mathbb{R}$, we write

$$B_f := \left\| x \mapsto \left(\mathcal{C}_{L_{\text{tar}},\mathrm{P}(\,\cdot\,|x),x}(f(x)) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{P}(\,\cdot\,|x),x} \right) \right\|_\infty , \tag{3.24}$$

i.e., B_f is the supremum of the excess inner target risk with respect to f. Note that in the following considerations we do *not* require $B_f < \infty$.

Our first two results on inequalities between excess risks will only assume that the involved distribution P is of some type $\mathcal{Q}$. In this case, the following notion of calibration will be crucial.

Definition 3.21. *Let $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be two losses and $\mathcal{Q}$ be a set of distributions on Y. Then the* ***uniform calibration function*** *with respect to $\mathcal{Q}$ is defined by*

$$\delta_{\max}(\varepsilon, \mathcal{Q}) := \inf_{\substack{\mathrm{Q}\in\mathcal{Q} \\ x\in X}} \delta_{\max}\left(\varepsilon, \mathrm{Q}, x\right) , \qquad \varepsilon \in [0,\infty].$$

Moreover, we say that L_{sur} is ***uniformly L_{tar}-calibrated*** *with respect to $\mathcal{Q}$ if $\delta_{\max}(\varepsilon,\mathcal{Q}) > 0$ for all $\varepsilon \in (0,\infty]$.*

Obviously, every uniformly calibrated loss function is calibrated; however, the converse implication does not hold in general. Since we will see important examples of the latter statement in Section 3.7, we do not present such an example here. Finally, note that an alternative definition of $\delta_{\max}(\varepsilon,\mathcal{Q})$ can be found in Exercise 3.5.

Now we are well-prepared to formulate our first result, which establishes inequalities between excess risks of different loss functions.

Theorem 3.22 (Uniform calibration inequalities). *Let X be a complete measurable space, $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be losses, and $\mathcal{Q}$ be a set of distributions on Y. Moreover, let $\delta : [0,\infty] \to [0,\infty]$ be an increasing function such that*

$$\delta_{\max}(\varepsilon, \mathcal{Q}) \;\geq\; \delta(\varepsilon)\,, \qquad\qquad \varepsilon \in [0,\infty]. \tag{3.25}$$

Then, for all distributions P of type $\mathcal{Q}$ satisfying $\mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} < \infty$ and all measurable $f : X \to \mathbb{R}$, we have

$$\delta^{**}_{B_f}\big(\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}\big) \;\leq\; \mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}\,, \tag{3.26}$$

where $\delta^{**}_{B_f} : [0, B_f] \to [0,\infty]$ is the Fenchel-Legendre biconjugate of $\delta_{|[0,B_f]}$ and B_f is defined by (3.24).

Proof. Inequalities (3.16) and (3.25) together with $\mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} < \infty$ give

$$\delta\big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big) \;\leq\; \mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x} \tag{3.27}$$

for P_X-almost all $x \in X$ and all $t \in \mathbb{R}$. For a measurable function $f : X \to \mathbb{R}$ with $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) < \infty$, Jensen's inequality together with the definition of B_f, $\delta^{**}_{B_f}(\,\cdot\,) \leq \delta(\,\cdot\,)$, and (3.27) now yields

$$\begin{aligned}
&\delta^{**}_{B_f}\big(\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}\big) \\
&\leq \int_X \delta^{**}_{B_f}\Big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\Big) d\mathrm{P}_X(x) \\
&\leq \int_X \mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\, d\mathrm{P}_X(x) \\
&= \mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}\,.
\end{aligned}$$

Finally, for $f : X \to \mathbb{R}$ with $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) = \infty$, we have $B_f = \infty$. If $\delta^{**}_\infty(\infty) = 0$, there is nothing to prove, and hence let us assume $\delta^{**}_\infty(\infty) > 0$. Then (3.25) implies $\delta(0) = 0$ and hence δ^{**}_∞ is increasing because of its convexity and $\delta^{**}_\infty(0) = \delta(0) = 0$. Consequently, if δ^{**}_∞ is finite on $[0,\infty)$, then there exists a $t_0 \geq 0$ and a $c_0 > 0$ such that the (Lebesgue)-almost surely defined derivative of δ^{**}_∞ satisfies $(\delta^{**}_\infty)'(t) \geq c_0$ for almost all $t \geq t_0$. By Lebesgue's version of the fundamental theorem of calculus, see Theorem A.6.6, we then find constants $c_1, c_2 \in (0,\infty)$ with $t \leq c_1\delta^{**}_\infty(t) + c_2$ for all $t \in [0,\infty]$. On the other hand, if there is a $t_0 > 0$ with $\delta^{**}_\infty(t_0) = \infty$, we have $t \leq c_1\delta^{**}_\infty(t) + c_2$ for $c_1 := 1$, $c_2 := t_0$, and all $t \in [0,\infty]$. In both cases, (3.27) now yields

$$\begin{aligned}
\infty &= \int_X \Big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\Big) d\mathrm{P}_X(x) \\
&\leq c_1 \int_X \delta^{**}_\infty\Big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\Big) d\mathrm{P}_X(x) + c_2 \\
&\leq c_1\big(\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}\big) + c_2
\end{aligned}$$

and hence we have $\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} = \infty$. □

Note that if L_{sur} is uniformly L_{tar}-calibrated with respect to $\mathcal{Q}$ and the function $f : X \to \mathbb{R}$ satisfies $B_f < \infty$, then Lemma A.6.20 shows that the bi-conjugate of $\delta_{\max}(\,\cdot\,, \mathcal{Q})_{|[0,B_f]}$ is strictly positive on $(0, B_f]$. Consequently, Theorem 3.22 gives a *non-trivial* inequality in this case.

Let us now illustrate the theory we have developed so far by a simple example dealing with the least squares and the hinge loss.

Example 3.23. Let L be either the **least squares loss** or the **hinge loss**, $\mathcal{Q}_Y$ be the set of all distributions on $Y := \{-1, 1\}$, and L_{class} be the binary classification loss. Using Example 3.16, we then obtain

$$\delta_{\max,L_{\text{class}},L}(\varepsilon, \mathcal{Q}_Y) = \inf_{\eta\in[0,1]} \delta_{\max,L_{\text{class}},L}(\varepsilon, \eta) = \inf_{|2\eta-1|\geq\varepsilon} \delta_{\max,L_{\text{class}},L}(\varepsilon, \eta)$$

for all $\varepsilon > 0$. For the least squares loss, this yields

$$\delta_{\max,L_{\text{class}},L}(\varepsilon, \mathcal{Q}_Y) = \varepsilon^2\,, \qquad\qquad \varepsilon > 0,$$

which by Theorem 3.22 implies that, for all measurable $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \;\leq\; \sqrt{\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}}\,.$$

On the other hand, for the hinge loss, we find $\delta_{\max,L_{\text{class}},L}(\varepsilon, \mathcal{Q}_Y) = \varepsilon$ for all $\varepsilon > 0$, and hence Theorem 3.22 recovers Zhang's inequality. ◁

The following result shows that uniform calibration is also *necessary* to establish non-trivial inequalities that hold for *all* distributions of some type.

Theorem 3.24. *Let X be a complete measurable space, $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be two losses, and $\mathcal{Q}$ be a set of distributions on Y such that $\mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x} < \infty$ and $\mathcal{C}^*_{L_{\text{sur}},\mathrm{Q},x} < \infty$ for all $x \in X$ and $\mathrm{Q} \in \mathcal{Q}$. Furthermore, let $\delta : [0,\infty] \to [0,\infty]$ be increasing with $\delta(0) = 0$ and $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$. If for all distributions P of type $\mathcal{Q}$ satisfying $\mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\text{sur}},\mathrm{P}} < \infty$ and all measurable $f : X \to \mathbb{R}$ we have*

$$\delta\big(\mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}}\big) \;\leq\; \mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}}\,,$$

then L_{sur} is uniformly L_{tar}-calibrated with respect to $\mathcal{Q}$.

Proof. Let us fix an $x \in X$ and a $\mathrm{Q} \in \mathcal{Q}$. Furthermore, let P be the distribution on $X \times Y$ with $\mathrm{P}_X = \delta_{\{x\}}$ and $\mathrm{P}(\,\cdot\,|x) = \mathrm{Q}$. Then P is of type $\mathcal{Q}$, and we have both $\mathcal{R}_{L_i,\mathrm{P}}(f) = \mathcal{C}_{L_i,\mathrm{Q},x}(f(x))$ and $\mathcal{R}^*_{L_i,\mathrm{P}} = \mathcal{C}^*_{L_i,\mathrm{Q},x} < \infty$ for $i = \{\text{tar},\text{sur}\}$ and all measurable $f : X \to \mathbb{R}$. Consequently, our assumption yields

$$\delta\big(\mathcal{C}_{L_{\text{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x}\big) \;\leq\; \mathcal{C}_{L_{\text{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{sur}},\mathrm{Q},x}\,, \qquad\qquad t \in \mathbb{R}.$$

Now let $\varepsilon > 0$ and $t \in \mathcal{M}_{L_{\text{sur}},\mathrm{Q},x}(\delta(\varepsilon))$. Then we have $\mathcal{C}_{L_{\text{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{sur}},\mathrm{Q},x} < \delta(\varepsilon)$, and hence the inequality above yields $\delta(\mathcal{C}_{L_{\text{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x}) < \delta(\varepsilon)$. Since δ is monotonically increasing, the latter shows $\mathcal{C}_{L_{\text{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x} < \varepsilon$, i.e., we have found $\mathcal{M}_{L_{\text{sur}},\mathrm{Q},x}(\delta(\varepsilon)) \subset \mathcal{M}_{L_{\text{tar}},\mathrm{Q},x}(\varepsilon)$. Equation (3.15) then shows $\delta(\varepsilon) \leq \delta_{\max}(\varepsilon, \mathrm{Q}, x)$, and hence L_{sur} is uniformly L_{tar}-calibrated with respect to $\mathcal{Q}$. □

It will turn out in Sections 3.7 and 3.9, for example, that in many situations we have calibrated losses that are not uniformly calibrated. We have just seen that in such cases we need assumptions on P stronger than the $\mathcal{Q}$-type to establish inequalities. The following theorem presents a result in this direction.

Theorem 3.25 (General calibration inequalities). *Let X be a complete measurable space, $L_{\rm tar} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\rm sur} : X \times Y \times \mathbb{R} \to [0,\infty)$ be two losses, and P be a distribution on $X \times Y$ such that $\mathcal{R}^*_{L_{\rm tar},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\rm sur},\mathrm{P}} < \infty$. Assume that there exist a $p \in (0,\infty]$ and functions $b : X \to [0,\infty]$ and $\delta : [0,\infty) \to [0,\infty)$ such that*

$$\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) \;\geq\; b(x)\,\delta(\varepsilon)\,, \qquad\qquad \varepsilon \geq 0,\, x \in X, \tag{3.28}$$

and $b^{-1} \in L_p(\mathrm{P}_X)$. Then, for $\bar{\delta} := \delta^{\frac{p}{p+1}} : [0,\infty) \to [0,\infty)$ and all measurable $f : X \to \mathbb{R}$, we have

$$\bar{\delta}^{**}_{B_f}\big(\mathcal{R}_{L_{\rm tar},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\rm tar},\mathrm{P}}\big) \;\leq\; \|b^{-1}\|^{\frac{p}{p+1}}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L_{\rm sur},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\rm sur},\mathrm{P}}\big)^{\frac{p}{p+1}}\,,$$

where $\bar{\delta}^{**}_{B_f} : [0, B_f] \to [0,\infty]$ is the Fenchel-Legendre biconjugate of $\bar{\delta}_{|[0,B_f]}$ and B_f is defined by (3.24).

Proof. Let us first consider the case $\mathcal{R}_{L_{\rm tar},\mathrm{P}}(f) < \infty$. Since $\bar{\delta}^{**}_{B_f}$ is convex and satisfies $\bar{\delta}^{**}_{B_f}(\varepsilon) \leq \bar{\delta}(\varepsilon)$ for all $\varepsilon \in [0, B_f]$, we see by Jensen's inequality that

$$\bar{\delta}^{**}_{B_f}\big(\mathcal{R}_{L_{\rm tar},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\rm tar},\mathrm{P}}\big) \;\leq\; \int_X \bar{\delta}\big(\mathcal{C}_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}\big)\,d\mathrm{P}_X(x)\,.$$

Moreover, using (3.28) and (3.16), we obtain

$$b(x)\,\delta\big(\mathcal{C}_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}\big) \;\leq\; \mathcal{C}_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}$$

for P_X-almost all $x \in X$ and all $t \in \mathbb{R}$. Now note that for $q := (1+p)/p$ the conjugate exponent satisfies $q' = 1 + p = pq$. By the definition of $\bar{\delta}$ and Hölder's inequality in the form of $\mathbb{E}|hg|^{1/q} \leq (\mathbb{E}|h|^{q'/q})^{1/q'}(\mathbb{E}|g|)^{1/q}$, we thus find

$$\begin{aligned}
&\int_X \bar{\delta}\big(\mathcal{C}_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\rm tar},\mathrm{P}(\,\cdot\,|x),x}\big)\,d\mathrm{P}_X(x)\\
&\leq \int_X \big(b(x)\big)^{-\frac{1}{q}}\Big(\mathcal{C}_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}\Big)^{\frac{1}{q}}\,d\mathrm{P}_X(x)\\
&\leq \left(\int_X b^{-p} d\mathrm{P}_X\right)^{\frac{1}{qp}}\left(\int_X \mathcal{C}_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\rm sur},\mathrm{P}(\,\cdot\,|x),x}\,d\mathrm{P}_X(x)\right)^{1/q}\\
&= \|b^{-1}\|^{\frac{1}{q}}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L_{\rm sur},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\rm tar},\mathrm{P}}\big)^{1/q}\,.
\end{aligned}$$

Combining this estimate with our first estimate then gives the assertion in the case $\mathcal{R}_{L_{\rm tar},\mathrm{P}}(f) < \infty$. On the other hand, if $\mathcal{R}_{L_{\rm tar},\mathrm{P}}(f) = \infty$, we have

$B_f = \infty$. If $\bar{\delta}^{**}_{\infty}(\infty) = 0$, there is nothing to prove and hence we restrict our considerations to the case where $\bar{\delta}^{**}_{\infty}(\infty) > 0$. In this case, the proof of Theorem 3.22 has already shown that then there exist constants $c_1, c_2 \in (0, \infty)$ such that $t \leq c_1 \bar{\delta}^{**}_{\infty}(t) + c_2$ for all $t \in [0, \infty]$. From this we obtain

$$\begin{aligned}
\infty &= \mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} \\
&\leq c_1 \int_X \bar{\delta}^{**}_{\infty}\big(\mathcal{C}_{L_{\text{tar}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{P}(\,\cdot\,|x),x}\big)\, d\mathrm{P}_X(x) + c_2 \\
&\leq c_1 \int_X \big(b(x)\big)^{-\frac{1}{q}} \Big(\mathcal{C}_{L_{\text{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\text{sur}},\mathrm{P}(\,\cdot\,|x),x}\Big)^{\frac{1}{q}} d\mathrm{P}_X(x) + c_2\,,
\end{aligned}$$

where the last step is analogous to our considerations in the case $\mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) < \infty$. Using $b^{-1} \in L_p(\mathrm{P}_X)$ and Hölder's inequality, we then conclude from the estimate above that $\mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}} = \infty$. □

The condition $b^{-1} \in L_p(\mathrm{P}_X)$ in the preceding theorem measures how much the calibration function $\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)$ violates a uniform lower bound of the form $\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) \geq c\,\delta(\varepsilon)$, $\varepsilon \in [0, \infty]$. Indeed, the larger we can choose p in condition (3.28), the more the function b is away from the critical level 0, and thus the closer condition (3.28) is to a uniform lower bound. In the extremal case $p = \infty$, condition (3.28) actually becomes a uniform bound, and the inequality of Theorem 3.25 equals the inequality of Theorem 3.22. Finally, for $\delta(\varepsilon) := \varepsilon^r$, $\varepsilon \geq 0$, the function $\bar{\delta}(\varepsilon) := \delta^{\frac{p}{p+1}}(\varepsilon) = \varepsilon^{\frac{rp}{p+1}}$ is convex if $r \geq 1 + 1/p$. In this case, we can thus omit the Fenchel-Legendre biconjugate in Theorem 3.25 and obtain the simpler inequality

$$\mathcal{R}_{L_{\text{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} \;\leq\; \|b^{-1}\|^{1/r}_{L_p(\mathrm{P}_X)} \big(\mathcal{R}_{L_{\text{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\mathrm{P}}\big)^{1/r}\,.$$

Here, the condition $r \geq 1 + 1/p$ means that we have to increase the convexity of δ if we wish to weaken the uniformity of the calibration.

Our last goal in this section is to improve the inequalities above for the following type of loss, which will be of great utility in the next sections.

Definition 3.26. *Let $A \subset X \times \mathbb{R}$ and $h : X \to [0, \infty)$ be measurable. Then we call $L : X \times \mathbb{R} \to [0, \infty)$ a **detection loss** with respect to (A, h) if*

$$L(x, t) = \mathbf{1}_A(x, t)\, h(x)\,, \qquad\qquad x \in X\,, t \in \mathbb{R}.$$

Every detection loss function is obviously measurable and hence an unsupervised loss function. In addition, for $x \in X$ and $t \in \mathbb{R}$, we have

$$\mathcal{C}_{L,x}(t) - \mathcal{C}^*_{L,x} = \begin{cases} 0 & \text{if } A(x) := \{t' \in \mathbb{R} : (x, t') \in A\} = \mathbb{R} \\ \mathbf{1}_A(x, t) h(x) & \text{otherwise.} \end{cases} \tag{3.29}$$

Since detection losses will play an important role for *both* supervised and unsupervised learning scenarios let us now establish some specific results for this class of target loss function. We begin with the following theorem, whose proof is similar to the proof of Corollary 3.19 and hence is left as Exercise 3.7.

Theorem 3.27 (Asymptotic calibration for detection losses). *Let X be a complete measurable space and $L_{\mathrm{tar}} : X \times \mathbb{R} \to [0,\infty)$ be a detection loss with respect to some (A,h). Moreover, let $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss and $\mathcal{Q}$ be a set of distributions on Y. Then the following statements are equivalent:*

i) L_{sur} is L_{tar}-calibrated with respect to $\mathcal{Q}$.

ii) For all distributions P of type $\mathcal{Q}$ that satisfy $h \in L_1(\mathrm{P}_X)$ and $\mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} < \infty$ and all $\varepsilon \in (0,\infty]$, there exists a $\delta \in (0,\infty]$ such that for all measurable $f : X \to \mathbb{R}$ we have

$$\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} + \delta \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) < \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} + \varepsilon\,.$$

If the target loss is a detection loss, then we can, of course, establish calibration inequalities by Theorems 3.22 and 3.25. However, using the specific form of detection losses, one can often improve the resulting inequalities, as we will discuss after the following rather general theorem.

Theorem 3.28 (Calibration inequalities for detection losses). *Let X be a complete measurable space, $L_{\mathrm{tar}} : X \times \mathbb{R} \to [0,\infty)$ be a detection loss with respect to (A,h), $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} < \infty$. For $s > 0$, we write*

$$B(s) := \Big\{ x \in X : A(x) \neq \mathbb{R} \text{ and } \delta_{\max}\big(h(x), \mathrm{P}(\,\cdot\,|x), x\big) < s\,h(x) \Big\}\,.$$

If there exist constants $c > 0$ and $\alpha \in (0,\infty]$ such that

$$\int_X \mathbf{1}_{B(s)} h \, d\mathrm{P}_X \;\leq\; (c\,s)^{\alpha}\,, \qquad\qquad s > 0, \tag{3.30}$$

then for all measurable functions $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \;\leq\; 2\,c^{\frac{\alpha}{\alpha+1}} \big(\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}\big)^{\frac{\alpha}{\alpha+1}}\,.$$

Proof. We write $\mathcal{C}_{\mathrm{tar},x}(f) := \mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(f(x)) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}$ for $x \in X$ and measurable $f : X \to \mathbb{R}$. Furthermore, for $s > 0$, we write

$$C(s) := \big\{ x \in X : A(x) \neq \mathbb{R}, \text{ and } \delta_{\max}(h(x), \mathrm{P}(\,\cdot\,|x), x) \geq s\,h(x) \big\}\,.$$

By (3.16) and (3.29), we then obtain

$$\begin{aligned}
&\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \\
&= \int_{B(s)} \mathbf{1}_A(x, f(x))\, h(x)\, d\mathrm{P}_X(x) + \int_{C(s)} \mathbf{1}_A(x, f(x))\, h(x)\, d\mathrm{P}_X(x) \\
&\leq \int_X \mathbf{1}_{B(s)} h \, d\mathrm{P}_X + s^{-1} \int_{C(s)} \delta_{\max}\big(h(x), \mathrm{P}(\,\cdot\,|x), x\big)\, \mathbf{1}_A(x, f(x))\, d\mathrm{P}_X(x) \\
&\leq (c\,s)^{\alpha} + s^{-1} \int_{C(s)} \delta_{\max}\big(\mathcal{C}_{\mathrm{tar},x}(f), \mathrm{P}(\,\cdot\,|x), x\big)\, d\mathrm{P}_X(x) \\
&\leq (c\,s)^{\alpha} + s^{-1} \big(\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}\big)\,.
\end{aligned}$$

If $\alpha < \infty$, we now choose $s := (\alpha c^\alpha)^{-\frac{1}{\alpha+1}} (\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}})^{\frac{1}{\alpha+1}}$. Using $\alpha^{-\frac{\alpha}{\alpha+1}} + \alpha^{\frac{1}{\alpha+1}} \leq 2$ then yields the assertion. Furthermore, for $\alpha = \infty$, the assertion follows by setting $s^{-1} := 2c$. □

The preceding theorem can improve the inequalities we obtained for general target losses in various cases. The following two remarks illustrate this.

Remark 3.29. For detection losses with $h = \mathbf{1}_X$, Theorem 3.28 yields an improvement over Theorem 3.25. Indeed, if (3.28) is satisfied for $\delta(\varepsilon) = \varepsilon^q$ and a $b : X \to [0, \infty]$ with $b^{-1} \in L_p(\text{P}_X)$ and $q \geq \frac{p+1}{p}$, then Theorem 3.25 gives

$$\mathcal{R}_{L_{\text{tar}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\text{P}} \;\leq\; \|b^{-1}\|^{1/q}_{L_p(\text{P}_X)} \left(\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}}\right)^{1/q}. \tag{3.31}$$

On the other hand, some calculations show $B(s) \subset \{x \in X : b(x) < s\}$, and since $b^{-1} \in L_p(\text{P}_X)$ implies

$$\text{P}_X\big(\{x \in X : b(x) < s\}\big) \leq \|b^{-1}\|^p_p \, s^p\,, \qquad\qquad s > 0,$$

we find (3.30) for $c := \|b^{-1}\|_{L_p(\text{P}_X)}$ and $\alpha := p$. Theorem 3.28 thus yields

$$\mathcal{R}_{L_{\text{tar}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\text{P}} \;\leq\; 2\,\|b^{-1}\|^{\frac{p}{p+1}}_{L_p(\text{P}_X)} \left(\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}}\right)^{\frac{p}{p+1}}. \tag{3.32}$$

Now note that for $q > \frac{p+1}{p}$, (3.32) is sharper than (3.31) whenever the excess risk $\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}}$ is sufficiently small. ◁

Remark 3.30. In some cases, Theorem 3.28 also improves the inequalities of Theorem 3.22. Indeed, if L_{sur} is uniformly L_{tar}-calibrated with respect to some class $\mathcal{Q}$ of distributions and the uniform calibration function satisfies $\delta_{\max}(\,\cdot\,, \mathcal{Q}) \geq c_q \varepsilon^q$ for some $q > 1$, $c_q > 0$, and all $\varepsilon \geq 0$, then Theorem 3.22 gives

$$\mathcal{R}_{L_{\text{tar}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\text{P}} \;\leq\; c_q^{-1/q} \big(\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}}\big)^{1/q} \tag{3.33}$$

for all measurable functions $f : X \to \mathbb{R}$. However, an easy calculation shows that the assumptions above imply $B(s) \subset \{x \in X : 0 < h(x) < (s/c_q)^{1/(q-1)}\}$. Consequently, if we have constants $C > 0$ and $\beta \in (0, \infty]$ such that

$$\text{P}_X\big(\big\{x \in X : 0 < h(x) < s\big\}\big) \;\leq\; (C\,s)^\beta\,, \qquad\qquad s > 0, \tag{3.34}$$

then it is easy to check that (3.30) is satisfied for $c = c_q^{-1} C^{\frac{\beta q - \beta}{\beta+1}}$ and $\alpha := \frac{\beta+1}{q-1}$. Theorem 3.28 thus yields

$$\mathcal{R}_{L_{\text{tar}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{tar}},\text{P}} \;\leq\; 2\, c_q^{-\frac{\beta+1}{\beta+q}} C^{\frac{\beta q - \beta}{\beta+q}} \left(\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}}\right)^{\frac{\beta+1}{\beta+q}}. \tag{3.35}$$

Now note that for $q > 1$, we have $\frac{\beta+1}{\beta+q} > \frac{1}{q}$, and thus (3.35) is sharper than (3.33) whenever $\mathcal{R}_{L_{\text{sur}},\text{P}}(f) - \mathcal{R}^*_{L_{\text{sur}},\text{P}}$ is sufficiently small. ◁

3.4 Surrogates for Unweighted Binary Classification

In this section, we apply the general theory on surrogate loss functions developed in the previous sections to the standard binary classification scenario. The result of this section will be important for Section 8.5, where we investigate SVMs for classification that do not use the hinge loss as a surrogate.

Let us first recall (see Example 2.4) that in binary classification we consider the label space $Y := \{-1, 1\}$ together with the supervised loss L_{class}. In the following, we write $\mathcal{Q}_Y$ for the set of all distributions on Y. Moreover, recall that any distribution $\mathrm{Q} \in \mathcal{Q}_Y$ can be uniquely described by an $\eta \in [0,1]$ using the identification $\eta = \mathrm{Q}(\{1\})$. If $L : Y \times \mathbb{R} \to [0,\infty)$ is a supervised loss, we therefore use the notation

$$\begin{aligned} \mathcal{C}_{L,\eta}(t) &:= \mathcal{C}_{L,\mathrm{Q}}(t)\,, \qquad & t \in \mathbb{R}, \\ \mathcal{C}^*_{L,\eta} &:= \mathcal{C}^*_{L,\mathrm{Q}}\,, \end{aligned} \tag{3.36}$$

as well as $\mathcal{M}_{L,\eta}(0^+) := \mathcal{M}_{L,\mathrm{Q}}(0^+)$, $\mathcal{M}_{L,\eta}(\varepsilon) := \mathcal{M}_{L,\mathrm{Q}}(\varepsilon)$, and $\delta_{\max}(\varepsilon,\eta) := \delta_{\max}(\varepsilon,\mathrm{Q})$ for $\varepsilon \in [0,\infty]$. Note that, by the special structure of margin-based losses and the distributions $\mathrm{Q} \in \mathcal{Q}_Y$, we have the following symmetries:

$$\begin{aligned} \mathcal{C}_{L,\eta}(t) &= \mathcal{C}_{L,1-\eta}(-t) \qquad && \text{and} \qquad && \mathcal{C}^*_{L,\eta} = \mathcal{C}^*_{L,1-\eta}\,, \\ \mathcal{M}_{L,\eta}(\varepsilon) &= -\mathcal{M}_{L,1-\eta}(\varepsilon) \qquad && \text{and} \qquad && \mathcal{M}_{L,\eta}(0^+) = -\mathcal{M}_{L,1-\eta}(0^+)\,. \end{aligned}$$

Furthermore, it is interesting to note that the quantity $2\eta-1$, which will occur at many places in the following results, is the *expectation* of the corresponding Q, i.e., $\mathbb{E}\mathrm{Q} := \mathbb{E}_{\mathrm{Q}}\,\mathrm{id}_Y = 2\eta-1$. Before we present our first results, let us finally simplify our nomenclature.

Definition 3.31. *A supervised loss function $L : Y \times \mathbb{R} \to [0,\infty)$ is said to be* ***(uniformly) classification calibrated*** *if it is (uniformly) L_{class}-calibrated with respect to $\mathcal{Q}_Y$.*

Now our first aim is to compute the calibration function $\delta_{\max,L_{\text{class}},L}(\,\cdot\,,\eta)$ for supervised surrogates L of L_{class}.

Lemma 3.32 (Calibration function). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss. Then, for all $\eta \in [0,1]$ and $\varepsilon \in (0,\infty]$, we have*

$$\delta_{\max,L_{\text{class}},L}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \varepsilon > |2\eta-1| \\ \inf_{t\in\mathbb{R}:(2\eta-1)\operatorname{sign} t \le 0} \bigl(\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta}\bigr) & \text{if } \varepsilon \le |2\eta-1|. \end{cases}$$

Proof. The assertion immediately follows from the formula

$$\mathcal{M}_{L_{\text{class}},\eta}(\varepsilon) = \begin{cases} \mathbb{R} & \text{if } \varepsilon > |2\eta-1| \\ \{t \in \mathbb{R} : (2\eta-1)\operatorname{sign} t > 0\} & \text{if } 0 < \varepsilon \le |2\eta-1|\,, \end{cases}$$

which we derived in Example 3.8, and $\inf\emptyset = \infty$. □

The formula for the calibration function presented in Lemma 3.32 implies that $\delta_{\max}(\,\cdot\,,\eta)$ is a step function that only attains one value different from 0 and ∞. This particular form of the calibration function is the key ingredient of the following considerations on the relation between classification calibration and uniform classification calibration. We begin with a preliminary lemma.

Lemma 3.33 (Alternative to the calibration function). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a margin-based loss and $H : [0,1] \to [0,\infty)$ be defined by*

$$H(\eta) := \inf_{t\in\mathbb{R}:(2\eta-1)t\le 0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta}, \qquad \eta \in [0,1]. \tag{3.37}$$

Then the following statements are true:

i) L is classification calibrated if and only if $H(\eta) > 0$ for all $\eta \neq 1/2$.
ii) If L is continuous, we have $\delta_{\max}(\varepsilon,\eta) = H(\eta)$ for all $0 < \varepsilon \le |2\eta - 1|$.
iii) H is continuous and satisfies $H(\eta) = H(1-\eta)$, $\eta \in [0,1]$, and $H(1/2) = 0$.

Proof. i). Let us first assume that L is classification calibrated. We fix an $\eta \neq 1/2$. Then Lemma 3.32 together with $\operatorname{sign} 0 = 1$ shows $\mathcal{C}_{L,\eta}(0) > \mathcal{C}^*_{L,\eta}$ if $\eta \in [0,1/2)$. Moreover, if $\eta \in (1/2,1]$, we find the same inequality by

$$\mathcal{C}_{L,\eta}(0) - \mathcal{C}^*_{L,\eta} = \mathcal{C}_{L,1-\eta}(0) - \mathcal{C}^*_{L,1-\eta} > 0\,.$$

Finally, Lemma 3.32 yields

$$\inf_{t\in\mathbb{R}:(2\eta-1)t<0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} \;\ge\; \delta_{\max}(\varepsilon,\eta) \;>\; 0 \tag{3.38}$$

for $0 < \varepsilon \le |2\eta - 1|$, and hence we find $H(\eta) > 0$. Conversely, Lemma 3.32 gives $\delta_{\max}(\varepsilon,\eta) \ge H(\eta)$ for all $0 < \varepsilon \le |2\eta - 1|$, and hence L is classification calibrated if $H(\eta) > 0$ for all $\eta \neq 1/2$.

ii). Since there is nothing to prove in the case $\eta = 1/2$, we assume $\eta \neq 1/2$. Now, if L is continuous, then $\mathcal{C}_{L,\eta}(\,\cdot\,)$ is continuous at 0, and hence we have

$$\delta_{\max}(\varepsilon,\eta) \le \inf_{t\in\mathbb{R}:(2\eta-1)t<0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \inf_{t\in\mathbb{R}:(2\eta-1)t\le 0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = H(\eta)$$

by (3.38). Moreover, for $0 < \varepsilon \le |2\eta - 1|$, we always have $\delta_{\max}(\varepsilon,\eta) \ge H(\eta)$.

iii). The equality $H(1/2) = 0$ is trivial, and $H(\eta) = H(1-\eta)$, $\eta \in [0,1]$, immediately follows from symmetries mentioned at the beginning of this section. In order to prove the continuity of H, we now define

$$\begin{aligned} h(\eta) &= \inf_{t\in\mathbb{R}:(2\eta-1)t\le 0} \mathcal{C}_{L,\eta}(t)\,, \\ g^+(\eta) &= \inf_{t\le 0} \mathcal{C}_{L,\eta}(t)\,, \\ g^-(\eta) &= \inf_{t\ge 0} \mathcal{C}_{L,\eta}(t)\,, \end{aligned}$$

for $\eta \in [0,1]$. Then the functions $g^+ : [0,1] \to [0,\infty)$ and $g^- : [0,1] \to [0,\infty)$ can be defined by suprema taken over affine linear functions in $\eta \in$

$\mathbb{R}$, and since g^+ and g^- are also finite for $\eta \in [0,1]$, we find by Lemma A.6.4 that g^+ and g^- are continuous at every $\eta \in [0,1]$. Moreover, we have $\mathcal{C}^*_{L,\eta} = \min\{g^+(\eta), g^-(\eta)\}$ for all $\eta \in [0,1]$, and hence $\eta \mapsto \mathcal{C}^*_{L,\eta}$ is continuous. Finally, we have $h(\eta) = g^-(\eta)$ for $\eta \in [0,1/2)$, $h(\eta) = g^+(\eta)$ for $\eta \in (1/2,1]$, and $h(1/2) = \min\{g^+(1/2), g^-(1/2)\} = g^-(1/2) = g^+(1/2)$. This shows that $h : [0,1] \to [0,\infty)$ is continuous, and by combining these results we then obtain the continuity of H. □

Now we can establish the main result of this section, which shows that classification calibrated, margin-based losses are *uniformly* classification calibrated. In addition, it provides a lower bound of the Fenchel-Legendre biconjugate (see Definition 3.20) of the uniform calibration function $\delta_{\max}(\cdot, \mathcal{Q}_Y)$.

Theorem 3.34 (Classification calibration). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a margin-based loss. Then the following statements are equivalent:*

i) L is classification calibrated.
ii) L is uniformly classification calibrated.

Furthermore, for H defined by (3.37) and $\delta : [0,1] \to [0,\infty)$ defined by

$$\delta(\varepsilon) := H\Big(\frac{1+\varepsilon}{2}\Big), \qquad \varepsilon \in [0,1],$$

the Fenchel-Legendre bi-conjugates of δ and $\delta_{\max}(\cdot, \mathcal{Q}_Y)$ satisfy

$$\delta^{**}(\varepsilon) \le \delta^{**}_{\max, L_{\text{class}}, L}(\varepsilon, \mathcal{Q}_Y), \qquad \varepsilon \in [0,1], \tag{3.39}$$

and both quantities are actually equal if L is continuous. Finally, if L is classification calibrated, we have $\delta^{**}(\varepsilon) > 0$ for all $\varepsilon \in (0,1]$.

Proof. We begin with a preliminary consideration. To this end, let us fix an $\varepsilon \in (0,1]$. Then, by Lemma 3.32 and the symmetry of H around $1/2$, we find

$$\delta_{\max}(\varepsilon, \mathcal{Q}_Y) = \inf_{|2\eta-1| \ge \varepsilon} \delta_{\max}(\varepsilon, \eta) \ge \inf_{|2\eta-1| \ge \varepsilon} H(\eta) = \inf_{\eta \ge \frac{\varepsilon+1}{2}} H(\eta) =: \tilde{\delta}(\varepsilon),$$

and with $\tilde{\delta}(0) := 0$ we also have $\delta_{\max}(0, \mathcal{Q}_Y) = \tilde{\delta}(0)$.

i) $\Leftrightarrow$ *ii)*. Since *ii)* $\Rightarrow$ *i)* is trivial, it suffices to show *i)* $\Rightarrow$ *ii)*. To this end, recall that H is continuous and strictly positive on all intervals $[\frac{\varepsilon+1}{2}, 1]$, $\varepsilon \in (0,1]$, by Lemma 3.33, and consequently we have $\tilde{\delta}(\varepsilon) > 0$ for all $\varepsilon > 0$. From this we find $\delta_{\max}(\varepsilon, \mathcal{Q}_Y) > 0$ for all $\varepsilon > 0$ by our preliminary consideration.

In order to show (3.39), recall that $\tilde{\delta}(\varepsilon) \le \delta_{\max}(\varepsilon, \mathcal{Q}_Y)$ holds for all $\varepsilon \in [0,1]$, and hence we find $\tilde{\delta}^{**}(\varepsilon) \le \delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)$ for all $\varepsilon \in [0,1]$. Furthermore, we obviously have $\tilde{\delta}(\varepsilon) = \inf_{\varepsilon' \ge \varepsilon} \delta(\varepsilon')$, and hence Lemma A.6.21 gives $\delta^{**} = \tilde{\delta}^{**}$. In addition, if L is continuous, then our preliminary consideration together with Lemma 3.33 actually yields $\tilde{\delta}(\varepsilon) = \delta_{\max}(\varepsilon, \mathcal{Q}_Y)$ for all $\varepsilon \in [0,1]$. Repeating the arguments above thus shows $\delta^{**}(\varepsilon) = \delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)$ for all $\varepsilon \in [0,1]$.

Table 3.1. Some common margin-based losses and the corresponding values for $H(\eta)$, $\eta \in [0,1]$, and $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)$, $\varepsilon \in [0,1]$. All results easily follow from Theorem 3.36. For the logistic loss, we used the abbreviation $\Lambda(x) := x\ln(x)$. Note that if one wants to derive inequalities for the logistic loss using the above form of $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)$, it is useful to know that $\varepsilon^2 \le \Lambda(1+\varepsilon) + \Lambda(1-\varepsilon) \le \varepsilon^2 \ln 4$ for all $\varepsilon \in [0,1]$.

Loss function	$H(\eta)$	$\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)$
Least squares	$(2\eta-1)^2$	ε^2
Hinge loss	$\lvert 2\eta-1\rvert$	ε
Squared hinge	$(2\eta-1)^2$	ε^2
Logistic loss	$\ln 2 + \Lambda(\eta) + \Lambda(1-\eta)$	$\frac{1}{2}\big(\Lambda(1+\varepsilon) + \Lambda(1-\varepsilon)\big)$

Finally, if L is classification calibrated, we have already seen $\tilde{\delta}(\varepsilon) > 0$ for all $\varepsilon \in (0,1]$, and hence $\tilde{\delta}^{**}(\varepsilon) > 0$, $\varepsilon \in (0,1]$, by Lemma A.6.20. Since we have also proved $\delta^{**} = \tilde{\delta}^{**}$, we finally find $\delta^{**}(\varepsilon) > 0$, $\varepsilon \in (0,1]$. □

For classification calibrated margin-based losses L, the preceding theorem shows that using δ^{**} in Theorem 3.22 always gives non-trivial inequalities between the excess L-risk and the excess classification risk. Furthermore, Theorem 3.34 shows that in order to establish such inequalities it suffices to compute the function $H(\,\cdot\,)$ defined by (3.37), and as we will see later in Theorem 3.36, this computation is rather simple if L is convex. For the margin-based losses considered in the examples of Section 2.3, the functions H and $\delta^{**}_{\max}(\,\cdot\,, \mathcal{Q}_Y)$ are summarized in Table 3.1. Establishing the resulting inequalities is left as an exercise (see Exercise 3.9). However, note that for some losses these inequalities can be improved if the considered P satisfies an additional assumption, as the following remark shows (see also Theorem 8.29).

Remark 3.35. It is important to note that (3.9) can be used to describe the classification scenario by a detection loss. Indeed, if for a given distribution P on $X \times Y$ with $\eta(x) := \mathrm{P}(y=1|x)$, $x \in X$, we define

$$L_{\mathrm{P}}(x,t) := |2\eta(x)-1| \cdot \mathbf{1}_{(-\infty,0]}\big((2\eta(x)-1)\operatorname{sign} t\big)\,, \qquad x \in X\,, t \in \mathbb{R},$$

then $L_{\mathrm{P}} : X \times \mathbb{R} \to [0,\infty)$ is obviously a detection loss with respect to $A := \{(x,t) \in X \times \mathbb{R} : (2\eta(x)-1)\operatorname{sign} t \le 0\}$ and $h(x) = |2\eta(x)-1|$, $x \in X$. Furthermore, (3.9) states that

$$\mathcal{C}_{L_{\mathrm{class}},\eta(x)}(t) - \mathcal{C}^*_{L_{\mathrm{class}},\eta(x)} \;=\; \mathcal{C}_{L_{\mathrm{P}},x}(t) - \mathcal{C}^*_{L_{\mathrm{P}},x}$$

for all $x \in X$, $t \in \mathbb{R}$, i.e., for the distribution P, both losses describe the same learning goal. Now, condition (3.34) becomes

$$\mathrm{P}_X\big(\{x \in X : 0 < |2\eta(x) - 1| < s\}\big) \;\le\; (c\,s)^\beta\,, \qquad s > 0, \tag{3.40}$$

which, in a slightly stronger form, will be very important condition on P when establishing fast learning rates for SVMs in Section 8.3. For now, however, we would only like to mention that, assuming (3.40), we can immediately improve the inequalities that we would obtain by combining Theorem 3.34 with Theorem 3.22 for most of the margin-based losses considered in the examples. For more details, we refer to Remark 3.30 and Exercise 3.9. ◁

Up to now, we only know that the few examples listed in Table 3.1 are classification calibrated. The following theorem gives a powerful yet easy tool to check whether a *convex* margin-based loss is classification calibrated or not.

Theorem 3.36 (Test for classification calibration). *Let L be a convex, margin-based loss represented by $\varphi : \mathbb{R} \to [0,\infty)$. Then the following statements are equivalent:*

i) L is classification calibrated.
ii) φ is differentiable at 0 and $\varphi'(0) < 0$.

Furthermore, if L is classification calibrated, then the Fenchel-Legendre biconjugate of the uniform calibration function $\delta_{\max}(\,\cdot\,, \mathcal{Q}_Y)$ satisfies

$$\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) = \varphi(0) - \mathcal{C}^*_{L,\frac{\varepsilon+1}{2}}\,, \qquad \varepsilon \in [0,1]. \tag{3.41}$$

Proof. ii) $\Rightarrow$ i). Since φ is differentiable at 0, the map $t \to \mathcal{C}_{L,\eta}(t)$ is differentiable at 0 and its derivative is $\mathcal{C}'_{L,\eta}(0) = (2\eta-1)\varphi'(0)$. Consequently, we have $\mathcal{C}'_{L,\eta}(0) < 0$ for $\eta \in (1/2, 1]$. Now recall that the convexity of $\mathcal{C}_{L,\eta}(\,\cdot\,)$ implies that its derivative is almost everywhere defined and increasing by Theorem A.6.6 and Proposition A.6.12. Therefore, $\mathcal{C}_{L,\eta}(\,\cdot\,)$ is decreasing on $(-\infty, 0]$ and for $\eta \in (1/2, 1]$ we thus have

$$H(\eta) = \inf_{\substack{t\in\mathbb{R}:\\ (2\eta-1)t\le 0}} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \inf_{t\le 0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \mathcal{C}_{L,\eta}(0) - \mathcal{C}^*_{L,\eta}\,. \tag{3.42}$$

Furthermore, $\mathcal{C}'_{L,\eta}(0) < 0$ shows that $\mathcal{C}_{L,\eta}(\,\cdot\,)$ does not have a minimum at 0 and thus we find $H(\eta) > 0$ for all $\eta \in (1/2, 1]$. Lemma 3.33 then gives the classification calibration.

i) $\Rightarrow$ ii). Recall the basic facts on subdifferentials listed in Section A.6.2. Let us begin with assuming that φ is not differentiable at 0. Then there exist $w_1, w_2 \in \partial\varphi(0)$ with $w_1 < w_2$ and $w_1 \neq -w_2$. Let us fix an η with

$$\frac{1}{2} < \eta < \frac{1}{2} + \frac{w_2 - w_1}{2|w_1 + w_2|}\,.$$

Obviously, this choice implies $\frac{1}{2}(w_2 - w_1) > |w_1 + w_2|(\eta - \frac{1}{2})$, and by the definition of the subdifferential, we further have $\varphi(t) \ge w_i t + \varphi(0)$ for $t \in \mathbb{R}$ and $i = 1, 2$. For $t > 0$, we consequently find

$$\begin{aligned}\mathcal{C}_{L,\eta}(t) = \eta\varphi(t) + (1-\eta)\varphi(-t) &\geq \eta\big(w_2 t + \varphi(0)\big) + (1-\eta)\big(-w_1 t + \varphi(0)\big) \\ &= \left(\frac{1}{2}(w_2 - w_1) + (w_1 + w_2)\Big(\eta - \frac{1}{2}\Big)\right) t + \varphi(0) \\ &> \left(\Big(|w_1 + w_2| + (w_1 + w_2)\Big)\Big(\eta - \frac{1}{2}\Big)\right) t + \mathcal{C}_{L,\eta}(0) \\ &\geq \mathcal{C}_{L,\eta}(0)\,. \end{aligned} \tag{3.43}$$

Furthermore, since L is classification calibrated, we have $H(\eta) > 0$, and thus we find $\inf_{t>0} \mathcal{C}_{L,\eta}(t) = \mathcal{C}^*_{L,\eta}$. Together with (3.43), this shows $\mathcal{C}^*_{L,\eta} \geq \mathcal{C}_{L,\eta}(0)$. However, the latter yields $H(\eta) \leq 0$ by (3.42), and thus φ must be differentiable at 0. Let us now assume that $\varphi'(0) \geq 0$. We then obtain

$$\mathcal{C}_{L,1}(t) = \varphi(t) \geq \varphi'(0)t + \varphi(0) \geq \mathcal{C}_{L,1}(0)$$

for all $t > 0$. Again this contradicts the classification calibration of L.

In order to show (3.41), we first observe $\mathcal{C}'_{L,1/2}(0) = \frac{1}{2}\varphi'(0) - \frac{1}{2}\varphi'(0) = 0$. This immediately gives $\mathcal{C}_{L,1/2}(0) = \mathcal{C}^*_{L,1/2}$, and consequently we have

$$H(\eta) = \varphi(0) - \mathcal{C}^*_{L,\eta}\,, \qquad \eta \in [1/2, 1], \tag{3.44}$$

by (3.42) and $\mathcal{C}_{L,\eta}(0) = \varphi(0)$. Now recall that $\eta \to \mathcal{C}^*_{L,\eta}$ is defined by an infimum taken over affine linear functions, and hence it is a concave function. Consequently, H is convex on $[1/2, 1]$ and therefore (3.44) together with Theorem 3.34 and the continuity of L shows (3.41). □

3.5 Surrogates for Weighted Binary Classification

In this section, we investigate surrogate loss functions for the weighted binary classification scenario introduced in Example 2.5. To this end, recall that this scenario is characterized by the loss function

$$L_{\alpha\text{-class}}(y,t) = \begin{cases} 1-\alpha & \text{if } y = 1 \text{ and } t < 0 \\ \alpha & \text{if } y = -1 \text{ and } t \geq 0 \\ 0 & \text{otherwise}\,, \end{cases}$$

where $\alpha \in (0,1)$ was a fixed weighting parameter and $Y := \{-1, 1\}$. Adopting the notations around (3.36), we begin by computing $\delta_{\max}(\varepsilon, \eta)$.

Lemma 3.37 (Calibration function). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss. Then, for all $\alpha \in (0,1)$, $\eta \in [0,1]$, and $\varepsilon \in (0,\infty]$, we have*

$$\delta_{\max, L_{\alpha\text{-class}}, L}(\varepsilon, \eta) = \begin{cases} \infty & \text{if } \varepsilon > |\eta - \alpha| \\ \inf_{t \in \mathbb{R}:(\eta-\alpha)\operatorname{sign} t \leq 0} \big(\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta}\big) & \text{if } \varepsilon \leq |\eta - \alpha|. \end{cases}$$

Proof. For $t \in \mathbb{R}$, we have $\mathcal{C}_{L_{\alpha\text{-class}},\eta}(t) = (1-\alpha)\eta \mathbf{1}_{(-\infty,0)}(t) + \alpha(1-\eta)\mathbf{1}_{[0,\infty)}(t)$ and $\mathcal{C}^*_{L_{\alpha\text{-class}},\eta} = \min\{(1-\alpha)\eta, \alpha(1-\eta)\}$. From this we easily deduce

$$\mathcal{C}_{L_{\alpha\text{-class}},\eta}(t) - \mathcal{C}^*_{L_{\alpha\text{-class}},\eta} = |\eta - a| \cdot \mathbf{1}_{(-\infty,0]}\big((\eta-\alpha)\operatorname{sign} t\big)\,.$$

Now the assertion follows as in the proof of Lemma 3.32. □

In the following, we investigate how margin-based losses must be modified to make them $L_{\alpha\text{-class}}$-calibrated. To this end, let L be a margin-based loss represented by some $\varphi : \mathbb{R} \to [0,\infty)$. For $\alpha \in (0,1)$, we define the **α-weighted version** L_α of L by

$$L_\alpha(y,t) := \begin{cases} (1-\alpha)\varphi(t) & \text{if } y = 1 \\ \alpha\varphi(-t) & \text{if } y = -1, \end{cases} \qquad t \in \mathbb{R}.$$

Our next goal is to translate the results from the previous section for the unweighted classification scenario into results for the weighted case. To this end, we will frequently use the quantities

$$w_\alpha(\eta) := (1-\alpha)\eta + \alpha(1-\eta) \tag{3.45}$$

and

$$\vartheta_\alpha(\eta) := \frac{(1-\alpha)\eta}{(1-\alpha)\eta + \alpha(1-\eta)}\,, \tag{3.46}$$

which are defined for $\eta \in [0,1]$. Moreover, we need the following lemma, which describes the relation between the inner risks of L_α and L.

Lemma 3.38 (Weighted inner risks). *Let L be a margin-based loss. Then for $\alpha \in (0,1)$ and $\eta \in [0,1]$ the following statements are true:*

i) $\mathcal{C}_{L_\alpha,\eta}(t) = w_\alpha(\eta)\,\mathcal{C}_{L,\vartheta_\alpha(\eta)}(t)$ *for all* $t \in \mathbb{R}$, *and* $\mathcal{C}^*_{L_\alpha,\eta} = w_\alpha(\eta)\,\mathcal{C}^*_{L,\vartheta_\alpha(\eta)}$.
ii) $\min\{\alpha, 1-\alpha\} \le w_\alpha(\eta) \le \max\{\alpha, 1-\alpha\}$.
iii) If L is classification calibrated and $\eta \ne \alpha$, then $\mathcal{C}_{L_\alpha,\eta}(0) > \mathcal{C}^*_{L_\alpha,\eta}$.

Proof. i). A straightforward calculation shows $1-\vartheta_\alpha(\eta) = \frac{\alpha(1-\eta)}{(1-\alpha)\eta+\alpha(1-\eta)}$, and hence we obtain

$$\begin{aligned} \mathcal{C}_{L_\alpha,\eta}(t) &= (1-\alpha)\eta\varphi(t) + \alpha(1-\eta)\varphi(-t) \\ &= \big((1-\alpha)\eta + \alpha(1-\eta)\big)\big(\vartheta_\alpha(\eta)\varphi(t) + (1-\vartheta_\alpha(\eta))\varphi(-t)\big) \\ &= w_\alpha(\eta)\,\mathcal{C}_{L,\vartheta_\alpha(\eta)}(t)\,. \end{aligned}$$

ii). This follows from $w_\alpha(\eta) = (1-2\alpha)\eta + \alpha$.

iii). We have $\eta \ne \alpha$ if and only if $\vartheta_\alpha(\eta) \ne 1/2$. Furthermore, Lemma 3.33 showed $H(\eta) > 0$ for $\eta \ne 1/2$, where H is defined by (3.37), and hence we have $\mathcal{C}_{L,\eta}(0) > \mathcal{C}^*_{L,\eta}$ for $\eta \ne 1/2$. Therefore, the assertion follows from

$$\mathcal{C}_{L_\alpha,\eta}(0) = w_\alpha(\eta)\,\mathcal{C}_{L,\vartheta_\alpha(\eta)}(0) > w_\alpha(\eta)\,\mathcal{C}^*_{L,\vartheta_\alpha(\eta)} = \mathcal{C}^*_{L_\alpha,\eta}\,.$$

□

With the help of the preceding lemma, we can now characterize when α-weighted versions of margin-based loss functions are $L_{\alpha\text{-class}}$-calibrated.

Theorem 3.39 (Weighted classification calibration). *Let L be a margin-based loss function and $\alpha \in (0,1)$. We define $H_\alpha : [0,1] \to [0,\infty)$ by*

$$H_\alpha(\eta) := \inf_{t\in\mathbb{R}:(\eta-\alpha)t\le 0} \mathcal{C}_{L_\alpha,\eta}(t) - \mathcal{C}^*_{L_\alpha,\eta}, \qquad \eta \in [0,1]. \tag{3.47}$$

Then the following statements are equivalent:

i) L_α is uniformly $L_{\alpha\text{-class}}$-calibrated with respect to $\mathcal{Q}_Y$.
ii) L_α is $L_{\alpha\text{-class}}$-calibrated with respect to $\mathcal{Q}_Y$.
iii) L is classification calibrated.
iv) $H_\alpha(\eta) > 0$ for all $\eta \in [0,1]$ with $\eta \neq \alpha$.

Furthermore, if H is defined by (3.37) then, for all $\eta \in [0,1]$, we have

$$H_\alpha(\eta) = w_\alpha(\eta) H\big(\vartheta_\alpha(\eta)\big)\,. \tag{3.48}$$

Proof. ii) $\Leftrightarrow$ iii). An easy calculation shows $2\vartheta_\alpha(\eta) - 1 = \frac{\eta-\alpha}{(1-\alpha)\eta+\alpha(1-\eta)}$, and hence we find $\operatorname{sign}(\eta-\alpha) = \operatorname{sign}(2\vartheta_\alpha(\eta)-1)$. For $\varepsilon \le |\eta-\alpha|$, this gives

$$\begin{aligned}
\delta_{\max,L_{\alpha\text{-class}},L_\alpha}(\varepsilon,\eta) &= \inf_{\substack{t\in\mathbb{R}\\ (\eta-\alpha)\operatorname{sign} t\le 0}} \mathcal{C}_{L_\alpha,\eta}(t) - \mathcal{C}^*_{L_\alpha,\eta} \\
&= w_\alpha(\eta) \inf_{\substack{t\in\mathbb{R}\\ (2\vartheta_\alpha(\eta)-1)\operatorname{sign} t\le 0}} \mathcal{C}_{L,\vartheta_\alpha(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\alpha(\eta)} \\
&= w_\alpha(\eta)\, \delta_{\max,L_{\text{class}},L}(\varepsilon,\vartheta_\alpha(\eta))\,.
\end{aligned} \tag{3.49}$$

Since $w_\alpha(\eta) > 0$ and $\vartheta_\alpha([0,1]) = [0,1]$, we then obtain the equivalence. The proof of (3.48) is analogous to (3.49).

i) $\Rightarrow$ ii). Trivial.

iii) $\Rightarrow$ i). Recall that L is uniformly classification calibrated by Theorem 3.34. Then the implication follows from using $w_\alpha(\eta) \ge \min\{\alpha, 1-\alpha\}$ in (3.49).

ii) $\Rightarrow$ iv). Part *iii)* of Lemma 3.38 together with Lemma 3.37 implies $H_\alpha(\eta) > 0$ for all $\eta \neq \alpha$.

iv) $\Rightarrow$ ii). By Lemma 3.37, we have $\delta_{\max,\alpha}(\varepsilon,\eta) \ge H_\alpha(\eta) > 0$ for $\eta \neq \alpha$ and $0 < \varepsilon \le |\eta-\alpha|$. This gives the assertion. □

With the help of the results above we can now establish our main theorem of this section, which describes an easy way to establish inequalities for $L_{\alpha\text{-class}}$-calibrated loss functions.

Theorem 3.40 (Weighted uniform calibration function). *Let L be a margin-based loss and $\alpha \in (0,1)$. For $\alpha_{\max} := \max\{\alpha, 1-\alpha\}$, we define*

$$\delta_\alpha(\varepsilon) := \inf_{\substack{\eta\in[0,1]\\ |\eta-\alpha|\ge\varepsilon}} H_\alpha(\eta)\,, \qquad \varepsilon \in [0,\alpha_{\max}],$$

Table 3.2. The functions H, H_α, and δ_α^{**} for some common margin-based losses. The values for δ_α^{**} are only for α with $0 < \alpha \le 1/2$. Note that, for the hinge loss, the function δ_α^{**} is actually independent of α. Furthermore, the formulas for the logistic loss for classification do not fit into the table but can be easily computed.

Loss function	$H(\eta)$	$H_\alpha(\eta)$	$\delta_\alpha^{**}(\varepsilon)$
Least squares	$(2\eta-1)^2$	$\frac{(\eta-\alpha)^2}{\alpha+\eta-2\alpha\eta}$	$\frac{\varepsilon^2}{2\alpha(1-\alpha)+\varepsilon(1-2\alpha)}$
Hinge loss	$\lvert 2\eta-1\rvert$	$\lvert\eta-\alpha\rvert$	ε
Squared hinge	$(2\eta-1)^2$	$\frac{(\eta-\alpha)^2}{\alpha+\eta-2\alpha\eta}$	$\frac{\varepsilon^2}{2\alpha(1-\alpha)+\varepsilon(1-2\alpha)}$

where $H_\alpha(\,\cdot\,)$ is defined by (3.47). Then, for all $\varepsilon \in [0, \alpha_{\max}]$, we have

$$\delta_\alpha^{**}(\varepsilon) \le \delta^{**}_{\max, L_{\alpha\text{-class}}, L}(\varepsilon, \mathcal{Q}_Y)\,,$$

and if L is continuous, both quantities are actually equal.

Proof. Let $\varepsilon \in [0, \alpha_{\max}]$. Then Lemma 3.37 together with $\inf \emptyset = \infty$ yields

$$\inf_{Q\in\mathcal{Q}_Y} \delta_{\max}(\varepsilon, Q) = \inf_{\substack{\eta\in[0,1]\\ |\eta-\alpha|\ge\varepsilon}} \inf_{\substack{t\in\mathbb{R}\\ (\eta-\alpha)\operatorname{sign} t\le 0}} \mathcal{C}_{L_\alpha,\eta}(t) - \mathcal{C}^*_{L_\alpha,\eta} \ge \inf_{\substack{\eta\in[0,1]\\ |\eta-\alpha|\ge\varepsilon}} H_\alpha(\eta)\,.$$

□

Obviously, we can use the identity $H_\alpha(\eta) = w_\alpha(\eta) H\big(\vartheta_\alpha(\eta)\big)$ in order to compute the function $\delta_\alpha(\varepsilon)$ of the preceding theorem. Doing so, we see that δ_α is a continuous function that is strictly positive on $(0, \alpha_{\max}]$ if L is classification calibrated. Consequently, Theorem 3.40 together with Theorem 3.22 yields non-trivial inequalities. Furthermore, for some important loss functions, we already know $H(\eta)$, $\eta \in [0,1]$, and hence the computation of $\delta_\alpha^{**}(\varepsilon)$ is straightforward. The corresponding results are summarized in Table 3.2.

Up to now, we have only investigated the $L_{\alpha\text{-class}}$-calibration of α-weighted versions of classification calibrated loss functions. We finally show that other weighted versions are *not* $L_{\alpha\text{-class}}$-calibrated.

Theorem 3.41 (Using the correct weights). *Let $\alpha, \beta \in (0,1)$, L be a margin-based, classification calibrated loss, and L_β be its β-weighted version. Then L_β is $L_{\alpha\text{-class}}$-calibrated if and only if $\beta = \alpha$.*

Proof. We already know that L_α is $L_{\alpha\text{-class}}$-calibrated, and hence we assume $\alpha \ne \beta$. Without loss of generality, we only consider the case $\beta > \alpha$. For a fixed $\eta \in (\alpha, \beta)$, an easy computation then shows that $\vartheta_\beta(\eta)$ defined in (3.46) satisfies $\vartheta_\beta(\eta) < 1/2 < \vartheta_\alpha(\eta)$, and hence for $\varepsilon > 0$ with $\varepsilon \le |\eta - \alpha|$ we obtain

$$\begin{aligned}\delta_{\max, L_{\alpha\text{-class}}, L_\beta}(\varepsilon, \eta) &= \inf_{(n-\alpha)\operatorname{sign} t\le 0} \mathcal{C}_{L_\beta,\eta}(t) - \mathcal{C}^*_{L_\beta,\eta} \\ &= w_\beta(\eta) \inf_{t<0} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)}\,. \qquad (3.50)\end{aligned}$$

The classification calibration of L implies $\inf_{t\geq 0} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)} > 0$, and since $\inf_{t\in\mathbb{R}} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)} = 0$, we find $\inf_{t<0} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)} = 0$. Together with (3.50), this shows that L_β is not $L_{\alpha\text{-class}}$-calibrated. □

The preceding theorem in particular shows that an α-weighted version of a classification calibrated margin-based loss function is classification calibrated if and only if $\alpha = 1/2$. In other words, using a weighted margin-based loss for an unweighted classification problem may lead to methodical errors.

3.6 Template Loss Functions

Sometimes an unsupervised loss function explicitly depends on the data-generating distribution. For example, if we have a distribution P on $X \times \mathbb{R}$ with $|\mathrm{P}|_1 < \infty$ and we wish to estimate the conditional mean function $x \mapsto \mathbb{E}_\mathrm{P}(Y|x)$, we could describe this learning goal by the loss function

$$L(x,t) := \big|\mathbb{E}_\mathrm{P}(Y|x) - t\big|\,, \qquad\qquad x \in X,\, t \in \mathbb{R}.$$

Now note that when we change the distribution we have to change the loss function, though the learning goal remains the same. In view of our analysis on surrogate losses, this fact is at least annoying. The goal of this section is to resolve this issue by introducing a new type of "loss function" that may depend on distributions Q. Let us begin with a precise definition.

Definition 3.42. *Let $\mathcal{Q}$ be a set of distributions on a closed subset $Y \subset \mathbb{R}$. Then we call a function $L : \mathcal{Q} \times \mathbb{R} \to [0,\infty)$ a* ***template loss*** *if, for all complete measurable spaces X and all distributions* P *of type $\mathcal{Q}$, the* P***-instance*** *L_P of L defined by*

$$\begin{aligned} L_\mathrm{P} : X \times \mathbb{R} &\to [0,\infty) \\ (x,t) &\mapsto L\big(\mathrm{P}(\,\cdot\,|x), t\big) \end{aligned} \tag{3.51}$$

is measurable.

Note that the key condition of this definition is the *measurability*, which enables us to interpret P-instances as unsupervised losses. In particular, we can define the risk of a template loss L by the risk of its P-instance, i.e., by

$$\mathcal{R}_{L,\mathrm{P}}(f) := \mathcal{R}_{L_\mathrm{P},\mathrm{P}}(f) = \int_X L\big(\mathrm{P}(\,\cdot\,|x), f(x)\big)\, d\mathrm{P}_X(x)\,,$$

where $f : X \to \mathbb{R}$ is measurable. This motivates us to define the inner risks of a template loss $L : \mathcal{Q} \times \mathbb{R} \to [0,\infty)$ analogously to the inner risks of unsupervised losses, i.e., we write

$$\begin{aligned} \mathcal{C}_{L,\mathrm{Q}}(t) &:= L(\mathrm{Q},t), \\ \mathcal{C}^*_{L,\mathrm{Q}} &:= \inf_{t'\in\mathbb{R}} L(\mathrm{Q},t') \end{aligned}$$

for $\mathrm{Q} \in \mathcal{Q}$ and $t \in \mathbb{R}$. Note that the right-hand sides of these definitions have the form we used for unsupervised losses in the sense that no integrals occur while the left-hand sides have the form we obtained for supervised losses in the sense that the inner risks are independent of x. Having defined the inner risks, we write, as usual,

$$\mathcal{M}_{L,\mathrm{Q}}(\varepsilon) := \left\{ t \in \mathbb{R} : \mathcal{C}_{L,\mathrm{Q}}(t) < \mathcal{C}^*_{L,\mathrm{Q}} + \varepsilon \right\}, \qquad \mathrm{Q} \in \mathcal{Q},\ \varepsilon \in [0,\infty],$$

for the corresponding sets of approximate minimizers. Moreover, given a *supervised* surrogate loss $L_{\mathrm{sur}} : Y \times \mathbb{R} \to [0,\infty)$, we define the calibration function $\delta_{\max}(\,\cdot\,, \mathrm{Q}) : [0,\infty] \to [0,\infty]$ of (L, L_{sur}) by

$$\delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathrm{Q}) := \inf_{\substack{t \in \mathbb{R} \\ t \notin \mathcal{M}_{L,\mathrm{Q}}(\varepsilon)}} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q}}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q}}, \qquad \varepsilon \in [0,\infty],$$

if $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q}} < \infty$ and by $\delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathrm{Q}) := \infty$ otherwise. Since in the proof of Lemma 3.14 we did not use that the inner risks are defined by integrals, it is then not hard to see that this lemma also holds for the calibration function above. Consequently, we say that L_{sur} is **L-calibrated** with respect to $\mathcal{Q}$ if

$$\delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathrm{Q}) > 0$$

for all $\varepsilon > 0$ and $\mathrm{Q} \in \mathcal{Q}$. Analogously, we say that L_{sur} is **uniformly L-calibrated** with respect to $\mathcal{Q}$ if

$$\delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathcal{Q}) := \inf_{\mathrm{Q} \in \mathcal{Q}} \delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathrm{Q}) > 0$$

for all $\varepsilon > 0$. If we now consider a P-instance L_{P} of L, we immediately obtain

$$\delta_{\max,L_{\mathrm{P}},L_{\mathrm{sur}}}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = \delta_{\max,L,L_{\mathrm{sur}}}(\varepsilon, \mathrm{P}(\,\cdot\,|x)) \tag{3.52}$$

for all $\varepsilon \in [0,\infty]$ and $x \in X$, where $\delta_{\max,L_{\mathrm{P}},L_{\mathrm{sur}}}(\,\cdot\,,\,\cdot\,,\,\cdot\,)$ denotes the calibration function of $(L_{\mathrm{P}}, L_{\mathrm{sur}})$. In other words, L-calibration of L_{sur} can be investigated analogously to *supervised* losses, i.e., in terms of $\mathcal{Q}$ and independent of x, while the corresponding results can be used to determine the relation between the excess L_{sur}-risk and the excess risk of the *unsupervised* loss L_{P}. In the following sections, we will extensively make use of template losses, mainly because of this technical merit.

3.7 Surrogate Losses for Regression Problems

In regression, the goal is to predict a real-valued output y given an input x. The discrepancy between the prediction $f(x)$ and the observation y is often measured by the least squares loss, but we have already seen in Section 2.4 that there are various alternatives. In this section, we investigate the relation

of these alternatives to the least squares loss. These considerations will be important for Chapters 9 and 10 on regression and robustness, respectively.

Let us begin by introducing some notation. To this end let, $\mathcal{Q}$ be a set of distributions on $\mathbb{R}$ and $L : \mathbb{R} \times \mathbb{R} \to [0, \infty]$ be a supervised loss. Since in our general results on calibration the assumption $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$ was crucial, we define

$$\mathcal{Q}(L) := \left\{ \mathrm{Q} \in \mathcal{Q} : \mathcal{C}^*_{L,\mathrm{Q}} < \infty \right\}.$$

Recall that for distance-based losses we have investigated the condition $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$ in Lemma 2.36. In the following, $\mathcal{Q}_{\mathbb{R}}$ denotes the set of distributions on $\mathbb{R}$, and more generally, $\mathcal{Q}_I$ denotes the set of all distributions whose support is contained in the subset $I \subset \mathbb{R}$. In addition, for $p \in (0, \infty]$, the set of distributions on $\mathbb{R}$ with p-th finite moment is denoted by

$$\mathcal{Q}^{(p)}_{\mathbb{R}} := \left\{ \mathrm{Q} : \mathrm{Q} \text{ distribution on } \mathbb{R} \text{ with } |\mathrm{Q}|_p < \infty \right\},$$

whereas the set of all distributions with bounded support is denoted by

$$\mathcal{Q}_{\text{bounded}} := \mathcal{Q}^{(\infty)}_{\mathbb{R}} = \bigcup_{M>0} \mathcal{Q}_{[-M,M]}.$$

Note that $\mathcal{Q}_I \subset \mathcal{Q}_{\text{bounded}} \subset \mathcal{Q}^{(1)}_{\mathbb{R}}$ holds for all bounded intervals I, and if L is a continuous, distance-based loss, we actually have $\mathcal{Q}_{\text{bounded}} \subset \mathcal{Q}^{(1)}_{\mathbb{R}}(L)$.

Now let Q be a distribution on $\mathbb{R}$ such that $|\mathrm{Q}|_1 < \infty$. Then the **mean** of Q is denoted by

$$\mathbb{E}\mathrm{Q} := \int_{\mathbb{R}} y \, d\mathrm{Q}(y).$$

We call Q **symmetric** around some $c \in \mathbb{R}$ if $\mathrm{Q}(c + A) = \mathrm{Q}(c - A)$ for all measurable $A \subset [0, \infty)$. Furthermore, we say that Q is symmetric if it is symmetric around some $c \in \mathbb{R}$. Obviously, Q is symmetric around c if and only if its **centered version** $\mathrm{Q}^{(c)}$ defined by $\mathrm{Q}^{(c)}(A) := \mathrm{Q}(c + A)$, $A \subset \mathbb{R}$ measurable, is centered around 0. In the following, the set of all symmetric distributions with p-finite moment is denoted by $\mathcal{Q}^{(p)}_{\mathbb{R},\text{sym}}$. Finally, the sets $\mathcal{Q}_{I,\text{sym}}$, for $I \subset \mathbb{R}$, and $\mathcal{Q}_{\text{bounded,sym}}$ are defined in the obvious way.

Let us now assume that Q is symmetric around c. For a measurable function $h : \mathbb{R} \to \mathbb{R}$, we then have

$$\begin{aligned} \int_{\mathbb{R}} h(y - c) d\mathrm{Q}(y) \;&=\; \int_{\mathbb{R}} h(y) d\mathrm{Q}^{(c)}(y) = \int_{\mathbb{R}} h(-y) d\mathrm{Q}^{(c)}(y) \\ &= \int_{\mathbb{R}} h(c - y) d\mathrm{Q}(y) \qquad (3.53) \end{aligned}$$

whenever one (and then all) of the integrals exists. In particular, for $h(y) := y + c$, $y \in \mathbb{R}$, and Q satisfying $|\mathrm{Q}|_1 < \infty$, this equation yields

$$\mathbb{E}\mathrm{Q} = \int_{\mathbb{R}} y d\mathrm{Q}(y) = \int_{\mathbb{R}} h(y - c) d\mathrm{Q}(y) = c + \int_{\mathbb{R}} y d\mathrm{Q}^{(c)}(y) = c,$$

i.e., the center c is unique and equals the mean $\mathbb{E}\mathrm{Q}$.

Let us get back to our main goal, which is identifying L_{LS}-calibrated losses, where L_{LS} denotes the least squares loss. To this end, recall that for $\mathrm{Q} \in \mathcal{Q}_{\mathbb{R}}(L_{\mathrm{LS}}) = \mathcal{Q}_{\mathbb{R}}^{(2)}$ we have already seen in Example 2.6 that

$$\mathcal{M}_{L_{\mathrm{LS}},\mathrm{Q}}(0^+) = \{\mathbb{E}\mathrm{Q}\}\,.$$

Consequently, if L is a supervised, L_{LS}-calibrated loss function, we must have $\mathcal{M}_{L,\mathrm{Q}}(0^+) \subset \{\mathbb{E}\mathrm{Q}\}$ for all $\mathrm{Q} \in \mathcal{Q}_{\mathbb{R}}^{(2)}(L)$. This observation motivates the following two propositions in which we investigate the sets $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ for distance-based losses.

Proposition 3.43 (Exact minimizers for distance-based losses I). *Let L be a distance-based loss whose representing function $\psi : \mathbb{R} \to [0,\infty)$ satisfies $\lim_{r\to\pm\infty} \psi(r) = \infty$. Moreover, let $\mathrm{Q} \in \mathcal{Q}_{\mathbb{R}}$ be a distribution with $\mathcal{C}_{L,\mathrm{Q}}(t) < \infty$ for all $t \in \mathbb{R}$. Then the following statements are true:*

i) If ψ is convex, then $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t)$ is convex and continuous. Moreover, we have $\lim_{t\to\pm\infty} \mathcal{C}_{L,\mathrm{Q}}(t) = \infty$ and $\mathcal{M}_{L,\mathrm{Q}}(0^+) \neq \emptyset$.

ii) If ψ is strictly convex, then $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t)$ is strictly convex and $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ contains exactly one element.

Proof. Our first goal is to show that $\lim_{t\to\pm\infty} \mathcal{C}_{L,\mathrm{Q}}(t) = \infty$. To this end, we fix a $B > 0$ and let $(t_n) \subset \mathbb{R}$ be a sequence with $t_n \to -\infty$. Since $\lim_{r\to\pm\infty} \psi(r) = \infty$, there then exists an $r_0 > 0$ such that $\psi(r) \geq 2B$ for all $r \in \mathbb{R}$ with $|r| \geq r_0$. Since $\mathrm{Q}(\mathbb{R}) = 1$, there exists also an $M > 0$ with $\mathrm{Q}([-M,M]) \geq 1/2$. Finally, there exists an $n_0 \geq 1$ with $t_n \leq -M - r_0$ for all $n \geq n_0$. For $y \in [-M,M]$, this yields $y - t_n \geq r_0$, and hence we find $\psi(y - t_n) \geq 2B$ for all $n \geq n_0$. From this we easily conclude

$$\mathcal{C}_{L,\mathrm{Q}}(t_n) \geq \int_{[-M,M]} \psi(y - t_n)\, d\mathrm{Q}(y) \geq 2B\, \mathrm{Q}\big([-M,M]\big) = B\,,$$

i.e., we have shown $\mathcal{C}_{L,\mathrm{Q}}(t_n) \to \infty$. Analogously we can show $\lim_{t\to\infty} \mathcal{C}_{L,\mathrm{Q}}(t) = \infty$, and consequently we have $\lim_{t\to\pm\infty} \mathcal{C}_{L,\mathrm{Q}}(t) = \infty$. This shows that

$$\big\{t \in \mathbb{R} : \mathcal{C}_{L,\mathrm{Q}}(t) \leq \mathcal{C}_{L,\mathrm{Q}}(0)\big\}$$

is a non-empty and bounded subset of $\mathbb{R}$. Furthermore, the convexity of ψ implies that $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t)$ is convex and hence this map is continuous by Lemma A.6.2. Now the assertions follow from Theorem A.6.9. □

Note that for distributions $\mathrm{Q} \in \mathcal{Q}_{\mathrm{bounded}}$ we automatically have $\mathcal{C}_{L,\mathrm{Q}}(t) < \infty$ for all $t \in \mathbb{R}$ and all distance-based losses L. Furthermore, if L is of some growth type $p \in (0,\infty)$, then Lemma 2.36 shows $\mathcal{C}_{L,\mathrm{Q}}(t) < \infty$ for all $t \in \mathbb{R}$ and all distributions Q having finite p-th moment. Consequently, the preceding proposition gives $\mathcal{M}_{L,\mathrm{Q}}(0^+) \neq \emptyset$ in both cases.

The following proposition compares $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ with the mean $\mathbb{E}\mathrm{Q}$.

Proposition 3.44 (Exact minimizers for distance-based losses II). *Let L be a distance-based loss whose representing function ψ is locally Lipschitz continuous, and let $M > 0$. Then the following statements are true:*

i) If $\mathbb{E}\mathrm{Q} \in \mathcal{M}_{L,\mathrm{Q}}(0^+)$ for all $\mathrm{Q} \in \mathcal{Q}_{[-M,M],\mathrm{sym}}$, then L is symmetric.
ii) If $\mathbb{E}\mathrm{Q} \in \mathcal{M}_{L,\mathrm{Q}}(0^+)$ for all $\mathrm{Q} \in \mathcal{Q}_{\mathrm{bounded}}$, then there exists a constant $c \geq 0$ with $\psi(t) = ct^2$ for all $t \in \mathbb{R}$.

Proof. Recall that the fundamental theorem of calculus for Lebesgue integrals (see Theorem A.6.6) shows that the derivative ψ' is (Lebesgue)-almost surely defined and integrable on every bounded interval.

i). Let us fix a $y \in [-M, M]$ such that ψ is differentiable at y and $-y$. We define $\mathrm{Q} := \frac{1}{2}\delta_{\{-y\}} + \frac{1}{2}\delta_{\{y\}}$. Then we have $\mathrm{Q} \in \mathcal{Q}_{\mathbb{R},\mathrm{sym}}$ with $\mathbb{E}\mathrm{Q} = 0$, and $\mathcal{C}_{L,\mathrm{Q}}(t) = \frac{1}{2}\psi(-y-t)+\frac{1}{2}\psi(y-t)$. Consequently, the derivative of $\mathcal{C}_{L,\mathrm{Q}}(\,\cdot\,)$ exists at 0 and can be computed by $\mathcal{C}'_{L,\mathrm{Q}}(0) = -\frac{1}{2}\psi'(-y) - \frac{1}{2}\psi'(y)$. Furthermore, our assumption shows that $\mathcal{C}_{L,\mathrm{Q}}(\,\cdot\,)$ has a minimum at 0, and hence we have $0 = \mathcal{C}'_{L,\mathrm{Q}}(0)$, i.e., $\psi'(-y) = -\psi'(y)$. According to our preliminary remark, the latter relation holds for almost all y, and hence Theorem A.6.6 shows that, for all $y_0 \in \mathbb{R}$, we have

$$\begin{aligned}\psi(y_0) \;=\; \psi(0) + \int_0^{y_0} \psi'(t)dt \;=\; \psi(0) - \int_0^{y_0} \psi'(-t)dt = \psi(0) - \int_{-y_0}^{0} \psi'(t)dt \\ = \psi(-y_0)\,.\end{aligned}$$

ii). Let $y \neq 0$ and $\alpha > 0$ be real numbers such that ψ is differentiable at y, $-y$, and αy. We define $\mathrm{Q} := \frac{\alpha}{1+\alpha}\delta_{\{0\}} + \frac{1}{1+\alpha}\delta_{\{(1+\alpha)y\}}$, so that we obtain $\mathbb{E}\mathrm{Q} = y$ and $\mathcal{C}_{L,\mathrm{Q}}(t) = \frac{\alpha}{1+\alpha}\psi(-t) + \frac{1}{1+\alpha}\psi(y+\alpha y-t)$ for all $t \in \mathbb{R}$. This shows that the derivative of $\mathcal{C}_{L,\mathrm{Q}}(\,\cdot\,)$ exists at y and can be computed by

$$\mathcal{C}'_{L,\mathrm{Q}}(y) = -\frac{\alpha}{1+\alpha}\psi'(-y) - \frac{1}{1+\alpha}\psi'(\alpha y) = \frac{\alpha}{1+\alpha}\psi'(y) - \frac{1}{1+\alpha}\psi'(\alpha y)\,,$$

where in the last step we used *i)*. Now, our assumption $\mathbb{E}\mathrm{Q} \in \mathcal{M}_{L,\mathrm{Q}}(0^+)$ gives $\mathcal{C}'_{L,\mathrm{Q}}(y) = 0$, and hence we find $\alpha\psi'(y) = \psi'(\alpha y)$. Obviously, the latter relation holds for almost all $\alpha > 0$, and thus we obtain

$$\psi(ty) = \psi(0) + \int_0^t \psi'(sy)y\,ds = \int_0^t s\psi'(y)y\,ds = \frac{\psi'(y)}{2y}(ty)^2$$

for all $t > 0$. From this we easily obtain the assertion for $c := \frac{\psi'(y)}{2y}$. □

Proposition 3.44 shows that there is basically *no* distance-based surrogate for the least squares loss L_{LS} if one is interested in the entire class

$$\mathcal{Q}_{\mathbb{R}}(L_{\mathrm{LS}}) = \mathcal{Q}_{\mathbb{R}}^{(2)} = \big\{\mathrm{Q} \in \mathcal{Q}_{\mathbb{R}} : |\mathrm{Q}|_2 < \infty\big\}\,.$$

Furthermore, it shows that the least squares loss is essentially the only distance-based loss function whose minimizer is the mean for all distributions

in $\mathcal{Q}_{\mathbb{R}}^{(2)}$. In other words, if we are actually interested in finding the **regression function** $x \mapsto \mathbb{E}_P(Y|x)$, and we just know $|P|_2 < \infty$, then the least squares loss is the *only* suitable distance-based loss for this task. However, if we cannot ensure the tail assumption $|P|_2 < \infty$ but know instead that the conditional distributions $P(\cdot\,|x)$ are *symmetric*, then Proposition 3.44 suggests that we may actually have alternatives to the least squares loss. In order to investigate this conjecture systematically, we first need a target loss that describes the goal of estimating the mean. To this end, let us consider the **mean distance** template loss $L_{\text{mean}} : \mathcal{Q}_{\mathbb{R}}^{(1)} \times \mathbb{R} \to [0,\infty)$, which is defined by

$$L_{\text{mean}}(Q,t) := |\mathbb{E}Q - t|\,, \qquad t \in \mathbb{R}\,, Q \in \mathcal{Q}_{\mathbb{R}}^{(1)}\,.$$

Note that this indeed defines a template loss, since given a $\mathcal{Q}_{\mathbb{R}}^{(1)}$-type distribution P on $X \times \mathbb{R}$, it is easy to see that

$$(x,t) \mapsto L_{\text{mean}}(P(\cdot\,|x),t) = \bigl|\mathbb{E}_P(Y|x) - t\bigr|$$

is measurable. Moreover, we have

$$L_{\text{mean}}^2(Q,t) = \bigl(\mathbb{E}Q - t\bigr)^2 = \mathcal{C}_{L_{\text{LS}},Q}(t) - \mathcal{C}^*_{L_{\text{LS}},Q}\,, \qquad Q \in \mathcal{Q}_{\mathbb{R}}^{(2)},\, t \in \mathbb{R},$$

and since the minimal L_{mean}-risks equal 0, we thus obtain $\mathcal{M}_{L_{\text{mean}},Q}(\sqrt{\varepsilon}) = \mathcal{M}_{L_{\text{LS}},Q}(\varepsilon)$ for all $\varepsilon > 0$. From this we immediately find

$$\delta_{\max,L_{\text{mean}},L}(\sqrt{\varepsilon},Q) = \delta_{\max,L_{\text{LS}},L}(\varepsilon,Q)\,, \qquad \varepsilon \in [0,\infty], \tag{3.54}$$

for all distance-based losses L and all $Q \in \mathcal{Q}_{\mathbb{R}}^{(2)} \cap \mathcal{Q}_{\mathbb{R}}(L)$. In other words, by considering L_{mean}-calibration, we simultaneously obtain results on L_{LS}-calibration.

We saw in Section 3.1 that the inner risks are the key quantities for computing calibration functions. The following lemma presents a way to compute the inner risks $\mathcal{C}_{L,Q}(\cdot\,)$ when both L and Q are symmetric.

Lemma 3.45 (Inner risks of symmetric losses). *Let L be a symmetric loss with representing function ψ and $Q \in \mathcal{Q}_{\mathbb{R},\text{sym}}^{(1)}(L)$. Then we have*

$$\mathcal{C}_{L,Q}(\mathbb{E}Q + t) = \mathcal{C}_{L,Q}(\mathbb{E}Q - t) = \frac{1}{2}\int_{\mathbb{R}} \psi(y - \mathbb{E}Q - t) + \psi(y - \mathbb{E}Q + t)\,dQ(y)$$

for all $t \in \mathbb{R}$. In addition, if L is convex, we have

$$\mathcal{C}_{L,Q}(\mathbb{E}Q) = \mathcal{C}^*_{L,Q}\,,$$

and if L is strictly convex, we also have $\mathcal{C}_{L,Q}(\mathbb{E}Q + t) > \mathcal{C}^*_{L,Q}$ for all $t \neq 0$.

Proof. Let us write $m := \mathbb{E}Q$. Recalling that the centered version $Q^{(m)}$ of Q is symmetric around 0, the symmetry of ψ and (3.53) then yield

$$\begin{aligned}\mathcal{C}_{L,\mathrm{Q}}(m+t) = \int_{\mathbb{R}} \psi(y-t)d\mathrm{Q}^{(m)}(y) &= \int_{\mathbb{R}} \psi(-y-t)d\mathrm{Q}^{(m)}(y) \\ &= \int_{\mathbb{R}} \psi(y+t)d\mathrm{Q}^{(m)}(y) \\ &= \mathcal{C}_{L,\mathrm{Q}}(m-t)\,.\end{aligned}$$

Since this yields $\mathcal{C}_{L,\mathrm{Q}}(m+t) = \frac{1}{2}(\mathcal{C}_{L,\mathrm{Q}}(m+t) + \mathcal{C}_{L,\mathrm{Q}}(m-t))$, we also obtain the second equation. Furthermore, if ψ is convex, we can easily conclude that

$$\mathcal{C}_{L,\mathrm{Q}}(m+t) = \frac{1}{2}\int_{\mathbb{R}} \psi(y-t)+\psi(y+t)d\mathrm{Q}^{(m)}(y) \geq \int_{\mathbb{R}} \psi(y)d\mathrm{Q}^{(m)}(y) = \mathcal{C}_{L,\mathrm{Q}}(m)$$

for all $t \in \mathbb{R}$. This shows the second assertion. The third assertion can be shown analogously. □

With the help of the preceding lemma, we can derive a simple formula for the calibration function $\delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q})$ if L is convex.

Lemma 3.46 (Calibration function for symmetric losses). *Let L be a symmetric, convex loss and $\mathrm{Q} \in \mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}}(L)$. Then, for all $\varepsilon \geq 0$, we have*

$$\delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q}) = \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+\varepsilon) - \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q})\,. \tag{3.55}$$

Consequently, $\varepsilon \mapsto \delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q})$ is convex and the following statements are equivalent:

i) $\delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q}) > 0$ for all $\varepsilon > 0$.
ii) $\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t) > \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q})$ for all $t \in \mathbb{R}$ with $t \neq 0$.

Proof. Obviously, it suffices to prove (3.55). To this end, observe that $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t)$ is a convex function on $\mathbb{R}$, and Lemma 3.45 shows that it is also symmetric in the sense of $\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t) = \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}-t)$ for all $t \in \mathbb{R}$. Therefore, $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t)$ is decreasing on $(\infty,0]$ and increasing on $[0,\infty)$, and hence we find

$$\delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q}) = \inf_{t\notin(\varepsilon,\varepsilon)} \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t) - \mathcal{C}^*_{L,\mathrm{Q}} = \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+\varepsilon) - \mathcal{C}^*_{L,\mathrm{Q}}\,.$$

Since we already know that $\mathcal{C}^*_{L,\mathrm{Q}} = \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q})$ by Lemma 3.45, we then obtain the assertion. □

Our next result is a technical lemma that will be used to establish *upper* bounds on $\delta_{\max,L_{\mathrm{mean}},L}(\varepsilon,\mathrm{Q})$. For its formulation, we need the set

$$\mathcal{Q}^*_{\mathbb{R},\mathrm{sym}} := \Big\{\mathrm{Q} \in \mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}} : \mathrm{Q}\big([\mathbb{E}\mathrm{Q}-\rho,\mathbb{E}\mathrm{Q}+\rho]\big) > 0 \text{ for all } \rho > 0\Big\}\,,$$

which contains all symmetric distributions on $\mathbb{R}$ that do not vanish around their means. Moreover, we also need the sets $\mathcal{Q}^*_{I,\mathrm{sym}} := \mathcal{Q}_I \cap \mathcal{Q}^*_{\mathbb{R},\mathrm{sym}}$, for $I \subset \mathbb{R}$, and $\mathcal{Q}^*_{\mathrm{bounded,sym}} := \mathcal{Q}_{\mathrm{bounded}} \cap \mathcal{Q}^*_{\mathbb{R},\mathrm{sym}}$. Now the result reads as follows.

Lemma 3.47 (Upper bound on excess risks). *Let L be a symmetric, continuous loss with representing function ψ. Assume that there exist a $\delta_0 \in \mathbb{R}$, $s_1, s_2 \in \mathbb{R}$ with $s_1 \neq s_2$, and an $\varepsilon_0 > 0$ such that for all $\varepsilon \in [0, \varepsilon_0]$ we have*

$$\frac{\psi(s_1+\varepsilon)+\psi(s_2+\varepsilon)}{2} - \psi\Big(\frac{s_1+s_2}{2}+\varepsilon\Big) \leq \delta_0\,. \tag{3.56}$$

Let us write $M := |\frac{s_1+s_2}{2}| + \varepsilon_0$ and $t := \frac{s_2-s_1}{2}$. Then, for all $\delta > 0$, there exists a Lebesgue absolutely continuous $\mathrm{Q} \in \mathcal{Q}^*_{[-M,M],\mathrm{sym}}$ with $\mathbb{E}\mathrm{Q} = 0$ and

$$\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t) - \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}) \leq \delta_0 + \delta\,.$$

Moreover, there exists a Lebesgue absolutely continuous $\mathrm{Q} \in \mathcal{Q}_{[-M,M],\mathrm{sym}}$ with $\mathbb{E}\mathrm{Q} = 0$ and $\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}+t) - \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}) \leq \delta_0$.

Proof. In the following, $\mu_{[a,b]}$ denotes the uniform distribution on the interval $[a,b]$. We write $y_0 := \frac{s_1+s_2}{2}$. Furthermore, if $y_0 = 0$, we define $\mathrm{Q} := \mu_{[-\varepsilon_0,\varepsilon_0]}$, and otherwise we define

$$\mathrm{Q} := \alpha\mu_{[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]} + \frac{1-\alpha}{2}\mu_{[-y_0-\varepsilon_0,-y_0]} + \frac{1-\alpha}{2}\mu_{[y_0,y_0+\varepsilon_0]}\,,$$

where $\alpha \in (0,1)$ is a real number satisfying

$$\sup_{y\in[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]} \Big|\frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)\Big| \leq \frac{\delta}{\alpha}\,.$$

Now we obviously have $\mathbb{E}\mathrm{Q} = 0$ in both cases. Moreover, if $y_0 \neq 0$, the construction together with Lemma 3.45 yields

$$\begin{aligned}
&\mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(0)\\
&= \int_{\mathbb{R}} \frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)d\mathrm{Q}(y)\\
&= \alpha\int_{[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]} \frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)\,d\mu_{[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]}(y)\\
&\qquad +(1-\alpha)\int_{[y_0,y_0+\varepsilon_0]} \frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)\,d\mu_{[y_0,y_0+\varepsilon_0]}(y)\\
&\leq \delta + (1-\alpha)\int_{[0,\varepsilon_0]} \frac{\psi(s_1+\varepsilon)+\psi(s_2+\varepsilon)}{2} - \psi\Big(\frac{s_1+s_2}{2}+\varepsilon\Big)\,d\mu_{[0,\varepsilon_0]}(\varepsilon)\\
&\leq \delta_0 + \delta\,.
\end{aligned}$$

Furthermore, the case $y_0 = 0$ can be shown analogously, since $y_0 = 0$ implies $y - t = s_1 + y$ and $y + t = s_2 + y$. The last assertion follows if we repeat the construction above with $\alpha = 0$. □

Let us now establish our first two main results, which characterize losses L that are L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\mathbb{R},\mathrm{sym}}(L)$ and $\mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}}$, respectively.

Theorem 3.48 (Mean calibration I). *Let $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a symmetric and continuous loss. Then the following statements are equivalent:*

i) L is L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\mathbb{R},\text{sym}}(L)$.
ii) L is L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\text{bounded,sym}}$.
iii) L is L_{mean}-calibrated with respect to $\mathcal{Q}^*_{[-M,M],\text{sym}}$ for all $M > 0$.
iv) L is convex, and its representing function ψ has its only minimum at 0.

Proof. *i)* $\Rightarrow$ *ii)* $\Rightarrow$ *iii)*. Trivial.

iv) $\Rightarrow$ *i)*. Assume that L is not L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\mathbb{R},\text{sym}}(L)$. By Lemma 3.46, there then exist a $\mathrm{Q} \in \mathcal{Q}^*_{\mathbb{R},\text{sym}}(L)$ and a $t \neq 0$ with $\mathcal{C}_{L,\mathrm{Q}}(m+t) = \mathcal{C}^*_{L,\mathrm{Q}}$, where $m := \mathbb{E}\mathrm{Q}$. Using $\mathcal{C}_{L,\mathrm{Q}}(m) = \mathcal{C}^*_{L,\mathrm{Q}}$, which we know from Lemma 3.45, then yields

$$\int_{\mathbb{R}} \frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)\, d\mathrm{Q}^{(m)}(y) = \mathcal{C}_{L,\mathrm{Q}}(m+t) - \mathcal{C}_{L,\mathrm{Q}}(m) = 0\,,$$

and hence the convexity of ψ shows $\frac{\psi(y-m-t)+\psi(y-m+t)}{2} - \psi(y-m) = 0$ for Q-almost all $y \in \mathbb{R}$. The continuity of ψ and the assumption $\mathrm{Q}(m + [-\rho, \rho]) > 0$ for all $\rho > 0$, then guarantee that $\frac{\psi(y-m-t)+\psi(y-m+t)}{2} - \psi(y-m) = 0$ holds for $y := m$. However, by the symmetry of ψ, this implies $\psi(t) = \psi(0)$.

iii) $\Rightarrow$ *iv)*. Assume that ψ is not convex. Then Lemma A.6.17 shows that there exist $s_1, s_2 \in \mathbb{R}$ with $s_1 \neq s_2$ and $\frac{\psi(s_1)+\psi(s_2)}{2} - \psi(\frac{s_1+s_2}{2}) < 0$. By the continuity of ψ, we then find (3.56) for some suitable $\delta_0 < 0$ and $\varepsilon_0 > 0$, and hence Lemma 3.47 gives an $M > 0$, a $\mathrm{Q} \in \mathcal{Q}^*_{[-M,M],\text{sym}}$, and a $t^* \neq 0$ with $\mathcal{C}_{L,\mathrm{Q}}(t^*) < \mathcal{C}_{L,\mathrm{Q}}(0)$ and $\mathbb{E}\mathrm{Q} = 0$. Now observe that since ψ is continuous and Q has bounded support, the map $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t)$ is continuous on $\mathbb{R}$ by Lemma A.6.2. Let $(t_n) \subset \mathbb{R}$ be a sequence with $\mathcal{C}_{L,\mathrm{Q}}(t_n) \to \mathcal{C}^*_{L,\mathrm{Q}}$ for $n \to \infty$. Since our previous considerations showed $\mathcal{C}_{L,\mathrm{Q}}(0) \neq \mathcal{C}^*_{L,\mathrm{Q}}$, there must exist an $\varepsilon > 0$ and an $n_0 \in \mathbb{N}$ such that $|t_n| \geq \varepsilon$ for all $n \geq n_0$. Since $\mathbb{E}\mathrm{Q} = 0$, this shows

$$\delta_{\max,L_{\text{mean}},L}(\varepsilon, \mathrm{Q}) = \inf_{t \notin (-\varepsilon,\varepsilon)} \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}} \;\leq\; \mathcal{C}_{L,\mathrm{Q}}(t_n) - \mathcal{C}^*_{L,\mathrm{Q}}$$

for all $n \geq n_0$. For $n \to \infty$, we hence find $\delta_{\max}(\varepsilon, \mathrm{Q}) = 0$, and consequently L is convex. Finally, assume that there exists a $t \neq 0$ with $\psi(t) = \psi(0)$. Then we find $\mathcal{C}_{L,\mathrm{Q}}(t) = \mathcal{C}^*_{L,\mathrm{Q}}$ for the distribution Q defined by $\mathrm{Q}(\{0\}) = 1$, and hence we obtain $\delta_{\max}(|t|, \mathrm{Q}) = 0$. Therefore ψ has its only minimum at 0. □

Theorem 3.49 (Mean calibration II). *Let $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a symmetric and continuous loss. Then the following statements are equivalent:*

i) L is L_{mean}-calibrated with respect to $\mathcal{Q}^{(1)}_{\mathbb{R},\text{sym}}(L)$.
ii) L is L_{mean}-calibrated with respect to $\mathcal{Q}_{\text{bounded,sym}}$.
iii) L is L_{mean}-calibrated with respect to $\mathcal{Q}_{[-M,M],\text{sym}}$ for all $M > 0$.
iv) L is strictly convex.

Proof. *i)* $\Rightarrow$ *ii)* $\Rightarrow$ *iii)*. Trivial.

iv) $\Rightarrow$ *i)*. It immediately follows from Lemma 3.45 and Lemma 3.46.

iii) $\Rightarrow$ *iv)*. If L is L_{mean}-calibrated with respect to $\mathcal{Q}_{[-M,M],\text{sym}}$ for all $M > 0$, then Theorem 3.48 shows that L is convex. Let us suppose that its representing function ψ is *not* strictly convex. Then there are $r_1, r_2 \in \mathbb{R}$ with $r_1 \neq r_2$ and

$$\psi\Big(\frac{1}{2}r_1 + \frac{1}{2}r_2\Big) = \frac{1}{2}\psi(r_1) + \frac{1}{2}\psi(r_2)\,.$$

From this and Lemma A.6.17, we find (3.56) for $\delta_0 = 0$ and some suitable $s_1 \neq s_2$ and $\varepsilon_0 > 0$. Lemma 3.47 then gives an $M > 0$, a $\mathrm{Q} \in \mathcal{Q}_{[-M,M],\text{sym}}$, and a $t_0 \neq 0$, with $\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q} + t_0) = \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q})$, and hence Lemma 3.46 shows that L is not L_{mean}-calibrated with respect to $\mathrm{Q} \in \mathcal{Q}_{[-M,M],\text{sym}}(L)$. □

Our next aim is to estimate the function $\varepsilon \mapsto \delta_{\max}(\varepsilon, \mathcal{Q})$ for some classes of distributions $\mathcal{Q} \subset \mathcal{Q}_{\mathbb{R},\text{sym}}$. To this end, we define the **modulus of convexity** of a function $f : I \to \mathbb{R}$ defined on some interval I by

$$\delta_f(\varepsilon) := \inf\left\{\frac{f(x_1) + f(x_2)}{2} - f\Big(\frac{x_1 + x_2}{2}\Big) : x_1, x_2 \in I \text{ with } |x_1 - x_2| \geq \varepsilon\right\},$$

where $\varepsilon > 0$. In addition we say that f is **uniformly convex** if $\delta_f(\varepsilon) > 0$ for all $\varepsilon > 0$. We refer to Section A.6.3 for some properties of the modulus of convexity and uniformly convex functions.

With the help of the modulus of convexity, we can now formulate the following theorem that estimates $\delta_{\max,L_{\text{mean}},L}(\varepsilon, \mathcal{Q})$ and characterizes uniform L_{mean}-calibration.

Theorem 3.50 (Uniform mean calibration). *Let L be a symmetric, convex loss with representing function ψ. Then the following statements are true:*

i) For all $M > 0$, $\varepsilon > 0$, and $\mathcal{Q}^*_{[-M,M],\text{sym}} \subset \mathcal{Q} \subset \mathcal{Q}_{[-M,M],\text{sym}}$, we have

$$\delta_{\psi_{|[-(2M+\varepsilon),2M+\varepsilon]}}(2\varepsilon) \;\leq\; \delta_{\max,L_{\text{mean}},L}(\varepsilon, \mathcal{Q}) \;\leq\; \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon)\,. \quad (3.57)$$

Moreover, the following statements are equivalent:

a) L is uniformly L_{mean}-calibrated w.r.t. $\mathcal{Q}^*_{[-M,M],\text{sym}}$ for all $M > 0$.
b) L is uniformly L_{mean}-calibrated w.r.t. $\mathcal{Q}_{[-M,M],\text{sym}}$ for all $M > 0$.
c) The function ψ is strictly convex.

ii) For all $\varepsilon > 0$, we have

$$\delta_\psi(2\varepsilon) = \delta_{\max,L_{\text{mean}},L}(\varepsilon, \mathcal{Q}_{\mathbb{R},\text{sym}}(L)) = \delta_{\max}(\varepsilon, \mathcal{Q}^*_{\text{bounded,sym}})\,. \quad (3.58)$$

Moreover, the following statements are equivalent:

a) L is uniformly L_{mean}-calibrated with respect to $\mathcal{Q}^{(1)}_{\mathbb{R},\text{sym}}(L)$.
b) L is uniformly L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\text{bounded,sym}}$.
c) The function ψ is uniformly convex.

Proof. i). Let $\mathrm{Q} \in \mathcal{Q}_{[-M,M],\mathrm{sym}}$. Then we have $\mathbb{E}\mathrm{Q} \in [-M, M]$, and hence Lemmas 3.45 and 3.46 yield

$$\begin{aligned}\delta_{\max}(\varepsilon, \mathrm{Q}) &= \int_{[-M,M]} \frac{\psi\big(y - \mathbb{E}\mathrm{Q} - \varepsilon\big) + \psi\big(y - \mathbb{E}\mathrm{Q} + \varepsilon\big)}{2} - \psi(y - \mathbb{E}\mathrm{Q})\, d\mathrm{Q}(y) \\ &\geq \delta_{\psi_{|[-(2M+\varepsilon),2M+\varepsilon]}}(2\varepsilon).\end{aligned}$$

This shows the first inequality of (3.57). To prove the second inequality, we observe that it suffices to consider the case $\varepsilon \leq M/2$ since for $\varepsilon > M/2$ we have $\delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) = \infty$. Let us now fix an $n \geq 1$. Then there exist $s_1, s_2 \in [-M/2, M/2]$ with $s_1 - s_2 \geq 2\varepsilon$ and

$$\frac{\psi(s_1) + \psi(s_2)}{2} - \psi\left(\frac{s_1 + s_2}{2}\right) < \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) + \frac{1}{n} =: \delta_0 < \infty\,.$$

By the continuity of ψ, there thus exists an $\varepsilon_0 \in (0, M/2]$ such that (3.56) is satisfied for δ_0, and consequently Lemma 3.47 gives a $\mathrm{Q} \in \mathcal{Q}^*_{[-M,M],\mathrm{sym}}$ with

$$\mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q} + t) - \mathcal{C}_{L,\mathrm{Q}}(\mathbb{E}\mathrm{Q}) \leq \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) + \frac{2}{n}\,,$$

where $t := \frac{s_1 - s_2}{2}$. Using $t \geq \varepsilon$ and Lemma 3.46, we hence find

$$\delta_{\max}(\varepsilon, \mathcal{Q}^*_{[-M,M],\mathrm{sym}}) \leq \delta_{\max}(\varepsilon, \mathrm{Q}) \leq \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) + \frac{2}{n}.$$

Since this holds for all $n \geq 1$, the second inequality of (3.57) follows. Finally, from Lemma A.6.17, we know that ψ is strictly convex if and only if $\delta_{\psi_{|[-B,B]}}(\varepsilon) > 0$ for all B and $\varepsilon > 0$, and hence the characterization follows.

ii). Analogously to the proof of the first inequality in (3.57), we find

$$\delta_\psi(2\varepsilon) \leq \delta_{\max}(\varepsilon, \mathcal{Q}_{\mathbb{R},\mathrm{sym}}(L))\,, \qquad \varepsilon > 0.$$

Furthermore, analogously to the proof of the second inequality in (3.57), we obtain

$$\delta_{\max}(\varepsilon, \mathcal{Q}^*_{\mathrm{bounded,sym}}) \leq \delta_\psi(2\varepsilon)\,, \qquad \varepsilon > 0,$$

and hence (3.58) is proved. Finally, the characterization is a trivial consequence of (3.58). □

The preceding theorem shows that the modulus of convexity completely determines whether a symmetric loss is uniformly L_{mean}-calibrated with respect to $\mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}}(L)$ or $\mathcal{Q}^*_{\mathrm{bounded,sym}}$. Unfortunately, Lemma A.6.19 shows that, for all distance-based losses of upper growth type $p < 2$, we have $\delta_\psi(\varepsilon) = 0$ for *all* $\varepsilon > 0$. In particular, Lipschitz continuous, distance-based losses, which are of special interest for robust regression methods (see Chapter 10), are *not* uniformly calibrated with respect to $\mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}}(L)$ or $\mathcal{Q}^*_{\mathrm{bounded,sym}}$, and consequently we cannot establish *distribution independent* relations between the

Table 3.3. Some symmetric loss functions and corresponding upper and lower bounds for the moduli of convexity $\delta_{\psi_{|[-B,B]}}(2\varepsilon)$, $0 < \varepsilon \leq B$. The asymptotics for the L_p-loss, $1 < p < 2$, are computed in Exercise 3.12. For the L_p-loss, $p \geq 2$, and Huber's loss, the lower bounds can be found by Clarkson's inequality (see Lemma A.5.24), and the upper bounds can be found by picking suitable $t_1, t_2 \in [-B, B]$. The calculations for the logistic loss can be found in Example 3.51.

Loss Function	Lower Bound of $\delta_{\psi_{\|[-B,B]}}(2\varepsilon)$	Upper Bound of $\delta_{\psi_{\|[-B,B]}}(2\varepsilon)$
$L_{\text{1-dist}}$	0	0
$L_{p\text{-dist}}$, $p \in (1,2)$	$\frac{p(p-1)}{2} B^{p-2} \varepsilon^2$	$\frac{p}{2(p-1)^2} B^{p-2} \varepsilon^2$
$L_{p\text{-dist}}$, $p \in [2,\infty)$	ε^p	ε^p
$L_{\text{r-logist}}$	$\frac{1-e^{-\varepsilon}}{2} \ln \frac{e^B+e^{2\varepsilon}}{e^B+e^{\varepsilon}}$	$(1-e^{-\varepsilon}) \ln \frac{e^B+e^{2\varepsilon}}{e^B+e^{\varepsilon}}$
$L_{\alpha\text{-Huber}}$, $\alpha > 0$	$\frac{\varepsilon^2}{2}$ if $B \leq \alpha$ 0 else	$\frac{\varepsilon^2}{2}$ if $B \leq \alpha$ 0 else

excess L-risks and $\mathcal{R}_{L_{\text{mean}},\text{P}}(\,\cdot\,)$ in the sense of Question 3.2. On the other hand, symmetric, strictly convex losses L are L_{mean}-calibrated with respect to $\mathcal{Q}^{(1)}_{\mathbb{R},\text{sym}}(L)$, and hence we can show analogously to Theorem 3.61 below that $f_n \to \mathbb{E}(Y|\,\cdot\,)$ in probability P_X whenever $\mathcal{R}_{L,\text{P}}(f_n) \to \mathcal{R}^*_{L,\text{P}}$ and P is of type $\mathcal{Q}^{(1)}_{\mathbb{R},\text{sym}}(L)$. In addition, if we restrict our considerations to $\mathcal{Q}_{[-M,M],\text{sym}}$ or $\mathcal{Q}^*_{[-M,M],\text{sym}}$, then every symmetric, strictly convex loss becomes uniformly L_{mean}-calibrated, and in this case $\delta_{\psi_{|[-B,B]}}(\,\cdot\,)$, $B > 0$, can be used to describe the corresponding calibration function. For some important losses, we have listed the behavior of $\delta_{\psi_{|[-B,B]}}(\,\cdot\,)$ in Table 3.3. Furthermore, Lemma A.6.19 establishes a formula for the modulus of convexity that often helps to bound the modulus. The following example illustrates this.

Example 3.51. Recall from Example 2.40 that the **logistic loss for regression** is the symmetric loss represented by $\psi(t) := -\ln \frac{4e^t}{(1+e^t)^2}$, $t \in \mathbb{R}$. Let us show, that for $B > 0$ and $\varepsilon \in (0, B]$, we have

$$\frac{1-e^{-\varepsilon}}{2} \ln \frac{e^B+e^{2\varepsilon}}{e^B+e^{\varepsilon}} \;\leq\; \delta_{\psi_{|[-B,B]}}(2\varepsilon) \;\leq\; (1-e^{-\varepsilon}) \ln \frac{e^B+e^{2\varepsilon}}{e^B+e^{\varepsilon}}\,.$$

To see this, we first observe that $\psi'(t) = \frac{e^t-1}{e^t+1}$ for all $t \in \mathbb{R}$, and hence we obtain

$$\psi'(t) - \psi'(t-\varepsilon) = \frac{e^t-1}{e^t+1} - \frac{e^t-e^{\varepsilon}}{e^t+e^{\varepsilon}} = \frac{2e^t(e^{\varepsilon}-1)}{(e^t+1)(e^t+e^{\varepsilon})}$$

for all $t \in \mathbb{R}$ and $\varepsilon \geq 0$. Consequently, we have

$$\frac{e^\varepsilon - 1}{e^t + e^\varepsilon} \;\le\; \psi'(t) - \psi'(t-\varepsilon) \;\le\; 2\frac{e^\varepsilon - 1}{e^t + e^\varepsilon}$$

for all $t \ge 0$ and $\varepsilon \ge 0$. Furthermore, for $\varepsilon > 0$ an easy calculation gives

$$\begin{aligned}\inf_{x\in[0,B-\varepsilon]} \int_x^{x+\varepsilon} \frac{e^\varepsilon - 1}{e^t + e^\varepsilon} dt &= \int_{B-\varepsilon}^{B} \frac{e^\varepsilon - 1}{e^t + e^\varepsilon} dt = (1 - e^{-\varepsilon}) \Big(t - \ln\big(e^t + e^\varepsilon\big) \Big) \Big|_{t=B-\varepsilon}^{B} \\ &= (1 - e^{-\varepsilon}) \ln \frac{e^B + e^{2\varepsilon}}{e^B + e^\varepsilon} \,.\end{aligned}$$

Using Lemma A.6.19 then yields the assertion. ◁

In Theorem 3.48, we have seen that for $\mathrm{Q} \in \mathcal{Q}^*_{\mathbb{R},\mathrm{sym}}(L)$ we may have $\delta_{\max}(\varepsilon, \mathrm{Q}) > 0$, $\varepsilon > 0$, even if L is not strictly convex. The key reason for this possibility was the assumption that Q has some mass around its center. Now recall that in the proof of the upper bounds of Theorem 3.50 we used the fact that for general $\mathrm{Q} \in \mathcal{Q}^*_{\mathbb{R},\mathrm{sym}}$ this mass can be arbitrarily small. However, if we enforce lower bounds on this mass, the construction of this proof no longer works. Instead, it turns out that we can establish lower bounds on $\delta_{\max}(\varepsilon, \mathrm{Q})$, as the following example illustrates (see also Example 3.67).

Example 3.52. Recall that the **absolute distance loss** is the symmetric loss represented by $\psi(t) = |t|$, $t \in \mathbb{R}$. Then, for all $\mathrm{Q} \in \mathcal{Q}^{(1)}_{\mathbb{R},\mathrm{sym}}$ and $\varepsilon > 0$, we have

$$\delta_{\max, L_{\mathrm{mean}}, L_{1\text{-dist}}}(\varepsilon, \mathrm{Q}) = \int_0^\varepsilon \mathrm{Q}^{(\mathbb{E}\mathrm{Q})}\big((-s, s)\big)\, ds \,. \tag{3.59}$$

To see this, recall that for symmetric distributions the mean equals the median, i.e., the 1/2-quantile. Now (3.59) follows from Proposition 3.9. ◁

The results in this section showed that using symmetric surrogate losses for regression problems requires some care: for example, let us suppose that the primary goal of the regression problem is to estimate the conditional mean. If we only know that the conditional distributions $\mathrm{P}(\,\cdot\,|x)$, $x \in X$, have finite variances (and expect these distributions to be rather asymmetric), then the least squares loss is the only reasonable, distance-based choice by Proposition 3.44. However, if we know that these distributions are (almost) symmetric, then symmetric, strictly convex, and Lipschitz continuous losses such as the logistic loss can be reasonable alternatives. In addition, if we are confident that these conditional distributions are also rather concentrated around their mean, e.g., in the form of $\mathrm{Q}^{(\mathbb{E}\mathrm{Q})}((-s, s)) > c_{\mathrm{Q}} s^q$ for small $s > 0$, then even the absolute distance loss can be a good choice. Finally, if we additionally expect that the data set contains extreme outliers, then the logistic loss or the absolute distance loss may actually be a better choice than the least squares loss. However, recall that such a decision only makes sense if the noise distribution is (almost) symmetric.

3.8 Surrogate Losses for the Density Level Problem

In this section, our goal is to find supervised loss functions that are calibrated with respect to the density level detection loss L_{DLD} introduced in Example 2.9. To this end, let us first recall that in the density level detection scenario our learning goal was to identify the ρ-level set $\{g > \rho\}$ of an unknown density $g : X \to [0, \infty)$ whose reference distribution μ on X is known. Unfortunately, the loss L_{DLD} formalizing this learning goal does depend on the unknown density g, and thus we cannot compute its associated risks. Consequently, our goal in this section is to find *supervised* surrogates for L_{DLD} that do not depend on g. At first glance, this goal seems to be rather impossible since supervised losses require labels that do not exist in the description of the DLD learning scenario. Therefore, our first goal is to resolve this issue by introducing *artificial* labels. To this end, we need the following definition.

Definition 3.53. *Let μ be a distribution on some X and $Y := \{-1, 1\}$. Furthermore, let $g : X \to [0, \infty)$ be measurable with $\|g\|_{L_1(\mu)} = 1$. Then, for $\rho > 0$, we write $g\mu \ominus_\rho \mu$ for the distribution* P *on $X \times Y$ that is defined by*

$$\begin{aligned} \mathrm{P}_X &:= \frac{g + \rho}{1 + \rho}\,\mu\,, \\ \mathrm{P}(y = 1|x) &:= \frac{g(x)}{g(x) + \rho}\,, \qquad x \in X. \end{aligned}$$

An elementary calculation shows that for measurable $A \subset X \times Y$ we have

$$g\mu \ominus_\rho \mu\,(A) = \frac{1}{1 + \rho}\mathbb{E}_{x\sim g\mu}\mathbf{1}_A(x, 1) + \frac{\rho}{1 + \rho}\mathbb{E}_{x\sim\mu}\mathbf{1}_A(x, -1)\,, \tag{3.60}$$

and hence $\mathrm{P} := g\mu \ominus_\rho \mu$ describes a binary classification problem in which the negative samples are drawn from the distribution μ with probability $\frac{\rho}{1+\rho}$ and in which the positive samples are drawn from the distribution $g\mu$ with probability $\frac{1}{1+\rho}$.

We have already mentioned in Example 2.9 that we are primarily interested in the quantity $\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f)$, which describes the discrepancy of the estimated level set $\{f \geq 0\}$ to the true ρ-level set. Now observe that, for $\mathrm{P} := g\mu \ominus_\rho \mu$ and measurable $f : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f) = \int_X L_{\mathrm{DLD}}(x, f(x))\,d\mu(x) = \int_X L_{\mathrm{DLD}}(x, f(x))\frac{1 + \rho}{g(x) + \rho}\,d\mathrm{P}_X(x)\,,$$

and consequently we can describe the DLD learning scenario by P and the detection loss $\bar{L} : X \times \mathbb{R} \to [0, \infty)$ defined by

$$\bar{L}(x, t) := L_{\mathrm{DLD}}(x, t)\frac{1 + \rho}{g(x) + \rho}\,, \qquad x \in X,\ t \in \mathbb{R}. \tag{3.61}$$

The first benefit of this reformulation is that our new target risk $\mathcal{R}_{\bar{L},\mathrm{P}}(\,\cdot\,)$ is defined by a distribution P, which produces labels, and consequently it makes

sense to look for supervised surrogates for $\bar{L}$. Furthermore, we have access to P via (3.60) in the sense that *a)* the distribution $g\mu$ can be estimated from the unlabeled samples given in the DLD scenario, see Example 2.9, and *b)* both μ and ρ are *known*. This makes it possible to construct an empirical approximation of P that can then lead to learning algorithms based on this approximation. For some literature in this direction, we refer to Section 2.5 and to the end of Section 8.6. The second benefit of considering the $\bar{L}$-risk is that P describes a *classification problem*, and hence it seems natural to *a)* investigate $\bar{L}$-calibration with the help of classification calibration and *b)* use classification algorithms for the DLD learning scenario. In order to confirm this intuition, let us consider the function $\bar{L}_{\mathrm{DLD}} : [0,1] \times \mathbb{R} \to [0,\infty)$ defined by

$$\bar{L}_{\mathrm{DLD}}(\eta, t) := (1-\eta)\mathbf{1}_{(-\infty,0)}\big((2\eta-1)\operatorname{sign} t\big)\,. \tag{3.62}$$

Using the identification $\eta = \mathrm{Q}(\{1\})$ between $\eta \in [0,1]$ and $\mathrm{Q} \in \mathcal{Q}_Y$, where $Y := \{-1,1\}$, we can regard the function $\bar{L}_{\mathrm{DLD}}$ as a template loss. For $\mathrm{P} = g\mu \ominus_\rho \mu$, the P-instance $\bar{L}_{\mathrm{DLD,P}}$ of $\bar{L}_{\mathrm{DLD}}$ then becomes

$$\begin{aligned}
\bar{L}_{\mathrm{DLD,P}}(x,t) = \bar{L}_{\mathrm{DLD}}\big(\mathrm{P}(\,\cdot\,|x), t\big) &= (1-\eta(x))\mathbf{1}_{(-\infty,0)}\big((2\eta(x)-1)\operatorname{sign} t\big) \\
&= \frac{\rho}{g(x)+\rho}\mathbf{1}_{(-\infty,0)}\big((g(x)-\rho)\operatorname{sign} t\big) \\
&= \frac{\rho}{1+\rho}\bar{L}(x,t)\,,
\end{aligned}$$

where we used $\eta(x) := \mathrm{P}(y=1|x) = \frac{g(x)}{g(x)+\rho}$. In other words, the P-instance $\bar{L}_{\mathrm{DLD,P}}$ of $\bar{L}_{\mathrm{DLD}}$ equals our detection loss $\bar{L}$ up to the constant $\frac{\rho}{1+\rho}$, and hence we obtain

$$\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f) = \mathcal{R}_{\bar{L},\mathrm{P}}(f) = \frac{1+\rho}{\rho}\mathcal{R}_{\bar{L}_{\mathrm{DLD,P}},\mathrm{P}}(f) \tag{3.63}$$

for $\mathrm{P} = g\mu \ominus_\rho \mu$ and all measurable functions $f : X \to \mathbb{R}$. Consequently, suitable supervised surrogates for the DLD problem are exactly the losses that are $\bar{L}_{\mathrm{DLD}}$-calibrated in the following sense.

Definition 3.54. *Let $Y := \{-1,1\}$ and $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss. We say that L is **(uniformly) density level detection calibrated** if L is (uniformly) $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to $\mathcal{Q}_Y$.*

In order to identify DLD-calibrated losses, we need to know the corresponding calibration function. This function is computed in the next lemma.

Lemma 3.55 (Calibration function for DLD). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function. Then, for all $\eta \in [0,1]$ and $\varepsilon \in (0,\infty]$, we have*

$$\delta_{\max,\bar{L}_{\mathrm{DLD}},L}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \varepsilon > 1-\eta \\ \inf_{t\in\mathbb{R}:(2\eta-1)\operatorname{sign} t<0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} & \text{if } \varepsilon \le 1-\eta. \end{cases}$$

Proof. A simple calculation shows $\mathcal{C}^*_{\bar{L}_{\mathrm{DLD}},\eta} = 0$, and consequently we obtain $\mathcal{M}_{\bar{L}_{\mathrm{DLD}},\eta}(\varepsilon) = \mathbb{R}$ if $\varepsilon > 1-\eta$, and $\mathcal{M}_{\bar{L}_{\mathrm{DLD}},\eta}(\varepsilon) = \{t \in \mathbb{R} : (2\eta-1)\operatorname{sign} t \geq 0\}$ otherwise. From this we immediately find the assertion. □

With the help of the preceding lemma, we now obtain the first main result, which compares classification calibration with $\bar{L}_{\mathrm{DLD}}$-calibration.

Theorem 3.56 (DLD-calibration). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss and $\eta \in [0,1]$. Then, for all $0 \leq \varepsilon \leq \min\{1-\eta, |2\eta-1|\}$, we have*

$$\delta_{\max,\bar{L}_{\mathrm{DLD}},L}(\varepsilon,\eta) \geq \delta_{\max,L_{\mathrm{class}},L}(\varepsilon,\eta)\,,$$

and consequently L is DLD-calibrated if L is classification calibrated. Moreover, if L is continuous, then the inequality above becomes an equality and L is classification calibrated if and only if L is DLD-calibrated.

Proof. Combining Lemma 3.55 with Lemma 3.32 yields

$$\begin{aligned}\delta_{\max,\bar{L}_{\mathrm{DLD}},L}(\varepsilon,\eta) = \inf_{\substack{t\in\mathbb{R}:\\ (2\eta-1)\operatorname{sign} t<0}} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} \;&\geq \inf_{\substack{t\in\mathbb{R}:\\ (2\eta-1)\operatorname{sign} t\leq 0}} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} \\ &= \delta_{\max,L_{\mathrm{class}},L}(\varepsilon,\eta)\,.\end{aligned}$$

Now assume that L is continuous. Since there is nothing to prove for $\eta = 1/2$, we additionally assume $\eta \neq 1/2$. Then the assertion can be found by using the continuity of $t \mapsto \mathcal{C}_{L,\eta}(t)$ in the estimate above. □

By the results on classification calibrated, margin-based losses from Section 3.4, we immediately obtain a variety of DLD-calibrated losses. Furthermore, the P-instances of $\bar{L}_{\mathrm{DLD}}$ are bounded and hence Theorem 3.27 yields

$$\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}^*_{L,\mathrm{P}} \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\mathrm{DLD}},\mu}(f_n) \to 0$$

whenever $\mathrm{P} = g\mu \ominus_\rho \mu$ and L is classification calibrated. In addition, one can show that for $L := L_{\mathrm{class}}$ the converse implication is also true. For details, we refer to Exercise 3.13.

Our next goal is to identify *uniformly* DLD-calibrated losses. The following theorem gives a complete, though rather disappointing, solution.

Theorem 3.57 (No uniform DLD-calibration). *There exists no supervised loss $L : Y \times \mathbb{R} \to [0,\infty)$ that is uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to both $\{\mathrm{Q} \in \mathcal{Q}_Y : \mathrm{Q}(\{1\}) \in [0,1/2)\}$ and $\{\mathrm{Q} \in \mathcal{Q}_Y : \mathrm{Q}(\{1\}) \in (1/2,3/4]\}$. In particular, there exists no uniform DLD-calibrated supervised loss.*

Proof. Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss. For $\eta \in [0,1]$, we define

$$g^+(\eta) = \inf_{t<0} \mathcal{C}_{L,\eta}(t) \qquad \text{and} \qquad g^-(\eta) = \inf_{t\geq 0} \mathcal{C}_{L,\eta}(t)\,.$$

Then the functions $g^+ : [0,1] \to [0,\infty)$ and $g^- : [0,1] \to [0,\infty)$ can be defined by suprema taken over affine linear functions in $\eta \in \mathbb{R}$, and since g^+ and g^-

are also finite for $\eta \in [0,1]$, we find by Lemma A.6.4 that g^+ and g^- are continuous at every $\eta \in [0,1]$. Moreover, we have $\mathcal{C}^*_{L,\eta} = \min\{g^+(\eta), g^-(\eta)\}$ for all $\eta \in [0,1]$, and hence $\mathcal{C}^*_{L,\eta}$ is continuous in η. Let us first consider the case $\mathcal{C}^*_{L,1/2} = g^+(1/2)$. To this end, we first observe that there exists a sequence $(t_n) \subset (-\infty, 0)$ with

$$g^+\bigl(1/2 + 1/n\bigr) \;\leq\; \mathcal{C}_{L,1/2+1/n}(t_n) \;\leq\; g^+\bigl(1/2 + 1/n\bigr) + 1/n \tag{3.64}$$

for all $n \geq 1$. Moreover, our assumption $\mathcal{C}^*_{L,1/2} = g^+(1/2)$ yields

$$\begin{aligned}\bigl|\mathcal{C}_{L,1/2+1/n}(t_n) - \mathcal{C}^*_{L,1/2+1/n}\bigr| \leq{}& \bigl|\mathcal{C}_{L,1/2+1/n}(t_n) - g^+(1/2+1/n)\bigr| \\ &+ \bigl|g^+\bigl(1/2+1/n\bigr) - g^+(1/2)\bigr| \\ &+ \bigl|\mathcal{C}^*_{L,1/2} - \mathcal{C}^*_{L,1/2+1/n}\bigr|\end{aligned}$$

for all $n \geq 1$. By (3.64) and the continuity of g^+ and $\eta \mapsto \mathcal{C}^*_{L,\eta}$, we hence find

$$\lim_{n\to\infty} \bigl|\mathcal{C}_{L,1/2+1/n}(t_n) - \mathcal{C}^*_{L,1/2+1/n}\bigr| = 0\,.$$

For $\mathcal{Q} := \bigl\{\mathrm{Q} \in \mathcal{Q}_Y : \mathrm{Q}(\{1\}) \in (1/2, 3/4]\bigr\}$, Lemma 3.55, the definition g^+, and (3.64) then yield

$$\begin{aligned}\delta_{\max,\bar{L}_{\mathrm{DLD}},L}(\varepsilon,\mathcal{Q}) = \inf_{\eta\in(\frac{1}{2},\frac{3}{4}]} g^+(\eta) - \mathcal{C}^*_{L,\eta} &\leq \inf_{n\geq 1} \mathcal{C}_{L,1/2+1/n}(t_n) - \mathcal{C}^*_{L,1/2+1/n} \\ &= 0\,.\end{aligned}$$

Consequently, L is not uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to $\mathcal{Q}$. Finally, in the case $\mathcal{C}^*_{L,1/2} = g^-(1/2)$, we can analogously show that L is not uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to $\{\mathrm{Q} \in \mathcal{Q}_Y : \mathrm{Q}(\{1\}) \in [0, 1/2)\}$. □

The preceding theorem shows that there exists no uniformly DLD-calibrated, supervised loss. Now recall that Theorem 3.24 showed that uniform calibration is *necessary* to establish inequalities between excess risks if essentially no assumptions on the data-generating distribution are imposed.[1] Together with Theorem 3.57, we consequently see that it is *impossible* to find a supervised loss $L : Y \times \mathbb{R} \to [0,\infty)$ and an increasing function $\delta : [0,\infty] \to [0,\infty]$ such that $\delta(0) = 0$, $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$, and

$$\delta\bigl(\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f)\bigr) \;\leq\; \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \tag{3.65}$$

for *all* μ, g, ρ, f, and $\mathrm{P} := g\mu \ominus_\rho \mu$. However, in the DLD learning scenario, we actually *know* μ and ρ, and hence the question remains whether for certain *fixed* μ and ρ there exists a non-trivial function δ satisfying (3.65). Unfortunately, Steinwart (2007) showed that the answer is again no.

[1] Formally, the result only holds for loss functions and *not template losses*. However, it is quite straightforward to see that the proof of Theorem 3.24 can be easily modified to establish an analogous result for instances of template losses.

3.9 Self-Calibrated Loss Functions

Given a loss L and a distribution P such that an *exact* minimizer $f^*_{L,\mathrm{P}}$ of $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$ exists, one may ask whether, and in which sense, approximate minimizers f of $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$ approximate $f^*_{L,\mathrm{P}}$. For example, in binary classification, one often wants to find a decision function f that not only has a small classification error but also estimates the conditional probability $\mathrm{P}(y=1|x)$. Now assume that we have found an f whose excess L-risk is small for a suitable surrogate L of the classification loss (recall Section 3.4 for examples of such surrogates). Assume further that the L-risk has a *unique* minimizer $f^*_{L,\mathrm{P}}$ that, in addition, has a one-to-one correspondence to the conditional probability. If we have a positive answer to the question above, we can then use a suitable transformation of $f(x)$ to estimate $\mathrm{P}(y=1|x)$. An important example of such a loss, namely the logistic loss for classification, is discussed in Example 3.66. Moreover, we will discuss how the pinball loss can be used to estimate quantiles. The main goal of this section is, however, to provide some general answers to the question above.

Let us begin by introducing some notation. To this end, let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function for some $Y \subset \mathbb{R}$ closed. We write

$$\begin{aligned}\mathcal{Q}_{\min}(L) &:= \bigl\{\mathrm{Q} : \mathrm{Q} \text{ is a distribution on } Y \text{ with } \mathcal{M}_{L,\mathrm{Q}}(0^+) \neq \emptyset\bigr\}\,,\\ \mathcal{Q}_{1\text{-}\min}(L) &:= \bigl\{\mathrm{Q} \in \mathcal{Q}_{\min}(L) : \exists\, t^*_{L,\mathrm{Q}} \in \mathbb{R} \text{ such that } \mathcal{M}_{L,\mathrm{Q}}(0^+) = \{t^*_{L,\mathrm{Q}}\}\bigr\}\,,\end{aligned}$$

i.e., $\mathcal{Q}_{\min}(L)$ contains the distributions on Y whose inner L-risks have an exact minimizer, while $\mathcal{Q}_{1\text{-}\min}(L)$ contains the distributions on Y whose inner L-risks have exactly one exact minimizer. Obviously, $\mathcal{Q}_{1\text{-}\min}(L) \subset \mathcal{Q}_{\min}(L)$ holds, and for strictly convex losses L, both sets actually coincide. Moreover, note that by Lemma 3.10 we have $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$ for all $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$. For $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$, we now define the **self-calibration loss** of L by

$$\breve{L}(\mathrm{Q},t) \;:=\; \operatorname{dist}\bigl(t, \mathcal{M}_{L,\mathrm{Q}}(0^+)\bigr) := \inf_{t' \in \mathcal{M}_{L,\mathrm{Q}}(0^+)} |t-t'|\,, \qquad t \in \mathbb{R}, \tag{3.66}$$

i.e., $\breve{L}(\mathrm{Q},t)$ measures the distance of t to the set of elements minimizing $\mathcal{C}_{L,\mathrm{Q}}$. The next lemma shows that the self-calibration loss is a template loss.

Lemma 3.58. *Let $Y \subset \mathbb{R}$ be closed and $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss. Then $\breve{L} : \mathcal{Q}_{\min}(L) \times \mathbb{R} \to [0,\infty)$ defined by (3.66) is a template loss.*

Proof. Let X be a complete measurable space and P be a distribution on $X \times Y$ with $\mathrm{P}(\,\cdot\,|x) \in \mathcal{Q}_{\min}(L)$ for all $x \in X$. We write $\bar{X} := X \times \mathbb{R}$ and $Z := \mathbb{R}$. Furthermore, for $\bar{x} = (x,t) \in \bar{X}$ and $t' \in Z$, we define

$$\begin{aligned}h(\bar{x},t') &:= \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x)}(t') - \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x)}\,,\\ F(\bar{x}) &:= \{t^* \in \mathbb{R} : h(\bar{x},t^*) = 0\}\,,\end{aligned}$$

and $\varphi(\bar{x},t') := |t-t'|$. For the P-instance $\breve{L}_{\mathrm{P}}$ of $\breve{L}$, we then have

$$\breve{L}_{\mathrm{P}}(x,t) = \inf_{t' \in \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x)}(0^+)} |t-t'| = \inf_{t' \in F(\bar{x})} \varphi(\bar{x},t')\,,$$

and consequently we obtain the assertion by part *iii)* of Lemma A.3.18. □

It is almost needless to say that the main statement of the preceding lemma is the *measurability* of the instances of $\breve{L}$. Now note that the definition of $\breve{L}$ immediately gives $\mathcal{C}^*_{\breve{L},\mathrm{Q}} = 0$, and therefore we have

$$\mathcal{M}_{\breve{L},\mathrm{Q}}(\varepsilon) = \bigl\{t \in \mathbb{R} : \breve{L}(\mathrm{Q},t) < \varepsilon\bigr\} = \bigl\{t \in \mathbb{R} \colon \exists t' \in \mathcal{M}_{L,\mathrm{Q}}(0^+) \text{ with } |t-t'| < \varepsilon\bigr\}$$

for all $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$ and $\varepsilon \in [0,\infty]$. Moreover, we have already mentioned in Section 3.6 that the results of Lemma 3.14 remain true for template losses. By (3.16), the **self-calibration function** $\delta_{\max,\breve{L},L}(\,\cdot\,,\mathrm{Q})$, which can be computed by

$$\delta_{\max,\breve{L},L}(\varepsilon,\mathrm{Q}) \;=\; \inf_{\substack{t\in\mathbb{R}\\ \operatorname{dist}(t,\mathcal{M}_{L,\mathrm{Q}}(0^+))\geq\varepsilon}} \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}} \tag{3.67}$$

for all $\varepsilon \in [0,\infty]$, thus satisfies

$$\delta_{\max,\breve{L},L}\bigl(\operatorname{dist}(t,\mathcal{M}_{L,\mathrm{Q}}(0^+)),\mathrm{Q}\bigr) \;\leq\; \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}\,, \qquad t \in \mathbb{R},$$

for all $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$. Note that for $\mathrm{Q} \in \mathcal{Q}_{1\text{-}\min}(L)$ this inequality becomes

$$\delta_{\max,\breve{L},L}\bigl(|t-t^*_{L,\mathrm{Q}}|,\mathrm{Q}\bigr) \;\leq\; \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}\,, \qquad t \in \mathbb{R},$$

where $\mathcal{M}_{L,\mathrm{Q}}(0^+) = \{t^*_{L,\mathrm{Q}}\}$. Consequently, the self-calibration function indeed quantifies how well an approximate $\mathcal{C}_{L,\mathrm{Q}}$-minimizer t approximates the exact minimizer $t^*_{L,\mathrm{Q}}$. This motivates the following, main definition of this section.

Definition 3.59. *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function and $\mathcal{Q} \subset \mathcal{Q}_{\min}(L)$. We say that L is **(uniformly) self-calibrated** with respect to $\mathcal{Q}$ if L is (uniformly) $\breve{L}$-calibrated with respect to $\mathcal{Q}$.*

Fortunately, convex loss functions are always self-calibrated, as the following lemma shows.

Lemma 3.60 (Self-calibration of convex losses). *Every convex loss function $L : Y \times \mathbb{R} \to [0,\infty)$ is self-calibrated with respect to $\mathcal{Q}_{\min}(L)$.*

Proof. For a fixed distribution $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$, we write $t_{\min} := \inf \mathcal{M}_{L,\mathrm{Q}}(0^+)$ and $t_{\max} := \sup \mathcal{M}_{L,\mathrm{Q}}(0^+)$. Now the map $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}$ is convex, and thus it is decreasing on $(-\infty,t_{\min}]$ and increasing on $[t_{\max},\infty)$. Furthermore, the convexity shows that $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ is an interval and hence we find $\mathcal{M}_{\breve{L},\mathrm{Q}}(\varepsilon) = \{t \in \mathbb{R} : \breve{L}(\mathrm{Q},t) < \varepsilon\} = (t_{\min}-\varepsilon, t_{\max}+\varepsilon)$, $\varepsilon > 0$. This gives

$$\begin{aligned}
\delta_{\max,\breve{L},L}(\varepsilon,\mathrm{Q}) &= \inf_{t \notin \mathcal{M}_{\breve{L},\mathrm{Q}}(\varepsilon)} \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}} \\
&= \min\Bigl\{\mathcal{C}_{L,\mathrm{Q}}(t_{\min}-\varepsilon), \mathcal{C}_{L,\mathrm{Q}}(t_{\max}+\varepsilon)\Bigr\} - \mathcal{C}^*_{L,\mathrm{Q}} \qquad (3.68)\\
&> 0\,,
\end{aligned}$$

where we used the convention $\mathcal{C}_{L,\mathrm{Q}}(\pm\infty) := \infty$. □

It is easy to see by the results of Section 3.7 that in general convex losses are *not* uniformly self-calibrated. Therefore, we usually cannot expect strong inequalities in the sense of Theorem 3.22 for the self-calibration problem. However, the following theorem shows that for general self-calibrated losses, approximate risk minimizers approximate the Bayes decision functions in a weak sense. Its consequences for convex losses are discussed in Corollary 3.62.

Theorem 3.61 (Asymptotic self-calibration). *Let X be a complete measurable space, $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss that is self-calibrated with respect to some $\mathcal{Q} \subset \mathcal{Q}_{\min}(L)$, and P be a distribution of type $\mathcal{Q}$ with $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Then, for all $\varepsilon > 0$ and $\rho > 0$, there exists a $\delta > 0$ such that for all measurable $f : X \to \mathbb{R}$ satisfying $\mathcal{R}_{L,\mathrm{P}}(f) < \mathcal{R}^*_{L,\mathrm{P}} + \delta$ we have*

$$\mathrm{P}_X\Big(\big\{x \in X : \operatorname{dist}\big(f(x), \mathcal{M}_{L,\mathrm{P}(\,\cdot\,|x)}(0^+)\big) \geq \rho\big\}\Big) < \varepsilon\,.$$

Proof. For a fixed $\rho > 0$, we write $A_\rho = \{(\mathrm{Q},t) \in \mathcal{Q} \times \mathbb{R} : \breve{L}(\mathrm{Q},t) \geq \rho\}$. By Lemma 3.58, we then see that $\bar{L} := \mathbf{1}_{A_\rho}$ defines a template loss function whose P-instance $\bar{L}_\mathrm{P}$ is a detection loss with respect to $h := \mathbf{1}_X$ and $A := \{(x,t) \in X \times \mathbb{R} : \breve{L}(\mathrm{P}(\,\cdot\,|x),t) \geq \rho\}$. Furthermore, we have

$$\mathcal{M}_{\bar{L},\mathrm{Q}}(\varepsilon) = \big\{t \in \mathbb{R} : \breve{L}(\mathrm{Q},t) < \rho\big\} = \mathcal{M}_{\breve{L},\mathrm{Q}}(\rho)$$

for all $\varepsilon \in (0,1]$ and $\mathrm{Q} \in \mathcal{Q}$, and thus we obtain

$$\delta_{\max,\bar{L},L}(\varepsilon,\mathrm{Q}) = \delta_{\max,\breve{L},L}(\rho,\mathrm{Q}) \;>\; 0\,, \qquad\qquad \varepsilon \in (0,1],\, \mathrm{Q} \in \mathcal{Q}.$$

Since calibration functions are increasing, we then find that L is $\bar{L}_\mathrm{P}$-calibrated with respect to $\mathcal{Q}$. For $\varepsilon > 0$, Theorem 3.27 thus gives a $\delta > 0$ such that for $f : X \to \mathbb{R}$ with $\mathcal{R}_{L,\mathrm{P}}(f) < \mathcal{R}^*_{L,\mathrm{P}} + \delta$ we have

$$\mathrm{P}_X\big(\big\{x \in X : \breve{L}_\mathrm{P}(x, f(x)) \geq \rho\big\}\big) = \mathcal{R}_{\bar{L}_\mathrm{P},\mathrm{P}}(f) - \mathcal{R}^*_{\bar{L}_\mathrm{P},\mathrm{P}} < \varepsilon\,. \qquad \square$$

For convex losses L and distributions of $\mathcal{Q}_{1\text{-}\min}(L)$-type, we obtain the following consequence.

Corollary 3.62. *Let X be a complete measurable space, $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex, supervised loss, and P be a distribution of type $\mathcal{Q}_{1\text{-}\min}(L)$ with $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Then there exists a P_X-almost surely unique minimizer $f^*_{L,\mathrm{P}}$ of $\mathcal{R}_{L,\mathrm{P}}$, and for all sequences (f_n) of measurable $f_n : X \to \mathbb{R}$, we have*

$$\mathcal{R}_{L,\mathrm{P}}(f_n) - \mathcal{R}^*_{L,\mathrm{P}} \to 0 \qquad \Longrightarrow \qquad f_n \to f^*_{L,\mathrm{P}} \quad \text{in probability } \mathrm{P}_X.$$

Proof. Lemma 3.12 together with the definition of $\mathcal{Q}_{1\text{-}\min}(L)$ shows that there exists a P_X-almost surely unique minimizer $f^*_{L,\mathrm{P}}$, and we thus find

$$\breve{L}_\mathrm{P}(x,t) = |t - f^*_{L,\mathrm{P}}(x)|\,, \qquad\qquad x \in X,\, t \in \mathbb{R}.$$

Theorem 3.61 together with Lemma 3.60 then yields $f_n \to f^*_{L,\mathrm{P}}$ in probability whenever the sequence (f_n) satisfies $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}^*_{L,\mathrm{P}}$. $\square$

Let us complete this discussion by describing situations in which we can replace the convergence in probability by a stronger notion of convergence.

Theorem 3.63 (Self-calibration inequalities). *Let X be a complete measurable space, $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss that is self-calibrated with respect to some $\mathcal{Q} \subset \mathcal{Q}_{1\text{-}\min}(L)$, and P be a distribution of type $\mathcal{Q}$ such that $\mathcal{R}^*_{L,\mathrm{P}} < \infty$. Moreover, assume that there exist a $p \in (0,\infty]$ and functions $b : X \to [0,\infty]$ and $\delta : [0,\infty) \to [0,\infty)$ such that*

$$\delta_{\max,\breve{L},L}(\varepsilon, \mathrm{P}(\,\cdot\,|x)) \geq b(x)\,\delta(\varepsilon)\,, \qquad \varepsilon > 0,\, x \in X,$$

and $b^{-1} \in L_p(\mathrm{P}_X)$. For a fixed $q \in (0,\infty)$, we define $\bar{\delta} : [0,\infty) \to [0,\infty)$ by

$$\bar{\delta}(\varepsilon) := \delta^{\frac{p}{p+1}}(\varepsilon^{1/q})\,, \qquad \varepsilon \in [0,\infty].$$

Then, for all measurable $f : X \to \mathbb{R}$ and $B_f := \|f - f^*_{L,\mathrm{P}}\|^q_\infty$, we have

$$\bar{\delta}^{**}_{B_f}\big(\|f - f^*_{L,\mathrm{P}}\|^q_{L_q(\mathrm{P}_X)}\big) \;\leq\; \|b^{-1}\|^{\frac{p}{p+1}}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\big)^{\frac{p}{p+1}}\,,$$

where $\bar{\delta}^{**}_{B_f} : [0, B_f] \to [0,\infty]$ is the Fenchel-Legendre bi-conjugate of $\bar{\delta}_{|[0,B_f]}$.

Proof. We write $\hat{\delta}(\varepsilon) := \delta(\varepsilon^{1/q})$ for $\varepsilon \geq 0$, and $\bar{L} := \breve{L}^q$. Then $\bar{L}$ is a template loss by Lemma 3.58, and since $\mathcal{M}_{\breve{L},\mathrm{Q}}(\varepsilon) = \{t \in \mathbb{R} : \breve{L}(\mathrm{Q}, t) < \varepsilon\}$, we find

$$\delta_{\max,\bar{L},L}(\varepsilon, \mathrm{Q}) \;=\; \delta_{\max,\breve{L},L}(\varepsilon^{1/q}, \mathrm{Q}) \;\geq\; \hat{b}(x)\,\delta(\varepsilon) \qquad \varepsilon > 0,\, \mathrm{Q} \in \mathcal{Q}.$$

Moreover, we have $\hat{\delta}^{\frac{p}{p+1}} = \bar{\delta}$, and hence Theorem 3.25 applied to $\bar{L}$ and $\hat{\delta}$ yields the assertion. □

Note that if the function δ is of the form $\delta(\varepsilon) = \varepsilon^r$ for some $r > 0$ and we consider $q := \frac{pr}{p+1}$, then we obtain $\bar{\delta}(\varepsilon) = \varepsilon$. In this case, Theorem 3.63 yields

$$\|f - f^*_{L,\mathrm{P}}\|_{L_q(\mathrm{P}_X)} \;\leq\; \|b^{-1}\|^{1/r}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\big)^{1/r}\,. \tag{3.69}$$

Moreover, if in this case we can only ensure $b^{-1} \in L_{p,\infty}(\mathrm{P}_X)$, then the norm $\|\cdot\|_{L_q(\mathrm{P}_X)}$ can be replaced by the Lorentz-norm $\|\cdot\|_{L_{q,\infty}(\mathrm{P}_X)}$ defined in Section A.5.5. For more details, we refer to Exercise 3.14.

The rest of this section applies the theory developed to some examples of practical importance. We begin with the problem of estimating the conditional probability $\mathrm{P}(y = 1|x)$ in classification, which has already been mentioned in the introduction of this section and which will be revisited in Section 8.5. To this end, we assume $Y := \{-1, 1\}$ in the following. Our first goal is to characterize situations when $\mathcal{Q}_{\min}(L) = \mathcal{Q}_Y$ for margin-based losses L.

Lemma 3.64 (Minimizers of margin-based losses). *Let L be a convex, margin-based loss represented by $\varphi : \mathbb{R} \to [0,\infty)$. Then we have $\mathcal{Q}_{\min}(L) = \mathcal{Q}_Y$ if and only if φ has a global minimum.*

Proof. If φ does not have a minimum, $\mathcal{C}_{L,1}(\,\cdot\,) = \varphi$ does not have a minimum, i.e., $\mathcal{M}_{L,1}(0^+) = \emptyset$. Conversely, if φ has a minimum, the same argument shows that $\mathcal{M}_{L,0}(0^+) = -\mathcal{M}_{L,1}(0^+) \neq \emptyset$. Therefore, let us fix an $\eta \in (0,1)$. If φ is constant, there is nothing to prove and hence we additionally assume that φ is not constant. The convexity of φ then shows that we have $\lim_{t\to\infty} \varphi(t) = \infty$ or $\lim_{t\to-\infty} \varphi(t) = \infty$. From this we immediately find $\mathcal{C}_{L,\eta}(t) \to \infty$ for $t \to \pm\infty$, and since $\mathcal{C}_{L,\eta}(\,\cdot\,)$ is continuous and convex, it thus has a global minimum. □

Together with Lemma 3.60, the preceding lemma immediately gives the following corollary that will be important when considering sparseness properties of support vector machines for classification in Section 8.5.

Corollary 3.65 (Self-calibration of margin-based losses). *Let L be a convex, margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. Then L is self-calibrated with respect to $\mathcal{Q}_Y$.*

With the help of Corollary 3.65, we see that the least squares loss and the (squared) hinge loss are self-calibrated with respect to $\mathcal{Q}_Y$, whereas the logistic loss is not. Furthermore, a simple calculation using Example 3.6 shows that the least squares loss is actually *uniformly* self-calibrated with respect to $\mathcal{Q}_Y$ and that the corresponding uniform self-calibration function is

$$\delta_{\max,\breve{L}_{\text{lsquares}},L_{\text{LS}}}(\varepsilon,\mathcal{Q}_Y) = \varepsilon^2\,, \qquad\qquad \varepsilon > 0\,.$$

However, neither the truncated least squares loss nor the hinge loss are uniformly self-calibrated with respect to $\mathcal{Q}_Y$, as we discuss in Exercise 3.15.

Let us now return to the problem of estimating the conditional probability $\eta(x) = \mathrm{P}(y = 1|x)$, $x \in X$. If we have a margin-based loss function L for which there is a one-to-one transformation between the sets of minimizers $\mathcal{M}_{L,\eta}(0^+)$ and η, then it seems natural to use self-calibration properties of L to investigate whether suitably transformed approximate L-risk minimizers approximate η. This approach is discussed in the following example.

Example 3.66. Exercise 3.2 shows that the **logistic loss for classification** $L_{\text{c-logist}}$ satisfies

$$\mathcal{M}_{L_{\text{c-logist}},\eta}(0^+) = \Big\{\ln\Big(\frac{\eta}{1-\eta}\Big)\Big\}\,, \qquad\qquad \eta \in (0,1).$$

In other words, if t_η^* denotes the element contained in $\mathcal{M}_{L_{\text{c-logist}},\eta}(0^+)$, then we have $\eta = \frac{1}{1+e^{-t_\eta^*}}$. Consequently, if t approximately minimizes $\mathcal{C}_{L_{\text{c-logist}},\eta}(\,\cdot\,)$, then it is close to t_η^* by Lemma 3.60 and hence $\frac{1}{1+e^{-t}}$ can serve as an estimate of η. However, investigating the quality of this estimate by the self-calibration function of $L_{\text{c-logist}}$ causes some technical problems since $L_{\text{c-logist}}$ is only self-calibrated with respect to the distributions $\mathrm{Q} \in \mathcal{Q}_Y$ with $\mathrm{Q}(\{1\}) \notin \{0,1\}$. Consequently, we now assess the quality of the estimate above *directly*. To this end, we introduce a new loss $L : \mathcal{Q}_Y \times \mathbb{R} \to [0,\infty)$, which we define by

$$L(\eta, t) := \Big|\eta - \frac{1}{1+e^{-t}}\Big|, \qquad \eta \in [0,1],\, t \in \mathbb{R}.$$

Then L is a template loss that measures the distance between η and its estimate $\frac{1}{1+e^{-t}}$. Let us compute the calibration function of $(L, L_{\text{c-logist}})$. To this end, we first observe that $\mathcal{C}^*_{L,\eta} = 0$ for all $\eta \in [0,1]$, and hence for $\varepsilon > 0$ an elementary calculation shows that

$$\begin{aligned}\mathcal{M}_{L,\eta}(\varepsilon) &= \{t \in \mathbb{R} : L(\eta, t) < \varepsilon\} \\ &= \left\{t \in \mathbb{R} : \ln \frac{(\eta-\varepsilon)_+}{1-\eta+\varepsilon} < t < \ln \frac{\eta+\varepsilon}{(1-\eta-\varepsilon)_+}\right\},\end{aligned}$$

where $(x)_+ := \max\{0, x\}$ for $x \in \mathbb{R}$ and $\ln 0 := -\infty$. For $\mathcal{C}_\eta(\infty) := \mathcal{C}_\eta(-\infty) := \infty$ and $\mathcal{C}_\eta(t) := \mathcal{C}_{L_{\text{c-logist}},\eta}(t) - \mathcal{C}^*_{L_{\text{c-logist}},\eta}$, Lemma 3.15 thus shows that

$$\delta_{\max,L,L_{\text{c-logist}}}(\varepsilon,\eta) = \min\left\{\mathcal{C}_\eta\left(-\ln\Big(\frac{1-\eta-\varepsilon}{\eta+\varepsilon}\Big)_+\right), \mathcal{C}_\eta\left(\ln\Big(\frac{\eta-\varepsilon}{1-\eta+\varepsilon}\Big)_+\right)\right\}.$$

From this we can conclude that $\delta_{\max,L,L_{\text{c-logist}}}(\varepsilon,\eta) = \delta_{\max,L,L_{\text{c-logist}}}(\varepsilon, 1-\eta)$ for all $\varepsilon \geq 0$, $\eta \in [0,1]$. Moreover, using the formulas of Exercise 3.2, we find

$$\mathcal{C}_\eta\left(\ln\Big(\frac{\eta-\varepsilon}{1-\eta+\varepsilon}\Big)_+\right) = \begin{cases}\eta \ln \frac{\eta}{\eta-\varepsilon} + (1-\eta)\ln\frac{1-\eta}{1-\eta+\varepsilon} & \text{if } \varepsilon < \eta \\ \infty & \text{otherwise}\end{cases}$$

and

$$\mathcal{C}_\eta\left(-\ln\Big(\frac{1-\eta-\varepsilon}{\eta+\varepsilon}\Big)_+\right) = \begin{cases}\eta \ln \frac{\eta}{\eta+\varepsilon} + (1-\eta)\ln\frac{1-\eta}{1-\eta-\varepsilon} & \text{if } \varepsilon < 1-\eta \\ \infty & \text{otherwise.}\end{cases}$$

In order to compare these expressions, let us write $g(\eta) := \eta \ln \frac{\eta}{\eta-\varepsilon} - \eta \ln \frac{\eta}{\eta+\varepsilon}$ for a fixed $\varepsilon \in (0, 1/2)$ and all η with $\varepsilon < \eta < 1-\varepsilon$. Then we have

$$g(1-\eta) = (1-\eta)\ln\frac{1-\eta}{1-\eta-\varepsilon} - (1-\eta)\ln\frac{1-\eta}{1-\eta+\varepsilon}$$

and

$$g'(\eta) = \frac{(\eta^2-\varepsilon^2)\ln\frac{\eta+\varepsilon}{\eta-\varepsilon} - 2\varepsilon\eta}{\eta^2-\varepsilon^2} =: \frac{g_\eta(\varepsilon)}{\eta^2-\varepsilon^2}\,.$$

Now observe that $g_\eta(0) = 0$ and $g'_\eta(\varepsilon) < 0$ for all $\varepsilon > 0$, and hence we obtain $g'(\eta) < 0$. Consequently, we have $g(\eta) \geq g(1-\eta)$, or in other words

$$\eta \ln \frac{\eta}{\eta-\varepsilon} + (1-\eta)\ln\frac{1-\eta}{1-\eta+\varepsilon} \;\geq\; \eta \ln \frac{\eta}{\eta+\varepsilon} + (1-\eta)\ln\frac{1-\eta}{1-\eta-\varepsilon}\,,$$

if and only if $\eta \leq \frac{1}{2}$. Therefore, for $\eta \in [0, 1/2]$, we find

$$\delta_{\max,L,L_{\text{c-logist}}}(\varepsilon,\eta) = \begin{cases}\eta \ln \frac{\eta}{\eta+\varepsilon} + (1-\eta)\ln\frac{1-\eta}{1-\eta-\varepsilon} & \text{if } \varepsilon < 1-\eta \\ \infty & \text{otherwise.}\end{cases}$$

In order to investigate whether $L_{\text{c-logist}}$ is L-calibrated with respect to $\mathcal{Q}_Y$, let us now find a simple lower bound of the calibration function above. To this end, let $h_\varepsilon(\eta) := \eta \ln \frac{\eta}{\eta+\varepsilon}$ for $\eta \in [0,1/2]$ and $\varepsilon \geq 0$. Then its derivative satisfies

$$h'_\varepsilon(\eta) = \ln \frac{\eta}{\eta+\varepsilon} + \frac{\varepsilon}{\eta+\varepsilon} = \ln\Bigl(1 - \frac{\varepsilon}{\eta+\varepsilon}\Bigr) + \frac{\varepsilon}{\eta+\varepsilon} \;\leq\; -\frac{\varepsilon}{\eta+\varepsilon} + \frac{\varepsilon}{\eta+\varepsilon} = 0\,,$$

and hence we find $\eta \ln \frac{\eta}{\eta+\varepsilon} \geq \frac{1}{2} \ln \frac{1}{1+2\varepsilon}$ for all $\eta \in [0,1/2]$, $\varepsilon \geq 0$. Analogously, we obtain $(1-\eta) \ln \frac{1-\eta}{1-\eta-\varepsilon} \geq \ln \frac{1}{1-\varepsilon}$ for $\eta \in [0,1/2]$, $\varepsilon \in [0,1-\eta)$. Both estimates together then yield

$$\delta_{\max,L,L_{\text{c-logist}}}(\varepsilon,\eta) \;\geq\; \frac{1}{2} \ln \frac{1}{1+2\varepsilon} + \ln \frac{1}{1-\varepsilon} \;\geq\; \varepsilon^2$$

for all $\eta \in [0,1/2]$ and all $\varepsilon \in [0,1-\eta)$. Consequently, $L_{\text{c-logist}}$ is *uniformly* L-calibrated with respect to $\mathcal{Q}_Y$, and the calibration function satisfies $\delta_{\max}(\varepsilon,\mathcal{Q}_Y) \geq \varepsilon^2$ for all $\varepsilon \geq 0$. For the loss function L^2, we thus obtain

$$\delta_{\max,L^2,L_{\text{c-logist}}}(\varepsilon,\mathcal{Q}_Y) = \delta_{\max,L,L_{\text{c-logist}}}(\varepsilon^{1/2},\mathcal{Q}_Y) \geq \varepsilon\,, \qquad\qquad \varepsilon \geq 0.$$

By Theorem 3.22, we then see that for all measurable $f : X \to \mathbb{R}$ we have

$$\int_X \Bigl|\eta(x) - \frac{1}{1+e^{-f(x)}}\Bigr|^2 \, d\mathrm{P}_X(x) \;\leq\; \mathcal{R}_{L_{\text{c-logist}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{c-logist}},\mathrm{P}}\,,$$

i.e., we can assess the quality of the estimate $\frac{1}{1+e^{-f(x)}}$ in terms of $\|\cdot\|_2$. ◁

Our last goal is to investigate the self-calibration properties of the τ-pinball loss $L_{\tau\text{-pin}}$. Proposition 3.9 showed that the minimizer of this convex supervised loss was the τ-quantile, and consequently $L_{\tau\text{-pin}}$ can be used to estimate the conditional τ-quantile. However, so far we only have a rather weak justification in the sense of Theorem 3.61. The following example discusses some conditions on the distribution P, which provides a stronger justification.

Example 3.67. For fixed $\tau \in (0,1)$, let $L := L_{\tau\text{-pin}}$ be the **τ-pinball loss** defined in Example 2.43. Furthermore, let Q be a distribution on $\mathbb{R}$ such that $|\mathrm{Q}|_1 < \infty$ and let t^* be a τ-quantile of Q, i.e., we simultaneously have

$$\mathrm{Q}\bigl((-\infty,t^*]\bigr) \geq \tau \qquad\qquad \text{and} \qquad\qquad \mathrm{Q}\bigl([t^*,\infty)\bigr) \geq 1-\tau\,. \tag{3.70}$$

If t^* is the only τ-quantile of Q, i.e., t^* is uniquely defined by (3.70), then the formulas of Proposition 3.9 show

$$\delta_{\max,\check{L},L}(\varepsilon,\mathrm{Q}) = \min\biggl\{\varepsilon q_+ + \int_0^\varepsilon \mathrm{Q}\bigl((t^*,t^*+s)\bigr)\,ds,\, \varepsilon q_- + \int_0^\varepsilon \mathrm{Q}\bigl((t^*-s,t^*)\bigr)\,ds\biggr\}$$

for all $\varepsilon \geq 0$, where q_+ and q_- are the real numbers found in Proposition 3.9.

Let us now denote the set of all distributions Q for which the inequalities in (3.70) strictly hold by $\mathcal{Q}_\tau^{>0}$. For $\mathrm{Q} \in \mathcal{Q}_\tau^{>0}$, we then have $\min\{q_+, q_-\} > 0$ and hence t^* is uniquely determined. Moreover, the self-calibration function satisfies

$$\delta_{\max,\check{L},L}(\varepsilon, \mathrm{Q}) \geq c_\mathrm{Q}\, \varepsilon\,, \qquad \varepsilon \geq 0, \tag{3.71}$$

where $c_\mathrm{Q} := \min\{q_+, q_-\}$. For a fixed distribution P of $\mathcal{Q}_\tau^{>0}$-type, we now define the function $b : X \to [0, \infty)$ by $b(x) := c_{\mathrm{P}(\,\cdot\,|x)}$, $x \in X$, where $c_{\mathrm{P}(\,\cdot\,|x)}$ denotes the constant in (3.71), which belongs to the conditional distribution $\mathrm{P}(\,\cdot\,|x)$. If we have $b^{-1} \in L_p(\mathrm{P}_X)$, then Theorem 3.63, see also (3.69), shows

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_q(\mathrm{P}_X)} \leq \|b^{-1}\|_{L_p(\mathrm{P}_X)} \left(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\right) \tag{3.72}$$

for all measurable functions $f : X \to \mathbb{R}$, where $f^*_{\tau,\mathrm{P}}(x)$ denotes the τ-quantile of $\mathrm{P}(\,\cdot\,|x)$ and $q := \frac{p}{p+1}$.

Although (3.72) provides a nice relationship between the excess pinball risk and our goal of estimating the conditional quantile function $f^*_{\tau,\mathrm{P}}$, the distributions P of $\mathcal{Q}_\tau^{>0}$-type seem a bit unrealistic for practical situations. Therefore, let us finally consider a more realistic scenario. To this end, we fix an $\alpha > 0$ and say that a distribution Q with $|\mathrm{Q}|_1 < \infty$ is of type $\mathcal{Q}_\tau^\alpha$ if there exists a τ-quantile t^* of Q and a constant $c_\mathrm{Q} > 0$ such that

$$\mathrm{Q}\big((t^*, t^* + s)\big) \geq c_\mathrm{Q}\, s \qquad \text{and} \qquad \mathrm{Q}\big((t^* - s, t^*)\big) \geq c_\mathrm{Q}\, s \tag{3.73}$$

for all $s \in [0, \alpha]$. Obviously, for such distributions, the τ-quantile t^* is uniquely determined. Moreover, if Q has a density h_Q with respect to the Lebesgue measure and this density satisfies $h_\mathrm{Q}(t) \geq c_\mathrm{Q}$ for all $t \in [t^* - \alpha, t^* + \alpha]$, then Q is of type $\mathcal{Q}_\tau^\alpha$. Let us now define $\delta : [0, \infty) \to [0, \infty)$ by

$$\delta(\varepsilon) := \begin{cases} \varepsilon^2/2 & \text{if } \varepsilon \in [0, \alpha] \\ \alpha\varepsilon - \alpha^2/2 & \text{if } \varepsilon > \alpha\,. \end{cases}$$

Then a simple calculation yields

$$\delta_{\max,\check{L},L}(\varepsilon, \mathrm{Q}) \geq c_\mathrm{Q}\delta(\varepsilon)\,, \qquad \varepsilon \geq 0,$$

for all $\mathrm{Q} \in \mathcal{Q}_\tau^\alpha$, where c_Q is the constant satisfying (3.73). For fixed $p \in (0, \infty]$, we further define $\bar{\delta} : [0, \infty) \to [0, \infty)$ by $\bar{\delta}(\varepsilon) := \delta^{\frac{p}{p+1}}(\varepsilon^{\frac{p+1}{p}})$, $\varepsilon \geq 0$. In view of Theorem 3.63, we then need to find a convex function $\hat{\delta} : [0, \infty) \to [0, \infty)$ such that $\hat{\delta} \leq \bar{\delta}$. To this end, we define

$$\hat{\delta}(\varepsilon) := \begin{cases} s_p^p \varepsilon^2 & \text{if } \varepsilon \in \big[0, s_p a_p\big] \\ a_p\big(\varepsilon - s_p^{p+2} a_p\big) & \text{if } \varepsilon > s_p a_p\,, \end{cases}$$

where $a_p := \alpha^{p/(p+1)}$ and $s_p := 2^{-1/(p+1)}$. An easy calculation shows that $\hat{\delta} : [0, \infty) \to [0, \infty)$ is continuously differentiable with non-decreasing derivative.

Consequently, $\hat{\delta}$ is convex. Moreover, since $(\varepsilon^{\frac{p+1}{p}} - \alpha/2)^{-\frac{1}{p+1}} \varepsilon^{\frac{1}{p}} \geq 1$ we have $\hat{\delta}' \leq \bar{\delta}'$ and hence we find $\hat{\delta} \leq \bar{\delta}$ by the fundamental theorem of calculus.

For a distribution P of $\mathcal{Q}_\tau^\alpha$-type, we now define the function $b : X \to [0, \infty)$ by $b(x) := c_{\mathrm{P}(\cdot|x)}$, $x \in X$, where $c_{\mathrm{P}(\cdot|x)}$ is determined by (3.73). If b satisfies $b^{-1} \in L_p(\mathrm{P}_X)$ for some $p \in (0, \infty]$, Theorem 3.63 together with our considerations above shows

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_q(\mathrm{P}_X)} \leq \sqrt{2}\, \|b^{-1}\|^{1/2}_{L_p(\mathrm{P}_X)} \left(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\right)^{1/2} \tag{3.74}$$

for $q := \frac{p}{p+1}$ and all $f : X \to \mathbb{R}$ satisfying $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \leq 2^{-\frac{p+2}{p+1}} \alpha^{\frac{2p}{p+1}}$. $\triangleleft$

3.10 Further Reading and Advanced Topics

The idea of using a surrogate loss developed quite independently in statistics and machine learning. Indeed, in statistics, its development was mainly motivated by the search for more robust estimation methods (see, e.g., Huber, 1964), in particular for regression problems. On the other hand, in machine learning, surrogate losses were mainly considered as a trick to find faster classification algorithms. However, only very recently has the relation between the risks of these surrogates and the classification risk been investigated. The first observations on the set of minimizers were made by Lin (2002b). Later he (see Lin, 2004, Theorem 3.1 and Lemma 4.1) established a result somewhat similar to Theorem 3.36 and a bound on the excess classification risk that generalizes the widely known Theorem 2.2 from Devroye *et al.* (1996). Independently of Lin, Zhang (2004b) established the first general inequalities between the excess classification risk and the excess risks of margin-based surrogate losses. Furthermore, he mentioned that some applications also require estimating the conditional probability and concludes that some margin-based losses, including the hinge loss, are not suited for this task. Another independent result, established by Steinwart (2005), gives a sufficient condition for continuous, supervised losses L that ensures an asymptotic relation (in the sense of Question 3.1) between the excess classification risk and the excess L-risk. However, the big breakthrough in understanding surrogate margin-based losses was then made by Bartlett *et al.* (2006). In fact, all the main results on classification calibrated, margin-based losses presented in Section 3.4 were shown by these authors, though condition (3.40) was already investigated by Mammen and Tsybakov (1999), and Tsybakov (2004) in the context of density level detection. We refer to Steinwart *et al.* (2005) and Steinwart (2007), who translated their findings into the language of calibration inequalities.

Prior to Steinwart (2007), the only result for weighted classification (also known as *cost-sensitive classification*) that deals with calibration issues was presented by Lin *et al.* (2002), though weighted classification itself had been considered earlier by, e.g., Elkan (2001). The presentation in Section 3.5 closely

follows Steinwart's work. Furthermore, there are recent results on surrogates for multi-class classification that we have not presented here due to lack of space. For more information, we refer to Lee *et al.* (2004), Zhang (2004a), Tewari and Bartlett (2005), and the references therein.

Proposition 3.44, which shows the unique role of the least squares loss for estimating the regression function, was independently found by Caponnetto (2005) and Steinwart (2007). Besides the basic notions and examples, the rest of Section 3.7 is based on the work of Steinwart (2007). Finally, it is worth mentioning that the approach in Section 3.7 substantially differs from the traditional maximum-likelihood motivation for the least squares loss already used by Gauss. We refer to Schölkopf and Smola (2002) for a brief introduction to the maximum-likelihood motivation and to Kardaun *et al.* (2003) for a discussion on this motivation.

The asymptotic theory on surrogate losses developed in Section 3.2 is a generalization of the results of Steinwart (2005). Moreover, the inequalities for general surrogate losses established in Section 3.3 were deeply inspired by the work of Bartlett *et al.* (2006). However, the key results of this section, namely Theorem 3.22 and Theorem 3.25, can also be derived from Theorem 24 of Zhang (2004a). Finally, a self-calibration result for classification calibrated surrogates similar to Theorem 3.61 was already shown by Steinwart (2003). In the presentation of all of these results, we closely followed Steinwart (2007).

3.11 Summary

In this chapter, we developed a general theory that allows us to *a)* identify suitable surrogate loss functions and *b)* relate the excess risks of such surrogate losses with the excess risks of the original (target) loss function. The main concept of this theory was the *calibration function*, which compares the *inner* excess risks of the losses involved. With the help of the calibration function, we then introduced the notions of calibration and uniform calibration, which (essentially) characterize how the excess risks involved can be compared. We then applied the general theory to some important learning scenarios:

- **Classification.** Here we showed that, for margin-based losses, calibration and uniform calibration are equivalent concepts. Furthermore, we developed a way to establish inequalities between the excess classification risk and the excess risk of margin-based losses. We then established an easy test to check whether a given *convex*, margin-based loss function is classification calibrated. Finally, we further simplified the computation of the uniform calibration function for such losses.
- **Weighted classification.** We showed that a simple weighting method for classification calibrated, margin-based loss functions produces loss functions that are calibrated to the weighted classification scenario. With the help of this weighting method, we then translated the major results on

unweighted classification calibration into analogous results on weighted classification calibration.

- **Regression.** Here we first showed that the least squares loss is essentially the only distance-based loss that can be used to find the regression function if one only knows that the average second moment of the noise distributions is finite. For some large classes of *symmetric* noise distributions, we then characterized (uniformly) least squares calibrated losses. Here it turned out that the convexity and related stronger notions play a crucial role. In particular, we showed that for symmetric, unbounded noise every uniformly least squares calibrated and symmetric loss must grow at least as fast as the least squares loss, and consequently one cannot avoid assuming the finiteness of the second moments for such distributions and losses. Furthermore, we have seen that for slower-growing losses, such as the absolute distance loss, the latter requirement can be replaced by non-parametric assumptions on the concentration around the mean.
- **Density level detection.** We first showed that the DLD learning scenario can be treated as a supervised learning problem that is similar to a classification problem. It then turned out that every classification calibrated loss is DLD-calibrated. However, unlike for classification, there exists *no* uniformly DLD-calibrated supervised loss, and consequently it is impossible to establish inequalities between the DLD-risk and excess risks of supervised surrogates without further assumptions on the density.
- **Self-calibration.** It is of both theoretical and practical interest whether approximate risk minimizers approximate the true risk minimizer. In Section 3.9, we developed a general framework to investigate this issue. In particular, we showed that convex losses always guarantee a weak positive result. Finally, we applied the general theory to the logistic loss for classification and the pinball loss.

The theory developed and its consequences for the learning scenarios above will play an important role when we investigate the corresponding kernel-based learning procedures in later chapters. However, it is worth mentioning that the results of this chapter are *algorithm independent*, i.e., they can be used for any algorithm whose *surrogate* risk performance is understood.

3.12 Exercises

3.1. Inner risks of the squared hinge loss ($\star$)
Recall that in Example 2.28 we defined the squared hinge loss by $L(y,t) := (\max\{0, 1-yt\})^2$, $y = \pm 1$, $t \in \mathbb{R}$. Using the definitions in (3.8), show that for $\eta \in [0,1]$ we have $\mathcal{C}^*_{L,\eta} = 4\eta(1-\eta)$ and

$$\mathcal{M}_{L,\eta}(0^+) = \begin{cases} (-\infty, -1] & \text{if } \eta = 0 \\ \{2\eta - 1\} & \text{if } 0 < \eta < 1 \\ [1, \infty) & \text{if } \eta = 1\,. \end{cases}$$

Moreover, show that, for $\eta \in [1/2, 1]$ and $t \in \mathbb{R}$, the excess inner risk can be computed by

$$\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \begin{cases} 4\eta^2 - 3\eta - 2\eta t + \eta t^2 & \text{if } t \le -1 \\ (t - 2\eta + 1)^2 & \text{if } t \in [-1, 1] \\ (1-\eta)(1 + 2t + t^2 - 4\eta) & \text{if } t \ge 1\,. \end{cases}$$

3.2. Logistic loss for classification($\star$)
Recall that in Example 2.29 we defined the logistic loss for classification by $L(y,t) := \ln(1+\exp(-yt))$, $y = \pm 1$, $t \in \mathbb{R}$. Show the following formulas using the notations in (3.8) and the convention $0 \ln 0 := 0$:

$$\begin{aligned} \mathcal{C}^*_{L,\eta} &= -\eta \ln(\eta) - (1-\eta)\ln(1-\eta)\,, \\ \mathcal{M}_{L,\eta}(0^+) &= \big\{\ln(\eta) - \ln(1-\eta)\big\}\,, \qquad \text{if } \eta \ne 0, 1, \\ \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} &= \eta \ln\big(\eta(1+e^{-t})\big) + (1-\eta)\ln\big((1-\eta)(1+e^t)\big). \end{aligned}$$

3.3. Calibration function ($\star$)
Let $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be loss functions, Q be a distribution on Y, and $x \in X$ with $\mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x} < \infty$ and $\mathcal{C}^*_{L_{\text{sur}},\mathrm{Q},x} < \infty$. Assume that $\delta : [0,\infty] \to [0,\infty]$ is an increasing function with

$$\delta\big(\mathcal{C}_{L_{\text{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{tar}},\mathrm{Q},x}\big) \;\le\; \mathcal{C}_{L_{\text{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\text{sur}},\mathrm{Q},x}\,, \qquad t \in \mathbb{R}.$$

Show that $\delta(\varepsilon) \le \delta_{\max}(\varepsilon, \mathrm{Q}, x)$ for all $\varepsilon \in [0,\infty]$.

Hint: Assume the converse and use Lemma 3.14.

3.4. Characterization of calibration ($\star\star\star$)
Prove Corollary 3.19.

Hint for ii) $\Rightarrow$ *i):* Assume that L_{sur} is not L_{tar}-calibrated to construct a "simple" distribution P that violates *ii*). Furthermore, use that the condition $\mathcal{R}^*_{L_{\text{tar}},\mathrm{P}} < \infty$ is automatically satisfied since L_{tar} is bounded.

3.5. Uniform calibration function ($\star\star$)
Let $L_{\text{tar}} : X \times Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : X \times Y \times \mathbb{R} \to [0,\infty)$ be loss functions and $\mathcal{Q}$ be a set of distributions on Y. Show that for all $\varepsilon \in [0,\infty]$ we have

$$\delta_{\max}(\varepsilon, \mathcal{Q}) = \max\big\{\delta \ge 0 : \mathcal{M}_{L_{\text{sur}},\mathrm{Q},x}(\delta) \subset \mathcal{M}_{L_{\text{tar}},\mathrm{Q},x}(\varepsilon) \text{ for all } \mathrm{Q} \in \mathcal{Q},\, x \in X\big\}\,.$$

3.6. Uniformly calibrated supervised losses ($\star\star\star\star$)
Let $L_{\text{tar}} : Y \times \mathbb{R} \to [0,\infty)$ and $L_{\text{sur}} : Y \times \mathbb{R} \to [0,\infty)$ be supervised loss functions, X be a complete measurable space, and μ be a probability measure on X. Assume that there exist mutually disjoint measurable subsets $A_n \subset X$ with $\mu(A_n) > 0$ for all $n \in \mathbb{N}$. Finally, let $\mathcal{Q}$ be a set of distributions on Y such that $\mathcal{C}^*_{L_{\text{tar}},\mathrm{Q}} < \infty$ and $\mathcal{C}^*_{L_{\text{sur}},\mathrm{Q}} < \infty$ for all $\mathrm{Q} \in \mathcal{Q}$. Show that there exists a distribution P on $X \times Y$ of type $\mathcal{Q}$ such that $\mathrm{P}_X = \mu$ and

$$\delta_{\max}(\varepsilon, \mathcal{Q}) = \inf_{x \in X} \delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x)\,, \qquad \varepsilon \in [0, \infty]. \tag{3.75}$$

Hint: First show that $U := \{\varepsilon > 0 : \delta_{\max}(\,\cdot\,, \mathcal{Q}) \text{ not continuous at } \varepsilon\}$ is at most countable. Then show equation (3.75) for an enumeration $(\varepsilon_n)_{n \in \mathbb{N}}$ of $U \cup \{r \in \mathrm{Q} : r \geq 0\}$. Use this to conclude the general case.

3.7. Characterization of calibration for detection losses ($\star\star\star$)
Prove Theorem 3.27 using the same idea as in Exercise 3.4.

3.8. Some more margin-based losses ($\star\star$)
Determine the calibration function with respect to the classification loss for the *exponential loss* given by $\varphi(t) := \exp(-t)$, $t \in \mathbb{R}$, and the *sigmoid loss* given by $\varphi(t) := 1 - \tanh t$, $t \in \mathbb{R}$. Is the latter classification calibrated?

3.9. Inequalities for unweighted classification ($\star\star$)
Use Theorems 3.34 and 3.22 to establish inequalities between the excess classification risk and the excess L-risk for L being the least squares loss, the hinge loss, the squared hinge loss, and the logistic loss for classification. How do these inequalities change when we additionally assume (3.40)?

3.10. Another weighted classification scenario ($\star\star\star$)
Let $h : X \to [0, \infty)$ be measurable. For the loss $L : X \times Y \times \mathbb{R} \to [0, \infty)$ defined by $L(x, y, t) := h(x) L_{\text{class}}(y, t)$, perform the following tasks:

i) Investigate which margin-based losses are L-calibrated.
ii) When are L-calibrated margin-based losses uniformly L-calibrated?
iii) Given a margin-based loss represented by some φ, determine the calibration function for the loss $(x, y, t) \mapsto h(x)\varphi(yt)$. Compare the results with those for the unweighted version.
iv) Find some practical situations in which L may be of interest.

3.11. Asymptotic relation between excess risks revisited ($\star\star$)
Show that in general a strictly positive calibration function is *not* sufficient for the implication (3.18).

Hint: Assume that L_{LS} is the target loss and that $L_{p\text{-dist}}$ is the surrogate loss for some $p \in [1, 2)$. Furthermore, consider the distribution P on $[0, 1) \times \mathbb{R}$ with P_X being the uniform distribution and $\mathrm{P}(\,\cdot\,|x) = \delta_{\{0\}}$ for all $x \in [0, 1]$.

3.12. Modulus of convexity for p-th power distance loss ($\star\star\star$)
For $p \in (1, 2)$, define $\psi : \mathbb{R} \to [0, \infty)$ by $\psi(t) := |t|^p$, $t \in \mathbb{R}$. Show for all $B > 0$ and $\varepsilon \in [0, B]$ that

$$\frac{p(p-1)}{2} B^{p-2} \varepsilon^2 \;\leq\; \delta_{\psi|_{[-B,B]}}(2\varepsilon) \;\leq\; \frac{p}{2(p-1)^2} B^{p-2} \varepsilon^2\,.$$

Hint: First show a $s^{a-1} \leq s^a - (s-1)^a \leq s^{a-1}$ for all $0 < a < 1$ and all $s \geq 1$. Use this to estimate $\psi'(t) - \psi'(t - \varepsilon)$, and then apply Lemma A.6.19.

3.13. Reverse calibration for DLD ($\star\star\star$)
Let μ be a probability measure on a measurable space X and $Y := \{-1, 1\}$. Furthermore, let $\rho > 0$ and $g : X \to [0, \infty)$ be a measurable function with $\|g\|_{L_1(\mu)} = 1$. Then, for $\mathrm{P} := g\mu \ominus_\rho \mu$ and all sequences (f_n) of measurable functions $f_n : X \to \mathbb{R}$, we have

$$\mathcal{R}_{L_{\mathrm{DLD}},\mathrm{P}}(f_n) \to 0 \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\mathrm{class}},\mathrm{P}}(f_n) \to \mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}}\,.$$

Hint: Compute the calibration function $\delta_{\max,L_{\mathrm{class}},\bar{L}_{\mathrm{DLD}}}(\,\cdot\,,\,\cdot\,)$ using Lemma 3.32. Then observe that $\mu(\{x \in X : \eta(x) = 1\}) = 0$ and use Corollary 3.19.

3.14. Another inequality for self-calibrated losses ($\star\star\star$)
Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a supervised loss that is self-calibrated with respect to some $\mathcal{Q} \subset \mathcal{Q}_{\min}(L)$ and P be a distribution on $X \times Y$ that is of type $\mathcal{Q}$. Assume further that there exist $p > 0$, $q > 0$, and a function $b : X \to [0, \infty]$ with $b^{-1} \in L_{p,\infty}(\mathrm{P}_X)$ and

$$\delta_{\max,\check{L}_{\mathrm{P}},L}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) \geq \varepsilon^q\, b(x)\,, \qquad \varepsilon > 0,\ x \in X.$$

Show that for all measurable $f : X \to \mathbb{R}$ we have

$$\mathrm{P}_X\big(\big\{x \in X : \check{L}_{\mathrm{P}}\big(x, f(x)\big) \geq \rho\big\}\big) \;\leq\; 2\left(\frac{\|b^{-1}\|_p\,\big(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\big)}{\rho^q}\right)^{\frac{p}{p+1}}.$$

If in addition $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$ has an almost surely unique minimizer $f^*_{L,\mathrm{P}}$, interpret the result in terms of Lorentz norms and compare it with Theorem 3.63.

Hint: Use the set A_ρ from the proof of Theorem 3.61 and apply Theorem 3.28.

3.15. Self-calibration of the (squared) hinge loss ($\star\star\star$)
i) Show that the self-calibration function of the hinge loss is given by

$$\delta_{\max,\check{L}_{\mathrm{hinge}},L_{\mathrm{hinge}}}(\varepsilon, \eta) = \begin{cases} \varepsilon\,\min\{\eta, 1-\eta, 2\eta - 1\} & \text{if } \eta \neq 0, 1/2, 1 \\ \varepsilon & \text{if } \eta \in \{0, 1\} \\ \infty & \text{if } \eta = 1/2 \end{cases}$$

for all $\varepsilon \in (0, 2]$, $\eta \in [0, 1]$. Is the hinge loss uniformly self-calibrated?
ii) Show that, for all distributions P on $X \times Y$ and all $p \in (0, \infty)$ and $\varepsilon > 0$, there exists a $\delta > 0$ such that for all measurable $f : X \to \mathbb{R}$ we have

$$\mathcal{R}_{L_{\mathrm{hinge}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{hinge}},\mathrm{P}} \;\leq\; \delta \quad \Longrightarrow \quad \|x \mapsto \check{L}_{\mathrm{hinge},\mathrm{P}}(x, \widehat{f}(x))\|_{L_p(\mathrm{P}_X)} \;\leq\; \varepsilon,$$

where the clipping is at ± 1. Compare this with Theorem 3.61. Find conditions on P such that Theorem 3.25 gives inequalities for clipped functions.
iii) Use Exercise 3.1 and Equation (3.68) to show that the squared hinge loss is *not* uniformly self-calibrated.

Hint: For the first implication in *ii)* use Theorem 3.17.

4

Kernels and Reproducing Kernel Hilbert Spaces

Overview. *We saw in Section 1.3 that kernels and their feature spaces are the devices by which the linear SVM approach produces non-linear decision functions. However, so far we only have a vague notion of kernels and hence we investigate them in more detail in this chapter.*

Prerequisites. *This chapter requires basic knowledge in functional analysis, which is provided in Section A.5. Section 4.4 on Gaussian kernels also needs some complex analysis from Section A.7.*

Usage. *Sections 4.1, 4.2, 4.3, and 4.6, which deal with the core material on kernels, are essential for the rest of this book. Moreover, Section 4.4 is needed for binary classification discussed in Chapter 8.*

As we have described in the introduction, one of the major steps in constructing a support vector machine is mapping the input space X into a feature space H that is equipped with an inner product. The benefit of this step is that for non-linear feature maps $\Phi : X \to H$, support vector machines can produce non-linear decision functions, although SVMs are only based on a linear discriminant approach. Furthermore, we have seen that SVMs only require computing the inner products $k(x, x') := \langle \Phi(x), \Phi(x') \rangle_H$ rather than Φ itself. Thus, instead of first constructing Φ and then computing the inner products, one can use SVMs with *efficiently* computable functions $k : X \times X \to \mathbb{R}$ that realize inner products of (possibly unknown) feature maps. We called such functions k kernels, and the approach described was called the kernel trick. Of course, this trick immediately raises some questions:

- When is a function $k : X \times X \to \mathbb{R}$ a kernel?
- How can we construct kernels?
- Given a kernel k, can we find a feature map and a feature space of k in a constructive way?
- How much does the kernel trick increase the expressive power of support vector machines?

The aim of this chapter is to answer these questions. To this end, we formalize the definition of kernels in Section 4.1. Moreover, in this section we also present some simple but useful examples of kernels. Then, in Section 4.2 we describe a canonical form of feature spaces, the so-called reproducing kernel Hilbert spaces. Basic properties of the functions contained in these spaces are

presented in Section 4.3. Moreover, for an important type of kernel we determine these spaces explicitly in Section 4.4. In Section 4.5, we derive a specific series representation for continuous kernels on compact spaces. Finally, in Section 4.6 we describe a class of kernels for which SVMs have a large expressive power.

4.1 Basic Properties and Examples of Kernels

In this section, we introduce the notions *kernel*, *feature space*, and *feature map*. Then we show how to construct new kernels from given kernels and present some important examples of kernels that will be used frequently in this book. Finally, we establish a criterion that characterizes kernels with the help of positive definite matrices related to these kernels.

Although in the context of machine learning one is originally only interested in real-valued kernels, we will develop the basic theory on kernels also for complex-valued kernels. This more general approach does not require any additional technical effort, but it will enable us in Section 4.4 to discover some features of the Gaussian RBF kernels that are widely used in practice.

Before we begin with the basic definitions, let us recall that every complex number $z \in \mathbb{C}$ can be represented in the form $z = x + iy$, where $x, y \in \mathbb{R}$ and $i := \sqrt{-1}$. Both x and y are uniquely determined, and in the following we thus write $\operatorname{Re} z := x$ and $\operatorname{Im} z := y$. Moreover, the conjugate of z is defined by $\bar{z} := x - iy$ and its absolute value is $|z| := \sqrt{z\bar{z}} = \sqrt{x^2 + y^2}$. In particular, we have $\bar{x} = x$ and $|x| = \sqrt{x^2}$ for all $x \in \mathbb{R}$. Furthermore, we use the symbol $\mathbb{K}$ whenever we want to treat the real and the complex cases simultaneously. For example, a $\mathbb{K}$-Hilbert space H is a real Hilbert space when considering $\mathbb{K} = \mathbb{R}$ and a complex one when $\mathbb{K} = \mathbb{C}$. Recall from Definition A.5.8 that in the latter case the inner product $\langle \cdot, \cdot \rangle_H : H \times H \to \mathbb{C}$ is sesqui-linear, i.e., $\langle x, \alpha x' \rangle_H = \bar{\alpha} \langle x, x' \rangle_H$, and anti-symmetric, i.e., $\langle x, x' \rangle_H = \overline{\langle x', x \rangle}_H$.

With the help of these preliminary considerations, we can now formalize the notion of kernels.

Definition 4.1. *Let X be a non-empty set. Then a function $k : X \times X \to \mathbb{K}$ is called a* ***kernel*** *on X if there exists a $\mathbb{K}$-Hilbert space H and a map $\Phi : X \to H$ such that for all $x, x' \in X$ we have*

$$k(x, x') \;=\; \langle \Phi(x'), \Phi(x) \rangle \,. \tag{4.1}$$

We call Φ a ***feature map*** *and H a* ***feature space*** *of k.*

Note that in the real case condition (4.1) can be replaced by the more natural equation $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$. In the complex case, however, $\langle \cdot, \cdot \rangle$ is anti-symmetric and hence (4.1) is equivalent to $k(x, x') = \overline{\langle \Phi(x), \Phi(x') \rangle}$.

Given a kernel, neither the feature map nor the feature space are uniquely determined. Let us illustrate this with a simple example. To this end, let

$X := \mathbb{R}$ and $k(x, x') := xx'$ for all $x, x' \in \mathbb{R}$. Then k is a kernel since obviously the identity map $\mathrm{id}_{\mathbb{R}}$ on $\mathbb{R}$ is a feature map with feature space $H := \mathbb{R}$. However, the map $\Phi : X \to \mathbb{R}^2$ defined by $\Phi(x) := (x/\sqrt{2}, x/\sqrt{2})$ for all $x \in X$ is also a feature map of k since we have

$$\langle \Phi(x'), \Phi(x) \rangle = \frac{x'}{\sqrt{2}} \cdot \frac{x}{\sqrt{2}} + \frac{x'}{\sqrt{2}} \cdot \frac{x}{\sqrt{2}} = xx' = k(x, x')$$

for all $x, x' \in X$. Moreover, note that a similar construction can be made for arbitrary kernels, and consequently every kernel has many different feature spaces. Finally, a less trivial example for different feature maps and spaces is discussed in Exercise 4.9.

Let us now present some commonly used kernels. To this end, we need some methods to construct kernels from scratch. We begin with a simple but instructive and fundamental observation.

Lemma 4.2. *Let X be a non-empty set and $f_n : X \to \mathbb{K}$, $n \in \mathbb{N}$, be functions such that $(f_n(x)) \in \ell_2$ for all $x \in X$. Then*

$$k(x, x') := \sum_{n=1}^{\infty} f_n(x)\overline{f_n(x')}, \qquad x, x' \in X, \tag{4.2}$$

defines a kernel on X.

Proof. Using Hölder's inequality for the sequence spaces ℓ_1 and ℓ_2, we obtain

$$\sum_{n=1}^{\infty} |f_n(x) f_n(x')| \leq \|(f_n(x))\|_{\ell_2} \, \|(f_n(x'))\|_{\ell_2},$$

and hence the series in (4.2) converges absolutely for all $x, x' \in X$. Now, we write $H := \ell_2$ and define $\Phi : X \to H$ by $\Phi(x) := (\overline{f_n(x)})$, $x \in X$. Then (4.2) immediately gives the assertion. □

We will see in the following that almost all kernels we are interested in have a series representation of the form (4.2). However, before we present some examples of such kernels, we first need to establish some results that allow us to construct new kernels from given ones. We begin with the following simple lemma, whose proof is left as an exercise.

Lemma 4.3 (Restriction of kernels). *Let k be a kernel on X, $\tilde{X}$ be a set, and $A : \tilde{X} \to X$ be a map. Then $\tilde{k}$ defined by $\tilde{k}(x, x') := k(A(x), A(x'))$, $x, x' \in X$, is a kernel on $\tilde{X}$. In particular, if $\tilde{X} \subset X$, then $k_{|\tilde{X} \times \tilde{X}}$ is a kernel.*

For a kernel $k : \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}$, Lemma 4.3 shows that the restriction $k_{|\mathbb{R}^d \times \mathbb{R}^d}$ is a kernel in the *complex* sense. The following result shows that it is also a kernel in the *real* sense if it satisfies $k(x, x') \in \mathbb{R}$ for all $x, x' \in \mathbb{R}^d$.

Lemma 4.4 (Real vs. complex kernels). *Let $k : X \times X \to \mathbb{C}$ be a kernel, H be a $\mathbb{C}$-Hilbert space, and $\Phi : X \to H$ be a feature map of k. Assume that we have $k(x, x') \in \mathbb{R}$ for all $x, x' \in X$. Then $H_0 := H$ equipped with the inner product*

$$\langle w, w' \rangle_{H_0} := \operatorname{Re} \langle w, w' \rangle_H , \qquad w,\, w' \in H_0,$$

is an $\mathbb{R}$-Hilbert space, and $\Phi : X \to H_0$ is a feature map of k.

Proof. It is elementary to check that $\langle \cdot\,, \cdot \rangle_{H_0}$ is a real inner product. Furthermore, we obviously have

$$k(x, x') = \big\langle \Phi(x'), \Phi(x) \big\rangle_H = \operatorname{Re} \big\langle \Phi(x'), \Phi(x) \big\rangle_H + i \operatorname{Im} \big\langle \Phi(x'), \Phi(x) \big\rangle_H$$

for all $x, x' \in X$. Consequently, $k(x, x') \in \mathbb{R}$ shows $\operatorname{Im} \langle \Phi(x'), \Phi(x) \rangle_H = 0$ for all $x, x' \in X$, and hence we obtain the assertion. □

Let us now establish some algebraic properties of the set of kernels on X. We begin with a simple lemma, whose proof is again left as an exercise.

Lemma 4.5 (Sums of kernels). *Let X be a set, $\alpha \geq 0$, and k, k_1, and k_2 be kernels on X. Then αk and $k_1 + k_2$ are also kernels on X.*

The preceding lemma states that the set of kernels on X is a cone. It is, however, not a vector space since in general differences of kernels are *not* kernels. To see this, let k_1 and k_2 be two kernels on X such that $k_1(x, x) - k_2(x, x) < 0$ for some $x \in X$. Then $k_1 - k_2$ is not a kernel since otherwise we would have a feature map $\Phi : X \to H$ of $k_1 - k_2$ with $0 \leq \langle \Phi(x), \Phi(x) \rangle = k_1(x, x) - k_2(x, x) < 0$. Let us now consider products of kernels.

Lemma 4.6 (Products of kernels). *Let k_1 be a kernel on X_1 and k_2 be a kernel on X_2. Then $k_1 \cdot k_2$ is a kernel on $X_1 \times X_2$. In particular, if $X_1 = X_2$, then $k(x, x') := k_1(x, x') k_2(x, x')$, $x, x' \in X$, defines a kernel on X.*

Proof. Let H_i be a feature space and $\Phi_i : X_i \to H_i$ be a feature map of k_i, $i = 1, 2$. Using the definition of the inner product in the tensor product space $H_1 \otimes H_2$ and its completion $H_1 \hat{\otimes} H_2$, see Appendix A.5.2, we obtain

$$\begin{aligned} k_1(x_1, x_1') \cdot k_2(x_2, x_2') &= \big\langle \Phi_1(x_1'), \Phi_1(x_1) \big\rangle_{H_1} \cdot \big\langle \Phi_2(x_2'), \Phi_2(x_2) \big\rangle_{H_2} \\ &= \big\langle \Phi_1(x_1') \otimes \Phi_2(x_2'),\, \Phi(x_1) \otimes \Phi_2(x_2) \big\rangle_{H_1 \hat{\otimes} H_2} , \end{aligned}$$

which shows that $\Phi_1 \otimes \Phi_2 : X_1 \times X_2 \to H_1 \hat{\otimes} H_2$ is a feature map of $k_1 \cdot k_2$. For the second assertion, we observe that k is a restriction of $k_1 \cdot k_2$. □

With the lemmas above, it is easy to construct non-trivial kernels. To illustrate this, let us assume for simplicity that $X := \mathbb{R}$. Then, for every integer $n \geq 0$, the map k_n defined by $k_n(x, x') := (xx')^n$, $x, x' \in X$, is a kernel by Lemma 4.2. Consequently, if $p : X \to \mathbb{R}$ is a polynomial of the form $p(t) = a_m t^m + \cdots + a_1 t + a_0$ with non-negative coefficients a_i, then $k(x, x') := p(xx')$,

$x, x' \in X$, defines a kernel on X by Lemma 4.5. In general, computing this kernel needs its feature map $\Phi(x) := (\sqrt{a}_m x^m, \ldots, \sqrt{a}_1 x, \sqrt{a}_0)$, $x \in X$, and consequently the computational requirements are determined by the degree m. However, for some polynomials, these requirements can be substantially reduced. Indeed, if for example we have $p(t) = (t+c)^m$ for some $c > 0$ and all $t \in \mathbb{R}$, then the time needed to compute k is independent of m. The following lemma, whose proof is left as an exercise, generalizes this idea.

Lemma 4.7 (Polynomial kernels). *Let $m \geq 0$ and $d \geq 1$ be integers and $c \geq 0$ be a real number. Then k defined by $k(z, z') := (\langle z, z' \rangle + c)^m$, $z, z' \in \mathbb{C}^d$, is a kernel on $\mathbb{C}^d$. Moreover, its restriction to $\mathbb{R}^d$ is an $\mathbb{R}$-valued kernel.*

Note that the polynomial kernels defined by $m = 1$ and $c = 0$ are called **linear kernels**. Instead of using polynomials for constructing kernels, one can use functions that can be represented by Taylor series. This is done in the following lemma.

Lemma 4.8. *Let $\mathring{B}_{\mathbb{C}}$ and $\mathring{B}_{\mathbb{C}^d}$ be the open unit balls of $\mathbb{C}$ and $\mathbb{C}^d$, respectively. Moreover, let $r \in (0, \infty]$ and $f : r\mathring{B}_{\mathbb{C}} \to \mathbb{C}$ be holomorphic with Taylor series*

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \qquad z \in r\mathring{B}_{\mathbb{C}}\,. \tag{4.3}$$

If $a_n \geq 0$ for all $n \geq 0$, then

$$k(z, z') := f(\langle z, z' \rangle)_{\mathbb{C}^d} = \sum_{n=0}^{\infty} a_n \langle z, z' \rangle_{\mathbb{C}^d}^n\,, \qquad z, z' \in \sqrt{r}\mathring{B}_{\mathbb{C}^d},$$

*defines a kernel on $\sqrt{r}\mathring{B}_{\mathbb{C}^d}$ whose restriction to $X := \{x \in \mathbb{R}^d : \|x\|_2 < \sqrt{r}\}$ is a real-valued kernel. We say that k is a kernel of **Taylor type**.*

Proof. For $z, z' \in \sqrt{r}\mathring{B}_{\mathbb{C}^d}$, we have $|\langle z, z' \rangle| \leq \|z\|_2 \|z'\|_2 < r$ and thus k is well-defined. Let z_i denote the i-th component of $z \in \mathbb{C}^d$. Since (4.3) is absolutely convergent, the multinomial formula (see Lemma A.1.2) then yields

$$\begin{aligned} k(z, z') = \sum_{n=0}^{\infty} a_n \left(\sum_{j=1}^{d} z_j \bar{z}'_j \right)^n &= \sum_{n=0}^{\infty} a_n \sum_{\substack{j_1, \ldots, j_d \geq 0 \\ j_1 + \cdots + j_d = n}} c_{j_1, \ldots, j_d} \prod_{i=1}^{d} (z_i \bar{z}'_i)^{j_i} \\ &= \sum_{j_1, \ldots, j_d \geq 0} a_{j_1 + \cdots + j_d} c_{j_1, \ldots, j_d} \prod_{i=1}^{d} (\bar{z}'_i)^{j_i} \prod_{i=1}^{d} z_i^{j_i}\,, \end{aligned}$$

where $c_{j_1, \ldots, j_d} := \frac{n!}{\prod_{i=1}^{d} j_i!}$. Let us define $\Phi : X \to \ell_2(\mathbb{N}_0^d)$ by

$$\Phi(z) := \left(\sqrt{a_{j_1 + \cdots + j_d} c_{j_1, \ldots, j_d}} \prod_{i=1}^{d} \bar{z}_i^{j_i} \right)_{j_1, \ldots, j_d \geq 0}, \qquad z \in \sqrt{r}\mathring{B}_{\mathbb{C}^d}.$$

Then we have $k(z, z') = \langle \Phi(z'), \Phi(z) \rangle_{\ell_2(\mathbb{N}_0^d)}$ for all $z, z' \in \sqrt{r}\mathring{B}_{\mathbb{C}^d}$, and hence k is a kernel. The assertion for the restriction is obvious. □

With the help of Lemma 4.8 we can now present some more examples of commonly used kernels.

Example 4.9. For $d \in \mathbb{N}$ and $x, x' \in \mathbb{K}^d$, we define $k(x, x') := \exp(\langle x, x' \rangle)$. Then k is a $\mathbb{K}$-valued kernel on $\mathbb{K}^d$ called the **exponential kernel**. $\triangleleft$

Example 4.9 can be used to introduce the following kernel that is often used in practice and will be considered in several parts of the book.

Proposition 4.10. *For $d \in \mathbb{N}$, $\gamma > 0$, $z = (z_1, \ldots, z_d) \in \mathbb{C}^d$, and $z' = (z'_1, \ldots, z'_d) \in \mathbb{C}^d$, we define*

$$k_{\gamma,\mathbb{C}^d}(z, z') := \exp\biggl(-\gamma^{-2} \sum_{j=1}^{d} (z_j - \bar{z}'_j)^2\biggr).$$

*Then $k_{\gamma,\mathbb{C}^d}$ is a kernel on $\mathbb{C}^d$, and its restriction $k_\gamma := (k_{\gamma,\mathbb{C}^d})_{|\mathbb{R}^d \times \mathbb{R}^d}$ is an $\mathbb{R}$-valued kernel, which is called the **Gaussian RBF kernel** with width γ. Moreover, k_γ can be computed by*

$$k_\gamma(x, x') = \exp\biggl(-\frac{\|x - x'\|_2^2}{\gamma^2}\biggr), \qquad x, x' \in \mathbb{R}^d.$$

Proof. Let us fix $z, z' \in \mathbb{C}^d$. Decomposing $k_{\gamma,\mathbb{C}^d}$ into

$$k_{\gamma,\mathbb{C}^d}(z, z') = \frac{\exp(2\gamma^{-2}\langle z, z' \rangle)}{\exp\bigl(\gamma^{-2} \sum_{j=1}^{d} z_j^2\bigr) \exp\bigl(\gamma^{-2} \sum_{j=1}^{d} (\bar{z}'_j)^2\bigr)}$$

and applying Lemmas 4.3 and 4.6, and Example 4.9 then yields the first assertion. The second assertion is trivial. $\square$

Besides the Gaussian RBF kernel, many other $\mathbb{R}$-valued kernels can be constructed using Lemma 4.8. Here we only give one more example and refer to Exercise 4.1 for some more examples.

Example 4.11. Let $X := \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$ and $\alpha > 0$. Then $k(x, x') := (1 - \langle x, x' \rangle)^{-\alpha}$ defines a kernel on X called a **binomial kernel**. Indeed, the binomial series $(1 - t)^{-\alpha} = \sum_{n=0}^{\infty} \binom{-\alpha}{n} (-1)^n t^n$ holds for all $|t| < 1$, where $\binom{\beta}{n} := \prod_{i=1}^{n} (\beta - i + 1)/i$. Now the assertion follows from $\binom{-\alpha}{n} (-1)^n > 0$. $\triangleleft$

The results above are based on Taylor series expansions. Instead of these expansions, one can also employ Fourier series expansions for constructing kernels. In the case $\mathbb{K} = \mathbb{R}$, the corresponding result reads as follows.

Lemma 4.12. *Let $f : [-2\pi, 2\pi] \to \mathbb{R}$ be a continuous function that can be expanded in a pointwise convergent Fourier series of the form*

$$f(t) = \sum_{n=0}^{\infty} a_n \cos(nt). \tag{4.4}$$

*If $a_n \geq 0$ holds for all $n \geq 0$, then $k(x, x') := \prod_{i=1}^{d} f(x_i - x'_i)$ defines a kernel on $[0, 2\pi)^d$. We say that k is a kernel of **Fourier type**.*

Proof. By induction and Lemma 4.6, we may restrict ourselves to $d = 1$. Then, letting $t = 0$ in (4.4), we get $(a_n)_{n\geq 0} \in \ell_1$, and thus the definition of k yields

$$k(x, x') = a_0 + \sum_{n=1}^{\infty} a_n \sin(nx) \sin(nx') + \sum_{n=1}^{\infty} a_n \cos(nx) \cos(nx')$$

for all $x, x' \in [0, 2\pi)$. Now the assertion follows from Lemma 4.2. □

The following two examples can be treated with the help of Lemma 4.12.

Example 4.13. For fixed $0 < q < 1$ and all $t \in [-2\pi, 2\pi]$, we define

$$f(t) := \frac{1 - q^2}{2 - 4q \cos t + 2q^2} .$$

Then $k(x, x') := \prod_{i=1}^{d} f(x_i - x'_i)$, $x, x' \in [0, 2\pi]^d$, is a kernel since we have $f(t) = 1/2 + \sum_{n=1}^{\infty} q^n \cos(nt)$ for all $t \in [0, 2\pi]$. ◁

Example 4.14. For fixed $1 < q < \infty$ and all $t \in [-2\pi, 2\pi]$, we define

$$f(t) := \frac{\pi q \cosh(\pi q - qt)}{2 \sinh(\pi q)} .$$

Then $k(x, x') := \prod_{i=1}^{d} f(x_i - x'_i)$, $x, x' \in [0, 2\pi]^d$, is a kernel since we have $f(t) = 1/2 + \sum_{n=1}^{\infty} (1 + q^{-2} n^2)^{-1} \cos(nt)$ for all $t \in [0, 2\pi]$. ◁

Although we have already seen several techniques to construct kernels, in general we still have to find a feature space in order to decide whether a given function k is a kernel. Since this can sometimes be a difficult task, we will now present a criterion that characterizes $\mathbb{R}$-valued kernels in terms of *inequalities*. To this end, we need the following definition.

Definition 4.15. *A function $k : X \times X \to \mathbb{R}$ is called **positive definite** if, for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and all $x_1, \ldots, x_n \in X$, we have*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) \geq 0 . \tag{4.5}$$

*Furthermore, k is said to be **strictly positive definite** if, for mutually distinct $x_1, \ldots, x_n \in X$, equality in (4.5) only holds for $\alpha_1 = \cdots = \alpha_n = 0$. Finally, k is called **symmetric** if $k(x, x') = k(x', x)$ for all $x, x' \in X$.*

Unfortunately, there is no common use of the preceding definitions in the literature. Indeed, some authors call positive definite functions *positive semi-definite*, and strictly positive definite functions are sometimes called positive definite. Moreover, for fixed $x_1, \ldots, x_n \in X$, the $n \times n$ matrix $K := (k(x_j, x_i))_{i,j}$ is often called the **Gram matrix**. Note that (4.5) is equivalent to saying that the Gram matrix is positive definite.

Obviously, if k is an $\mathbb{R}$-valued kernel with feature map $\Phi : X \to H$, then k is symmetric since the inner product in H is symmetric. Moreover, k is also positive definite since for $n \in \mathbb{N}$, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, and $x_1, \dots, x_n \in X$ we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) = \left\langle \sum_{i=1}^{n} \alpha_i \Phi(x_i), \sum_{j=1}^{n} \alpha_j \Phi(x_j) \right\rangle_H \geq 0 \,. \tag{4.6}$$

Now the following theorem shows that symmetry and positive definiteness are not only necessary for k to be a kernel but also sufficient.

Theorem 4.16 (Symmetric, positive definite functions are kernels). *A function $k : X \times X \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.*

Proof. In view of the discussion above, it suffices to show that a symmetric and positive definite function k is a kernel. To this end, we write

$$H_{\text{pre}} := \left\{ \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i) : n \in \mathbb{N},\, \alpha_1, \dots, \alpha_n \in \mathbb{R},\, x_1, \dots, x_n \in X \right\},$$

and for $f := \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i) \in H_{\text{pre}}$ and $g := \sum_{j=1}^{m} \beta_j k(\,\cdot\,, x'_j) \in H_{\text{pre}}$, we define

$$\langle f, g \rangle := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x'_j, x_i) \,.$$

Note that this definition is independent of the representation of f since we have $\langle f, g \rangle = \sum_{j=1}^{m} \beta_j f(x'_j)$. Furthermore, by the symmetry of k, we find $\langle f, g \rangle = \sum_{i=1}^{n} \alpha_i g(x_i)$, i.e., the definition is also independent of the representation of g. Moreover, the definition immediately shows that $\langle \,\cdot\,, \cdot\, \rangle$ is bilinear and symmetric, and since k is positive definite, $\langle \,\cdot\,, \cdot\, \rangle$ is also positive, i.e., $\langle f, f \rangle \geq 0$ for all $f \in H_{\text{pre}}$. Conequently (see Exercise 4.3), $\langle \,\cdot\,, \cdot\, \rangle$ satisfies the Cauchy-Schwarz inequality, i.e.,

$$|\langle f, g \rangle|^2 \leq \langle f, f \rangle \cdot \langle g, g \rangle \,, \qquad f, g \in H_{\text{pre}} \,.$$

Now let $f := \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i) \in H_{\text{pre}}$ with $\langle f, f \rangle = 0$. Then, for all $x \in X$, we have

$$|f(x)|^2 = \left| \sum_{i=1}^{n} \alpha_i k(x, x_i) \right|^2 = \left| \langle f, k(\,\cdot\,, x) \rangle \right|^2 \leq \langle k(\,\cdot\,, x), k(\,\cdot\,, x) \rangle \cdot \langle f, f \rangle = 0 \,,$$

and hence we find $f = 0$. Therefore we have shown that $\langle \,\cdot\,, \cdot\, \rangle$ is an inner product on H_{pre}. Let H be a completion of H_{pre} and $I : H_{\text{pre}} \to H$ be the corresponding isometric embedding. Then H is a Hilbert space and we have

$$\langle Ik(\,\cdot\,, x'), Ik(\,\cdot\,, x) \rangle_H = \langle k(\,\cdot\,, x'), k(\,\cdot\,, x) \rangle_{H_{\text{pre}}} = k(x, x')$$

for all $x, x' \in X$, i.e., $x \mapsto Ik(\,\cdot\,, x)$, $x \in X$, defines a feature map of k. □

The characterization above is often useful for checking whether a given function is a kernel. Let us illustrate this with the following example.

Corollary 4.17 (Limits of kernels are kernels). *Let (k_n) be a sequence of kernels on the set X that converges pointwise to a function $k : X \times X \to \mathbb{R}$, i.e., $\lim_{n\to\infty} k_n(x, x') = k(x, x')$ for all $x, x' \in X$. Then k is a kernel on X.*

Proof. Every k_n is symmetric and satisfies (4.5). Therefore, the same is true for the pointwise limit k. □

4.2 The Reproducing Kernel Hilbert Space of a Kernel

In this section, we will introduce reproducing kernel Hilbert spaces (RKHSs) and describe their relation to kernels. In particular, it will turn out that the RKHS of a kernel is in a certain sense the smallest feature space of this kernel and consequently can serve as a canonical feature space.

Let us begin with the following fundamental definitions.

Definition 4.18. *Let $X \neq \emptyset$ and H be a $\mathbb{K}$-Hilbert function space over X, i.e., a $\mathbb{K}$-Hilbert space that consists of functions mapping from X into $\mathbb{K}$.*

*i) A function $k : X \times X \to \mathbb{K}$ is called a **reproducing kernel** of H if we have $k(\cdot, x) \in H$ for all $x \in X$ and the **reproducing property***

$$f(x) = \langle f, k(\cdot, x)\rangle$$

holds for all $f \in H$ and all $x \in X$.

*ii) The space H is called a **reproducing kernel Hilbert space (RKHS)** over X if for all $x \in X$ the Dirac functional $\delta_x : H \to \mathbb{K}$ defined by*

$$\delta_x(f) := f(x), \qquad f \in H,$$

is continuous.

Note that $L_2(\mathbb{R}^d)$ does *not* consist of functions and consequently it is not an RKHS. For a generalization of this statement, we refer to Exercise 4.2.

Reproducing kernel Hilbert spaces have the remarkable and, as we will see later, important property that norm convergence implies pointwise convergence. More precisely, let H be an RKHS, $f \in H$, and $(f_n) \subset H$ be a sequence with $\|f_n - f\|_H \to 0$ for $n \to \infty$. Then, for all $x \in X$, we have

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \delta_x(f_n) = \delta_x(f) = f(x) \tag{4.7}$$

by the assumed continuity of the Dirac functionals. Furthermore, reproducing kernels are actually kernels in the sense of Definition 4.1, as the following lemma shows.

Lemma 4.19 (Reproducing kernels are kernels). *Let H be a Hilbert function space over X that has a reproducing kernel k. Then H is an RKHS and H is also a feature space of k, where the feature map $\Phi : X \to H$ is given by*

$$\Phi(x) = k(\cdot\,, x)\,, \qquad x \in X.$$

We call Φ the ***canonical feature map****.*

Proof. The reproducing property says that each Dirac functional can be represented by the reproducing kernel, and consequently we obtain

$$|\delta_x(f)| = |f(x)| = |\langle f, k(\cdot\,, x)\rangle| \leq \|k(\cdot\,, x)\|_H \, \|f\|_H \tag{4.8}$$

for all $x \in X$, $f \in H$. This shows the continuity of the functionals δ_x, $x \in X$.

In order to show the second assertion, we fix an $x' \in X$ and write $f := k(\cdot\,, x')$. Then, for $x \in X$, the reproducing property yields

$$\langle \Phi(x'), \Phi(x)\rangle = \langle k(\cdot\,, x'), k(\cdot\,, x)\rangle = \langle f, k(\cdot\,, x)\rangle = f(x) = k(x, x')\,. \qquad \square$$

We have just seen that every Hilbert function space with a reproducing kernel is an RKHS. The following theorem now shows that, conversely, every RKHS has a (unique) reproducing kernel, and that this kernel can be determined by the Dirac functionals.

Theorem 4.20 (Every RKHS has a unique reproducing kernel). *Let H be an RKHS over X. Then $k : X \times X \to \mathbb{K}$ defined by*

$$k(x, x') := \langle \delta_x, \delta_{x'}\rangle_H\,, \qquad x, x' \in X,$$

is the only reproducing kernel of H. Furthermore, if $(e_i)_{i\in I}$ is an orthonormal basis of H, then for all $x, x' \in X$ we have

$$k(x, x') = \sum_{i\in I} e_i(x)\overline{e_i(x')}\,. \tag{4.9}$$

Proof. We first show that k is a reproducing kernel. To this end, let $I : H' \to H$ be the isometric anti-linear isomorphism derived from Theorem A.5.12 that assigns to every functional in H' the representing element in H, i.e., $g'(f) = \langle f, Ig'\rangle$ for all $f \in H$, $g' \in H'$. Then, for all $x, x' \in X$, we have

$$k(x, x') = \langle \delta_x, \delta_{x'}\rangle_{H'} = \langle I\delta_{x'}, I\delta_x\rangle_H = \delta_x(I\delta_{x'}) = I\delta_{x'}(x)\,,$$

which shows $k(\cdot\,, x') = I\delta_{x'}$ for all $x' \in X$. From this we immediately obtain

$$f(x') = \delta_{x'}(f) = \langle f, I\delta_{x'}\rangle = \langle f, k(\cdot\,, x')\rangle\,,$$

i.e., k has the reproducing property. Now let $\tilde{k}$ be an *arbitrary* reproducing kernel of H. For $x' \in X$, the basis representation of $\tilde{k}(\cdot\,, x')$ then yields

$$\tilde{k}(\,\cdot\,,x') = \sum_{i\in I} \langle \tilde{k}(\,\cdot\,,x'), e_i\rangle e_i = \sum_{i\in I} \overline{e_i(x')} e_i\,,$$

where the convergence is with respect to $\|\cdot\|_H$. Using (4.7), we thus obtain (4.9) for $\tilde{k}$. Finally, since $\tilde{k}$ and $(e_i)_{i\in I}$ were arbitrarily chosen, we find $\tilde{k} = k$, i.e., k is the only reproducing kernel of H. □

Theorem 4.20 shows that an RKHS uniquely determines its reproducing kernel, which is actually a kernel by Lemma 4.19. The following theorem now shows that, conversely, every kernel has a unique RKHS. Consequently, we have a one-to-one relation between kernels and RKHSs. In addition, the following theorem shows that the RKHS of a kernel is in some sense the smallest feature space, and hence it can be considered as the "natural" feature space.

Theorem 4.21 (Every kernel has a unique RKHS). *Let $X \neq \emptyset$ and k be a kernel over X with feature space H_0 and feature map $\Phi_0 : X \to H_0$. Then*

$$H := \big\{ f : X \to \mathbb{K} \,\big|\, \exists\, w \in H_0 \text{ with } f(x) = \langle w, \Phi_0(x)\rangle_{H_0} \text{ for all } x \in X \big\} \tag{4.10}$$

equipped with the norm

$$\|f\|_H := \inf\big\{ \|w\|_{H_0} : w \in H_0 \text{ with } f = \langle w, \Phi_0(\,\cdot\,)\rangle_{H_0} \big\} \tag{4.11}$$

is the only RKHS for which k is a reproducing kernel. Consequently, both definitions are independent of the choice of H_0 and Φ_0. Moreover, the operator $V : H_0 \to H$ defined by

$$Vw := \langle w, \Phi_0(\,\cdot\,)\rangle_{H_0}\,, \qquad\qquad w \in H_0,$$

is a metric surjection, i.e. $V\mathring{B}_{H_0} = \mathring{B}_H$, where $\mathring{B}_{H_0}$ and $\mathring{B}_H$ are the open unit balls of H_0 and H, respectively. Finally, the set

$$H_{\text{pre}} := \left\{ \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i) : n \in \mathbb{N},\, \alpha_1, \ldots, \alpha_n \in \mathbb{K},\, x_1, \ldots, x_n \in X \right\} \tag{4.12}$$

is dense in H, and for $f := \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i) \in H_{\text{pre}}$ we have

$$\|f\|_H^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \overline{\alpha_j} k(x_j, x_i)\,. \tag{4.13}$$

Proof. Let us first show that H is a Hilbert function space over X. To this end, observe that H is obviously a vector space of functions from X to $\mathbb{K}$, and V is a surjective linear operator. Furthermore, for all $f \in H$, we have

$$\|f\|_H = \inf_{w \in V^{-1}(\{f\})} \|w\|_{H_0}\,,$$

where $V^{-1}(\{f\})$ denotes the pre-image of f under V. Let us show that $\|\cdot\|_H$ is a Hilbert space norm on H. To this end, let $(w_n) \subset \ker V$ be a convergent sequence in the null space $\ker V := \{w \in H_0 : Vw = 0\}$ of V and $w \in H_0$ its limit. Then we have $\langle w, \Phi(x)\rangle = \lim_{n\to\infty}\langle w_n, \Phi(x)\rangle = 0$ for all $x \in X$. Since this shows $w \in \ker V$, the null space $\ker V$ is a closed subspace of H_0. Let $\hat{H}$ denote its orthogonal complement so that we have the orthogonal decomposition $H_0 = \ker V \oplus \hat{H}$. Then the restriction $V_{|\hat{H}} : \hat{H} \to H$ of V to $\hat{H}$ is injective by construction. Let us show that it is also surjective. To this end, let $f \in H$ and $w \in H_0$ with $f = Vw$. Since this w can be decomposed into $w = w_0 + \hat{w}$ for suitable $w_0 \in \ker V$ and $\hat{w} \in \hat{H}$, we get $f = V(w_0 + \hat{w}) = V_{|\hat{H}}\hat{w}$, which shows the surjectivity of $V_{|\hat{H}}$. Furthermore, a similar reasoning gives

$$\begin{aligned}\|f\|_H^2 = \inf_{\substack{w_0\in\ker V,\, \hat{w}\in\hat{H}\\ w_0+\hat{w}\in V^{-1}(\{f\})}} \|w_0+\hat{w}\|_{H_0}^2 &= \inf_{\substack{w_0\in\ker V,\, \hat{w}\in\hat{H}\\ w_0+\hat{w}\in V^{-1}(\{f\})}} \|w_0\|_{H_0}^2 + \|\hat{w}\|_{H_0}^2 \\ &= \left\| \left(V_{|\hat{H}}\right)^{-1} f \right\|_{\hat{H}}^2 ,\end{aligned}$$

where $(V_{|\hat{H}})^{-1}$ denotes the inverse operator of $V_{|\hat{H}}$. From the equation above and the fact that $\hat{H}$ is a Hilbert space, we can immediately deduce that $\|\cdot\|_H$ is a Hilbert space norm on H and that $V_{|\hat{H}} : \hat{H} \to H$ is an isometric isomorphism. Furthermore, from the definition of V and $\|\cdot\|_H$, we can easily conclude that V is a *metric* surjection.

Let us now show that k is a reproducing kernel of H. To this end, observe that for $x \in X$ we have $k(\cdot, x) = \langle \Phi_0(x), \Phi_0(\cdot)\rangle = V\Phi_0(x) \in H$. Furthermore, we have $\langle w, \Phi_0(x)\rangle = 0$ for all $w \in \ker V$, which shows $\Phi_0(x) \in (\ker V)^\perp = \hat{H}$. Since $V_{|\hat{H}} : \hat{H} \to H$ is isometric, we therefore obtain

$$f(x) = \left\langle (V_{|\hat{H}})^{-1} f, \Phi_0(x)\right\rangle_{H_0} = \left\langle f, V_{|\hat{H}}\Phi_0(x)\right\rangle_H = \langle f, k(\cdot, x)\rangle_H$$

for all $f \in H$, $x \in X$, i.e., k has the reproducing property. By Lemma 4.19 we conclude that H is an RKHS.

Let us now show that the assertions on H_{pre} are true for an *arbitrary* RKHS $\tilde{H}$ for which k is a reproducing kernel. To this end, we first observe that $k(\cdot, x) \in \tilde{H}$ for all $x \in X$ implies $H_{\mathrm{pre}} \subset \tilde{H}$. Now let us suppose that H_{pre} was not dense in $\tilde{H}$. This assumption yields $(H_{\mathrm{pre}})^\perp \neq \{0\}$, and hence there would exist an $f \in (H_{\mathrm{pre}})^\perp$ and an $x \in X$ with $f(x) \neq 0$. Since this would imply

$$0 = \langle f, k(\cdot, x)\rangle = f(x) \neq 0 ,$$

we see that H_{pre} is dense in $\tilde{H}$. Finally, for $f := \sum_{i=1}^n \alpha_i k(\cdot, x_i) \in H_{\mathrm{pre}}$, the reproducing property implies

$$\|f\|_{\tilde{H}}^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \overline{\alpha_j} \left\langle k(\cdot, x_i), k(\cdot, x_j)\right\rangle_{\tilde{H}} = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \overline{\alpha_j} k(x_j, x_i) .$$

Let us now prove that k has only one RKHS. To this end, let H_1 and H_2 be two RKHSs of k. We have seen in the previous step that H_{pre} is dense in both

H_1 and H_2 and that the norms of H_1 and H_2 coincide on H_{pre}. Let us choose an $f \in H_1$. Then there exists a sequence $(f_n) \subset H_{\mathrm{pre}}$ with $\|f_n - f\|_{H_1} \to 0$. Since $H_{\mathrm{pre}} \subset H_2$, the sequence (f_n) is also contained in H_2, and since the norms of H_1 and H_2 coincide on H_{pre}, the sequence (f_n) is a Cauchy sequence in H_2. Therefore, there exists a $g \in H_2$ with $\|f_n - g\|_{H_2} \to 0$. Since convergence with respect to an RKHS norm implies pointwise convergence, see (4.7), we then find $f(x) = g(x)$ for all $x \in X$, i.e., we have shown $f \in H_2$. Furthermore, $\|f_n - f\|_{H_1} \to 0$ and $\|f_n - f\|_{H_2} \to 0$ imply

$$\|f\|_{H_1} = \lim_{n\to\infty} \|f_n\|_{H_1} = \lim_{n\to\infty} \|f_n\|_{H_{\mathrm{pre}}} = \lim_{n\to\infty} \|f_n\|_{H_2} = \|f\|_{H_2} \,,$$

i.e., H_1 is isometrically included in H_2. Since the converse inclusion $H_2 \subset H_1$ can be shown analogously, we obtain $H_1 = H_2$ with equal norms. □

Theorem 4.21 describes the RKHS H of a given kernel k as the "smallest" feature space of k in the sense that there is a canonical metric surjection V from any other feature space H_0 of k onto H. However, for kernelized algorithms, it is more the specific *form* (4.10) that makes the RKHS important. To illustrate this, recall from the introduction that the soft margin SVM produces decision functions of the form $x \mapsto \langle w, \Phi_0(x)\rangle$, where $\Phi_0 : X \to H_0$ is a feature map of k and $w \in H_0$ is a suitable weight vector. Now, (4.10) states that the RKHS associated to k consists exactly of all possible functions of this form. Moreover, (4.10) shows that this set of functions does not change if we consider different feature spaces or feature maps of k.

Theorem 4.21 can often be used to determine the RKHS of a given kernel and its modifications such as restrictions and normalization (see Exercise 4.4 for more details). To illustrate this, let us recall that every $\mathbb{C}$-valued kernel on X that is actually $\mathbb{R}$-valued has an $\mathbb{R}$-feature space by Lemma 4.4. The following corollary of Theorem 4.21 describes the corresponding $\mathbb{R}$-RKHS.

Corollary 4.22. *Let $k : X \times X \to \mathbb{C}$ be a kernel and H its corresponding $\mathbb{C}$-RKHS. If we actually have $k(x, x') \in \mathbb{R}$ for all $x, x' \in X$, then*

$$H_{\mathbb{R}} := \bigl\{ f : X \to \mathbb{R} \bigm| \exists\, g \in H \text{ with } \operatorname{Re} g = f \bigr\}$$

equipped with the norm

$$\|f\|_{H_{\mathbb{R}}} := \inf\bigl\{ \|g\|_H : g \in H \text{ with } \operatorname{Re} g = f \bigr\}\,, \qquad f \in H_{\mathbb{R}},$$

is the $\mathbb{R}$-RKHS of the $\mathbb{R}$-valued kernel k.

Proof. We have already seen in Lemma 4.4 that $H_0 := H$ equipped with the inner product

$$\langle f, f'\rangle_{H_0} := \operatorname{Re}\langle f, f'\rangle_H\,, \qquad f,\, f' \in H_0,$$

is an $\mathbb{R}$-feature space of the $\mathbb{R}$-valued kernel k. Moreover, for $f \in H_0$ and $x \in X$, we have

$$f(x) = \langle f, \Phi(x)\rangle_H = \operatorname{Re}\langle f, \Phi(x)\rangle_H + i \operatorname{Im}\langle f, \Phi(x)\rangle_H = \langle f, \Phi(x)\rangle_{H_0} + i \operatorname{Im} f(x),$$

i.e., we have found $\langle f, \Phi(x)\rangle_{H_0} = \operatorname{Re} f(x)$. Now, the assertion follows from Theorem 4.21. □

Let us finally assume that we have an RKHS H with kernel $k : \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}$ whose restriction to $\mathbb{R}^d$ is $\mathbb{R}$-valued, i.e., $k_{|\mathbb{R}^d \times \mathbb{R}^d} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Then combining the preceding corollary with Exercise 4.4 shows that

$$H_{\mathbb{R}} := \bigl\{ f : \mathbb{R}^d \to \mathbb{R} \,\big|\, \exists\, g : \mathbb{C}^d \to \mathbb{C} \text{ with } g \in H \text{ and } \operatorname{Re} g_{|\mathbb{R}^d} = f \bigr\}$$

equipped with the norm

$$\|f\|_{H_{\mathbb{R}}} := \inf\bigl\{ \|g\|_H : g \in H \text{ with } \operatorname{Re} g_{|\mathbb{R}^d} = f \bigr\}, \qquad f \in H_{\mathbb{R}},$$

is the $\mathbb{R}$-RKHS of the restriction $k_{|\mathbb{R}^d \times \mathbb{R}^d}$.

4.3 Properties of RKHSs

Usually, a kernel has additional properties such as measurability, continuity, or differentiability. In this section, we investigate whether the functions of its associated RKHS share these properties.

Let us begin by observing that for a kernel k on X with RKHS H the Cauchy-Schwarz inequality and the reproducing property imply

$$\begin{aligned} |k(x,x')|^2 = \bigl|\bigl\langle k(\cdot\,,x'), k(\cdot\,,x)\bigr\rangle_H\bigr|^2 &\le \|k(\cdot\,,x')\|_H^2 \cdot \|k(\cdot\,,x)\|_H^2 \\ &= k(x',x') \cdot k(x,x) \end{aligned} \tag{4.14}$$

for all $x, x' \in X$. This yields $\sup_{x,x' \in X} |k(x,x')| = \sup_{x \in X} k(x,x)$, and hence k is **bounded** if and only if

$$\|k\|_\infty := \sup_{x \in X} \sqrt{k(x,x)} < \infty\,. \tag{4.15}$$

Now, let $\Phi : X \to H_0$ be a feature map of k. Then we find $\|\Phi(x)\|_{H_0} = \sqrt{k(x,x)}$ for all $x \in X$, and hence Φ is bounded if and only if k is. The following lemma provides another important characterization of bounded kernels.

Lemma 4.23 (RKHSs of bounded kernels). *Let X be a set and k be a kernel on X with RKHS H. Then k is bounded if and only if every $f \in H$ is bounded. Moreover, in this case the inclusion* $\operatorname{id} : H \to \ell_\infty(X)$ *is continuous and we have* $\|\operatorname{id} : H \to \ell_\infty(X)\| = \|k\|_\infty$.

Proof. Let us first assume that k is bounded. Then the Cauchy-Schwarz inequality and the reproducing property imply

$$|f(x)| = |\langle f, k(\cdot\,,x)\rangle| \le \|f\|_H \sqrt{k(x,x)} \le \|f\|_H \|k\|_\infty$$

for all $f \in H$, $x \in X$. Hence we obtain $\|f\|_\infty \leq \|k\|_\infty \|f\|_H$ for all $f \in H$, which shows that $\operatorname{id} : H \to \ell_\infty(X)$ is well-defined and $\|\operatorname{id} : H \to \ell_\infty(X)\| \leq \|k\|_\infty$.

Conversely, let us now assume that every $f \in H$ is bounded. Then the inclusion $\operatorname{id} : H \to \ell_\infty(X)$ is well-defined. Our first goal is to show that the inclusion is continuous. To this end, we fix a sequence $(f_n) \subset H$ for which there exist an $f \in H$ and a $g \in \ell_\infty(X)$ such that $\lim_{n\to\infty} \|f_n - f\|_H = 0$ and $\lim_{n\to\infty} \|\operatorname{id} f_n - g\|_\infty = 0$. Then the first convergence implies $f_n(x) \to f(x)$ for all $x \in X$, while the second convergence implies $f_n(x) \to g(x)$ for all $x \in X$. We conclude $f = g$ and hence $\operatorname{id} : H \to \ell_\infty(X)$ is continuous by the closed graph theorem, see Theorem A.5.4. For $x \in X$, we then have

$$|k(x,x)| \leq \|k(\cdot\,,x)\|_\infty \leq \|\operatorname{id} : H \to \ell_\infty(X)\| \cdot \|k(\cdot\,,x)\|_H = \|\operatorname{id}\| \sqrt{k(x,x)}\,,$$

i.e., $\sqrt{k(x,x)} \leq \|\operatorname{id}\|$. This shows $\|k\|_\infty \leq \|\operatorname{id} : H \to \ell_\infty(X)\|$. □

Our next goal is to investigate measurable kernels and their integrability. We begin with the following lemma.

Lemma 4.24 (RKHSs of measurable kernels). *Let X be a measurable space and k be a kernel on X with RKHS H. Then all $f \in H$ are measurable if and only if $k(\cdot\,,x) : X \to \mathbb{R}$ is measurable for all $x \in X$.*

Proof. If all $f \in H$ are measurable, then $k(\cdot\,,x) \in H$ is measurable for all $x \in X$. Conversely, if $k(\cdot\,,x)$ is measurable for all $x \in X$, then all functions $f \in H_{\text{pre}}$ are measurable, where H_{pre} is defined by (4.12). Let us now fix an $f \in H$. By Theorem 4.21, there then exists a sequence $(f_n) \subset H_{\text{pre}}$ with $\lim_{n\to\infty} \|f_n - f\|_H = 0$, and since all Dirac functionals are continuous, we obtain $\lim_{n\to\infty} f_n(x) = f(x)$, $x \in X$. This gives the measurability of f. □

The next lemma investigates the measurability of canonical feature maps.

Lemma 4.25 (Measurability of the canonical feature map). *Let X be a measurable space and k be a kernel on X such that $k(\cdot\,,x) : X \to \mathbb{R}$ is measurable for all $x \in X$. If the RKHS H of k is separable, then both the canonical feature map $\Phi : X \to H$ and $k : X \times X \to \mathbb{R}$ are measurable.*

Proof. Let $w \in H'$ be a bounded linear functional. By the Fréchet-Riesz representation theorem (see Theorem A.5.12) there then exists an $f \in H$ with

$$\langle w, \Phi(x)\rangle_{H',H} = \langle f, \Phi(x)\rangle_H = f(x)\,, \qquad x \in X,$$

and hence $\langle w, \Phi(\cdot\,)\rangle_{H',H} : X \to \mathbb{R}$ is measurable by Lemma 4.24. By Petti's measurability theorem (see Theorem A.5.19), we then obtain the measurability of Φ. The second assertion now follows from $k(x,x') = \langle \Phi(x'), \Phi(x)\rangle$ and the fact that the inner product is continuous. □

Our next goal is to investigate under which assumptions on the kernel k the functions of its RKHS are integrable. To this end, recall that $x \mapsto k(x,x)$ is a non-negative function, and hence its integral is always defined, though in general it may not be finite.

Theorem 4.26 (Integral operators of kernels I). *Let X be a measurable space, μ be a σ-finite measure on X, and H be a separable RKHS over X with measurable kernel $k : X \times X \to \mathbb{R}$. Assume that there exists a $p \in [1, \infty)$ such that*

$$\|k\|_{L_p(\mu)} := \left(\int_X k^{p/2}(x,x) d\mu(x) \right)^{1/p} < \infty \,. \tag{4.16}$$

Then H consists of p-integrable functions and the inclusion $\operatorname{id} : H \to L_p(\mu)$ is continuous with $\|\operatorname{id} : H \to L_p(\mu)\| \leq \|k\|_{L_p(\mu)}$. Moreover, the adjoint of this inclusion is the operator $S_k : L_{p'}(\mu) \to H$ defined by

$$S_k g(x) := \int_X k(x,x') g(x') d\mu(x') \,, \qquad g \in L_{p'}(\mu),\, x \in X, \tag{4.17}$$

where p' is defined by $\frac{1}{p} + \frac{1}{p'} = 1$. Finally, the following statements are true:

i) H is dense in $L_p(\mu)$ if and only if $S_k : L_{p'}(\mu) \to H$ is injective.
ii) $S_k : L_{p'}(\mu) \to H$ has a dense image if and only if $\operatorname{id} : H \to L_p(\mu)$ is injective.

Proof. Let us fix an $f \in H$. Using $\|k(\,\cdot\,, x)\|_H = \sqrt{k(x,x)}$, we then find

$$\int_X |f(x)|^p d\mu(x) = \int_X \bigl|\langle f, k(\,\cdot\,, x)\rangle\bigr|^p d\mu(x) \leq \|f\|_H^p \int_X k^{p/2}(x,x) d\mu(x) \,,$$

which shows the first two assertions. Furthermore, for $g \in L_{p'}(\mu)$, inequality (4.14) together with Hölder's inequality yields

$$\begin{aligned} \int_X \bigl|k(x,x')g(x')\bigr| \, d\mu(x') &\leq \sqrt{k(x,x)} \int_X \sqrt{k(x',x')} \, |g(x')| \, d\mu(x') \\ &\leq \sqrt{k(x,x)} \, \|k\|_{L_p(\mu)} \, \|g\|_{L_{p'}(\mu)} \,, \end{aligned} \tag{4.18}$$

and hence $x' \mapsto k(x,x')g(x')$ is integrable. In other words, the integral defining $S_k g(x)$ exists for all $x \in X$. Moreover, since $\sqrt{k(x',x')} = \|\Phi(x')\|_H$, the second inequality in (4.18) shows $(x' \mapsto \|g(x')\Phi(x')\|_H) \in L_1(\mu)$, i.e., this function is Bochner integrable and

$$\bar{g} := \int_X g(x')\Phi(x') \, d\mu(x') \in H \,.$$

Moreover, (A.32) applied to the bounded linear operator $\langle\,\cdot\,, \Phi(x)\rangle : H \to \mathbb{R}$ yields

$$S_k g(x) = \int_X \langle \Phi(x'), \Phi(x)\rangle_H \, g(x') \, d\mu(x') = \left\langle \int_X g(x')\Phi(x') \, d\mu(x'), \Phi(x) \right\rangle_H$$

for all $x \in X$, and hence we conclude that $S_k g = \bar{g} \in H$. For $f \in H$, another application of (A.32) yields

$$\langle g, \mathrm{id}\, f\rangle_{L_{p'}(\mu),L_p(\mu)} = \int_X g(x)\big\langle f, k(\,\cdot\,,x)\big\rangle_H \, d\mu(x) = \left\langle f, \int_X g(x)k(\,\cdot\,,x)\, d\mu(x)\right\rangle_H$$
$$= \langle f, S_k g\rangle_H$$
$$= \langle \iota S_k g, f\rangle_{H',H}\,,$$

where $\iota : H \to H'$ is the isometric isomorphism described in Theorem A.5.12. By identifying H' with H via ι, we then find $\mathrm{id}' = S_k$. Finally, the last two assertions follow from the fact that a bounded linear operator has a dense image if and only if its adjoint is injective, as mentioned in Section A.5.1 around (A.19). □

One may be tempted to think that the "inclusion" $\mathrm{id} : H \to L_p(\mu)$ is always injective. However, since this map assigns every f to its *equivalence class* $[f]_\sim$ in $L_p(\mu)$, see (A.33), the opposite is true. To see this, consider for example an infinite-dimensional RKHS (see Section 4.6 for examples of such spaces) and an empirical measure μ. Then $L_p(\mu)$ is finite-dimensional and hence the map $\mathrm{id} : H \to L_p(\mu)$ cannot be injective. For a simple condition ensuring that $\mathrm{id} : H \to L_p(\mu)$ *is* injective, we refer to Exercise 4.6.

Let us now have a closer look at the case $p = 2$ in the preceding theorem. The following theorem shows that in this case the Hilbert space structure of $L_2(\mu)$ provides some additional features of the operator S_k which will be of particular interest in Chapter 7.

Theorem 4.27 (Integral operators of kernels II). *Let X be a measurable space with σ-finite measure μ and H be a separable RKHS over X with measurable kernel $k : X \times X \to \mathbb{R}$ satisfying $\|k\|_{L_2(\mu)} < \infty$. Then $S_k : L_2(\mu) \to H$ defined by (4.17) is a Hilbert-Schmidt operator with*

$$\|S_k\|_{\mathsf{HS}} = \|k\|_{L_2(\mu)}\,. \tag{4.19}$$

Moreover, the integral operator $T_k = S_k^* S_k : L_2(\mu) \to L_2(\mu)$ is compact, positive, self-adjoint, and nuclear with $\|T_k\|_{\mathsf{nuc}} = \|S_k\|_{\mathsf{HS}} = \|k\|_{L_2(\mu)}$.

Proof. Let us first show that $S_k^* : H \to L_2(\mu)$ is a Hilbert-Schmidt operator. To this end, let $(e_i)_{i\geq 1}$ be an ONB of H. By Theorem 4.20, we then find

$$\sum_{i=1}^{\infty} \|S_k^* e_i\|_{L_2(\mu)}^2 = \int_X \sum_{i=1}^{\infty} |S_k^* e_i(x)|^2 \, d\mu(x) = \int_X \sum_{i=1}^{\infty} e_i^2(x)\, d\mu(x) = \|k\|_{L_2(\mu)}^2 < \infty,$$

i.e., S_k^* is indeed Hilbert-Schmidt with the desired norm. Consequently, S_k is Hilbert-Schmidt, too. Now the remaining assertions follow from the spectral theory recalled around Theorem A.5.13. □

Since $S_k^* = \mathrm{id} : H \to L_2(\mu)$, one may be tempted to think that the operators T_k and S_k are the same modulo their image space. However, recall that in general $L_2(\mu)$ does not consist of functions, and hence $S_k f(x)$ is defined, while $T_k f(x)$ is *not*.

Our next goal is to investigate continuity properties of kernels. To this end, we say that a kernel k on a topological space X is **separately continuous** if $k(\cdot\,, x) : X \to \mathbb{R}$ is continuous for all $x \in X$. Now, our first lemma characterizes RKHSs consisting of continuous functions.

Lemma 4.28 (RKHSs consisting of continuous functions). *Let X be topological space and k be a kernel on X with RKHS H. Then k is bounded and separately continuous if and only if every $f \in H$ is a bounded and continuous function. In this case, the inclusion* $\operatorname{id} : H \to C_b(X)$ *is continuous and we have* $\|\operatorname{id} : H \to C_b(X)\| = \|k\|_\infty$.

Proof. Let us first assume that k is bounded and separately continuous. Then H_{pre} only contains continuous functions since k is separately continuous. Let us fix an arbitrary $f \in H$. By Theorem 4.21, there then exists a sequence $(f_n) \subset H_{\text{pre}}$ with $\lim_{n\to\infty} \|f_n - f\|_H = 0$. Since k is bounded, this implies $\lim_{n\to\infty} \|f_n - f\|_\infty = 0$ by Lemma 4.23 and hence f, as a uniform limit of continuous functions, is continuous. Finally, both the continuity of $\operatorname{id} : H \to C_b(X)$ and $\|\operatorname{id} : H \to C_b(X)\| = \|k\|_\infty$ follow from Lemma 4.23, too.

Conversely, let us now assume that H only contains continuous functions. Then $k(\cdot\,, x) : X \to \mathbb{K}$ is continuous for all $x \in X$, i.e., k is separately continuous. Furthermore, the boundedness of k follows from Lemma 4.23. □

Lemma 4.28 in particular applies to continuous kernels. Let us now discuss these kernels in more detail. To this end, let k be a kernel on X with feature map $\Phi : X \to H$. Then the **kernel metric** d_k on X is defined by

$$d_k(x, x') := \|\Phi(x) - \Phi(x')\|_H\,, \qquad x, x' \in X. \tag{4.20}$$

Obviously, d_k is a pseudo-metric on X, and if Φ is injective it is even a metric. Moreover, since

$$d_k(x, x') = \sqrt{k(x, x) - 2k(x, x') + k(x', x')}\,, \tag{4.21}$$

the definition of d_k is actually *independent* of the choice of Φ. Furthermore, the kernel metric can be used to characterize the continuity of k.

Lemma 4.29 (Characterization of continuous kernels). *Let (X, τ) be a topological space and k be a kernel on X with feature space H and feature map $\Phi : X \to H$. Then the following statements are equivalent:*

i) k is continuous.
ii) k is separately continuous and $x \mapsto k(x, x)$ is continuous.
iii) Φ is continuous.
iv) $\operatorname{id} : (X, \tau) \to (X, d_k)$ *is continuous.*

Proof. i) ⇒ ii). Trivial.

ii) ⇒ iv). By (4.21) and the assumptions, we see that $d_k(\cdot\,, x) : (X, \tau) \to \mathbb{R}$ is continuous for every $x \in X$. Consequently, $\{x' \in X : d_k(x', x) < \varepsilon\}$ is open with respect to τ and therefore $\operatorname{id} : (X, \tau) \to (X, d_k)$ is continuous.

iv) ⇒ iii). This implication follows from the fact that $\Phi : (X, d_k) \to H$ is continuous.

iii) ⇒ i). Let us fix $x_1, x_1' \in X$ and $x_2, x_2' \in X$. Then we have

$$\begin{aligned} |k(x_1, x_1') - k(x_2, x_2')| &\leq |\langle \Phi(x_1'), \Phi(x_1) - \Phi(x_2)\rangle| + |\langle \Phi(x_1') - \Phi(x_2'), \Phi(x_2)\rangle| \\ &\leq \|\Phi(x_1')\| \cdot \|\Phi(x_1) - \Phi(x_2)\| + \|\Phi(x_2)\| \cdot \|\Phi(x_1') - \Phi(x_2')\|, \end{aligned}$$

and from this we can easily deduce the assertion. □

As discovered by Lehto (1952), separately continuous, bounded kernels are not necessarily continuous, even if one only considers $X = [-1, 1]$. However, since his example is out of the scope of this book, we do not present it here.

We have seen in Lemma 4.23 that an RKHS over X is continuously included in $\ell_\infty(X)$ if its kernel is bounded. The next proposition gives a condition that ensures that this inclusion is even compact. This compactness will play an important role when we investigate the statistical properties of support vector machines in Chapter 6.

Proposition 4.30 (RKHSs compactly included in $\ell_\infty(\mathbf{X})$). *Let X be a set and k be a kernel on X with RKHS H and canonical feature map $\Phi : X \to H$. If $\Phi(X)$ is compact in H, then the inclusion* $\mathrm{id} : H \to \ell_\infty(X)$ *is compact.*

Proof. Since $\Phi(X)$ is compact, k is bounded and the space (X, d_k) is compact, where d_k is the semi-metric defined in (4.20). We write $C(X, d_k)$ for the space of functions from X to $\mathbb{R}$ that are continuous with respect to d_k. Obviously, $C(X, d_k)$ is a subspace of $\ell_\infty(X)$. Moreover, for $x, x' \in X$ and $f \in H$, we obtain

$$|f(x) - f(x')| = |\langle f, \Phi(x) - \Phi(x')\rangle| \leq \|f\|_H \cdot d_k(x, x')\,,$$

i.e., f is Lipschitz continuous on (X, d_k) with Lipschitz constant not larger than $\|f\|_H$. In particular, the unit ball B_H of H is equicontinuous, and since B_H is also $\|\cdot\|_\infty$-bounded by the boundedness of k, the theorem of Arzelà-Ascoli shows that $\overline{B_H}$ is compact in $C(X, d_k)$ and thus in $\ell_\infty(X)$. □

Obviously, the proposition above remains true if one only assumes the compactness of $\Phi(X)$ for an *arbitrary* feature map Φ. Furthermore, continuous images of compact sets are compact, and hence Proposition 4.30 has the following direct consequence.

Corollary 4.31. *Let X be a compact topological space and k be a continuous kernel on X with RKHS H. Then the inclusion* $\mathrm{id} : H \to C(X)$ *is compact.*

We emphasize that in general one cannot expect compactness of the inclusion $\mathrm{id} : H \to C_b(X)$ if k is bounded and continuous but X is *not* compact. The following example illustrates this.

Example 4.32. Let k_γ be the **Gaussian RBF kernel** on $\mathbb{R}$ with width $\gamma > 0$ and RKHS $H_\gamma(\mathbb{R})$. Obviously, k_γ is bounded and continuous, and hence the inclusion $\mathrm{id} : H_\gamma(\mathbb{R}) \to C_b(\mathbb{R})$ is well-defined and continuous. Moreover, since $\|k_\gamma\|_\infty = 1$, we also have $k_\gamma(\cdot, x) \in B_{H_\gamma(\mathbb{R})}$ for all $x \in \mathbb{R}$. However, for all $n, m \in \mathbb{N}$ with $n \neq m$, we obtain

$$\|k_\gamma(\cdot, n) - k_\gamma(\cdot, m)\|_\infty \;\geq\; |k_\gamma(n,n) - k_\gamma(n,m)| \;\geq\; |1 - \exp(-\gamma^{-2})|\,,$$

and thus $B_{H_\gamma(\mathbb{R})}$ cannot be relatively compact in $C_b(\mathbb{R})$.

Let us end the discussion on continuous kernels with the following lemma that gives a sufficient condition for the separability of RKHSs.

Lemma 4.33 (Separable RKHSs). *Let X be a separable topological space and k be a continuous kernel on X. Then the RKHS of k is separable.*

Proof. By Lemma 4.29, the canonical feature map $\Phi : X \to H$ into the RKHS H of k is continuous and hence $\Phi(X)$ is separable. Consequently, H_{pre}, considered in Theorem 4.21, is also separable, and hence so is H by Theorem 4.21. □

Our last goal in this section is to investigate how the differentiability of a kernel is inherited by the functions of its RKHS. In order to formulate the next lemma, which to some extent is of its own interest, we need to recall Banach space valued differentiation from Section A.5.3. Moreover, note that we can interpret a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as a function $\tilde{k} : \mathbb{R}^{2d} \to \mathbb{R}$. Consequently, considering the mixed partial derivative of the kernel $k(x, x')$ with respect to the i-th coordinates in x *and* x' is the same as considering the mixed partial derivative $\partial_i \partial_{i+d} \tilde{k}$ at (x, x'). In the following, we make this identification implicitly by writing $\partial_i \partial_{i+d} k := \partial_i \partial_{i+d} \tilde{k}$. Moreover, we extend this notation to kernels defined on open subsets of $\mathbb{R}^d$ in the obvious way.

Lemma 4.34 (Differentiability of feature maps). *Let $X \subset \mathbb{R}^d$ be an open subset, k be a kernel on X, H be a feature space of k, and $\Phi : X \to H$ be a feature map of k. Let $i \in \{1, \ldots, d\}$ be an index such that the mixed partial derivative $\partial_i \partial_{i+d} k$ of k with respect to the coordinates i and $i + d$ exists and is continuous. Then the partial derivative $\partial_i \Phi$ of $\Phi : X \to H$ with respect to the i-th coordinate exists, is continuous, and for all $x, x' \in X$ we have*

$$\big\langle \partial_i \Phi(x), \partial_i \Phi(x') \big\rangle_H = \partial_i \partial_{i+d} k(x, x') = \partial_{i+d} \partial_i k(x, x')\,. \tag{4.22}$$

Proof. Without loss of generality, we may assume $X = \mathbb{R}^d$. For $h \in \mathbb{R}$ and $e_i \in \mathbb{R}^d$ being the i-th vector of the canonical basis of $\mathbb{R}^d$, we then define $\Delta_h \Phi(x) := \Phi(x + h e_i) - \Phi(x)$, $x \in X$. In order to show that $\partial_i \Phi(x)$ exists for an arbitrary $x \in X$, it obviously suffices to show that $h_n^{-1} \Delta_{h_n} \Phi(x)$ converges for all sequences $(h_n) \subset \mathbb{R}^d \setminus \{0\}$ with $h_n \to 0$. Since a feature space is complete, it thus suffices to show that $(h_n^{-1} \Delta_{h_n} \Phi(x))$ is a Cauchy sequence. To this end, we first observe that

$$\begin{aligned}\left\| h_n^{-1}\Delta_{h_n}\Phi(x) - h_m^{-1}\Delta_{h_m}\Phi(x) \right\|_H^2 &= \left\langle h_n^{-1}\Delta_{h_n}\Phi(x), h_n^{-1}\Delta_{h_n}\Phi(x)\right\rangle_H \\ &\quad + \left\langle h_m^{-1}\Delta_{h_m}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(x)\right\rangle_H \\ &\quad - 2\left\langle h_n^{-1}\Delta_{h_n}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(x)\right\rangle_H\end{aligned}$$

for all $x \in X$ and $n, m \in \mathbb{N}$. For the function $K(x') := k(x+h_n e_i, x') - k(x, x')$, $x' \in X$, we further have $\langle \Delta_{h_n}\Phi(x), \Delta_{h_m}\Phi(x')\rangle_H = K(x' + h_m e_i) - K(x')$, and hence the mean value theorem yields a $\xi_{m,n} \in [-|h_m|, |h_m|]$ such that

$$\begin{aligned}&\left\langle \Delta_{h_n}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(x')\right\rangle_H \\ &= \partial_i K(x' + \xi_{m,n} e_i) \\ &= \partial_{i+d} k(x+h_n e_i, x'+\xi_{m,n} e_i) - \partial_{i+d} k(x, x'+\xi_{m,n} e_i).\end{aligned}$$

Another application of the mean value theorem yields an $\eta_{m,n} \in [-|h_n|, |h_n|]$ such that

$$\left\langle h_n^{-1}\Delta_{h_n}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(x')\right\rangle_H = \partial_i \partial_{i+d} k(x+\eta_{m,n} e_i, x'+\xi_{m,n} e_i)\,.$$

By the continuity of $\partial_i \partial_{i+d} k$, we conclude that for a given $\varepsilon > 0$ there exists an $n_0 \in \mathbb{N}$ such that for all $n, m \geq n_0$ we have

$$\left| \left\langle h_n^{-1}\Delta_{h_n}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(x')\right\rangle_H - \partial_i \partial_{i+d} k(x, x') \right| \leq \varepsilon\,. \tag{4.23}$$

Applying (4.23) for $x = x'$ to the three terms on the right-hand side of our first equation, we see that $(h_n^{-1}\Delta_{h_n}\Phi(x))$ is a Cauchy sequence. By definition, its limit is $\partial_i \Phi$, and the first equality in (4.22) is then a direct consequence of (4.23). The second equality follows from the symmetry of k. □

A direct consequence of the result above is that $\partial_i \partial_{i+d} k$ is a kernel on $X \times X$ with feature map $\partial_i \Phi$. Now assume that even $\partial_j \partial_{j+d} \partial_i \partial_{i+d} k$ exists and is continuous. Then an iterated application of the preceding lemma shows that $\partial_j \partial_i \Phi$ exists, is continuous, and is a feature map of the kernel $\partial_j \partial_{j+d} \partial_i \partial_{i+d} k$. Obviously, we can further iterate this procedure if even higher-order derivatives exist. In order to describe such situations, we write $\partial^{\alpha,\alpha} := \partial_1^{\alpha_1} \ldots \partial_d^{\alpha_d} \partial_{1+d}^{\alpha_1} \ldots \partial_{2d}^{\alpha_d}$, where $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ is a multi-index and arbitrary reorderings of the partial derivatives are allowed.

Definition 4.35. *Let k be a kernel on an open $X \subset \mathbb{R}^d$. For $m \geq 0$, we say that k is m-**times continuously differentiable** if $\partial^{\alpha,\alpha} k : X \times X \to \mathbb{R}$ exists and is continuous for all multi-indexes $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq m$.*

Iteratively applying Lemma 4.34 to an m-times continuously differentiable kernel yields the following result.

Corollary 4.36 (RKHSs of differentiable kernels). *Let $X \subset \mathbb{R}^d$ be an open subset, $m \geq 0$, and k be an m-times continuously differentiable kernel on X with RKHS H. Then every $f \in H$ is m-times continuously differentiable, and for $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq m$ and $x \in X$ we have*

$$\left|\partial^\alpha f(x)\right| \;\leq\; \|f\|_H \cdot \left(\partial^{\alpha,\alpha} k(x,x)\right)^{1/2}. \tag{4.24}$$

Proof. Let us write $\Phi : X \to H$ for the canonical feature map of k. By iteratively applying Lemma 4.34, we see that $\partial^\alpha \Phi : X \to H$ is a feature map of the kernel $\partial^{\alpha,\alpha} k : X \times X \to \mathbb{R}$. By the continuity of $\langle f, \cdot \rangle_H : H \to \mathbb{R}$, we then conclude that $\langle f, \partial^\alpha \Phi(x) \rangle_H = \partial^\alpha \langle f, \Phi(x) \rangle_H = \partial^\alpha f(x)$, i.e, the latter partial derivative exists and is continuous. Finally, (4.24) follows from

$$\left|\partial^\alpha f(x)\right| = \left|\langle f, \partial^\alpha \Phi(x) \rangle_H\right| \leq \|f\|_H \cdot \sqrt{\langle \partial^\alpha \Phi(x), \partial^\alpha \Phi(x) \rangle_H} \tag{4.25}$$

and an iterated application of (4.22) to the right-hand side of (4.25). □

4.4 Gaussian Kernels and Their RKHSs

The goal of this section is to use the developed theory on RKHSs to investigate the Gaussian RBF kernels and their RKHSs in more detail. In particular, we will present two representations of these RKHSs and discuss some consequences. We begin, however, with a simple result that describes the effect of the kernel parameter γ on the input domain.

Proposition 4.37. *Let $X \subset \mathbb{R}^d$ be a non-empty subset and $s, \gamma > 0$ be real numbers. Given a function $f : sX \to \mathbb{R}$, we define $\tau_s f(x) := f(sx)$ for $x \in X$. Then, for all $f \in H_{s\gamma}(sX)$, we have $\tau_s f \in H_\gamma(X)$, and the corresponding linear operator $\tau_s : H_{s\gamma}(sX) \to H_\gamma(X)$ is an isometric isomorphism.*

Proof. We define $\Phi : X \to H_{s\gamma}(sX)$ by $\Phi(x) := \Phi_{s\gamma}(sx)$, where $x \in X$ and $\Phi_{s\gamma} : sX \to H_{s\gamma}(sX)$ is the canonical feature map of $k_{s\gamma}$, i.e., $\Phi_{s\gamma}(y) = k_{s\gamma}(\cdot\,, y)$ for all $y \in sX$. For $x, x' \in X$, we then have

$$\begin{aligned}\langle \Phi(x'), \Phi(x) \rangle_{H_{s\gamma}(sX)} = \langle \Phi_{s\gamma}(sx'), \Phi_{s\gamma}(sx) \rangle_{H_{s\gamma}(sX)} &= k_{s\gamma}(sx', sx) \\ &= \exp\bigl(-(s\gamma)^{-2} \|sx - sx'\|_2^2\bigr) \\ &= k_\gamma(x, x')\,,\end{aligned}$$

and hence $\Phi : X \to H_{s\gamma}(sX)$ is a feature map of $k_\gamma : X \times X \to \mathbb{R}$. Let us now fix an $f \in H_{s\gamma}(sX)$. By Theorem 4.21, we then know that $\langle f, \Phi(\cdot) \rangle_{H_{s\gamma}(sX)} \in H_\gamma(X)$ and

$$\left\| \langle f, \Phi(\cdot) \rangle_{H_{s\gamma}(sX)} \right\|_{H_\gamma(sX)} \leq \|f\|_{H_{s\gamma}(X)}\,.$$

Moreover, for $x \in X$, the reproducing property in $H_{s\gamma}(sX)$ yields

$$\langle f, \Phi(x) \rangle_{H_{s\gamma}(sX)} = \langle f, \Phi_{s\gamma}(sx) \rangle_{H_{s\gamma}(sX)} = f(sx) = \tau_s f(x)\,,$$

and hence we have found $\tau_s f \in H_\gamma(X)$ with $\|\tau_s f\|_{H_\gamma(X)} \leq \|f\|_{H_{s\gamma}(sX)}$. Finally, we obtain the converse inequality by applying the results above to the dilation operator $\tau_{1/s}$. □

Portions of Section 4.4 are based on material originally published in "I. Steinwart, D. Hush, and C. Scovel (2006), 'An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels.' *IEEE Trans. Inf. Theory*, **52**, 4635–4643"

Roughly speaking, the preceding proposition states that scaling the kernel parameter has the same effect on the RKHSs as scaling the input space. Considering the definition of the Gaussian RBF kernels, this is not really surprising.

Our next goal is to determine an explicit formula for the RKHSs of Gaussian RBF kernels. To this end, let us fix $\gamma > 0$ and $d \in \mathbb{N}$. For a given holomorphic function $f : \mathbb{C}^d \to \mathbb{C}$, we define

$$\|f\|_{\gamma,\mathbb{C}^d} := \Big(\frac{2^d}{\pi^d \gamma^{2d}} \int_{\mathbb{C}^d} |f(z)|^2 e^{\gamma^{-2} \sum_{j=1}^d (z_j - \bar{z}_j)^2} dz\Big)^{1/2}, \tag{4.26}$$

where z_j is the j-th component of $z \in \mathbb{C}^d$, $\bar{z}_j$ its conjugate, and dz stands for the complex Lebesgue measure on $\mathbb{C}^d$. Furthermore, we write

$$H_{\gamma,\mathbb{C}^d} := \big\{ f : \mathbb{C}^d \to \mathbb{C} \,|\, f \text{ holomorphic and } \|f\|_{\gamma,\mathbb{C}^d} < \infty \big\}. \tag{4.27}$$

Obviously, $H_{\gamma,\mathbb{C}^d}$ is a $\mathbb{C}$-vector space with pre-Hilbert norm $\|\cdot\|_{\gamma,\mathbb{C}^d}$. Now, our first result shows that $H_{\gamma,\mathbb{C}^d}$ is the RKHS of the *complex* Gaussian RBF kernel $k_{\gamma,\mathbb{C}^d}$ defined in Proposition 4.10.

Theorem 4.38 (RKHS of the complex Gaussian RBF). *Let $\gamma > 0$ and $d \in \mathbb{N}$. Then $(H_{\gamma,\mathbb{C}^d}, \|\cdot\|_{H_{\gamma,\mathbb{C}^d}})$ is an RKHS and $k_{\gamma,\mathbb{C}^d}$ is its reproducing kernel. Furthermore, for $n \in \mathbb{N}_0$, let $e_n : \mathbb{C} \to \mathbb{C}$ be defined by*

$$e_n(z) := \sqrt{\frac{2^n}{\gamma^{2n} n!}}\, z^n e^{-\gamma^{-2} z^2}, \qquad z \in \mathbb{C}. \tag{4.28}$$

Then the system $(e_{n_1} \otimes \cdots \otimes e_{n_d})_{n_1,\ldots,n_d \geq 0}$ of functions $e_{n_1} \otimes \cdots \otimes e_{n_d} : \mathbb{C}^d \to \mathbb{C}$ defined by

$$e_{n_1} \otimes \cdots \otimes e_{n_d}(z_1, \ldots, z_d) := \prod_{j=1}^d e_{n_j}(z_j), \qquad (z_1, \ldots, z_d) \in \mathbb{C}^d,$$

is an orthonormal basis of $H_{\gamma,\mathbb{C}^d}$.

For the proof of Theorem 4.38, we need the following technical lemma.

Lemma 4.39. *For all $d \in \mathbb{N}$, all holomorphic functions $f : \mathbb{C}^d \to \mathbb{C}$, all $r_1, \ldots, r_d \in [0, 1)$, and all $z \in \mathbb{C}^d$, we have*

$$|f(z)|^2 \leq \frac{1}{(2\pi)^d} \int_0^{2\pi} \cdots \int_0^{2\pi} \big| f(z_1 + r_1 e^{i\theta_1}, \ldots, z_d + r_d e^{i\theta_d}) \big|^2 d\theta_1 \cdots d\theta_d, \tag{4.29}$$

where $i := \sqrt{-1}$ denotes the imaginary unit.

Proof. We proceed by induction over d. For $d = 1$, Hardy's convexity theorem (see Theorem A.7.3) states that the function

$$r \mapsto \frac{1}{2\pi} \int_0^{2\pi} \left| f(z + re^{i\theta}) \right|^2 d\theta$$

is non-decreasing on $[0, 1)$, and hence we obtain the assertion in this case.

Now let us suppose that we have already shown the assertion for $d \in \mathbb{N}$. Let $f : \mathbb{C}^{d+1} \to \mathbb{C}$ be a holomorphic function, and choose $r_1, \ldots, r_{d+1} \in [0, 1)$. Since for fixed $(z_1, \ldots, z_d) \in \mathbb{C}^d$ the function $z_{d+1} \mapsto f(z_1, \ldots, z_d, z_{d+1})$ is holomorphic, we already know that

$$|f(z_1, \ldots, z_{d+1})|^2 \leq \frac{1}{2\pi} \int_0^{2\pi} \left| f(z_1, \ldots, z_d, z_{d+1} + r_{d+1} e^{i\theta_{d+1}}) \right|^2 d\theta_{d+1} .$$

Now applying the induction hypothesis to the holomorphic functions

$$(z_1, \ldots, z_d) \mapsto f(z_1, \ldots, z_d, z_{d+1} + r_{d+1} e^{i\theta_{d+1}})$$

on $\mathbb{C}^d$ gives the assertion for $d + 1$. □

Proof (of Theorem 4.38). We first prove that $H_{\gamma,\mathbb{C}}$ is an RKHS. To this end, we begin by showing that for all compact subsets $K \subset \mathbb{C}^d$ there exists a constant $c_K > 0$ with

$$|f(z)| \leq c_K \, \|f\|_{\gamma,\mathbb{C}^d} , \qquad z \in K, \, f \in H_{\gamma,\mathbb{C}^d}. \tag{4.30}$$

In order to establish (4.30), we define

$$c := \max\{e^{-\gamma^{-2} \sum_{j=1}^d (z_j - \bar{z}_j)^2} : (z_1, \ldots, z_d) \in K + (B_{\mathbb{C}})^d\} ,$$

where $B_{\mathbb{C}}$ denotes the closed unit ball of $\mathbb{C}$. Now, by Lemma 4.39, we have

$$2^d r_1 \cdots r_d |f(z)|^2 \leq \frac{r_1 \cdots r_d}{\pi^d} \int_0^{2\pi} \cdots \int_0^{2\pi} \left| f(z_1 + r_1 e^{i\theta_1}, \ldots, z_d + r_d e^{i\theta_d}) \right|^2 d\theta_1 \cdots d\theta_d,$$

and integrating this inequality with respect to $r = (r_1, \ldots, r_d)$ over $[0, 1)^d$ then yields

$$\begin{aligned} |f(z)|^2 \leq \frac{1}{\pi^d} \int\limits_{z+(B_{\mathbb{C}})^d} |f(z')|^2 dz' &\leq \frac{c}{\pi^d} \int\limits_{z+(B_{\mathbb{C}})^d} |f(z')|^2 e^{\gamma^{-2} \sum_{j=1}^d (z'_j - \bar{z}'_j)^2} dz' \\ &\leq \frac{c\gamma^{2d}}{2^d} \, \|f\|^2_{\gamma,\mathbb{C}^d} , \qquad z \in K, \end{aligned}$$

by the continuity of f. This means that we have established (4.30). Now, (4.30) obviously shows that the Dirac functionals are bounded on $H_{\gamma,\mathbb{C}^d}$. Furthermore, (4.30) also shows that convergence in $\|\cdot\|_{\gamma,\mathbb{C}}$ implies *compact convergence*, i.e., uniform convergence on every compact subset. Using the fact

that a compactly convergent sequence of holomorphic functions has a holomorphic limit (see, e.g., Theorem A.7.2), we then immediately find that $H_{\gamma,\mathbb{C}^d}$ is complete. Therefore $H_{\gamma,\mathbb{C}^d}$ is an RKHS.

To show that the system $(e_{n_1} \otimes \cdots \otimes e_{n_d})_{n_1,\ldots,n_d \geq 0}$ is an ONB of $H_{\gamma,\mathbb{C}^d}$, we first consider the case $d = 1$. To this end, we observe that for $n \in \mathbb{N}_0$ we have

$$\begin{aligned}
\int_{\mathbb{C}} z^n (\bar{z})^n e^{-2\gamma^{-2} z\bar{z}} dz &= \int_0^\infty r \int_0^{2\pi} r^{2n} e^{-2\gamma^{-2} r^2} d\theta dr \\
&= 2\pi \int_0^\infty r^{2n+1} e^{-2\gamma^{-2} r^2} dr \\
&= \frac{\pi \gamma^{2(n+1)}}{2^{n+1}} \int_0^\infty t^n e^{-t} dt \\
&= \frac{\pi \gamma^{2(n+1)} n!}{2^{n+1}}\,, \qquad (4.31)
\end{aligned}$$

where in the last step we used the gamma function, see Section A.1. Furthermore, for $n, m \in \mathbb{N}_0$ with $n \neq m$, a simple calculation gives

$$\int_{\mathbb{C}} z^n (\bar{z})^m e^{-2\gamma^{-2} z\bar{z}} dz = \int_0^\infty r \int_0^{2\pi} r^{n+m} e^{i(n-m)\theta} e^{-2\gamma^{-2} r^2} d\theta dr = 0\,. \quad (4.32)$$

In addition, for $z, \bar{z} \in \mathbb{C}$ and $n, m \geq 0$, we have

$$\begin{aligned}
e_n(z)\overline{e_m(z)} e^{\gamma^{-2}(z-\bar{z})^2} &= \sqrt{\frac{2^{n+m}}{n!\, m!\, \gamma^{2(n+m)}}}\, z^n (\bar{z})^m e^{-\gamma^{-2} z^2 - \gamma^{-2} \bar{z}^2} e^{\gamma^{-2}(z-\bar{z})^2} \\
&= \sqrt{\frac{2^{n+m}}{n!\, m!\, \gamma^{2(n+m)}}}\, z^n (\bar{z})^m e^{-2\gamma^{-2} z\bar{z}}\,,
\end{aligned}$$

and consequently we obtain

$$\langle e_n, e_m \rangle = \frac{2}{\pi\gamma^2} \int_{\mathbb{C}} e_n(z)\overline{e_m(z)} e^{\gamma^{-2}(z-\bar{z})^2} dz = \begin{cases} 1 & \text{if } n = m \\ 0 & \text{otherwise} \end{cases}$$

by (4.31) and (4.32), i.e., $(e_n)_{n \geq 0}$ is an ONS. Now, let us show that this system is actually an ONB. To this end, let $f \in H_{\gamma,\mathbb{C}}$. Then $z \mapsto e^{\gamma^{-2} z^2} f(z)$ is an entire function, and therefore there exists a sequence $(a_n) \subset \mathbb{C}$ such that

$$f(z) = \sum_{n=0}^\infty a_n z^n e^{-\gamma^{-2} z^2} = \sum_{n=0}^\infty a_n \sqrt{\frac{\gamma^{2n} n!}{2^n}}\, e_n(z) \qquad (4.33)$$

for all $z \in \mathbb{C}$. Obviously, it suffices to show that the convergence above also holds with respect to $\|\cdot\|_{\gamma,\mathbb{C}}$. To prove this, we first recall from complex analysis that the series in (4.33) converges absolutely and compactly. Therefore, for $n \geq 0$ equations (4.31), (4.32), and (4.33) yield

$$\begin{aligned}\langle f, e_n\rangle &= \frac{2}{\pi\gamma^2}\int_{\mathbb{C}} f(z)\overline{e_n(z)}e^{\gamma^{-2}(z-\bar{z})^2}dz\\ &= \frac{2}{\pi\gamma^2}\sum_{m=0}^{\infty} a_m \int_{\mathbb{C}} z^m e^{-\gamma^{-2}z^2}\overline{e_n(z)}e^{\gamma^{-2}(z-\bar{z})^2}dz\\ &= \frac{2}{\pi\gamma^2}\sqrt{\frac{2^n}{\gamma^{2n}n!}}\sum_{m=0}^{\infty} a_m \int_{\mathbb{C}} z^m(\bar{z})^n e^{-2\gamma^{-2}z\bar{z}}dz\\ &= a_n\sqrt{\frac{\gamma^{2n}n!}{2^n}}\,. \end{aligned} \tag{4.34}$$

Furthermore, since (e_n) is an ONS, there exists a function $g \in H_{\gamma,\mathbb{C}}$ with $g = \sum_{n=0}^{\infty}\langle f, e_n\rangle e_n$, where the convergence takes place in $H_{\sigma,\mathbb{C}}$. Now, using (4.33), (4.34), and the fact that norm convergence in RKHSs implies pointwise convergence, we find $g = f$, i.e., the series in (4.33) converges with respect to $\|\cdot\|_{\sigma,\mathbb{C}}$.

Now, let us briefly treat the general, d-dimensional case. In this case, a simple calculation shows

$$\langle e_{n_1}\otimes\cdots\otimes e_{n_d}, e_{m_1}\otimes\cdots\otimes e_{m_d}\rangle_{H_{\gamma,\mathbb{C}^d}} = \prod_{j=1}^{d}\langle e_{n_j}, e_{m_j}\rangle_{H_{\gamma,\mathbb{C}}}\,,$$

and hence we find the orthonormality of $(e_{n_1}\otimes\cdots\otimes e_{n_d})_{n_1,\ldots,n_d\geq 0}$. In order to check that this orthonormal system is an ONB, let us fix an $f \in H_{\sigma,\mathbb{C}^d}$. Then $z \mapsto f(z)\exp(\sigma^2\sum_{i=1}^{d} z_i^2)$ is an entire function, and hence there exist $a_{n_1,\ldots,n_d} \in \mathbb{C}$, $(n_1,\ldots,n_d) \in \mathbb{N}_0^d$, such that

$$\begin{aligned} f(z) &= \sum_{(n_1,\ldots,n_d)\in\mathbb{N}_0^d} a_{n_1,\ldots,n_d}\prod_{i=1}^{d} z_i^{n_i}e^{-\sigma^2 z_i^2}\\ &= \sum_{(n_1,\ldots,n_d)\in\mathbb{N}_0^d} a_{n_1,\ldots,n_d}\prod_{i=1}^{d}\sqrt{\frac{n_i!}{(2\sigma^2)^{n_i}}}e_{n_i}(z) \end{aligned}$$

for all $z = (z_1,\ldots,z_d) \in \mathbb{C}^d$. From this we easily derive

$$\langle f, e_{n_1}\otimes\cdots\otimes e_{n_d}\rangle = a_{n_1,\ldots,n_d}\prod_{i=1}^{d}\sqrt{\frac{n_i!}{(2\sigma^2)^{n_i}}}\,.$$

Now we see that $(e_{n_1}\otimes\cdots\otimes e_{n_d})_{n_1,\ldots,n_d\geq 0}$ is an ONB as in the one-dimensional case.

Finally, let us show that $k_{\gamma,\mathbb{C}^d}$ is the reproducing kernel of $H_{\gamma,\mathbb{C}^d}$. To this end, we write k for the reproducing kernel of $H_{\gamma,\mathbb{C}^d}$. Then (4.9) and the Taylor series expansion of the exponential function yield

$$\begin{aligned}
k(z,z') &= \sum_{n_1,\ldots,n_d=0}^{\infty} e_{n_1}\otimes\cdots\otimes e_{n_d}(z)\overline{e_{n_1}\otimes\cdots\otimes e_{n_d}(z')} \\
&= \sum_{n_1,\ldots,n_d=0}^{\infty}\prod_{j=1}^{d}\frac{2^{n_j}}{\gamma^{2n_j}n_j!}(z_j\bar{z}'_j)^{n_j}e^{-\gamma^{-2}z_j^2-\gamma^{-2}(\bar{z}'_j)^2} \\
&= \prod_{j=1}^{d}\sum_{n_j=0}^{\infty}\frac{2^{n_j}}{\gamma^{2n_j}n_j!}(z_j\bar{z}'_j)^{n_j}e^{-\gamma^{-2}z_j^2-\gamma^{-2}(\bar{z}'_j)^2} \\
&= \prod_{j=1}^{d} e^{-\gamma^{-2}z_j^2-\gamma^{-2}(\bar{z}'_j)^2+2\gamma^{-2}z_j\bar{z}'_j} \\
&= e^{-\gamma^{-2}\sum_{j=1}^{d}(z_j-\bar{z}'_j)^2}\,.
\end{aligned}$$

□

With the help of Theorem 4.38, we can obtain some interesting information on the RKHSs of the *real-valued* Gaussian RBF kernels k_γ. Let us begin with the following corollary that describes their RKHSs.

Corollary 4.40 (RKHS of Gaussian RBF). *For $X \subset \mathbb{R}^d$ and $\gamma > 0$, the RKHS $H_\gamma(X)$ of the real-valued Gaussian RBF kernel k_γ on X is*

$$H_\gamma(X) = \bigl\{f : X \to \mathbb{R} \,|\, \exists\, g \in H_{\gamma,\mathbb{C}^d} \text{ with } \operatorname{Re} g_{|X} = f\bigr\}\,,$$

and for $f \in H_\gamma(X)$ the norm $\|\cdot\|_{H_\gamma(X)}$ in $H_\gamma(X)$ can be computed by

$$\|f\|_{H_\gamma(X)} = \inf\bigl\{\|g\|_{\gamma,\mathbb{C}^d} : g \in H_{\gamma,\mathbb{C}^d} \text{ with } \operatorname{Re} g_{|X} = f\bigr\}\,.$$

Proof. The assertion directly follows from Theorem 4.38, Proposition 4.10, and the discussion following Corollary 4.22. □

The preceding corollary shows that every $f \in H_\gamma(X)$ of the Gaussian RBF kernel k_γ originates from the complex RKHS $H_{\gamma,\mathbb{C}^d}$, which consists of entire functions. Consequently, every $f \in H_\gamma(X)$ can be represented by a power series that converges on $\mathbb{R}^d$. This observation suggests that there may be an intimate relationship between $H_\gamma(X)$ and $H_\gamma(\mathbb{R}^d)$ if X contains an open set. In order to investigate this conjecture, we need some additional notation. For a multi-index $\nu := (n_1,\ldots,n_d) \in \mathbb{N}_0^d$, we write $|v| := n_1+\cdots+n_d$. Furthermore, for $X \subset \mathbb{R}$ and $n \in \mathbb{N}_0$, we define $e_n^X : X \to \mathbb{R}$ by

$$e_n^X(x) := \sqrt{\frac{2^n}{\gamma^{2n}n!}}\,x^n e^{-\gamma^{-2}x^2}\,, \qquad x \in X\,, \tag{4.35}$$

i.e., we have $e_n^X = (e_n)_{|X} = (\operatorname{Re} e_n)_{|X}$, where $e_n : \mathbb{C} \to \mathbb{C}$ is an element of the ONB of $H_{\gamma,\mathbb{C}}$ defined by (4.28). Furthermore, for a multi-index $\nu := (n_1,\ldots,n_d) \in \mathbb{N}_0^d$, we write

$$e_\nu^X := e_{n_1}^X \otimes\cdots\otimes e_{n_d}^X$$

and $e_\nu := e_{n_1} \otimes \cdots \otimes e_{n_d}$. Given an $x := (x_1, \ldots, x_d) \in \mathbb{R}^d$, we also adopt the notation $x^\nu := x_1^{n_1} \cdot \ldots \cdot x_d^{n_d}$. Finally, recall that $\ell_2(\mathbb{N}_0^d)$ denotes the set of all *real*-valued square-summable families, i.e.,

$$\ell_2(\mathbb{N}_0^d) := \left\{ (a_\nu)_{\nu \in \mathbb{N}_0^d} : a_\nu \in \mathbb{R} \text{ for all } \nu \in \mathbb{N}_0^d \text{ and } \|(a_\nu)\|_2^2 := \sum_{\nu \in \mathbb{N}_0^d} a_\nu^2 < \infty \right\}.$$

With the help of these notations, we can now show an intermediate result.

Proposition 4.41. *Let $\gamma > 0$, $X \subset \mathbb{R}^d$ be a subset with non-empty interior, i.e., $\mathring{X} \neq \emptyset$, and $f \in H_\gamma(X)$. Then there exists a unique element $(b_\nu) \in \ell_2(\mathbb{N}_0^d)$ such that*

$$f(x) \;=\; \sum_{\nu \in \mathbb{N}_0^d} b_\nu e_\nu^X(x)\,, \qquad\qquad x \in X, \tag{4.36}$$

where the convergence is absolute. Furthermore, for all functions $g : \mathbb{C}^d \to \mathbb{C}$, the following two statements are equivalent:

i) We have $g \in H_{\gamma,\mathbb{C}^d}$ and $\operatorname{Re} g_{|X} = f$.
ii) There exists an element $(c_\nu) \in \ell_2(\mathbb{N}_0^d)$ with

$$g \;=\; \sum_{\nu \in \mathbb{N}_0^d} (b_\nu + ic_\nu) e_\nu\,. \tag{4.37}$$

Finally, we have the identity $\|f\|_{H_\gamma(X)}^2 = \sum_{\nu \in \mathbb{N}_0^d} b_\nu^2$.

Proof. *i)* $\Rightarrow$ *ii)*. Let us fix a $g \in H_{\gamma,\mathbb{C}^d}$ with $\operatorname{Re} g_{|X} = f$. Since (e_ν) is an ONB of $H_{\gamma,\mathbb{C}^d}$, we then have

$$g = \sum_{\nu \in \mathbb{N}_0^d} \langle g, e_\nu \rangle\, e_\nu\,,$$

where the convergence is with respect to $H_{\gamma,\mathbb{C}^d}$. In addition, recall that the family of Fourier coefficients is square-summable and satisfies Parseval's identity, see Lemma A.5.11,

$$\|g\|_{H_{\gamma,\mathbb{C}^d}}^2 \;=\; \sum_{\nu \in \mathbb{N}_0^d} \big|\langle g, e_\nu \rangle\big|^2\,.$$

Since convergence in $H_{\gamma,\mathbb{C}^d}$ implies pointwise convergence, we then obtain

$$f(x) = \operatorname{Re} g_{|X}(x) = \operatorname{Re}\left(\sum_{\nu \in \mathbb{N}_0^d} \langle g, e_\nu \rangle\, e_\nu(x) \right) = \sum_{\nu \in \mathbb{N}_0^d} \operatorname{Re}\big(\langle g, e_\nu \rangle\big) e_\nu^X(x)$$

for all $x \in X$, where in the last step we used $e_\nu(x) \in \mathbb{R}$ for $x \in X$. In order to prove *ii)*, it consequently remains to show that $b_\nu := \operatorname{Re}\langle g, e_\nu \rangle$ only depends on f but not on g. To this end, let $\tilde{g} \in H_{\gamma,\mathbb{C}^d}$ be another function with $\operatorname{Re} \tilde{g}_{|X} = f$. By repeating the argument above for $\tilde{g}$, we then find

$$f(x) \;=\; \sum_{\nu\in\mathbb{N}_0^d} \operatorname{Re}\big(\langle\tilde{g}, e_\nu\rangle\big) e_\nu^X(x)\,, \qquad\qquad x \in X.$$

Using the definition (4.35) of e_n^X, we then obtain

$$\sum_{\nu\in\mathbb{N}_0^d} \operatorname{Re}\big(\langle\tilde{g}, e_\nu\rangle\big)\, a_\nu\, x^\nu \;=\; \sum_{\nu\in\mathbb{N}_0^d} \operatorname{Re}\big(\langle g, e_\nu\rangle\big)\, a_\nu\, x^\nu\,, \qquad x \in X,$$

where $a_\nu := a_{n_1}\cdots a_{n_d}$ and $a_n := \big(\frac{2^n}{\gamma^{2n}n!}\big)^{1/2}$. Since X has a non-empty interior, the identity theorem for power series and $a_\nu \neq 0$ then give $\operatorname{Re}\langle\tilde{g}, e_\nu\rangle = \operatorname{Re}\langle g, e_\nu\rangle$ for all $\nu \in \mathbb{N}_0^d$. This shows both (4.36) and (4.37). Finally, Corollary 4.40 and Parseval's identity give

$$\begin{aligned}
\|f\|^2_{H_\gamma(X)} &= \inf\big\{\|g\|_{\gamma,\mathbb{C}^d} : g \in H_{\gamma,\mathbb{C}^d} \text{ with } \operatorname{Re} g_{|X} = f\big\} \\
&= \inf\left\{ \sum_{\nu\in\mathbb{N}_0^d} b_\nu^2 + c_\nu^2 : (c_\nu) \in \ell_2(\mathbb{N}_0^d) \right\} \\
&= \sum_{\nu\in\mathbb{N}_0^d} b_\nu^2\,.
\end{aligned}$$

ii) $\Rightarrow$ *i)*. Since $(b_\nu) \in \ell_2(\mathbb{N}_0^d)$ and $(c_\nu) \in \ell_2(\mathbb{N}_0^d)$ imply $\big(|b_\nu + ic_\nu|\big) \in \ell_2(\mathbb{N}_0^d)$, we have $g \in H_{\gamma,\mathbb{C}^d}$. Furthermore, $\operatorname{Re} g_{|X} = f$ follows from

$$\operatorname{Re} g(x) = \operatorname{Re} \sum_{\nu\in\mathbb{N}_0^d} (b_\nu + ic_\nu) e_\nu(x) = \sum_{\nu\in\mathbb{N}_0^d} b_\nu e_\nu^X(x) = f(x)\,, \qquad\qquad x \in X.$$

□

With the help of the preceding proposition, we can now establish our main result on $H_\gamma(X)$ for input spaces X having a non-empty interior. Roughly speaking, this result states that $H_\gamma(X)$ is isometrically embedded into $H_{\gamma,\mathbb{C}^d}$ via a canonical extension procedure based on a specific ONB of $H_\gamma(X)$.

Theorem 4.42 (ONB of real Gaussian RKHS). *Let $\gamma > 0$ and $X \subset \mathbb{R}^d$ be a subset with a non-empty interior. Furthermore, for an $f \in H_\gamma(X)$ represented by (4.36), we define*

$$\hat{f} := \sum_{\nu\in\mathbb{N}_0^d} b_\nu e_\nu\,.$$

Then the extension operator $\hat{}\colon H_\gamma(X) \to H_{\gamma,\mathbb{C}^d}$ defined by $f \mapsto \hat{f}$ satisfies

$$\begin{aligned}
\operatorname{Re} \hat{f}_{|X} &= f\,, \\
\|\hat{f}\|_{H_{\gamma,\mathbb{C}^d}} &= \|f\|_{H_\gamma(X)}
\end{aligned}$$

for all $f \in H_\gamma(X)$. Moreover, (e_ν^X) is an ONB of $H_\gamma(X)$, and for $f \in H_\gamma(X)$ having the representation (4.36), we have $b_\nu = \langle f, e_\nu^X\rangle$ for all $\nu \in \mathbb{N}_0^d$.

Proof. By (4.36), the extension operator is well-defined. The identities then follow from Proposition 4.41 and Parseval's identity. Moreover, the extension operator is obviously $\mathbb{R}$-linear and satisfies $\hat{e}_\nu^X = e_\nu$ for all $\nu \in \mathbb{N}_0^d$. Consequently, we obtain

$$\|e_{\nu_1}^X \pm e_{\nu_2}^X\|_{H_\gamma(X)} \;=\; \|\hat{e}_{\nu_1}^X \pm \hat{e}_{\nu_2}^X\|_{H_{\gamma,\mathbb{C}^d}} \;=\; \|e_{\nu_1} \pm e_{\nu_2}\|_{H_{\gamma,\mathbb{C}^d}}$$

for $\nu_1, \nu_2 \in \mathbb{N}_0^d$. Using the first polarization identity of Lemma A.5.9, we then see that (e_ν^X) is an ONS in $H_\gamma(X)$. To see that it actually is an ONB we fix an $f \in H_\gamma(X)$. Furthermore, let $(b_\nu) \in \ell_2(\mathbb{N}_0^d)$ be the family that satisfies (4.36). Then

$$\tilde{f} := \sum_{\nu \in \mathbb{N}_0^d} b_\nu e_\nu^X$$

converges in $H_\gamma(X)$. Since convergence in $H_\gamma(X)$ implies pointwise convergence, (4.36) then yields $\tilde{f}(x) = f(x)$ for all $x \in X$. Consequently, (e_ν^X) is an ONB of $H_\gamma(X)$. Finally, the identity $b_\nu = \langle f, e_\nu^X \rangle$, $\nu \in \mathbb{N}_0^d$, follows from the fact that the representation of f by (e_ν^X) is unique. □

In the following, we present some interesting consequences of the preceding theorem.

Corollary 4.43. *Let $X \subset \mathbb{R}^d$ be a subset with non-empty interior, $\gamma > 0$, and $\hat{} : H_\gamma(X) \to H_{\gamma,\mathbb{C}^d}$ be the extension operator defined in Theorem 4.42. Then the extension operator $I : H_\gamma(X) \to H_\gamma(\mathbb{R}^d)$ defined by $If := \operatorname{Re} \hat{f}_{|\mathbb{R}^d}$, $f \in H_\gamma(X)$, is an isometric isomorphism.*

Proof. For $f \in H_\gamma(X)$, we have $(\langle f, e_\nu^X \rangle) \in \ell_2(\mathbb{N}_0^d)$, and hence

$$\tilde{f} := \sum_{\nu \in \mathbb{N}_0^d} \langle f, e_\nu^X \rangle e_\nu^{\mathbb{R}^d}$$

is an element of $H_\gamma(\mathbb{R}^d)$. Moreover, for $\nu \in \mathbb{N}_0^d$, we have $(\operatorname{Re} e_\nu)_{|\mathbb{R}^d} = e_\nu^{\mathbb{R}^d}$ and $\langle f, e_\nu^X \rangle \in \mathbb{R}$, and hence we find $If = \tilde{f}$. Furthermore, $\|f\|_{H_\gamma(X)} = \|If\|_{H_\gamma(\mathbb{R}^d)}$ immediately follows from Parseval's identity. Consequently, I is isometric, linear, and injective. The surjectivity finally follows from the fact that, given an $\tilde{f} \in H_\gamma(\mathbb{R}^d)$, the function

$$f := \sum_{\nu \in \mathbb{N}_0^d} \langle f, e_\nu^{\mathbb{R}^d} \rangle e_\nu^X$$

obviously satisfies $f \in H_\gamma(X)$ and $If = \tilde{f}$. □

Roughly speaking, the preceding corollary means that $H_\gamma(\mathbb{R}^d)$ does not contain "more" functions than $H_\gamma(X)$ if X has a non-empty interior. Moreover, Corollary 4.43 in particular shows that $H_\gamma(X_1)$ and $H_\gamma(X_2)$ are isometrically isomorphic via a simple extension-restriction mapping going through

$H_\gamma(\mathbb{R}^d)$ whenever both input spaces $X_1, X_2 \subset \mathbb{R}^d$ have a non-empty interior. Consequently, we sometimes use the notation $H_\gamma := H_\gamma(X)$ and $\|\cdot\|_\gamma := \|\cdot\|_{H_\gamma(X)}$ if X has a non-empty interior and no confusion can arise.

Besides the isometry above, Theorem 4.42 also yields the following interesting observation.

Corollary 4.44 (Gaussian RKHSs do not contain constants). *Let $\gamma > 0$, $X \subset \mathbb{R}^d$ be a subset with a non-empty interior, and $f \in H_\gamma(X)$. If f is constant on a non-empty open subset A of X, then $f = 0$.*

Proof. Let $c \in \mathbb{R}$ be a constant with $f(x) = c$ for all $x \in A$. Let us define $a_n := (\frac{2^n}{\gamma^{2n} n!})^{1/2}$ for all $n \in \mathbb{N}_0$. Furthermore, for a multi-index $\nu := (n_1, \ldots, n_d) \in \mathbb{N}_0^d$, we write $b_\nu := \langle f, e_\nu^X \rangle$ and $a_\nu := a_{n_1} \cdot \ldots \cdot a_{n_d}$. For $x := (x_1, \ldots, x_d) \in A$, the definition (4.35) and the representation (4.36) then yield

$$c \exp\Bigl(\gamma^{-2} \sum_{j=1}^d x_j^2\Bigr) = f(x) \exp\Bigl(\gamma^{-2} \sum_{j=1}^d x_j^2\Bigr) = \sum_{\nu \in \mathbb{N}_0^d} b_\nu a_\nu x^\nu \,. \tag{4.38}$$

Moreover, for $x \in \mathbb{R}^d$, a simple calculation shows

$$\begin{aligned} \exp\Bigl(\gamma^{-2} \sum_{j=1}^d x_j^2\Bigr) = \prod_{j=1}^d e^{\gamma^{-2} x_j^2} &= \prod_{j=1}^d \Bigl(\sum_{n_j=0}^\infty \frac{x_j^{2n_j}}{n_j! \, \gamma^{2n_j}} \Bigr) \\ &= \sum_{n_1, \ldots, n_d = 0}^\infty \prod_{j=1}^d \frac{x_j^{2n_j}}{n_j! \, \gamma^{2n_j}} \,. \end{aligned}$$

Using (4.38) and the identity theorem for power series, we hence obtain

$$b_\nu a_\nu = \begin{cases} c \gamma^{-|\nu|} \prod_{j=1}^d \frac{1}{n_j!} & \text{if } \nu = (2n_1, \ldots, 2n_d) \text{ for some } (n_1, \ldots, n_d) \in \mathbb{N}_0^d \\ 0 & \text{otherwise} \,, \end{cases}$$

or in other words

$$b_\nu = \begin{cases} c \prod_{j=1}^d \frac{\sqrt{(2n_j)!}}{n_j!} \, 2^{-n_j} & \text{if } \nu = (2n_1, \ldots, 2n_d) \text{ for some } (n_1, \ldots, n_d) \in \mathbb{N}_0^d \\ 0 & \text{otherwise} \,. \end{cases}$$

Consequently, Parseval's identity yields

$$\begin{aligned} \|f\|_{H_\gamma(X)}^2 \;=\; \sum_{\nu \in \mathbb{N}_0^d} b_\nu^2 &= \sum_{n_1, \ldots, n_d = 0}^\infty c^2 \prod_{j=1}^d \frac{(2n_j)!}{(n_j!)^2} \, 2^{-2n_j} \\ &= \prod_{j=1}^d \Bigl(\sum_{n_j=0}^\infty c^{2/d} \, \frac{(2n_j)!}{(n_j!)^2} \, 2^{-2n_j} \Bigr) \\ &= \Bigl(\sum_{n=0}^\infty c^{2/d} \, \frac{(2n)!}{(n!)^2} \, 2^{-2n} \Bigr)^d \,. \end{aligned}$$

Let us write $\alpha_n := \frac{(2n)!}{(n!)^2}\, 2^{-2n}$ for $n \in \mathbb{N}_0$. By an easy calculation, we then obtain

$$\frac{\alpha_{n+1}}{\alpha_n} = \frac{(2(n+1))!\ (n!)^2\ 2^n}{(2n)!\ ((n+1)!)^2\ 2^{2(n+1)}} = \frac{(2n+1)(2n+2)}{4(n+1)^2} = \frac{2n+1}{2n+2} \geq \frac{n}{n+1}$$

for all $n \geq 1$. In other words, $(n\alpha_n)$ is an increasing, positive sequence. Consequently there exists an $\alpha > 0$ with $\alpha_n \geq \frac{\alpha}{n}$ for all $n \geq 1$, and hence we find $\sum_{n=0}^{\infty} \alpha_n = \infty$. Therefore, $\|f\|^2_{H_\gamma(X)} < \infty$ implies $c = 0$, and thus we have $f = 0$. □

The preceding corollary shows in particular that $\mathbf{1}_A \notin H_\gamma(X)$ for all open subsets $A \subset X$. Some interesting consequences of this observation with respect to the hinge loss used in classification are discussed in Exercise 4.8.

Let us now compare the norms $\|\cdot\|_\gamma$ for different values of γ. To this end, we first observe that the weight function in the definition of $\|\cdot\|_{\gamma,\mathbb{C}^d}$ satisfies

$$e^{\gamma^{-2}\sum_{j=1}^d (z_j - \bar{z}_j)^2} = e^{-4\gamma^{-2}\sum_{j=1}^d y_j^2},$$

where $y_j := \operatorname{Im} z_j$, $j = 1, \ldots, d$. For $\gamma_1 \leq \gamma_2$, we hence find $H_{\gamma_2,\mathbb{C}^d} \subset H_{\gamma_1,\mathbb{C}^d}$ and

$$\|f\|_{H_{\gamma_1,\mathbb{C}^d}} \ \leq\ \left(\frac{\gamma_2}{\gamma_1}\right)^d \|f\|_{H_{\gamma_2,\mathbb{C}^d}}, \qquad\qquad f \in H_{\gamma_2,\mathbb{C}^d}.$$

This suggests that a similar relation holds for the RKHSs of the real Gaussian kernels. In order to investigate this conjecture, let us now present another feature space and feature map for k_γ. To this end, recall that $L_2(\mathbb{R}^d)$ denotes the space of Lebesgue square-integrable functions $\mathbb{R}^d \to \mathbb{R}$ equipped with the usual norm $\|\cdot\|_2$. Our first result shows that $L_2(\mathbb{R}^d)$ is a feature space of k_γ.

Lemma 4.45. *For $0 < \gamma < \infty$ and $X \subset \mathbb{R}^d$, we define $\Phi_\gamma : X \to L_2(\mathbb{R}^d)$ by*

$$\Phi_\gamma(x) \ :=\ \frac{2^{\frac{d}{2}}}{\pi^{\frac{d}{4}}\gamma^{\frac{d}{2}}} e^{-2\gamma^{-2}\|x - \cdot\|_2^2}, \qquad\qquad x \in X.$$

Then $\Phi_\gamma : X \to L_2(\mathbb{R}^d)$ is a feature map of k_γ.

Proof. Let us first recall that, using the density of the normal distribution, we have

$$\int_{\mathbb{R}^d} e^{-t^{-1}\|z-x\|_2^2} dz = (\pi t)^{\frac{d}{2}} \tag{4.39}$$

for all $t > 0$ and $x \in \mathbb{R}^d$. Moreover, for $\alpha \geq 0$, an elementary calculation shows that

$$\|y - x\|_2^2 + \alpha\|y - x'\|_2^2 = \frac{\alpha}{1+\alpha}\|x - x'\|_2^2 + (1+\alpha)\left\| y - \frac{x + \alpha x'}{1+\alpha}\right\|_2^2 \tag{4.40}$$

for all $y, x, x' \in \mathbb{R}^d$. By using (4.39) and setting $\alpha := 1$ in (4.40), we now obtain

$$\begin{aligned}\langle \Phi_\gamma(x), \Phi_\gamma(x')\rangle_{L_2(\mathbb{R}^d)} &= \frac{2^d}{\pi^{\frac{d}{2}}\gamma^d}\int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-z\|_2^2} e^{-2\gamma^{-2}\|x'-z\|_2^2} dz \\ &= \frac{2^d}{\pi^{\frac{d}{2}}\gamma^d} e^{-\gamma^{-2}\|x-x'\|_2^2} \int_{\mathbb{R}^d} e^{-4\gamma^{-2}\|z-\frac{x+x'}{2}\|_2^2} dz \\ &= \frac{2^d}{\pi^{\frac{d}{2}}\gamma^d} \cdot e^{-\gamma^{-2}\|x-x'\|_2^2} \Big(\frac{\pi\gamma^2}{4}\Big)^{\frac{d}{2}} \\ &= k_\gamma(x, x')\,,\end{aligned}$$

and hence Φ_γ is a feature map and $L_2(\mathbb{R}^d)$ is a feature space of k_γ. □

Having the feature map $\Phi_\gamma : X \to L_2(\mathbb{R}^d)$ of k_γ, we can now give another description of the RKHS of k_γ. To this end, we need the integral operators $W_t : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$, $t > 0$, defined by

$$W_t g(x) := (\pi t)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{-t^{-1}\|y-x\|_2^2} g(y) dy\,, \qquad g \in L_2(\mathbb{R}^d),\, x \in \mathbb{R}^d. \quad (4.41)$$

Note that W_t is actually a convolution operator, i.e., for $g \in L_2(\mathbb{R}^d)$ we have $W_t g = k * g$, where $k := (\pi t)^{-\frac{d}{2}} e^{-t^{-1}\|\cdot\|_2^2}$. Moreover, we have $\|k\|_1 = 1$ by (4.39), and hence Young's inequality (see Theorem A.5.23) that shows

$$\|W_t g\|_2 \le \|g\|_2\,, \qquad g \in L_2(\mathbb{R}^d),\, t > 0. \quad (4.42)$$

In other words, we have $\|W_t : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d)\| \le 1$ for all $t > 0$.

With the help of the operator family $(W_t)_{t>0}$, we can now give another description of the spaces $H_\gamma(X)$.

Proposition 4.46. *For $0 < \gamma_1 < \gamma_2 < \infty$, we define $t := \frac{1}{2}(\gamma_2^2 - \gamma_1^2)$. Then, for all non-empty $X \subset \mathbb{R}^d$, we obtain a commutative diagram*

$$\begin{array}{ccc} H_{\gamma_2}(X) & \xrightarrow{\ \mathrm{id}\ } & H_{\gamma_1}(X) \\ \uparrow{\scriptstyle V_{\gamma_2}} & & \uparrow{\scriptstyle V_{\gamma_1}} \\ L_2(\mathbb{R}^d) & \xrightarrow[(\frac{\gamma_2}{\gamma_1})^{\frac{d}{2}} W_t]{} & L_2(\mathbb{R}^d) \end{array}$$

where the vertical maps V_{γ_1} and V_{γ_2} are the metric surjections of Theorem 4.21. Moreover, these metric surjections are of the form

$$V_\gamma g(x) = \frac{2^{\frac{d}{2}}}{\gamma^{\frac{d}{2}}\pi^{\frac{d}{4}}} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-y\|_2^2} g(y) dy\,, \qquad g \in L_2(\mathbb{R}^d),\, x \in X, \quad (4.43)$$

where $\gamma \in \{\gamma_1, \gamma_2\}$. Finally, we have

$$\|\,\mathrm{id} : H_{\gamma_2}(X) \to H_{\gamma_1}(X)\| \le \left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{d}{2}}. \quad (4.44)$$

Proof. For $\gamma > 0$, let $V_\gamma : L_2(\mathbb{R}^d) \to H_\gamma(X)$ be the metric surjection of Theorem 4.21. Furthermore, let Φ_γ be the feature map defined in Lemma 4.45. For $g \in L_2(\mathbb{R}^d)$ and $x \in X$, we then have $V_\gamma g(x) = \langle g, \Phi_\gamma(x) \rangle_{L_2(\mathbb{R}^d)}$, and hence we obtain (4.43). In order to establish the diagram, let us first consider the case $X = \mathbb{R}^d$. Then (4.41) together with (4.43) gives the relation

$$V_\gamma g = (\pi\gamma^2)^{\frac{d}{4}} W_{\frac{\gamma^2}{2}} g\,, \qquad g \in L_2(\mathbb{R}^d). \tag{4.45}$$

Let us now show that the operator family $(W_t)_{t>0}$ is a semi-group, i.e., it satisfies

$$W_{t_1+t_2} = W_{t_1} W_{t_2}\,, \qquad t_1, t_2 > 0. \tag{4.46}$$

To this end, let us fix a $g \in L_2(\mathbb{R}^d)$ and an $x_0 \in \mathbb{R}^d$. Then, for $\alpha := \frac{t_1}{t_2}$, equations (4.40) and (4.39) yield

$$\begin{aligned}
W_{t_1} W_{t_2} g(x_0) &= (\pi t_1)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{-t_1^{-1}\|x_0 - y\|_2^2} W_{t_2} g(y) dy \\
&= (\pi^2 t_1 t_2)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-t_1^{-1}\|x_0 - y\|_2^2} e^{-t_2^{-1}\|x - y\|_2^2} g(x) dx\, dy \\
&= (\pi^2 t_1 t_2)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-\frac{\|x_0 - x\|_2^2}{t_1 + t_2} - \frac{t_1 + t_2}{t_1 t_2}\|y - \frac{x_0 + \alpha x}{1+\alpha}\|_2^2} g(x) dy\, dx \\
&= W_{t_1+t_2} g(x_0)\,,
\end{aligned}$$

i.e., (4.46) is verified. Combining (4.45) and (4.46) then gives the diagram in the case of $X = \mathbb{R}^d$. The general case $X \subset \mathbb{R}^d$ follows from the fact that the *computation* of V_γ in (4.43) is independent of X. Finally, since V_{γ_2} is a metric surjection, we obtain

$$\|\operatorname{id} \circ V_{\gamma_2} : L_2(\mathbb{R}^d) \to H_{\gamma_1}(X)\| = \|\operatorname{id} : H_{\gamma_2}(X) \to H_{\gamma_1}(X)\|\,,$$

and hence the commutativity of the diagram implies

$$\|\operatorname{id} : H_{\gamma_2}(X) \to H_{\gamma_1}(X)\| = \left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{d}{2}} \|V_{\gamma_1} \circ W_t\| \leq \left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{d}{2}} \|W_t\|\,.$$

Moreover, we have $\|W_t\| \leq 1$ by (4.42), and thus we find the assertion. □

If the set X in the preceding proposition has a non-empty interior, then the metric surjections V_{γ_1} and V_{γ_2} are actually isometric isomorphisms. This is a direct consequence of the following theorem, (4.43), and the fact that the restriction operator mapping $H_\gamma(\mathbb{R}^d)$ to $H_\gamma(X)$ is an isometric isomorphism.

Theorem 4.47 (Injectivity of Gaussian integral operators). *Let μ be either a finite measure on $\mathbb{R}^d$ or the Lebesgue measure on $\mathbb{R}^d$, and $p \in (1, \infty)$. Moreover, let k_γ be the Gaussian RBF kernel with width $\gamma > 0$. Then the operator $S_{k_\gamma} : L_p(\mu) \to H_\gamma(\mathbb{R}^d)$ defined by (4.17) is injective.*

Proof. Let us write $S_\gamma := S_{k_\gamma}$. We fix an $f \in L_p(\mu)$ with $S_\gamma f = 0$. Obviously, our goal is to show that $f = 0$. To this end, our first intermediate goal is to prove that the map $g : \mathbb{R}^d \times (0,\infty) \to \mathbb{R}$ defined by

$$g(x,t) := \int_{\mathbb{R}^d} e^{-t\|x-x'\|_2^2} f(x')\, d\mu(x')\,, \qquad x \in \mathbb{R}^d,\, t \in (0,\infty),$$

is real-analytic in t for all fixed $x \in \mathbb{R}^d$. Here we note that $e^{-t\|x-\cdot\|_2^2} \in L_{p'}(\mu)$ together with Hölder's inequality ensures that the integral above is defined and finite. To show the analyticity, we now fix a $t_0 \in (0,\infty)$ and define

$$a_i(x,x',t) := \frac{(-\|x-x'\|_2^2)^i e^{-t_0\|x-x'\|_2^2}}{i!}(t-t_0)^i\, f(x')$$

for all $x, x' \in \mathbb{R}^d$, $t \in (0,t)$, and $i \geq 0$. Obviously, we have

$$g(x,t) = \int_{\mathbb{R}^d} \sum_{i=0}^{\infty} a_i(x,x',t)\, d\mu(x') \tag{4.47}$$

for all $x \in \mathbb{R}^d$ and $t \in (0,\infty)$. Moreover, for $t \in (0,t_0]$, we find

$$\sum_{i=0}^{\infty} \bigl|a_i(x,x',t)\bigr| = \sum_{i=0}^{\infty} \frac{\|x-x'\|_2^{2i} e^{-t_0\|x-x'\|_2^2}}{i!}(t_0-t)^i f(x') = e^{-t\|x-x'\|_2^2} f(x')\,,$$

and hence Hölder's inequality yields

$$\int_{\mathbb{R}^d} \sum_{i=0}^{\infty} \bigl|a_i(x,x',t)\bigr|\, d\mu(x') < \infty\,. \tag{4.48}$$

On the other hand, for $t \in [t_0,\infty)$, we have

$$\begin{aligned}\sum_{i=0}^{\infty} \bigl|a_i(x,x',t)\bigr| &= \sum_{i=0}^{\infty} \frac{\|x-x'\|_2^{2i} e^{-t_0\|x-x'\|_2^2}}{i!}(t-t_0)^i f(x') \\ &= e^{-(2t_0-t)\|x-x'\|_2^2} f(x')\,,\end{aligned}$$

and from this it is easy to conclude by Hölder's inequality that (4.48) also holds for $t \in [t_0, 2t_0)$. By Fubini's theorem, we can then change the order of integration and summation in (4.47) to obtain

$$g(x,t) = \sum_{i=0}^{\infty} \biggl(\int_{\mathbb{R}^d} \frac{(-\|x-x'\|_2^2)^i e^{-t_0\|x-x'\|_2^2}}{i!}\, f(x')\, d\mu(x') \biggr) (t-t_0)^i$$

for all $t \in (0, 2t_0)$. In other words, $g(x,\cdot\,)$ can be locally expressed by a power series, i.e., it is real-analytic. Let us now define

$$u(x,t) := t^{-\frac{d}{2}} g\Big(x, \frac{1}{4t}\Big) = \int_{\mathbb{R}^d} t^{-\frac{d}{2}} e^{-\frac{\|x-x'\|_2^2}{4t}} f(x')\, d\mu(x')\,, \qquad x \in \mathbb{R}^d,\, t > 0\,.$$

Obviously, $u(x,\cdot\,)$ is again real-analytic for all $x \in \mathbb{R}^d$. Moreover, for fixed $x' := (x'_1, \ldots, x'_d) \in \mathbb{R}^d$, the map

$$u_0(x,t) := t^{-\frac{d}{2}} e^{-\frac{\|x-x'\|_2^2}{4t}}\,, \qquad x \in \mathbb{R}^d,\, t > 0,$$

which appears in the integral above, satisfies

$$\frac{\partial u_0}{\partial t}(x,t) = t^{-\frac{d}{2}-2} e^{-\frac{\|x-x'\|_2^2}{4t}} \left(\frac{\|x-x'\|_2^2}{4t} - \frac{d\,t}{2} \right),$$
$$\frac{\partial^2 u_0}{\partial^2 x_i}(x,t) = t^{-\frac{d}{2}-2} e^{-\frac{\|x-x'\|_2^2}{4t}} \left(\frac{(x_i - x'_i)^2}{4t} - \frac{t}{2} \right),$$

for all $t > 0$ and all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Consequently, u_0 satisfies the partial differential equation

$$\frac{\partial u_0}{\partial t} = \Delta u_0 := \sum_{i=1}^d \frac{\partial^2 u_0}{\partial^2 x_i}\,.$$

Moreover, as a function of x', all derivatives of u_0 are contained in $L_{p'}(\mu)$, and these derivatives are continuous with respect to the variables x and t. Another application of Hölder's inequality, together with Corollary A.3.7, shows that the function u satisfies the same partial differential equation. This leads to

$$\frac{\partial^2 u}{\partial^2 t} = \frac{\partial}{\partial t} \sum_{i=1}^d \frac{\partial^2 u}{\partial^2 x_i} = \sum_{i=1}^d \frac{\partial^3 u}{\partial^2 x_i \partial t} = \sum_{i=1}^d \sum_{j=1}^d \frac{\partial^4 u}{\partial^2 x_i \partial^2 x_j} = \Delta^2 u\,,$$

and by iterating this procedure we obtain $\frac{\partial^n u}{\partial^n t} = \Delta^n u$ for all $n \geq 1$. Let us now recall that our $f \in L_p(\mu)$ satisfies $S_\gamma f = 0$. For $t_0 := \gamma^2/4$, we then have $u(x,t_0) = (2/\gamma)^d\, S_\gamma f(x) = 0$ for all $x \in \mathbb{R}^d$, and hence we obtain

$$\frac{\partial^n u}{\partial^n t}(x,t_0) = \Delta^n u(x,t_0) = 0\,, \qquad x \in \mathbb{R}^d.$$

By the analyticity of $u(x,\cdot\,)$, we thus conclude that $u(x,t) = 0$ for all $x \in \mathbb{R}^d$ and all $t > 0$. Now let $h : \mathbb{R}^d \to \mathbb{R}$ be a continuous function with compact support. Then we obviously have $\|h\|_\infty < \infty$, $h \in L_p(\mu)$, and

$$0 = \int_{\mathbb{R}^d} h(x) u(x,t) dx = t^{-\frac{d}{2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} h(x) e^{-\frac{\|x-x'\|_2^2}{4t}} f(x')\, d\mu(x') dx \qquad (4.49)$$

for all $t > 0$. Now note that if μ is finite, we easily find that

$$(x,x') \mapsto h(x) e^{-\frac{\|x-x'\|_2^2}{4t}} f(x') \qquad (4.50)$$

is integrable with respect to the product of μ and the Lebesgue measure on $\mathbb{R}^d$. Moreover, if μ is the Lebesgue measure, its translation invariance yields

$$\begin{aligned}
&\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \bigl| h(x) e^{-\frac{\|x-x'\|_2^2}{4t}} f(x') \bigr| \, d\mu(x') dx \\
&\leq \int_{\mathbb{R}^d} \bigl| h(x) \bigr| \cdot \|f\|_{L_p(\mu)} \left(\int_{\mathbb{R}^d} e^{-\frac{p' \|x-x'\|_2^2}{4t}} \, d\mu(x') \right)^{1/p'} dx \\
&< \infty \,,
\end{aligned}$$

i.e., the function in (4.50) is integrable in this case, too. For

$$h_t(x') \;:=\; t^{-\frac{d}{2}} \int_{\mathbb{R}^d} h(x) e^{-\frac{\|x-x'\|_2^2}{4t}} \, dx, \qquad\qquad x' \in \mathbb{R}^d, \, t > 0,$$

Fubini's theorem and (4.49) then yield

$$0 = t^{-\frac{d}{2}} \int\limits_{\mathbb{R}^d} \int\limits_{\mathbb{R}^d} h(x) e^{-\frac{\|x-x'\|_2^2}{4t}} \, f(x') \, dx \, d\mu(x') = \int\limits_{\mathbb{R}^d} f(x') h_t(x') d\mu(x') \,. \quad (4.51)$$

Now fix an $x \in \mathbb{R}^d$ and an $\varepsilon > 0$. Then there exists a $\delta > 0$ such that, for all $x' \in \mathbb{R}^d$ with $\|x' - x\|_2 \leq \delta$, we have $|h(x') - h(x)| \leq (4\pi)^{-d/2} \epsilon$. Since

$$(4\pi t)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{-\frac{\|x-x'\|_2^2}{4t}} dx' = 1 \,, \qquad\qquad t > 0,$$

we hence obtain

$$\begin{aligned}
h_t(x) - (4\pi)^{\frac{d}{2}} h(x) &= t^{-\frac{d}{2}} \int_{\mathbb{R}^d} \bigl(h(x') - h(x) \bigr) e^{-\frac{\|x-x'\|_2^2}{4t}} dx' \\
&\leq \varepsilon + t^{-\frac{d}{2}} \int\limits_{\|x'-x\|_2 > \delta} \bigl(h(x') - h(x) \bigr) e^{-\frac{\|x-x'\|_2^2}{4t}} dx' \\
&\leq \varepsilon + 2 \|h\|_\infty \, t^{-\frac{d}{2}} \int\limits_{\|x'\|_2 > \delta} e^{-\frac{\|x'\|_2^2}{4t}} dx' \\
&\leq \varepsilon + 8 \pi^{d/2} \frac{\max\{1, d/2\}}{\Gamma(d/2)} \|h\|_\infty \, \delta^{d-2} t^{1-d/2} e^{-\frac{\delta^2}{4t}}
\end{aligned}$$

for all $0 < t \leq \delta^2/(2d)$, where in the last step we used (A.3) and (A.5). Since the last term of this estimate tends to 0 for $t \to 0$, we conclude that $\lim_{t \to 0} h_t(x) = (4\pi)^{\frac{d}{2}} h(x)$ for all $x \in \mathbb{R}^d$. Therefore the dominated convergence theorem and (4.51) yield

$$0 = \lim_{t \to 0} \int_{\mathbb{R}^d} f(x') h_t(x') d\mu(x') = \int_{\mathbb{R}^d} f(x') h(x') d\mu(x') = \langle f, h \rangle_{L_{p'}(\mu), L_p(\mu)}.$$

Since for finite measures it follows from Theorem A.3.15 and Theorem A.5.25 that the continuous functions with compact support are dense in $L_p(\mu)$, we find $f = 0$. Finally, the Lebesgue measure is also regular, and hence we find the assertion in this case analogously. □

Our last goal is to compute Sobolev norms for functions in $H_\gamma(X)$. This is done in the following theorem.

Theorem 4.48 (Sobolev norms for Gaussian RKHSs). *Let $X \subset \mathbb{R}^d$ be a bounded non-empty open set, $\gamma > 0$, and $m \geq 1$. Then there exists a constant $c_{m,d} > 0$ only depending on m and d such that for all $f \in H_\gamma(X)$ we have*

$$\|f\|_{W^m(X)} \leq c_{m,d}\sqrt{\mathrm{vol}(X)} \left(\sum_{\substack{\alpha \in \mathbb{N}_0^d \\ |\alpha| \leq m}} \gamma^{-2|\alpha|} \right)^{1/2} \|f\|_{H_\gamma(X)} .$$

Proof. Let us fix a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $|\alpha| = m$. Moreover, let $V_\gamma : L_2(X) \to H_\gamma(X)$ be the metric surjection defined by (4.43). For a fixed $f \in H_\gamma(X)$ and $\varepsilon > 0$, there then exists a $g \in L_2(\mathbb{R}^d)$ such that $V_\gamma g = f$ and $\|g\|_{L_2(\mathbb{R}^d)} \leq (1+\varepsilon)\|f\|_{H_\gamma(X)}$. By Hölder's inequality, we then have

$$\begin{aligned}
\left\| \partial^\alpha f \right\|^2_{\mathcal{L}_2(X)} &= \frac{2^d}{\gamma^d \pi^{\frac{d}{2}}} \int_X \left(\partial_x^\alpha \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-y\|_2^2} g(y) dy \right)^2 dx \\
&\leq \frac{2^d}{\gamma^d \pi^{\frac{d}{2}}} \int_X \left(\int_{\mathbb{R}^d} \partial_x^\alpha e^{-2\gamma^{-2}\|x-y\|_2^2} \, |g(y)| \, dy \right)^2 dx \\
&\leq \frac{2^d}{\gamma^d \pi^{\frac{d}{2}}} \|g\|^2_{L_2(\mathbb{R}^d)} \int_X \int_{\mathbb{R}^d} \left| \partial_x^\alpha e^{-2\gamma^{-2}\|x-y\|_2^2} \right|^2 dy \, dx . \qquad (4.52)
\end{aligned}$$

Now recall that the Hermite polynomials h_n, $n \geq 0$, defined in (A.1) satisfy

$$\frac{\partial^n}{\partial t^n} e^{-t^2} = (-1)^n e^{-t^2} h_n(t) , \qquad t \in \mathbb{R} ,$$

and hence we have

$$\frac{\partial^n}{\partial t^n} e^{-2\gamma^{-2}(t-s)^2} = \left(-\sqrt{2}\,\gamma^{-1}\right)^n e^{-2\gamma^{-2}(t-s)^2} h_n\left(\sqrt{2}\,\gamma^{-1}(t-s)\right)$$

for all $s, t \in \mathbb{R}$. Using the translation invariance of the Lebesgue measure, $h_n(-s) = (-1)^n h_n(s)$, a change of variables, and (A.2), we conclude that

$$\begin{aligned}
\int_{\mathbb{R}} \left| \frac{d^n}{dt^n} e^{-2\gamma^{-2}(t-s)^2} \right|^2 ds &= \left(2\gamma^{-2}\right)^n \int_{\mathbb{R}} e^{-4\gamma^{-2}(t-s)^2} h_n^2\left(\sqrt{2}\,\gamma^{-1}(t-s)\right) ds \\
&= \left(2\gamma^{-2}\right)^n \int_{\mathbb{R}} e^{-4\gamma^{-2}s^2} h_n^2\left(\sqrt{2}\,\gamma^{-1}s\right) ds \\
&= \left(\sqrt{2}\,\gamma^{-1}\right)^{2n-1} \int_{\mathbb{R}} e^{-2s^2} h_n^2(s) ds \\
&\leq \sqrt{\pi}\, 2^{2n-1/2} n! \, \gamma^{1-2n} .
\end{aligned}$$

Since $e^{-2\gamma^{-2}\|x-y\|_2^2} = \prod_{i=1}^d e^{-2\gamma^{-2}(x_i-y_i)^2}$, we hence find

$$\int_{\mathbb{R}^d} \left|\partial_x^\alpha e^{-2\gamma^{-2}\|x-y\|_2^2}\right|^2 dy \leq \pi^{m/2} 2^{2m-d/2} \alpha!\, \gamma^{d-2m}\,,$$

where $\alpha! := \alpha_1! \cdots \alpha_d!$. Combining this estimate with (4.52), we obtain

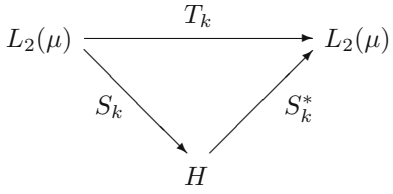
$$\left\| \partial^\alpha f \right\|_{\mathcal{L}_2(X)}^2 \leq (1+\varepsilon)\, 2^{2m+d/2} \pi^{(m-d)/2} \alpha!\ \operatorname{vol}(X) \gamma^{-2m} \|f\|_{H_\gamma(X)}^2\,.$$

Finally, since f is a restriction of an analytic function defined on $\mathbb{R}^d$, see Corollary 4.40, we have $\partial^{(\alpha)} f = \partial^\alpha f$, where $\partial^{(\alpha)} f$ denotes the weak α-derivative defined in Section A.5.5. From this we easily obtain the assertion. □

4.5 Mercer's Theorem (*)

In this section, we present Mercer's theorem, which provides a series representation for continuous kernels on compact domains. This series representation is then used to describe the corresponding RKHSs.

Let us begin with some preliminary considerations. To this end, let X be a measurable space, μ be a σ-finite measure on X, and k be a measurable kernel on X with $\|k\|_{L_2(\mu)} < \infty$. Moreover, recall the following factorization of the operators defined in Theorem 4.26 and Theorem 4.27:



Theorem 4.27 showed that $T_k = S_k^* S_k$ is compact, positive, and self-adjoint, and hence the Spectral Theorem A.5.13 shows that there exist an at most countable ONS $(e_i)_{i\in I}$ and a family $(\lambda_i)_{i\in I} \subset \mathbb{R}$ converging to 0 such that $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$ and

$$T_k f = \sum_{i\in I} \lambda_i \langle f, e_i\rangle e_i, \qquad\qquad f \in L_2(\mu)\,.$$

Moreover, $\{\lambda_i : i \in I\}$ is the set of non-zero eigenvalues of T_k. Let us write $\tilde{e}_i := \lambda_i^{-1} S_k e_i \in H$ for $i \in I$. Then the diagram shows $\tilde{e}_i = \lambda_i^{-1} T_k e_i$ almost surely, and hence we have $e_i = \lambda_i^{-1} T_k e_i = \tilde{e}_i$ almost surely. Consequently, we may assume without loss of generality that $e_i \in H$ and $\lambda_i e_i = S_k e_i$ for all $i \in I$. From this we conclude that

$$\begin{aligned}\lambda_i \lambda_j \langle e_i, e_j\rangle_H = \langle S_k e_i, S_k e_j\rangle_H = \langle e_i, S_k^* S_k e_j\rangle_{L_2(\mu)} &= \langle e_i, T_k e_j\rangle_{L_2(\mu)} \\ &= \lambda_j \langle e_i, e_j\rangle_{L_2(\mu)}\,.\end{aligned}$$

In other words, $(\sqrt{\lambda_i}e_i)_{i\in I}$ is an ONS in H. The goal of this section is to show that under certain circumstances this family is even an ONB. To this end, we need the following theorem, whose proof can be found, for example, in Riesz and Nagy (1990).

Theorem 4.49 (Mercer's theorem). *Let X be a compact metric space and $k : X \times X \to \mathbb{R}$ be a continuous kernel. Furthermore, let μ be a finite Borel measure with $\operatorname{supp}\mu = X$. Then, for $(e_i)_{i\in I}$ and $(\lambda_i)_{i\in I}$ as above, we have*

$$k(x, x') = \sum_{i\in I} \lambda_i e_i(x) e_i(x')\,, \qquad x, x' \in X, \tag{4.53}$$

where the convergence is absolute and uniform.

Note that (4.53) together with the proof of Lemma 4.2 shows that $\Phi : X \to \ell_2$ defined by $\Phi(x) := (\sqrt{\lambda_i}e_i(x))_{i\in I}$, $x \in X$, is a feature map of k. In order to show that $(\sqrt{\lambda_i}e_i)_{i\in I}$ is an ONB of H, we need the following corollary.

Corollary 4.50. *With the assumptions and notations of Theorem 4.49, the series $\sum_{i\in I} a_i\sqrt{\lambda_i}e_i(x)$ converges absolutely and uniformly for all $(a_i) \in \ell_2(I)$.*

Proof. For $x \in X$ and $J \subset I$, Hölder's inequality and Mercer's theorem imply

$$\sum_{i\in J} |a_i\sqrt{\lambda_i}e_i(x)| \le \left(\sum_{i\in J} a_i^2\right)^{1/2} \left(\sum_{i\in J} \lambda_i e_i^2(x)\right)^{1/2} = \|(a_i)\|_{\ell_2(I)} \cdot \sqrt{k(x,x)}\,.$$

From this the assertion easily follows. □

With the help of the Corollary 4.50, we can now give an explicit representation of the RKHSs of continuous kernels on a compact metric space X.

Theorem 4.51 (Mercer representation of RKHSs). *With the assumptions and notations of Theorem 4.49, we define*

$$H := \left\{ \sum_{i\in I} a_i\sqrt{\lambda_i}e_i \,:\, (a_i) \in \ell_2(I) \right\}.$$

Moreover, for $f := \sum_{i\in I} a_i\sqrt{\lambda_i}e_i \in H$ and $g := \sum_{i\in I} b_i\sqrt{\lambda_i}e_i \in H$, we write

$$\langle f, g\rangle_H := \sum_{i\in I} a_i b_i\ .$$

Then H equipped with inner product $\langle\cdot,\cdot\rangle_H$ is the RKHS of the kernel k. Furthermore, the operator $T_k^{1/2} : L_2(\mu) \to H$ is an isometric isomorphism.

Proof. Routine work shows that $\langle \cdot\,, \cdot \rangle$ is a well-defined inner product and hence H is a Hilbert function space. Now, for fixed $x \in X$, Mercer's theorem implies

$$k(\cdot\,, x) = \sum_{i \in I} \sqrt{\lambda_i} e_i(x) \sqrt{\lambda_i} e_i(\cdot\,),$$

and since Mercer's theorem also yields

$$\|(\sqrt{\lambda_i} e_i(x))\|^2_{\ell_2(I)} = \sum_{i \in I} \lambda_i e_i^2(x) = k(x,x) < \infty\,,$$

we find $k(\cdot\,, x) \in H$. Moreover, for $f := \sum_{i \in I} a_i \sqrt{\lambda_i} e_i \in H$, we have

$$\langle f, k(\cdot\,, x) \rangle_H = \sum_{i \in I} a_i \sqrt{\lambda_i} e_i(x) = f(x)\,, \qquad x \in X,$$

i.e., k is the reproducing kernel of H.

Let us now consider the operator $T_k^{1/2}$. To this end, let us fix an $f \in L_2(\mu)$. Since (e_i) is an orthonormal basis in $L_2(\mu)$, we then find $f = \sum_{i \in I} \langle f, e_i \rangle_{L_2(\mu)} e_i$, where the convergence takes place in $L_2(\mu)$. Consequently, we have

$$T_k^{1/2} f = \sum_{i \in I} \langle f, e_i \rangle_{L_2(\mu)} \sqrt{\lambda_i} e_i\,, \tag{4.54}$$

where the convergence is again with respect to the $L_2(\mu)$-norm. Now, Parseval's identity gives $(\langle f, e_i \rangle_{L_2(\mu)}) \in \ell_2(I)$, and hence we find $T_k^{1/2} f \in H$ for all $f \in L_2(\mu)$. Moreover, this also shows by Corollary 4.50 that the convergence in (4.54) is absolute and uniform and that

$$\|T_k^{1/2} f\|_H^2 = \sum_{i \in I} |\langle f, e_i \rangle_{L_2(\mu)}|^2 = \|f\|^2_{L_2(\mu)}\,.$$

In other words, $T_k^{1/2} : L_2(\mu) \to H$ is isometric. Finally, to check that the operator is surjective, we fix an $f \in H$. Then there exists an $(a_i) \in \ell_2$ such that $f(x) = \sum_{i \in I} a_i \sqrt{\lambda_i} e_i(x)$ for all $x \in X$. Now we obviously have $g := \sum_{i \in I} a_i e_i \in L_2(\mu)$ with convergence in $L_2(\mu)$, and thus $\langle g, e_i \rangle_{L_2(\mu)} = a_i$. Furthermore, we have already seen that the convergence in (4.54) is pointwise, and hence for all $x \in X$ we finally obtain

$$T_k^{1/2} g(x) = \sum_{i \in I} \langle g, e_i \rangle_{L_2(\mu)} \sqrt{\lambda_i} e_i(x) = \sum_{i \in I} a_i \sqrt{\lambda_i} e_i(x) = f(x)\,. \qquad \square$$

4.6 Large Reproducing Kernel Hilbert Spaces

We saw in Section 1.2 that SVMs are based on minimization problems over RKHSs. Moreover, we will see in the following chapters that the size of the

chosen RKHS has a twofold impact on the generalization ability of the SVM: on the one hand, a "small size" inhibits the learning machine to produce highly complex decision functions and hence can prevent the SVM from overfitting in the presence of noise. On the other hand, for complex distributions, a "small" RKHS may not be sufficient to provide an accurate decision function, so the SVM underfits. In this section, we thus investigate RKHSs that are rich enough to provide *arbitrarily accurate* decision functions for *all* distributions. The reason for introducing these RKHSs is that their flexibility is necessary to guarantee learning in the absence of assumptions on the data-generating distribution. However, as we have indicated above, this flexibility also carries the danger of overfitting. We will thus investigate in Chapters 6 and 7 how regularized learning machines such as SVMs use the regularizer to avoid this overfitting.

Let us now begin by introducing a class of particularly large RKHSs.

Definition 4.52. *A continuous kernel k on a compact metric space X is called* ***universal*** *if the RKHS H of k is dense in $C(X)$, i.e., for every function $g \in C(X)$ and all $\varepsilon > 0$ there exists an $f \in H$ such that*

$$\|f - g\|_\infty \leq \varepsilon \,.$$

Instead of using the RKHS in the preceding definition, one can actually consider an arbitrary feature space H_0 of k. Indeed, if $\Phi_0 : X \to H_0$ is a corresponding feature map, then the RKHS of k is given by (4.10) and hence k is universal if and only if for all $g \in C(X)$ and $\varepsilon > 0$ there exists a $w \in H_0$ such that $\|\langle w, \Phi_0(\,\cdot\,)\rangle - g\|_\infty \leq \varepsilon$. Although this is a rather trivial observation, we will see below that it is very useful for finding universal kernels.

One may wonder whether the preceding definition also makes sense for compact *topological* spaces. At first glance, this is indeed the case, but some further analysis shows that there exists no universal kernel if the topology is not generated by a metric (see Exercise 4.13).

Let us now discuss some of the surprising geometric properties of universal kernels. To this end, we need the following definition.

Definition 4.53. *Let k be a kernel on a metric space X with RKHS H. We say that k* ***separates the disjoint sets*** *$A, B \subset X$ if there exists an $f \in H$ with $f(x) > 0$ for all $x \in A$, and $f(x) < 0$ for all $x \in B$. Furthermore, we say that k* ***separates all finite (or compact) sets*** *if k separates all finite (or compact) disjoint sets $A, B \subset X$.*

It can be shown (see Exercise 4.11) that strictly positive definite kernels separate all finite sets. Furthermore, every kernel that separates all compact sets obviously also separates all finite sets, but in general the converse is not true (see Exercise 4.14). Moreover, every universal kernel separates all compact sets, as the following proposition shows.

Proposition 4.54. *Let X be a compact metric space and k be a universal kernel on X. Then k separates all compact sets.*

Proof. Let $A, B \subset X$ be disjoint compact subsets and d be the metric of X. Then, for all $x \in X$, we define

$$g(x) := \frac{\operatorname{dist}(x, B)}{\operatorname{dist}(x, A) + \operatorname{dist}(x, B)} - \frac{\operatorname{dist}(x, A)}{\operatorname{dist}(x, A) + \operatorname{dist}(x, B)},$$

where we used the distance function $\operatorname{dist}(x, C) := \inf_{x' \in C} \operatorname{dist}(x, x')$ for $x \in X$ and $C \subset X$. Since this distance function is continuous, we see that g is a continuous function. Furthermore, we have $g(x) = 1$ for all $x \in A$ and $g(x) = -1$ for all $x \in B$. Now, let H be the RKHS of k. Then there exists an $f \in H$ with $\|f - g\|_\infty \leq 1/2$, and by our previous considerations this f then satisfies $f(x) \geq 1/2$ for all $x \in A$ and $f(x) \leq 1/2$ for all $x \in B$. □

Although Proposition 4.54 easily follows from the notion of universality, it has surprising consequences for the geometric interpretation of the shape of the feature maps of universal kernels. Indeed, let k be a universal kernel on X with feature space H_0 and feature map $\Phi_0 : X \to H_0$. Furthermore, let us suppose that we have a finite subset $\{x_1, \ldots, x_n\}$ of X. Then Proposition 4.54 ensures that for *every* choice of signs $y_1, \ldots, y_n \in \{-1, 1\}$ we find a function f in the RKHS H of k with $y_i f(x_i) > 0$ for all $i = 1, \ldots, n$. By (4.10), this f can be represented by $f = \langle w, \Phi_0(\cdot) \rangle$ for a suitable $w \in H_0$. Consequently, the mapped training set $((\Phi_0(x_1), y_1), \ldots, (\Phi_0(x_n), y_n))$ can be correctly separated in H_0 by the hyperplane defined by w. Moreover, a closer look at the proof of Proposition 4.54 shows that this can even be done by a hyperplane that has almost the same distance to every point of $\Phi(x_i)$, $i = 1, \ldots, n$. Obviously, all these phenomena are impossible for general training sets in $\mathbb{R}^2$ or $\mathbb{R}^3$, and hence every two- or three-dimensional illustration of the feature space of universal kernels such as Figure 1.1 can be misleading. In particular, it seems to be very difficult to *geometrically* understand the learning mechanisms of both hard- and soft margin SVMs when these SVMs use universal kernels.

The geometric interpretation above raises the question of whether universal kernels can exist. As we will see below, the answer to this question is "yes" and in addition, many standard kernels, including the Gaussian RBF kernels, are universal. To establish these results, we need the following simple lemma.

Lemma 4.55 (Properties of universal kernels). *Let X be a compact metric space and k be a universal kernel on X. Then the following statements are true:*

i) Every feature map of k is injective.
ii) We have $k(x, x) > 0$ for all $x \in X$.
iii) Every restriction of k onto some compact $X' \subset X$ is universal.
iv) The ***normalized kernel*** *$k^* : X \times X \to \mathbb{R}$ defined by*

$$k^*(x, x') := \frac{k(x, x')}{\sqrt{k(x, x) k(x', x')}}, \qquad x, x' \in X,$$

is universal.

Proof. The first three assertions are direct consequences of Proposition 4.54 and the definition. To prove the fourth assertion, let $\Phi : X \to H$ be the canonical feature map of k into its RKHS H. Defining $\alpha(x) := k(x,x)^{-1/2}$ for all $x \in X$, we see that $\alpha\Phi : X \to H$ is a feature map of k^* and thus k^* is a kernel. To show that k^* is universal, we fix a function $g \in C(X)$ and an $\varepsilon > 0$. For $c := \|\alpha\|_\infty < \infty$, we then get an $f \in H$ with $\|f - \frac{g}{\alpha}\|_\infty \leq \frac{\varepsilon}{c}$. This yields

$$\left\| \left\langle f, \alpha(\,\cdot\,)\Phi(\,\cdot\,)\right\rangle - g \right\|_\infty \leq \|\alpha\|_\infty \left\| f - \frac{g}{\alpha} \right\|_\infty \leq \varepsilon\,,$$

and thus k^* is universal by the observation following Definition 4.52. □

Let us now investigate the existence of universal kernels. We begin by presenting a simple sufficient condition for the universality of kernels.

Theorem 4.56 (A test for universality). *Let X be a compact metric space and k be a continuous kernel on X with $k(x,x) > 0$ for all $x \in X$. Suppose that we have an injective feature map $\Phi : X \to \ell_2$ of k. We write $\Phi_n : X \to \mathbb{R}$ for its n-th component, i.e., $\Phi(x) = (\Phi_n(x))_{n\in\mathbb{N}}$, $x \in X$. If $\mathcal{A} := \operatorname{span}\{\Phi_n : n \in \mathbb{N}\}$ is an algebra, then k is universal.*

Proof. We will apply Stone-Weierstraß' theorem (see Theorem A.5.7). To this end, we first observe that the algebra $\mathcal{A}$ does not vanish since $\|(\Phi_n(x))\|_{\ell_2}^2 = k(x,x) > 0$ for all $x \in X$. Moreover, k is continuous and thus every $\Phi_n : X \to \mathbb{R}$ is continuous by Lemma 4.29. This shows that $\mathcal{A} \subset C(X)$. Moreover, the injectivity of Φ implies that $\mathcal{A}$ separates points, and thus Stone-Weierstraß' theorem shows that $\mathcal{A}$ is dense in $C(X)$. Now we fix a $g \in C(X)$ and an $\varepsilon > 0$. Then there exists a function $f \in \mathcal{A}$ of the form

$$f = \sum_{j=1}^{m} \alpha_j \Phi_{n_j}$$

with $\|f - g\|_\infty \leq \varepsilon$. For $n \in \mathbb{N}$, we define $w_n := \alpha_j$ if there is an index j with $n_j = n$ and $w_n := 0$ otherwise. This yields $w := (w_n) \in \ell_2$ and $f = \langle w, \Phi(\,\cdot\,)\rangle_{\ell_2}$, and thus k is universal by the observation following Definition 4.52. □

With the help of the preceding theorem, we are now in a position to give examples of universal kernels. Let us begin with kernels of Taylor type.

Corollary 4.57 (Universal Taylor kernels). *Fix an $r \in (0,\infty]$ and a C^∞-function $f : (-r,r) \to \mathbb{R}$ that can be expanded into its Taylor series at 0, i.e.,*

$$f(t) = \sum_{n=0}^{\infty} a_n t^n\,, \qquad t \in (-r,r).$$

Let $X := \{x \in \mathbb{R}^d : \|x\|_2 < \sqrt{r}\}$. If we have $a_n > 0$ for all $n \geq 0$, then k given by

$$k(x,x') := f(\langle x, x'\rangle)\,, \qquad x, x' \in X,$$

is a universal kernel on every compact subset of X.

Proof. We have already seen in Lemma 4.8 and its proof that k is a kernel with feature space $\ell_2(\mathbb{N}_0^d)$ and feature map $\Phi : X \to \ell_2(\mathbb{N}_0^d)$ defined by

$$\Phi(x) := \left(\sqrt{a_{j_1+\cdots+j_d} c_{j_1,\ldots,j_d}} \prod_{i=1}^{d} x_i^{j_i}\right)_{j_1,\ldots,j_d \geq 0}, \qquad x \in X.$$

Obviously, k is also continuous and $a_0 > 0$ implies $k(x,x) > 0$ for all $x \in X$. Furthermore, it is easy to see that Φ is injective. Finally, since polynomials form an algebra, $\operatorname{span}\{\Phi_{j_1,\ldots,j_d} : j_1, \ldots, j_d \geq 0\}$ is an algebra, and thus we obtain by Theorem 4.56 that k is universal. □

Recall that we presented some examples of Taylor kernels in Section 4.1. The following corollary shows that all these kernels are universal.

Corollary 4.58 (Examples of universal kernels). *Let X be a compact subset of $\mathbb{R}^d$, $\gamma > 0$, and $\alpha > 0$. Then the following kernels on X are universal:*

$$\begin{aligned} \textit{exponential kernel}: &\quad k(x,x') := \exp(\langle x, x' \rangle)\,, \\ \textit{Gaussian RBF kernel}: &\quad k_\gamma(x,x') := \exp(-\gamma^{-2}\|x - x'\|_2^2)\,, \\ \textit{binomial kernel}: &\quad k(x,x') := (1 - \langle x, x' \rangle)^{-\alpha}\,, \end{aligned}$$

where for the last kernel we additionally assume $X \subset \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$.

Proof. The assertion follows from Examples 4.9 and 4.11, Proposition 4.10, Corollary 4.57, and part *iv)* of Lemma 4.55. □

Note that a result similar to Corollary 4.57 can be established for Fourier type kernels (see Exercise 4.12 for details). Furthermore, it is obvious that polynomial kernels cannot be universal whenever $|X| = \infty$. By Proposition 5.41, it will thus be easy to show that there do exist learning problems that are extremely underfitted by these types of kernels.

We will see in Corollary 5.29 that the universality of a kernel with RKHS H guarantees

$$\inf_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}^*_{L,\mathrm{P}} \tag{4.55}$$

for all continuous P-integrable Nemitski losses. However, this result requires the input space X to be a compact metric space, and hence many interesting spaces, such as $\mathbb{R}^d$ and *infinite* discrete sets, are excluded. On the other hand, Theorem 5.31 will show that, for almost all interesting loss functions, it suffices to know that H is dense in $L_p(\mathrm{P}_X)$ for some $p \geq 1$ in order to establish (4.55). In the rest of this section, we will therefore investigate RKHSs that are dense in $L_p(\mathrm{P}_X)$. To this end, our main tool will be Theorem 4.26, which characterized this type of denseness by the injectivity of the associated integral operator $S_k : L_{p'}(\mathrm{P}_X) \to H$ defined by (4.17).

We begin by considering distributions P_X that are absolutely continuous with respect to a suitable reference measure μ.

Lemma 4.59. *Let X be a measurable space, μ be a measure on X, and k be a measurable kernel on X with RKHS H and $\|k\|_{L_p(\mu)} < \infty$ for some $p \in [1,\infty)$. Assume that the integral operator $S_k : L_{p'}(\mu) \to H$ is injective. Then H is dense in $L_q(h\mu)$ for all $q \in [1,p]$ and all measurable $h : X \to [0,\infty)$ with $h \in L_s(\mu)$, where $s := \frac{p}{p-q}$.*

Proof. Let us fix an $f \in L_{q'}(h\mu)$. Then we have $f|h|^{\frac{1}{q'}} \in L_{q'}(\mu)$ and, for r defined by $\frac{1}{q'} + \frac{1}{r} = \frac{1}{p'}$, Hölder's inequality and $\frac{r}{q} = s$ thus yield

$$\|fh\|_{L_{p'}(\mu)} \;=\; \big\| f|h|^{\frac{1}{q'}} |h|^{\frac{1}{q}} \big\|_{L_{p'}(\mu)} \;\le\; \big\| f|h|^{\frac{1}{q'}} \big\|_{L_{q'}(\mu)} \big\| |h|^{\frac{1}{q}} \big\|_{L_r(\mu)} \;<\; \infty\,.$$

Moreover, if $f \neq 0$ in $L_{q'}(h\mu)$, we have $fh \neq 0$ in $L_{p'}(\mu)$, and hence we obtain

$$0 \neq S_k(fh) = \int_X f(x)h(x)k(\,\cdot\,,x)\,d\mu(x) = \int_X f(x)k(\,\cdot\,,x)\,d(h\mu)(x)\,.$$

Since the latter integral describes the integral operator $L_{q'}(h\mu) \to H$, we then obtain the assertion by Theorem 4.26. □

Let us now investigate denseness properties of RKHSs over discrete spaces X. To this end, let us write $\ell_p(X) := L_p(\nu)$, where $p \in [1,\infty]$ and ν is the counting measure on X, which is defined by $\nu(\{x\}) = 1$, $x \in X$. Note that these spaces obviously satisfy the inclusion $\ell_p(X) \subset \ell_q(X)$ for $p \le q$, which is used in the proof of the following result.

Proposition 4.60 (Large RKHSs on discrete spaces I). *Let X be a countable set and k be a kernel on X with $\|k\|_{\ell_p(X)} < \infty$ for some $p \in [1,\infty)$. If k satisfies*

$$\sum_{x,x' \in X} k(x,x')f(x)f(x') > 0 \tag{4.56}$$

for all $f \in \ell_{p'}(X)$ with $f \neq 0$, then the RKHS of k is dense in $L_q(\mu)$ for all $q \in [1,\infty)$ and all distributions μ on X.

Proof. Recall that the counting measure ν is σ-finite since X is countable. Let us fix an $f \in \ell_{p'}(X)$ with $f \neq 0$. For the operator $S_k : \ell_{p'}(X) \to H$ defined by (4.17), we then have $S_k f \in H \subset \ell_p(X)$ and hence we obtain

$$\langle S_k f, f\rangle_{\ell_p(X),\ell_{p'}(X)} = \sum_{x,x' \in X} k(x,x')f(x)f(x') > 0\,.$$

This shows that $S_k : \ell_{p'}(X) \to H$ is injective. Now let μ be a distribution on X. Then there exists a function $h \in \ell_1(X)$ with $\mu = h\nu$. Since for $q \in [1,p]$ we have $s := \frac{p}{p-q} \ge 1$, we then find $h \in \ell_s(X)$ and hence we obtain the assertion by applying Lemma 4.59. In addition, for $q > p$, we have $\|k\|_{\ell_q(X)} \le \|k\|_{\ell_p(X)} < \infty$ and $\ell_{q'}(X) \subset \ell_{p'}(X)$, and consequently this case follows from the case $q = p$ already shown . □

Note that the case $p = \infty$ is excluded in Proposition 4.60. The reason for this is that the dual of $\ell_\infty(X)$ is *not* $\ell_1(X)$. However, if instead we consider the *pre*-dual of $\ell_1(X)$, namely the Banach **space of functions vanishing at infinity**,

$$c_0(X) := \left\{ f : X \to \mathbb{R} \,\middle|\, \forall \varepsilon > 0 \, \exists \text{ finite } A \subset X \; \forall x \in X \setminus A : |f(x)| \leq \varepsilon \right\},$$

which is equipped with the usual $\|\cdot\|_\infty$-norm, we obtain the following result.

Theorem 4.61 (Large RKHSs on discrete spaces II). *Let X be a countable set and k be a bounded kernel on X that satisfies both $k(\cdot, x) \in c_0(X)$ for all $x \in X$ and (4.56) for all $f \in \ell_1(X)$ with $f \neq 0$. Then the RKHS of k is dense in $c_0(X)$.*

Proof. Since $k(\cdot, x) \in c_0(X)$ for all $x \in X$, we see $H_{\text{pre}} \subset c_0(X)$, where H_{pre} is the space defined in (4.12). Let us write H for the RKHS of k. Since k is bounded, the inclusion $I : H \to \ell_\infty(X)$ is well-defined and continuous by Lemma 4.23. Now let us fix an $f \in H$. By Theorem 4.21, there then exists a sequence $(f_n) \subset H_{\text{pre}}$ with $\lim_{n\to\infty} \|f - f_n\|_H = 0$, and the continuity of $I : H \to \ell_\infty(X)$ then yields $\lim_{n\to\infty} \|f - f_n\|_\infty = 0$. Now the completeness of $c_0(X)$ shows that $c_0(X)$ is a closed subspace of $\ell_\infty(X)$, and since we already know $f_n \in c_0(X)$ for all $n \geq 1$, we can conclude that $f \in c_0(X)$. In other words, the inclusion $I : H \to c_0(X)$ is well-defined and continuous. Moreover, a simple calculation analogous to the one in the proof of Theorem 4.26 shows that its adjoint operator is the integral operator $S_k : \ell_1(X) \to H$. Since this operator is injective by (4.56), we see that H is dense in $c_0(X)$ by Theorem 4.26. □

One may be tempted to assume that condition (4.56) is already satisfied if it holds for all functions $f : X \to \mathbb{R}$ with $0 < |\{x \in X : f(x) \neq 0\}| < \infty$, i.e., for strictly positive definite kernels. The following result shows that this is not the case in a strong sense.

Theorem 4.62. *There exists a bounded, strictly positive definite kernel k on $X := \mathbb{N}_0$ with $k(\cdot, x) \in c_0(X)$ for all $x \in X$ such that for all finite measures μ on X with $\mu(\{x\}) > 0$, $x \in X$, and all $q \in [1, \infty]$, the RKHS H of k is not dense in $L_q(\mu)$.*

Proof. Let us write $p_n := \mu(\{n\})$, $n \in \mathbb{N}_0$. Moreover, let $(b_i)_{i\geq 1} \subset (0, 1)$ be a strictly positive sequence with $\|(b_i)\|_2 = 1$ and $(b_i) \in \ell_1$. Furthermore, let (e_n) be the canonical ONB of ℓ_2. We write $\Phi(0) := (b_i)$ and $\Phi(n) := e_n$, $n \geq 1$. Then we have $\Phi(n) \in \ell_2$ for all $n \in \mathbb{N}_0$, and hence

$$k(n, m) := \big\langle \Phi(n), \Phi(m) \big\rangle_{\ell_2}, \qquad n, m \geq 0,$$

defines a kernel. Moreover, an easy calculation shows $k(0, 0) = 1$, $k(n, m) = \delta_{n,m}$, and $k(n, 0) = b_n$ for $n, m \geq 1$. Since $b_n \to 0$, we hence find $k(\cdot, n) \in$

$c_0(X)$ for all $n \in \mathbb{N}_0$. Now let $n \in \mathbb{N}_0$ and $\alpha := (\alpha_0, \ldots, \alpha_n) \in \mathbb{R}^{n+1}$ be a vector with $\alpha \neq 0$. Then the definition of k yields

$$\begin{aligned} A := \sum_{i=0}^{n}\sum_{j=0}^{n} \alpha_i \alpha_j k(i,j) &= \alpha_0^2 k(0,0) + 2\sum_{i=1}^{n} \alpha_i \alpha_0 k(i,0) + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(i,j) \\ &= \alpha_0^2 + 2\alpha_0 \sum_{i=1}^{n} \alpha_i b_i + \sum_{i=1}^{n} \alpha_i^2 \\ &= \alpha_0^2 + \sum_{i=1}^{n} \alpha_i (2\alpha_0 b_i + \alpha_i)\,. \end{aligned}$$

If $\alpha_0 = 0$, we hence find $A = \sum_{i=1}^{n} \alpha_i^2 > 0$ since we assumed $\alpha \neq 0$. Moreover, if $\alpha_0 \neq 0$, we find $t(2\alpha_0 b_i + t) \geq -\alpha_0^2 b_i^2$ for all $t \in \mathbb{R}$ by simple calculus, and hence our assumptions $\|(b_i)\|_2 = 1$ and $b_i > 0$, $i \geq 1$, imply

$$A \;\geq\; \alpha_0^2 - \sum_{i=1}^{n} \alpha_0^2 b_i^2 \;=\; \alpha_0^2 \sum_{i=n+1}^{\infty} b_i^2 \;>\; 0\,.$$

Consequently, we have $A > 0$ in any case, and from this it is easy to see that k is strictly positive definite. Let us now define $f : \mathbb{N}_0 \to \mathbb{R}$ by $f(0) := 1$ and $f(n) := -\frac{b_n}{p_n} p_0$ for $n \geq 1$. Then we have $\|f\|_{L_1(\mu)} = p_0 + p_0 \|(b_n)\|_{\ell_1} < \infty$, and a simple calculation yields

$$S_k f(0) = k(0,0) f(0) p_0 + \sum_{n=1}^{\infty} k(0,n) f(n) p_n = p_0 - p_0 \sum_{n=1}^{\infty} b_n^2 = 0\,.$$

Furthermore, for $m \geq 1$, our construction yields

$$S_k f(m) = k(m,0) f(0) p_0 + \sum_{n=1}^{\infty} k(m,n) f(n) p_n = b_m f(0) p_0 - f(m) p_m = 0\,,$$

and hence we have $S_k f = 0$, i.e., $S_k : L_1(\mu) \to H$ is not injective. Moreover, by (A.34), the space $L_1(\mu)$ can be interpreted as a subspace of $L'_\infty(\mu)$, and we have $S''_k f = S_k f$ for all $f \in L_1(\mu)$ as we mention in (A.20). From this we conclude that $S''_k : L'_\infty(\mu) \to H$ is not injective, and hence $S'_k : H \to L_\infty(\mu)$ does not have a dense image. Repeating the proof of Theorem 4.26, we further see that $\mathrm{id} : H \to L_\infty(\mu)$ equals S'_k, and thus H is not dense in $L_\infty(\mu)$. From this we easily find the assertion for $q \in [1, \infty)$. □

Finally, let us treat the Gaussian RBF kernels yet another time.

Theorem 4.63 (Gaussian RKHS is large). *Let $\gamma > 0$, $p \in [1, \infty)$, and μ be a finite measure on $\mathbb{R}^d$. Then the RKHS $H_\gamma(\mathbb{R}^d)$ of the Gaussian RBF kernel k_γ is dense in $L_p(\mu)$.*

Proof. Since $L_p(\mu)$ is dense in $L_1(\mu)$, it suffices to consider the case $p > 1$. Moreover, by Theorem 4.26, it suffices to show that the integral operator $S_{k_\gamma} : L_{p'}(\mu) \to H_\gamma(\mathbb{R}^d)$ of k_γ is injective. However, the latter was already established in Theorem 4.47. □

4.7 Further Reading and Advanced Topics

The idea of using kernels for pattern recognition algorithms dates back to the 1960s, when Aizerman *et al.* (1964) gave a feature space interpretation of the potential function method. However, it took almost thirty years before Boser *et al.* (1992) combined this idea with another old idea, namely the generalized portrait algorithm of Vapnik and Lerner (1963), in the hard margin SVM. Shortly thereafter, Cortes and Vapnik (1995) added slack variables to this first type of SVM, which led to soft margin SVMs. In these papers on SVMs, the feature space interpretation was based on an informal version of Mercer's theorem, which may cause some misunderstandings, as discussed in Exercise 4.10. The RKHS interpretation for SVMs was first found in 1996 and then spread rapidly; see, e.g., the books by Schölkopf (1997) and Vapnik (1998). For more information, we refer to G. Wahba's talk on multi-class SVMs given at IPAM in 2005 (see `http://www.oid.ucla.edu/Webcast/ipam/`). Since the introduction of SVMs, many kernels for specific learning tasks have been developed; for an overview, we refer to Schölkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004). In addition, it was first observed by Schölkopf *et al.* (1998) that the "kernel trick", i.e., the idea of combining a linear algorithm with a kernel to obtain a non-linear algorithm, works not only for SVMs but actually for a variety of different algorithms. Many of these "kernelized" algorithms can be found in the books by Schölkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004).

As indicated above, the use of kernels for machine learning methods was discovered relatively recently. However, the theory of kernels and their applications to various areas of mathematics are much older. Indeed, Mercer's theorem has been known for almost a century (see Mercer, 1909), and based on older work by Moore (1935, 1939) and others, Aronszajn (1950) developed the theory of RKHSs in the 1940s. The latter article also provides a good overview of the early history and the first applications of kernels. Since then, many new applications have been discovered. We refer to the books by Berlinet and Thomas-Agnan (2004), Ritter (2000), and Wahba (1990) for a variety of examples.

We must admit that two important types of kernels have been almost completely ignored in this chapter. The first of these are the **translation-invariant kernels**, i.e., kernels $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{K}$ for which there exists a function $\kappa : \mathbb{R}^d \to \mathbb{K}$ such that

$$k(x, x') = \kappa(x - x')\,, \qquad x, x' \in \mathbb{R}^d. \tag{4.57}$$

Bochner (1932, 1959) showed that, given a continuous function $\kappa : \mathbb{R}^d \to \mathbb{C}$, equation (4.57) defines a kernel k if and only if there exists a unique finite Borel measure μ on $\mathbb{R}^d$ such that

$$\kappa(x) = \int_{\mathbb{R}^d} e^{i\langle x,y\rangle}\, d\mu(y)\,, \qquad x \in \mathbb{R}^d. \tag{4.58}$$

From this and Exercise 4.5, it is easy to conclude that for continuous functions $\kappa : \mathbb{R}^d \to \mathbb{R}$, equation (4.57) defines a kernel if there exists a unique finite Borel measure μ on $\mathbb{R}^d$ such that

$$\kappa(x) = \int_{\mathbb{R}^d} \cos\langle x, y\rangle \, d\mu(y) \,, \qquad x \in \mathbb{R}^d. \tag{4.59}$$

Note that this sufficient condition is a generalization of the Fourier kernels introduced in Lemma 4.12, and in fact one could prove this condition directly along the lines of the proof of Lemma 4.12. Finally, Cucker and Zhou (2007) showed in their Proposition 2.14 that k is a kernel if the Fourier transform of κ is non-negative. The second type of kernel we did not systematically consider are **radial kernels**, i.e., kernels $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for which there exists a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ such that

$$k(x, x') = \kappa(\|x - x'\|_2^2) \,, \qquad x, x' \in \mathbb{R}^d. \tag{4.60}$$

Schoenberg (1938), see also Section 5.2 in Berg *et al.* (1984), showed that, given a continuous function $\kappa : \mathbb{R} \to \mathbb{R}$, equation (4.60) defines a kernel k for *all* $d \geq 1$, if and only if there exists a unique finite Borel measure μ on $[0, \infty)$ such that

$$\kappa(t) = \int_{\mathbb{R}^d} e^{-ty} \, d\mu(y) \,, \qquad t \in [0, \infty). \tag{4.61}$$

Finally, it is known that if κ is completely monotonic, then (4.60) defines a kernel. For a proof, we refer to Proposition 2.18 of Cucker and Zhou (2007).

Most of the material presented in Sections 4.1, 4.2, and 4.3 is folklore and can be found in many other introductions to RKHSs (see, e.g., Hille, 1972; Meschkowski, 1962; Saitoh, 1988, 1997). Polynomial kernels were first used in the machine literature by Poggio (1975). The exponential kernel and its RKHS are closely related to the so-called *Fock space* considered in quantum mechanics (see, e.g., Bargmann, 1961; Folland, 1989). Furthermore, the binomial kernel is a generalization of the Bergmann kernel (see, e.g., Duren, 1970; Duren and Schuster, 2004; Hedenmalm *et al.*, 2000), and the examples of Fourier type kernels were considered by Vapnik (1998), who also presents some more examples of kernels of possible interest for machine learning. Finally, the notion of separately continuous kernels in Section 4.3 is taken from Hein and Bousquet (2004).

The description of $H_\gamma(X)$ follows Steinwart *et al.* (2006a), but some of the results can also be found in the book by Saitoh (1997). The operator W_t is known as the *Gauss-Weierstraß integral operator* and is used for the heat equation (see, e.g., Hille and Phillips, 1957). Since this integral operator is neither surjective nor compact, Theorem 4.47 can be used to show that the inclusion $\mathrm{id} : H_{\gamma_2}(X) \to H_{\gamma_1}(X)$ considered in Proposition 4.46 is neither surjective nor compact if X has a non-empty interior. In addition, the bound on its norm given in (4.44) turns out to be sharp for such X. We refer to Steinwart *et al.* (2006a) for more information.

The RKHS representation based on Mercer's theorem closely follows the presentation of Cucker and Smale (2002). This article also provides some other useful insights into the theory of RKHSs. For a proof of Mercer's theorem, we refer to Werner (1995) and Riesz and Nagy (1990).

The first part of Section 4.6 is taken almost completely from Steinwart (2001). It is not hard to see that Corollary 4.57 does *not* provide a necessary condition for universality. Indeed, if, for example, one only assumes $a_n > 0$ for all indexes n but one $n_0 \neq 0$, then k is still a universal kernel. This raises the question of how many non-vanishing coefficients are necessary for the universality. Surprisingly, this question was answered by Dahmen and Micchelli (1987) in a different context. Their result states that k is universal if and only if $a_0 > 0$ and

$$\sum_{a_{2n}>0} \frac{1}{2n} = \sum_{a_{2n+1}>0} \frac{1}{2n+1} = \infty \,.$$

Note that this condition implies that the sets $N_{\text{even}} := \{2n \in \mathbb{N} : a_{2n} > 0\}$ and $N_{\text{odd}} := \{2n+1 \in \mathbb{N} : a_{2n+1} > 0\}$ are infinite. Interestingly, Pinkus (2004) has recently shown that the latter characterize strictly positive definite kernels, i.e., he has shown that a kernel is strictly positive definite if and only if $a_0 > 0$ and $|N_{\text{odd}}| = |N_{\text{even}}| = \infty$. In particular, both results together show that not every strictly positive definite kernel is universal. An elementary proof of this latter observation can be found by combining Exercise 4.11 and Exercise 4.14. Moreover, it is interesting to note that this observation can also be deduced from Theorem 4.62. Recently, Micchelli *et al.* (2006) investigated under which conditions translation-invariant kernels and radial kernels are universal. Besides other results, they showed that *complex* translation-invariant kernels are universal if the support of the measure μ in (4.58) has a strictly positive Lebesgue measure. Using a feature map similar to that of the proof of Lemma 4.12, it is then easy to conclude that kernels represented by (4.59) are universal if $\operatorname{vol}(\operatorname{supp}\mu) > 0$. Moreover, Micchelli *et al.* (2006) showed that radial kernels are universal if the measure μ in (4.61) satisfies $\operatorname{supp}\mu \neq \{0\}$. Finally, the second part of Section 4.6, describing denseness results of H in $L_p(\mu)$, is taken from Steinwart *et al.* (2006b).

4.8 Summary

In this chapter, we gave an introduction to the mathematical theory of kernels. We first defined kernels via the existence of a feature map, but it then turned out that kernels can also be characterized by simple inequalities, namely the positive definiteness condition. Furthermore, we saw that certain representations of kernel functions lead directly to feature maps. This observation helped us to introduce several important kernels.

Although neither the feature map nor the feature space are uniquely determined for a given kernel, we saw in Section 4.2 that we can always construct

a canonical feature space consisting of functions. We called this feature space the reproducing kernel Hilbert space. One of our major results was that there is a one-to-one relation between kernels and RKHSs. Moreover, we showed in Section 4.3 that many properties of kernels such as measurability, continuity, or differentiability are inherited by the functions in the RKHS.

We then determined the RKHSs of Gaussian RBF kernels and gained some insight into their structure. In particular, we were able to compare the RKHS norms for different widths and showed that these RKHSs do not contain constant functions. We further investigated properties of their associated integral operators, showing, e.g., that in many cases these operators are injective.

For continuous kernels on compact input spaces, Mercer's theorem provided a series representation in terms of the eigenvalues and functions of the associated integral operators. This series representation was then used in Section 4.5 to give another characterization of the functions contained in the corresponding RKHSs.

In Section 4.6, we then considered kernels whose RKHS H is large in the sense that H is dense in either $C(X)$ or a certain Lebesgue space of p-integrable functions. In particular, we showed that, among others, the Gaussian RBF kernels belong to this class. As we will see in later chapters, this denseness is one of the key reasons for the universal learning ability of SVMs.

4.9 Exercises

4.1. Some more kernels of Taylor type ($\star$)
Use Taylor expansions to show that the following functions can be used to construct kernels by Lemma 4.8: $x \mapsto \cosh x$, $x \mapsto \operatorname{arcoth} x^{-1}$, $x \mapsto \ln\left(\frac{1+x}{1-x}\right)$, and $x \mapsto \operatorname{arctanh} x$. What are the corresponding (maximal) domains of these kernels? Are these kernels universal?

4.2. Many standard Hilbert spaces are not RKHSs ($\star$)
Let μ be a measure on the non-empty set X. Show that $L_2(\mu)$ is an RKHS if and only if for all non-empty $A \subset X$ we have $\mu(A) > 0$.

4.3. Cauchy-Schwarz inequality ($\star\star$)
Let E be an $\mathbb{R}$-vector space and $\langle \cdot, \cdot \rangle : E \to \mathbb{R}$ be a positive, symmetric bilinear form, i.e., it satisfies

i) $\langle x, x \rangle \geq 0$
ii) $\langle x, y \rangle = \langle y, x \rangle$
iii) $\langle \alpha x + y, z \rangle = \alpha \langle x, z \rangle + \langle y, z \rangle$

for all $x, y, z \in E$, $\alpha \in \mathbb{R}$. Show the Cauchy-Schwarz inequality

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle \,, \qquad x, y \in E.$$

Hint: Start with $0 \leq \langle x + \alpha y, x + \alpha y \rangle$ and consider the cases $\alpha = 1$ and $\alpha = -1$ if $\langle x, x \rangle = \langle y, y \rangle = 0$. Otherwise, if, e.g., $\langle y, y \rangle \neq 0$, use $\alpha := -\frac{\langle x, y \rangle}{\langle y, y \rangle}$.

4.4. The RKHSs of restricted and normalized kernels ($\star\star\star$)
Let k be a kernel on X with RKHS H. Using Theorem 4.21, show that:

i) For $X' \subset X$, the RKHS of the restricted kernel $k_{|X' \times X'}$ is

$$H_{|X'} := \bigl\{ f : X' \to \mathbb{R} \,|\, \exists \hat{f} \in H \text{ with } \hat{f}_{|X'} = f \bigr\}$$

with norm $\|f\|_{H_{|X'}} := \inf\{\|\hat{f}\|_H : \hat{f} \in H \text{ with } \hat{f}_{|X'} = f\}$.

ii) Suppose $k(x,x) > 0$ for all $x \in X$. Then the RKHS H^* of the normalized kernel k^* considered in Lemma 4.55 is

$$H^* = \bigl\{ f : X \to \mathbb{R} \,|\, (x \mapsto k(x,x) f(x)) \in H \bigr\}$$

and has norm $\|f\|_{H^*} := \|(x \mapsto k(x,x) f(x))\|_H$.

iii) Determine the RKHS of the exponential kernel with the help of $H_{\gamma,\mathbb{C}^d}$.

4.5. Real part of complex kernels ($\star\star$)
Let $k : X \times X \to \mathbb{C}$ be a kernel. Show that $\operatorname{Re} k : X \times X \to \mathbb{R}$ is a kernel.

Hint: Show that $\operatorname{Re} k$ is symmetric and positive definite. For the latter, use $k(x,x') + k(x',x) = 2\operatorname{Re} k(x,x')$.

4.6. Injectivity of $\mathrm{id} : H \to L_p(\mu)$ ($\star\star$)
Let X be a Polish space and μ be a Borel measure with $\operatorname{supp}\mu = X$. Moreover, let k be a continuous kernel on X with $\|k\|_{L_p(\mu)} < \infty$ for some $p \in [1,\infty]$. Show that $\mathrm{id} : H \to L_p(\mu)$ is injective.

4.7. Properties of functions contained in the Gaussian RKHSs ($\star\star$)
For $\gamma > 0$, show the following statements:

i) Every $f \in H_\gamma(\mathbb{R}^d)$ is infinitely many times differentiable.

ii) Every $f \in H_\gamma(\mathbb{R}^d)$ is 2-integrable, and the inclusion $\mathrm{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mathbb{R}^d)$ is continuous.

iii) Every $f \in H_\gamma(\mathbb{R}^d)$ is bounded, and the inclusion $\mathrm{id} : H_\gamma(\mathbb{R}^d) \to \ell_\infty(\mathbb{R}^d)$ is continuous.

Hint: For *ii)*, use that the integral operator $S_k : L_2(\mathbb{R}^d) \to H_\gamma(\mathbb{R}^d)$ is continuous. Then consider its adjoint.

4.8. Gaussian kernels and the hinge loss ($\star\star\star$)
Let P be a distribution on $X \times Y$, where $X \subset \mathbb{R}^d$ and $Y := \{-1,1\}$. Furthermore, let L_{hinge} be the hinge loss defined in Example 2.27 and $H_\gamma(X)$ be a Gaussian RKHS. Show that no minimizer $f^*_{L_{\text{hinge}},\mathrm{P}}$ of the L_{hinge}-risk is contained in $H_\gamma(X)$ if for $\eta(x) := \mathrm{P}(y = 1|x)$, $x \in X$, the set $\{x : \eta(x) \neq 0, 1/2, 1\}$ has a non-empty interior. Give some (geometric) examples for such distributions. Does a similar observation hold for P satisfying $\mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} = 0$?

4.9. Different feature spaces of the Gaussian kernels ($\star\star$)
Compare the different feature spaces and maps of the Gaussian RBF kernels we presented in Corollary 4.40 and Lemma 4.45.

4.10. Discussion of Mercer's theorem ($\star\star\star$)
Using inadequate versions of Mercer's theorem can lead to mistakes. Consider the following two examples:

i) Sometimes a version of Mercer's theorem is presented that holds not only for continuous kernels but also for bounded and measurable kernels. For these kernels, the relation (4.53) is only stated μ^2-almost surely. Now, one might think that by modifying the eigenfunctions on a zero set one can actually obtain (4.53) for *all* $x, x' \in X$. Show that in general such a modification does not exist.

ii) Show that if the assumption $\operatorname{supp}\mu = X$ of Theorem 4.49 is dropped, (4.53) holds at least for all $x, x' \in \operatorname{supp}\mu$. Furthermore, give an example that demonstrates that in general (4.53) does not hold for all $x, x' \in X$.

Hint: For *for i)* Use $[0,1]$ equipped with the Lebesgue measure and consider the kernel k defined by $k(x,x) := 1$ for $x \in X$ and $k(x,x') = 0$ otherwise.

4.11. Strictly positive definite kernels separate all finite subsets ($\star\star$)
Let $k : X \times X \to \mathbb{R}$ be a kernel. Show that k separates all finite subsets if and only if it is strictly positive definite.

Hint: Recall from linear algebra that a symmetric matrix is (strictly) positive definite if and only if its eigenvalues are all real and (strictly) positive. Then express the equations $f(x_i) = y_i$, $i = 1, \ldots, n$, $f \in H$, in terms of the Gram matrix $(k(x_j, x_i))_{i,j}$.

4.12. Universality of Fourier type kernels ($\star\star\star$)
Formulate and prove a condition for Fourier type kernels (see Lemma 4.12) that ensures universality. Then show that the kernels in Examples 4.13 and 4.14 are universal.

Hint: Use a condition similar to that of Corollary 4.57.

4.13. Existence of universal kernels ($\star\star\star\star$)
Let (X, τ) be a compact topological space. Show that the following statements are equivalent:

i) (X, τ) is metrizable, i.e., there exists a metric d on X such that the collection of the open subsets defined by d equals the topology τ.

ii) There exists a continuous kernel on X whose RKHS is dense in $C(X)$.

Hint: Use that X is metrizable if and only if $C(X)$ is separable (see, e.g., Theorem V.6.6 of Conway, 1990). Furthermore, for *i)* $\Rightarrow$ *ii)*, use a countable, dense subset of $C(X)$ to construct a universal kernel in the spirit of Lemma 4.2. For the other direction, use that every compact topological space is separable.

4.14. A kernel separating all finite but not all compact sets ($\star\star\star\star$)
Let $X := \{-1, 0\} \cup \{1/n : n \in \mathbb{N}\}$ and (e_n) be the canonical ONB of ℓ_2. Define the map $\Phi : X \to \ell_2 \oplus_2 \mathbb{R}$ by $\Phi(-1) := (\sum_{n=1}^{\infty} 2^{-n} e_n, 1)$, $\Phi(0) := (0, 1)$, and $\Phi(1/n) := (n^{-2} e_n, 1)$ for $n \in \mathbb{N}$. Then the kernel associated to the feature map Φ separates all finite sets but does not separate the compact sets $\{-1\}$ and $X \setminus \{-1\}$.

5

Infinite-Sample Versions of Support Vector Machines

Overview. *In this chapter, we show that interesting structural properties of SVMs can be discovered by considering the SVM formulation for "infinite" training sets. In particular, we will present a representation for the corresponding SVM solutions and discuss their dependence on the underlying probability measure. Moreover, we will investigate the behavior of SVMs for vanishing regularization parameter.*

Prerequisites. *Chapters 2 and 4 form the fundament of this chapter. In addition, we need the subdifferential calculus for convex functions, which is provided in Section A.6.2.*

Usage. *Section 5.1 is needed for computational aspects discussed in Chapter 11, while Section 5.2 is mainly used in Section 5.3. The latter section is required for the statistical analysis of SVMs conducted in Sections 6.4, 6.5, and 9.2. Finally, Sections 5.4 and 5.5 are important for the generalization performance of SVMs analyzed in Sections 6.4 and 6.5 and Chapters 8 and 9.*

We saw in the introduction that support vector machines obtain their decision functions by finding a minimizer $f_{\mathrm{D},\lambda}$ of the regularized empirical risk

$$\mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f) := \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)\,, \qquad f \in H\,. \tag{5.1}$$

The first questions that then arise are whether such minimizers exist and, if so, whether they are unique. Moreover, the data set D defining the empirical measure D is usually an i.i.d. sample drawn from a distribution P. The law of large numbers then shows that $\mathcal{R}_{L,\mathrm{D}}(f)$ is close to $\mathcal{R}_{L,\mathrm{P}}(f)$, and hence one may think of $\mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f)$ as an estimate of the *infinite-sample* regularized risk

$$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) := \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f)\,. \tag{5.2}$$

In this chapter, we will thus investigate such regularized risks for *arbitrary* distributions, so that we simultaneously obtain results for (5.1) and (5.2). In particular, we will consider the following questions:

- When does (5.2) have exactly one minimizer $f_{\mathrm{P},\lambda} \in H$?
- Is there a way to represent $f_{\mathrm{P},\lambda}$ in a suitable form?
- How does $f_{\mathrm{P},\lambda}$ change if P or λ changes?
- How close is $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$?

Namely, we show in Section 5.1 that under rather general conditions there exists a unique minimizer $f_{\mathrm{P},\lambda}$, and in the following section we derive a

representation for this minimizer. This representation is then used in Section 5.3 to investigate the dependence of $f_{\mathrm{P},\lambda}$ on P. In Sections 5.4 and 5.5, we finally consider the question of whether and how $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ converges to the Bayes risk for $\lambda \to 0$.

5.1 Existence and Uniqueness of SVM Solutions

In this section, we first investigate under which conditions the general regularized risk (5.2) possesses a unique minimizer. Furthermore, we establish a representation of the finite sample solutions $f_{\mathrm{D},\lambda}$ of (5.1).

Let us begin by introducing some notions. To this end, let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, H be the RKHS of a measurable kernel on X, and P be a distribution on $X \times Y$. For $\lambda > 0$, we then call an $f_{\mathrm{P},\lambda,H} \in H$ that satisfies

$$\lambda \|f_{\mathrm{P},\lambda,H}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda,H}) = \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f) \tag{5.3}$$

a **general SVM solution** or a **general SVM decision function**. Moreover, in order to avoid notational overload, we usually use the shorthand $f_{\mathrm{P},\lambda} := f_{\mathrm{P},\lambda,H}$ if no confusion can arise. Now note that for such a function $f_{\mathrm{P},\lambda}$ we have

$$\lambda \|f_{\mathrm{P},\lambda}\|_H^2 \le \lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) = \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f) \le \mathcal{R}_{L,\mathrm{P}}(0)\,,$$

or in other words

$$\|f_{\mathrm{P},\lambda}\|_H \le \sqrt{\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda}}\,. \tag{5.4}$$

Let us now investigate under which assumptions there exists exactly one $f_{\mathrm{P},\lambda}$. We begin with the following result showing uniqueness for convex losses.

Lemma 5.1 (Uniqueness of SVM solutions). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss, H be the RKHS of a measurable kernel over X, and P be a distribution on $X \times Y$ with $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for some $f \in H$. Then for all $\lambda > 0$ there exists at most one general SVM solution $f_{\mathrm{P},\lambda}$.*

Proof. Let us assume that the map $f \mapsto \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f)$ has two minimizers $f_1, f_2 \in H$ with $f_1 \neq f_2$. By the last statement in Lemma A.5.9, we then find $\|\frac{1}{2}(f_1+f_2)\|_H^2 < \frac{1}{2}\|f_1\|_H^2 + \frac{1}{2}\|f_2\|_H^2$. The convexity of $f \mapsto \mathcal{R}_{L,\mathrm{P}}(f)$ together with $\lambda\|f_1\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_1) = \lambda\|f_2\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_2)$ then shows for $f^* := \frac{1}{2}(f_1+f_2)$ that

$$\lambda\|f^*\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f^*) < \lambda\|f_1\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_1)\,,$$

i.e., f_1 is not a minimizer of $f \mapsto \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f)$. Consequently, the assumption that there are two minimizers is false. □

Our next result shows that for convex, integrable Nemitski loss functions (see Definition 2.16) there always exists a general SVM solution.

Theorem 5.2 (Existence of SVM solutions). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss and P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss. Furthermore, let H be the RKHS of a bounded measurable kernel over X. Then, for all $\lambda > 0$, there exists a general SVM solution $f_{\mathrm{P},\lambda}$.*

Proof. Since the kernel k of H is measurable, H consists of measurable functions by Lemma 4.24. Moreover, k is bounded, and thus Lemma 4.23 shows that $\mathrm{id} : H \to L_\infty(\mathrm{P}_X)$ is continuous. In addition, we have $L(x,y,t) < \infty$ for all $(x,y,t) \in X \times Y \times \mathbb{R}$, and hence L is a continuous loss by the convexity of L and Lemma A.6.2. Therefore, Lemma 2.17 shows that $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to \mathbb{R}$ is continuous, and hence $\mathcal{R}_{L,\mathrm{P}} : H \to \mathbb{R}$ is continuous. In addition, Lemma 2.13 provides the convexity of this map. Furthermore, $f \mapsto \lambda\|f\|_H^2$ is also convex and continuous, and hence so is $f \mapsto \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f)$. Now consider the set

$$A := \left\{ f \in H : \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f) \le M \right\},$$

where $M := \mathcal{R}_{L,\mathrm{P}}(0)$. Then we obviously have $0 \in A$. In addition, $f \in A$ implies $\lambda\|f\|_H^2 \le M$, and hence $A \subset (M/\lambda)^{1/2} B_H$, where B_H is the closed unit ball of H. In other words, A is a non-empty and bounded subset and thus Theorem A.6.9 gives the existence of a minimizer $f_{\mathrm{P},\lambda}$. □

It is interesting to note that in Theorem 5.2 the convexity of L is *not necessary* for the existence of a general SVM solution (see Exercise 5.7).

We have presented some important classes of Nemitski losses in Chapter 2. The following corollaries specify Theorem 5.2 for these types of losses.

Corollary 5.3. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex, locally Lipschitz continuous loss, P be a distribution on $X \times Y$ with $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, and H be the RKHS of a bounded measurable kernel over X. Then, for all $\lambda > 0$, there exists a unique general SVM solution $f_{\mathrm{P},\lambda} \in H$. Furthermore, if L is actually a convex margin-based loss represented by $\varphi : \mathbb{R} \to [0,\infty)$, we have*

$$\|f_{\mathrm{P},\lambda}\|_H \;\le\; \left(\frac{\varphi(0)}{\lambda}\right)^{1/2}.$$

Proof. The first assertion follows from Lemma 5.1, Theorem 5.2, and the discussion around (2.11). Moreover, convex margin-based losses are locally Lipschitz continuous by Lemma 2.25, and (5.4) together with $\mathcal{R}_{L,\mathrm{P}}(0) = \varphi(0)$ shows the inequality for these losses. □

Corollary 5.4. *Let $L : \mathbb{R} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss of upper growth type $p \ge 1$, P be a distribution on $X \times Y$ with $|\mathrm{P}|_p < \infty$, and H be the RKHS of a bounded measurable kernel over X. Then, for all $\lambda > 0$, there exists a unique general SVM solution $f_{\mathrm{P},\lambda} \in H$. Moreover, there exists a constant $c_{L,p} > 0$ only depending on L and p such that*

$$\|f_{\mathrm{P},\lambda}\|_H \;\le\; c_{L,p}\left(\frac{|\mathrm{P}|_p^p + 1}{\lambda}\right)^{1/2}.$$

Proof. Combine Lemma 2.38, Lemma 5.1, Theorem 5.2, and (5.4). □

If H is the RKHS of a bounded measurable kernel on X and L is a convex, distance-based loss of growth type $p \geq 1$, then it is easy to see by Lemma 2.38 that, for distributions P on $X \times \mathbb{R}$ with $|\mathrm{P}|_p = \infty$, we have $\mathcal{R}_{L,\mathrm{P}}(f) = \infty$ for *all* $f \in H$. Consequently, the distributions P with $|\mathrm{P}|_p < \infty$ are in general the only distributions admitting a *non-trivial* SVM solution $f_{\mathrm{P},\lambda}$.

Let us now discuss **empirical SVM solutions** in some more detail. To this end, we denote, as usual, the empirical measure of a sequence of observations $D := ((x_1, y_1), \ldots, (x_n, y_n))$ by D, i.e., $\mathrm{D} := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$. The next result shows that $f_{\mathrm{D},\lambda}$ exists under somewhat minimal assumptions on L. Furthermore, it provides a simple representation of $f_{\mathrm{D},\lambda}$ in terms of the kernel.

Theorem 5.5 (Representer theorem). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss and $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$. Furthermore, let H be an RKHS over X. Then, for all $\lambda > 0$, there exists a unique empirical SVM solution, i.e., a unique $f_{\mathrm{D},\lambda} \in H$ satisfying*

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) = \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)\,. \tag{5.5}$$

In addition, there exist $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$f_{\mathrm{D},\lambda}(x) = \sum_{i=1}^n \alpha_i k(x, x_i)\,, \qquad x \in X. \tag{5.6}$$

Proof. Uniqueness. It follows by repeating the proof of Lemma 5.1.

Existence. Since convergence in H implies pointwise convergence, we obtain the continuity of $\mathcal{R}_{L,\mathrm{D}} : H \to [0,\infty)$ by the continuity of L. Now the existence can be shown as in the proof of Theorem 5.2.

Representation (5.6). Let us write $X' := \{x_i : i = 1, \ldots, n\}$ and

$$H_{|X'} := \operatorname{span}\bigl\{k(\,\cdot\,, x_i) : i = 1, \ldots, n\bigr\}\,.$$

Then $H_{|X'}$ is the RKHS of $k_{|X' \times X'}$ by (4.12), and consequently we already know that there exists an empirical SVM solution $f_{\mathrm{D},\lambda,H_{|X'}} \in H_{|X'}$. Now let $H_{|X'}^\perp$ be the orthogonal complement of $H_{|X'}$ in H. For $f \in H_{|X'}^\perp$, we then have

$$f(x_i) = \langle f, k(\,\cdot\,, x_i)\rangle = 0\,, \qquad i = 1, \ldots, n\,.$$

If $P_{X'} : H \to H$ denotes the orthogonal projection onto $H_{|X'}$ we thus find

$$\mathcal{R}_{L,\mathrm{D}}(P_{X'} f) = \mathcal{R}_{L,\mathrm{D}}(f)$$

and $\|P_{X'} f\|_H \leq \|f\|_H$ for all $f \in H$. Since this yields

$$\begin{aligned}
\inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f) &\leq \inf_{f \in H_{|X'}} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f) \\
&= \inf_{f \in H} \lambda \|P_{X'} f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(P_{X'} f) \\
&\leq \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)\,,
\end{aligned}$$

we actually have equality in the above chain of inequalities. Therefore, $f_{\mathrm{D},\lambda,H_{|X'}}$ minimizes $\lambda \mapsto \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)$ in H, and the uniqueness of $f_{\mathrm{D},\lambda,H}$ then shows $f_{\mathrm{D},\lambda,H_{|X'}} = f_{\mathrm{D},\lambda,H}$. In other words, we have $f_{\mathrm{D},\lambda} := f_{\mathrm{D},\lambda,H} \in H_{|X'}$, and hence (5.6) follows from the definition of $H_{|X'}$. □

Our last theorem in this section shows that in most situations the general SVM decision functions are non-trivial.

Theorem 5.6 (Non-trivial SVM solutions). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss and P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss. Furthermore, let H be the RKHS of a bounded measurable kernel over X such that $\mathcal{R}^*_{L,\mathrm{P},H} < \mathcal{R}_{L,\mathrm{P}}(0)$. Then, for all $\lambda > 0$, we have $f_{\mathrm{P},\lambda} \neq 0$.*

Proof. By our assumptions, there exists an $f^* \in H$ with $\mathcal{R}_{L,\mathrm{P}}(f^*) < \mathcal{R}_{L,\mathrm{P}}(0)$. For $\alpha \in [0,1]$, we then have

$$\lambda\|\alpha f^*\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\alpha f^*) \leq \lambda\alpha^2\|f^*\|_H^2 + \alpha\mathcal{R}_{L,\mathrm{P}}(f^*) + (1-\alpha)\mathcal{R}_{L,\mathrm{P}}(0) =: h(\alpha)$$

by the convexity of $\mathcal{R}_{L,\mathrm{P}}$. Now, $\mathcal{R}_{L,\mathrm{P}}(f^*) < \mathcal{R}_{L,\mathrm{P}}(0)$ together with the quadratic form of $\alpha \mapsto h(\alpha)$ implies that $h : [0,1] \to [0,\infty)$ is minimized at some $\alpha^* \in (0,1]$. Consequently, we have

$$\lambda\|\alpha^* f^*\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\alpha^* f^*) \leq h(\alpha^*) < h(0) = \lambda\|0\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(0)\,. \qquad \square$$

5.2 A General Representer Theorem

We have seen in Theorem 5.5 that *empirical* SVM solutions can be represented by linear combinations of the canonical feature map. However, this result does not provide information about the *values* of the coefficients $\alpha_1, \ldots, \alpha_n$, and, in addition, it also remains unclear whether a somewhat similar result can hold for *general* SVM solutions.

Let us begin with a simplified consideration. To this end, let X be a measurable space, P be a distribution on $X \times Y$, and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex, differentiable, and P-integrable Nemitski loss. Furthermore, assume that $|L'| : X \times Y \times \mathbb{R} \to [0,\infty)$ is also a P-integrable Nemitski loss. Then Lemma 2.21 shows that $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to [0,\infty)$ is Fréchet differentiable and that its derivative at $f \in L_\infty(\mathrm{P}_X)$ is the bounded linear operator $\mathcal{R}'_{L,\mathrm{P}}(f) : L_\infty(\mathrm{P}_X) \to \mathbb{R}$ given by

$$\mathcal{R}'_{L,\mathrm{P}}(f)g = \int_{X\times Y} g(x)L'\big(x,y,f(x)\big)\,d\mathrm{P}(x,y)\,, \qquad g \in L_\infty(\mathrm{P}_X).$$

Now let H be a separable RKHS with bounded and measurable kernel k and $\Phi : X \to H$ be the corresponding canonical feature map. Lemma 4.25 then shows that $\Phi : X \to H$ is measurable, and from Lemmas 4.23 and 4.24 we can

infer that $\mathrm{id} : H \to L_\infty(\mathrm{P}_X)$ is well-defined and continuous. For $f_0 \in H$, the chain rule (see Lemma A.5.15) thus yields

$$\bigl(\mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id}\bigr)'(f_0) = \mathcal{R}'_{L,\mathrm{P}}(\mathrm{id}\, f_0) \circ \mathrm{id}'(f_0) = \mathcal{R}'_{L,\mathrm{P}}(f_0) \circ \mathrm{id}\,,$$

and hence we find for $f \in H$ that

$$\begin{aligned}\bigl(\mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id}\bigr)'(f_0)f = \bigl(\mathcal{R}'_{L,\mathrm{P}}(f_0) \circ \mathrm{id}\bigr)f &= \int_{X\times Y} f(x)L'\bigl(x, y, f_0(x)\bigr)\, d\mathrm{P}(x, y) \\ &= \mathbb{E}_{(x,y)\sim\mathrm{P}} L'\bigl(x, y, f_0(x)\bigr)\bigl\langle f, \Phi(x)\bigr\rangle \\ &= \Bigl\langle f, \mathbb{E}_{(x,y)\sim\mathrm{P}} L'\bigl(x, y, f_0(x)\bigr)\Phi(x)\Bigr\rangle\,.\end{aligned}$$

Note that the last expectation is H-valued and hence a *Bochner integral* in the sense of Section A.5.4. Using the Fréchet-Riesz isomorphism $\iota : H \to H'$ described in Theorem A.5.12, we thus see that

$$\bigl(\mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id}\bigr)'(f_0) = \iota\mathbb{E}_{(x,y)\sim\mathrm{P}} L'\bigl(x, y, f_0(x)\bigr)\Phi(x)\,. \tag{5.7}$$

Moreover, a straightforward calculation shows that the function $G : H \to \mathbb{R}$ defined by $Gf := \|f\|_H^2$, $f \in H$, is Fréchet differentiable and its derivative at f_0 is $G'(f_0) = 2\iota f_0$. Now recall that $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(\,\cdot\,) = \lambda G + \mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id}$, and hence $f_{\mathrm{P},\lambda}$ minimizes the function $\lambda G + \mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id} : H \to \mathbb{R}$. Consequently, we obtain

$$0 = \bigl(\lambda G + \mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id}\bigr)'(f_{\mathrm{P},\lambda}) = \iota\Bigl(2\lambda f_{\mathrm{P},\lambda} + \mathbb{E}_{(x,y)\sim\mathrm{P}} L'\bigl(x, y, f_{\mathrm{P},\lambda}(x)\bigr)\Phi(x)\Bigr)\,,$$

and hence we have $2\lambda f_{\mathrm{P},\lambda} = -\mathbb{E}_{(x,y)\sim\mathrm{P}} L'(x, y, f_{\mathrm{P},\lambda}(x))\Phi(x)$ by the injectivity of ι. The reproducing property then yields

$$f_{\mathrm{P},\lambda}(x) = -\int_{X\times Y} \frac{L'(x', y, f_{\mathrm{P},\lambda}(x'))}{2\lambda} k(x, x')\, d\mathrm{P}(x', y)\,, \qquad x \in X. \tag{5.8}$$

This equation answers both questions we posed in the introduction of this section. Indeed, for a data set $D = ((x_1, y_1), \ldots, (x_n, y_n))$ with corresponding empirical distribution D, equation (5.8) becomes

$$f_{\mathrm{D},\lambda}(x) = -\frac{1}{2\lambda n} \sum_{i=1}^{n} L'\bigl(x_i, y_i, f_{\mathrm{D},\lambda}(x_i)\bigr) k(x, x_i)\,, \qquad x \in X,$$

i.e., possible coefficients in (5.6) are

$$\alpha_i := -\frac{L'(x_i, y_i, f_{\mathrm{D},\lambda}(x_i))}{2\lambda n}\,, \qquad i = 1, \ldots, n.$$

Moreover, note that by writing $h(x, y) := L'(x, y, f_{\mathrm{D},\lambda}(x))$, $(x, y) \in X \times Y$, the representation in (5.6) becomes

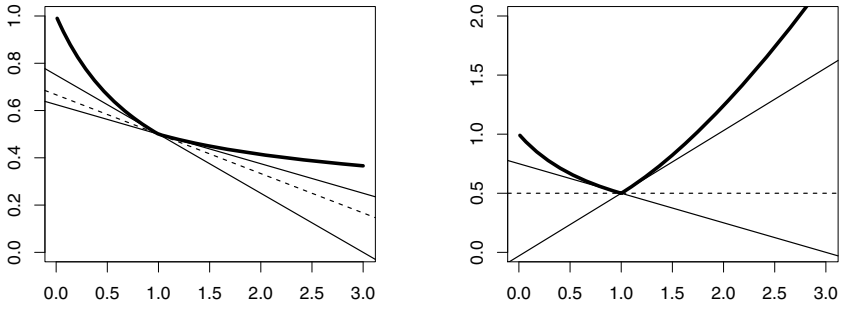


Fig. 5.1. Convex functions (bold lines) and some of their subdifferentials. Left: The function f is not differentiable at $x := 1$, so that the subdifferential $\partial f(x)$ contains the slopes of all tangents of f at x. In particular, it contains the left and the right derivatives of f at x (solid lines). Moreover, all slopes in between (dashed line) are contained. A formal statement of this illustration can be found in Lemma A.6.15. Right: The subdifferential at a minimum contains 0, i.e., the flat slope (dashed line).

$$f_{\mathrm{D},\lambda} = -\frac{1}{2\lambda n} \sum_{i=1}^{n} h(x_i, y_i) k(\,\cdot\,, x_i) = -\frac{1}{2\lambda} \mathbb{E}_{\mathrm{D}} h \Phi \,.$$

On the other hand, (5.8) can be rewritten as

$$f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda} \int_{X \times Y} h(x, y) k(\,\cdot\,, x)\, d\mathrm{P}(x, y) = -\frac{1}{2\lambda} \mathbb{E}_{\mathrm{P}} h \Phi \,.$$

This makes it clear that (5.8) can be viewed as a "continuous" and "quantified" version of the representer theorem.

The approach above yielded a quantified version of the representer theorem for *differentiable* losses. However, some important losses, such as the hinge loss and the pinball loss, are *not* differentiable, and since we are also interested in distributions P having point masses, this non-differentiability can cause problems even if it only occurs at one point. On the other hand, we are particularly interested in *convex* losses since these promise an efficient algorithmic treatment. Fortunately, for convex functions there exists a concept weaker than the derivative, for which the basic rules of calculus still hold. The following definition introduces this concept.

Definition 5.7. *Let E be a Banach space, $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex function and $w \in E$ be an element with $f(w) \neq \infty$. Then the* ***subdifferential*** *of f at w is defined by*

$$\partial f(w) := \big\{ w' \in E' : \langle w', v - w \rangle \leq f(v) - f(w) \text{ for all } v \in E \big\} \,.$$

Roughly speaking, the subdifferential $\partial f(w)$ contains all functionals describing the affine hyperplanes that are dominated by f and that are equal to

it at w. For an illustration of this interpretation, see Figure 5.1. In particular, if f is Gâteaux differentiable at w, then $\partial f(w)$ contains only the derivative of f at w. This and some other important properties of the subdifferential can be found in Section A.6.2.

If $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is a convex loss, we usually write $\partial L(x, y, t_0)$ for the subdifferential of the convex function $t \mapsto L(x, y, t)$ at the point $t_0 \in \mathbb{R}$. Moreover, we use analogous notation for supervised and unsupervised losses.

With these preparations, we can now state the main result of this section, which generalizes our considerations above to convex but not necessarily differentiable losses.

Theorem 5.8 (General representer theorem). *Let $p \in [1, \infty)$, P be a distribution on $X \times Y$, and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex, P-integrable Nemitski loss of order p. Furthermore, let k be a bounded and measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. Then, for all $\lambda > 0$, there exists an $h \in \mathcal{L}_{p'}(\mathrm{P})$ such that*

$$h(x, y) \in \partial L\big(x, y, f_{\mathrm{P},\lambda}(x)\big)\,, \qquad (x, y) \in X \times Y, \tag{5.9}$$

$$f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\,\mathbb{E}_{\mathrm{P}} h\Phi\,, \tag{5.10}$$

where p' is the conjugate exponent of p defined by $1/p' + 1/p = 1$.

Proof. Recall that L is a continuous loss since it is convex and finite. Moreover, L is a P-integrable Nemitski loss of order p, and hence we see by an almost literal repetition of the proof of Lemma 2.17 that $R : L_p(\mathrm{P}) \to [0, \infty)$ defined by

$$R(f) := \int_{X \times Y} L\big(x, y, f(x, y)\big)\, d\mathrm{P}(x, y)\,, \qquad f \in L_p(\mathrm{P}),$$

is well-defined and continuous.[1] Furthermore, Proposition A.6.13 shows that the subdifferential of R can be computed by

$$\partial R(f) = \big\{h \in L_{p'}(\mathrm{P}) : h(x, y) \in \partial L(x, y, f(x, y)) \text{ for P-almost all } (x, y)\big\}.$$

Now, we easily infer from Lemma 4.23 that the inclusion map $I : H \to L_p(\mathrm{P})$ defined by $(If)(x, y) := f(x)$, $f \in H$, $(x, y) \in X \times Y$, is a bounded linear operator. Moreover, for $h \in L_{p'}(\mathrm{P})$ and $f \in H$, the reproducing property yields

$$\langle h, If\rangle_{L_{p'}(\mathrm{P}), L_p(\mathrm{P})} = \mathbb{E}_{\mathrm{P}} hIf = \mathbb{E}_{\mathrm{P}} h\langle f, \Phi\rangle_H = \langle f, \mathbb{E}_{\mathrm{P}} h\Phi\rangle_H = \langle \iota \mathbb{E}_{\mathrm{P}} h\Phi, f\rangle_{H', H}\,,$$

where $\iota : H \to H'$ is the Fréchet-Riesz isomorphism described in Theorem A.5.12. Consequently, the adjoint operator I' of I is given by $I'h = \iota \mathbb{E}_{\mathrm{P}} h\Phi$,

[1] If we write $\bar{X} := X \times Y$, this can also be seen by interpreting R as the risk of the unsupervised, continuous loss $\bar{L} : \bar{X} \times \mathbb{R} \to [0, \infty)$, defined by $\bar{L}(\bar{x}, t) := L(x, y, t)$, $(x, y) := \bar{x} \in \bar{X}$, $t \in \mathbb{R}$.

$h \in L_{p'}(\mathrm{P})$. Moreover, the L-risk functional $\mathcal{R}_{L,\mathrm{P}} : H \to [0,\infty)$ restricted to H satisfies $\mathcal{R}_{L,\mathrm{P}} = R \circ I$, and hence the chain rule for subdifferentials (see Proposition A.6.12) yields $\partial\mathcal{R}_{L,\mathrm{P}}(f) = \partial(R \circ I)(f) = I'\partial R(If)$ for all $f \in H$. Applying the formula for $\partial R(f)$ thus yields

$$\begin{aligned}&\partial\mathcal{R}_{L,\mathrm{P}}(f)\\ &= \bigl\{\iota\mathbb{E}_{\mathrm{P}} h\Phi : h \in L_{p'}(\mathrm{P}) \text{ with } h(x,y) \in \partial L(x,y,f(x)) \text{ P-almost surely}\bigr\}\end{aligned}$$

for all $f \in H$. In addition, $f \mapsto \|f\|_H^2$ is Fréchet differentiable and its derivative at f is $2\iota f$ for all $f \in H$. By picking suitable representations of $h \in L_{p'}(\mathrm{P})$, Proposition A.6.12 thus gives

$$\begin{aligned}&\partial\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)\\ &= 2\lambda\iota f + \bigl\{\iota\mathbb{E}_{\mathrm{P}} h\Phi : h \in \mathcal{L}_{p'}(\mathrm{P}) \text{ with } h(x,y) \in \partial L(x,y,f(x)) \text{ for all } (x,y)\bigr\}\end{aligned}$$

for all $f \in H$. Now recall that $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(\,\cdot\,)$ has a minimum at $f_{\mathrm{P},\lambda}$, and therefore we have $0 \in \partial\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{\mathrm{P},\lambda})$ by another application of Proposition A.6.12. This together with the injectivity of ι yields the assertion. □

5.3 Stability of Infinite-Sample SVMs

Given a distribution P for which the general SVM solution $f_{\mathrm{P},\lambda}$ exists, one may ask how this solution changes if the underlying distribution P changes. The goal of this section is to answer this question with the help of the generalized representer theorem established in Section 5.2. The results we derive in this direction will be crucial for the stability-based statistical analysis in Sections 6.4 and 9.2. Moreover, it will be a key element in the robustness considerations of Sections 10.3 and 10.4.

Let us begin with the following theorem that, roughly speaking, provides the Lipschitz continuity of the map $\mathrm{P} \mapsto f_{\mathrm{P},\lambda}$.

Theorem 5.9. *Let $p \in [1,\infty)$ be a real number and $p' \in (1,\infty]$ be its conjugate defined by $\frac{1}{p} + \frac{1}{p'} = 1$. Furthermore, let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss function and P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss of order p. Furthermore, let k be a bounded and measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. Then, for all $\lambda > 0$, there exists an $h \in \mathcal{L}_{p'}(\mathrm{P})$ such that*

$$h(x,y) \in \partial L\bigl(x,y,f_{\mathrm{P},\lambda}(x)\bigr)\,, \qquad (x,y) \in X \times Y, \tag{5.11}$$

$$f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\,\mathbb{E}_{\mathrm{P}}\, h\Phi\,, \tag{5.12}$$

$$h \in \mathcal{L}_1(\bar{\mathrm{P}})\,, \tag{5.13}$$

$$\bigl\|\, f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\,\bigr\|_H \le \frac{1}{\lambda}\bigl\|\,\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\bar{\mathrm{P}}} h\Phi\,\bigr\|_H\,, \tag{5.14}$$

for all distributions $\bar{\mathrm{P}}$ on $X \times Y$ for which L is a $\bar{\mathrm{P}}$-integrable Nemitski loss.

Proof. By Theorem 5.8, there exists an $h \in \mathcal{L}_{p'}(\mathrm{P})$ satisfying (5.11) and (5.12). Let us first show that h is $\bar{\mathrm{P}}$-integrable, i.e., that (5.13) holds. To this end, observe that, since k is a bounded kernel, we have

$$\|f_{\mathrm{P},\lambda}\|_\infty \leq \|k\|_\infty \|f_{\mathrm{P},\lambda}\|_H \leq \|k\|_\infty \sqrt{\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda}} =: B_\lambda < \infty \tag{5.15}$$

by Lemma 4.23 and (5.4). Moreover, since L is a $\bar{\mathrm{P}}$-integrable Nemitski loss, there exist a $\bar{b} \in \mathcal{L}_1(\bar{\mathrm{P}})$ and an increasing function $\bar{h} : [0,\infty) \to [0,\infty)$ with

$$L(x,y,t) \leq \bar{b}(x,y) + \bar{h}(|t|)\,, \qquad (x,y,t) \in X \times Y \times \mathbb{R}\,.$$

Now (5.11) and Proposition A.6.11 with $\delta := 1$ yield

$$|h(x,y)| \leq \sup\bigl|\partial L(x,y,f_{\mathrm{P},\lambda}(x))\bigr| \leq \bigl|L(x,y,\,\cdot\,)_{|[-1+\|f_{\mathrm{P},\lambda}\|_\infty, 1+\|f_{\mathrm{P},\lambda}\|_\infty]}\bigr|_1\,,$$

and hence Lemma A.6.5 together with (5.15) shows

$$\begin{aligned} |h(x,y)| \leq \bigl|L(x,y,\,\cdot\,)_{|[-1+B_\lambda, 1+B_\lambda]}\bigr|_1 &\leq \frac{1}{1+B_\lambda}\,\bigl\|L(x,y,\,\cdot\,)_{|[-2+2B_\lambda, 2+2B_\lambda]}\bigr\|_\infty \\ &\leq \frac{\bar{b}(x,y) + \bar{h}(2+2B_\lambda)}{1+B_\lambda} \end{aligned} \tag{5.16}$$

for all $(x,y) \in X \times Y$. From this we deduce $h \in \mathcal{L}_1(\bar{\mathrm{P}})$.

Let us now establish (5.14). To this end, observe that by (5.11) and the definition of the subdifferential, we have

$$h(x,y)\bigl(f_{\bar{\mathrm{P}},\lambda}(x) - f_{\mathrm{P},\lambda}(x)\bigr) \leq L\bigl(x,y,f_{\bar{\mathrm{P}},\lambda}(x)\bigr) - L\bigl(x,y,f_{\mathrm{P},\lambda}(x)\bigr)$$

for all $(x,y) \in X \times Y$. By integrating with respect to $\bar{\mathrm{P}}$, we hence obtain

$$\langle f_{\bar{\mathrm{P}},\lambda} - f_{\mathrm{P},\lambda}\,,\, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi\rangle \;\leq\; \mathcal{R}_{L,\bar{\mathrm{P}}}(f_{\bar{\mathrm{P}},\lambda}) - \mathcal{R}_{L,\bar{\mathrm{P}}}(f_{\mathrm{P},\lambda})\,. \tag{5.17}$$

Moreover, an easy calculation shows

$$2\lambda\,\langle f_{\bar{\mathrm{P}},\lambda} - f_{\mathrm{P},\lambda}\,,\, f_{\mathrm{P},\lambda}\rangle + \lambda\|f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\|_H^2 = \lambda\|f_{\bar{\mathrm{P}},\lambda}\|_H^2 - \lambda\,\|f_{\mathrm{P},\lambda}\|_H^2\,. \tag{5.18}$$

By combining (5.17) and (5.18), we then find

$$\begin{aligned} \bigl\langle f_{\bar{\mathrm{P}},\lambda} - f_{\mathrm{P},\lambda}, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi + 2\lambda f_{\mathrm{P},\lambda}\bigr\rangle + \lambda\|f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\|_H^2 &\leq \mathcal{R}^{reg}_{L,\bar{\mathrm{P}},\lambda}(f_{\bar{\mathrm{P}},\lambda}) - \mathcal{R}^{reg}_{L,\bar{\mathrm{P}},\lambda}(f_{\mathrm{P},\lambda}) \\ &\leq 0\,, \end{aligned}$$

and consequently the representation $f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\,\mathbb{E}_{\mathrm{P}} h\Phi$ yields

$$\begin{aligned} \lambda\|f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\|_H^2 &\leq \bigl\langle f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\,,\, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi - \mathbb{E}_{\mathrm{P}} h\Phi\bigr\rangle \\ &\leq \|f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\|_H \cdot \bigl\|\,\mathbb{E}_{\bar{\mathrm{P}}} h\Phi - \mathbb{E}_{\mathrm{P}} h\Phi\,\bigr\|_H\,. \end{aligned}$$

From this we easily obtain (5.14). □

For the applications in later chapters, it is often necessary to have more information on the representing function h in (5.12). For two important types of losses, such additional information is presented in the following two corollaries.

Corollary 5.10. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex, locally Lipschitz continuous loss and P be a distribution on $X \times Y$ with $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$. Furthermore, let k be a bounded and measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. We write*

$$B_\lambda := \|k\|_\infty \left(\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda} \right)^{1/2}, \qquad \lambda > 0.$$

Then, for all $\lambda > 0$, there exists a bounded measurable function $h : X \times Y \to \mathbb{R}$ such that, for all distributions $\bar{\mathrm{P}}$ on $X \times Y$ with $\mathcal{R}_{L,\bar{\mathrm{P}}}(0) < \infty$, we have

$$h(x,y) \in \partial L\big(x, y, f_{\mathrm{P},\lambda}(x)\big)\,, \qquad (x,y) \in X \times Y, \tag{5.19}$$

$$f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\, \mathbb{E}_{\mathrm{P}}\, h\Phi\,, \tag{5.20}$$

$$\|h\|_\infty \le |L|_{B_\lambda,1}\,, \tag{5.21}$$

$$\big\| f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda} \big\|_H \le \frac{1}{\lambda} \big\| \mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\bar{\mathrm{P}}} h\Phi \big\|_H\,. \tag{5.22}$$

Proof. Let us fix a $\lambda > 0$ and write $B := B_\lambda$. By Lemma 4.23 and Corollary 5.3, we then know that

$$\|f_{\mathrm{P},\lambda}\|_\infty \;\le\; \|k\|_\infty \, \|f_{\mathrm{P},\lambda}\|_H \;\le\; \|k\|_\infty \left(\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda} \right)^{1/2} = B\,.$$

For $(x,y) \in X \times Y$, we further know that $L(x,y,\cdot\,) : \mathbb{R} \to [0,\infty)$ is convex, and hence Lemma A.6.16 shows that there exists a convex and Lipschitz continuous function $\tilde{L}(x,y,\cdot\,) : \mathbb{R} \to [0,\infty)$ with

$$\begin{aligned} \tilde{L}(x,y,\cdot\,)_{|[-B,B]} &= L(x,y,\cdot\,)_{|[-B,B]} \\ |\tilde{L}(x,y,\cdot\,)|_1 &= \big|L(x,y,\cdot\,)_{|[-B,B]}\big|_1 \\ \partial\tilde{L}(x,y,t) &\subset \partial L(x,y,t)\,, \qquad t \in [-B,B]. \end{aligned}$$

Moreover, considering the proof of Lemma A.6.16, we see that $\tilde{L} : X \times Y \times \mathbb{R} \to [0,\infty)$ can also be assumed to be measurable. Therefore, $\tilde{L}$ is a convex, Lipschitz continuous loss function with $|\tilde{L}|_1 = |L|_{B,1}$ and $\mathcal{R}_{\tilde{L},\mathrm{P}}(0) = \mathcal{R}_{L,\mathrm{P}}(0)$. Consequently, Corollary 5.3 shows that the general SVM solution

$$\tilde{f}_{\mathrm{P},\lambda} \in \arg\min_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{\tilde{L},\mathrm{P}}(f)$$

exists and satisfies $\|\tilde{f}_{\mathrm{P},\lambda}\|_\infty \le B$. Furthermore, we have $\mathcal{R}_{\tilde{L},\mathrm{P}}(f) = \mathcal{R}_{L,\mathrm{P}}(f)$ for all measurable $f : X \to [-B,B]$, and hence $\tilde{f}_{\mathrm{P},\lambda}$ is also a minimizer of

$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(\cdot)$. By the uniqueness of $f_{\mathrm{P},\lambda}$, we thus find $f_{\mathrm{P},\lambda} = \tilde{f}_{\mathrm{P},\lambda}$. Now recall that the Lipschitz continuity of $\tilde{L}$ gives

$$\tilde{L}(x,y,t) \le L(x,y,0) + |\tilde{L}|_1\,|t|\,, \qquad (x,y) \in X \times Y,\, t \in \mathbb{R},$$

i.e., $\tilde{L}$ is a Nemitski loss of order $p := 1$. Therefore, Theorem 5.8 gives a bounded measurable function $h : X \times Y \to \mathbb{R}$ such that

$$h(x,y) \in \partial\tilde{L}\big(x,y,\tilde{f}_{\mathrm{P},\lambda}(x)\big) \subset \partial L\big(x,y,f_{\mathrm{P},\lambda}(x)\big), \qquad (x,y) \in X \times Y, \quad (5.23)$$

and $-\frac{1}{2\lambda}\,\mathbb{E}_{\mathrm{P}} h\Phi = \tilde{f}_{\mathrm{P},\lambda} = f_{\mathrm{P},\lambda}$. In other words, we have shown (5.19) and (5.20). In addition, combining (5.23) with Proposition A.6.11 yields

$$\big|h(x,y)\big| \;\le\; \sup\big\{|t| : t \in \partial\tilde{L}(x,y,\tilde{f}_{\mathrm{P},\lambda}(x))\big\} \;\le\; |\tilde{L}|_1 \;=\; |L|_{B,1}$$

for all $(x,y) \in X \times Y$, i.e., we have shown (5.21). Finally, (5.22) can be shown as in the proof of Theorem 5.9. □

Recall that convex, margin-based losses are locally Lipschitz continuous, and hence Corollary 5.10 provides a generalized representer theorem together with the Lipschitz continuity of $\mathrm{P} \mapsto f_{\mathrm{P},\lambda}$ for this important class of loss functions. Let us now consider distance-based losses.

Corollary 5.11. *Let $L : \mathbb{R} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss of upper growth type $p \ge 1$ with representing function $\psi : \mathbb{R} \to [0,\infty)$. Furthermore, let P be a distribution on $X \times \mathbb{R}$ with $|\mathrm{P}|_p < \infty$ and k be a bounded and measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. Then there exists a constant $c_L > 0$ depending only on L such that for all $\lambda > 0$ there exists a measurable function $h : X \times Y \to \mathbb{R}$ such that*

$$h(x,y) \in -\partial\psi\big(y - f_{\mathrm{P},\lambda}(x)\big)\,, \qquad (x,y) \in X \times Y, \qquad (5.24)$$

$$f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\,\mathbb{E}_{\mathrm{P}}\, h\Phi\,, \qquad (5.25)$$

$$\|h\|_{L_s(\bar{\mathrm{P}})} \le 8^p c_L \left(1 + |\bar{\mathrm{P}}|_q^{p-1} + \|f_{\mathrm{P},\lambda}\|_\infty^{p-1}\right), \qquad (5.26)$$

$$\big\|\, f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\,\big\|_H \le \frac{1}{\lambda}\big\|\,\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\bar{\mathrm{P}}} h\Phi\,\big\|_H\,, \qquad (5.27)$$

for all $q \in [p,\infty]$, all distributions $\bar{\mathrm{P}}$ on $X \times Y$ with $|\bar{\mathrm{P}}|_q < \infty$, and $s := \frac{q}{p-1}$.

Proof. Recall that L is a P-integrable Nemitski loss of order p by Lemma 2.38, and by Theorem 5.9 there thus exists an $h \in \mathcal{L}_{p'}(\mathrm{P})$ satisfying (5.11), (5.12), and (5.14). Since (5.12) equals (5.25) and (5.14) equals (5.27), it remains to show (5.24) and (5.26). To this end, recall that ψ satisfies $L(y,t) = \psi(y - t)$, $y,t \in \mathbb{R}$, and hence it is convex. By Proposition A.6.12, we then have $\partial L(y,t) = -\partial\psi(y-t)$ for $y,t \in \mathbb{R}$, and hence (5.11) implies (5.24). Moreover, for $p = 1$, the loss L is Lipschitz continuous by Lemma 2.36, and consequently

(5.26) follows from Proposition A.6.11. Therefore let us now consider the case $p > 1$. For $(x, y) \in X \times Y$ with $r := |y - f_{\mathrm{P},\lambda}(x)| \geq 1$, Proposition A.6.11 with $\delta := r$ and Lemma 2.36 then yield

$$|h(x,y)| \leq |\psi_{|[-2r,2r]}|_1 \leq r^{-1} \, \|\psi_{|[4r,4r]}\|_\infty \leq cr^{-1} \left((4r)^p + 1\right) \leq c \, 4^{p+1} \, r^{p-1} \,,$$

where $c > 0$ is the constant arising in the upper growth type definition. Moreover, for $(x, y) \in X \times Y$ with $r := |y - f_{\mathrm{P},\lambda}(x)| \leq 1$, Proposition A.6.11 gives

$$|h(x,y)| \leq |\psi_{|[-2r,2r]}|_1 \leq |\psi_{|[-2,2]}|_1 \,.$$

Together, these estimates show that for all $(x, y) \in X \times Y$ we have

$$|h(x,y)| \leq 4^p \, c_L \max\bigl\{1, |y - f_{\mathrm{P},\lambda}(x)|^{p-1}\bigr\} \,, \tag{5.28}$$

where c_L is a suitable constant depending only on the loss function L. For $q = \infty$, we then easily find the assertion, and hence let us assume $q \in [p, \infty)$. In this case, the inequality above yields

$$|h(x,y)|^s \leq 4^{ps} c_L^s \max\bigl\{1, |y - f_{\mathrm{P},\lambda}(x)|^q\bigr\} \leq 4^{ps} 2^{q-1} c_L^s \bigl(1 + |y|^q + |f_{\mathrm{P},\lambda}(x)|^q\bigr) \,,$$

and using $4^p 2^{\frac{q-1}{s}} \leq 8^p$ we consequently find

$$\|h\|_{L_s(\bar{\mathrm{P}})} \leq 8^p c_L \bigl(1 + |\bar{\mathrm{P}}|_q^{p-1} + \|f_{\mathrm{P},\lambda}\|_\infty^{p-1}\bigr) \,. \qquad \square$$

Note that the larger we can choose the real number q in the preceding corollary, the larger the corresponding s becomes, i.e., the stronger the integrability condition on h becomes. In particular, the case $q = \infty$ yields $s = \infty$, i.e., the representing function h is bounded.

The following corollary shows that Theorem 5.9 holds under weaker assumptions if all distributions involved are empirical.

Corollary 5.12. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss function, $n \geq 1$, and $D \in (X \times Y)^n$. Furthermore, let H be the RKHS of a kernel on X and $\Phi : X \to H$ be its canonical feature map. Moreover, let $\lambda > 0$ and*

$$B_\lambda := \|k\|_\infty \left(\frac{\mathcal{R}_{L,\mathrm{D}}(0)}{\lambda}\right)^{1/2} . \tag{5.29}$$

Then there exists a function $h : X \times Y \to \mathbb{R}$ such that, for all $m \geq 1$ and all $\bar{D} \in (X \times Y)^m$, we have

$$\begin{aligned} h(x,y) &\in \partial L\bigl(x, y, f_{\mathrm{D},\lambda}(x)\bigr) \,, \qquad\qquad (x,y) \in D \cup \bar{D}, \\ f_{\mathrm{D},\lambda} &= -\frac{1}{2\lambda} \, \mathbb{E}_{\mathrm{D}} \, h\Phi \,, \\ \|h\|_\infty &\leq |L|_{B_\lambda,1} \,, \\ \|f_{\mathrm{D},\lambda} - f_{\bar{\mathrm{D}},\lambda}\|_H &\leq \frac{1}{\lambda} \bigl\| \, \mathbb{E}_{\mathrm{D}} h\Phi - \mathbb{E}_{\bar{\mathrm{D}}} h\Phi \, \bigr\|_H \,. \end{aligned}$$

Proof. We write $D = ((x_1, y_1), \ldots, (x_n, y_n))$ and $\bar{D} = ((\bar{x}_1, \bar{y}_1), \ldots, (\bar{x}_m, \bar{y}_m))$. Furthermore, we define

$$X' := \{x_i : i = 1, \ldots, n\} \cup \{\bar{x}_i : i = 1, \ldots, m\},$$
$$Y' := \{y_i : i = 1, \ldots, n\} \cup \{\bar{y}_i : i = 1, \ldots, m\},$$

and equip both sets with the discrete σ-algebra. Then $k_{|X' \times X'}$ is a bounded measurable kernel with finite-dimensional, and hence separable, RKHS. In addition, both D and $\bar{\mathrm{D}}$ are distributions on $X' \times Y'$. Furthermore, $L(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ is convex for all $(x, y) \in X' \times Y'$ and hence L is locally Lipschitz continuous. Since X' and Y' are finite sets, we then see that L restricted to $X' \times Y'$ is locally Lipschitz. Moreover, we obviously have both $\mathcal{R}_{L,\mathrm{D}}(0) < \infty$ and $\mathcal{R}_{L,\bar{\mathrm{D}}}(0) < \infty$, and hence the assertion follows from Corollary 5.10. □

Let us finally use the results above to show that, under some additional conditions, the map $D \mapsto f_{\mathrm{D},\lambda}$ is continuous.

Lemma 5.13. *Let X be a metric space and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a continuous function that is a convex and differentiable loss. Furthermore, let H be the RKHS of a continuous kernel on X, $n \geq 1$, and $\lambda > 0$. Then the map $(X \times Y)^n \to H$ defined by $D \mapsto f_{\mathrm{D},\lambda}$ is continuous.*

Proof. Let us fix two sample sets $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ and $\bar{D} := ((\bar{x}_1, \bar{y}_1), \ldots, (\bar{x}_m, \bar{y}_m)) \in (X \times Y)^n$. By Corollary 5.12, we then obtain

$$\left\| f_{\mathrm{D},\lambda} - f_{\bar{\mathrm{D}},\lambda} \right\|_H \leq \frac{1}{\lambda n} \left\| \sum_{i=1}^{n} L'\big(x_i, y_i, f_{\mathrm{D},\lambda}(x_i)\big)\Phi(x_i) - L'\big(\bar{x}_i, \bar{y}_i, f_{\mathrm{D},\lambda}(\bar{x}_i)\big)\Phi(\bar{x}_i) \right\|.$$

Now Lemma 4.29 shows that both $f_{\mathrm{D},\lambda}$ and $\Phi : X \to H$ are continuous. Moreover, every convex differentiable function is continuously differentiable by Proposition A.6.14, and hence we obtain the assertion. □

5.4 Behavior for Small Regularization Parameters

In this section, we investigate how the general SVM solution $f_{\mathrm{P},\lambda}$ and its associated risk $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ behaves for vanishing regularization parameter, i.e., for $\lambda \to 0$. In addition, we compare the behavior of the minimized regularized risk with the approximation error of the scaled unit balls $\lambda^{-1} B_H$.

Let us begin by introducing a new quantity. To this end, let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, k be a measurable kernel over X with RKHS H, and P be a distribution on $X \times Y$. Then we write

$$\mathcal{R}^*_{L,\mathrm{P},H} := \inf_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) \tag{5.30}$$

for the smallest possible L-risk on H. Moreover, we say that an element $f^* \in H$ **minimizes the L-risk in H** if it satisfies $\mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}^*_{L,\mathrm{P},H}$. Finally, we need the following fundamental definition, which is closely related to the minimization of the regularized risk (5.2).

Definition 5.14. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, H be the RKHS of a measurable kernel on X, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Then we define the* ***approximation error function*** *$A_2 : [0,\infty) \to [0,\infty)$ by*

$$A_2(\lambda) := \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H}\,, \qquad \lambda \geq 0.$$

Our first lemma collects some simple properties of the function A_2.

Lemma 5.15 (Properties of the approximation error function). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, H be the RKHS of a measurable kernel on X, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Then $A_2 : [0,\infty) \to [0,\infty)$ is increasing, concave, and continuous. Moreover, we have $A_2(0) = 0$ and*

$$\begin{aligned} \frac{A_2(\kappa)}{\kappa} &\leq \frac{A_2(\lambda)}{\lambda}\,, && 0 < \lambda \leq \kappa, \qquad (5.31)\\ A_2(\lambda) &\leq \mathcal{R}_{L,\mathrm{P}}(0) - \mathcal{R}^*_{L,\mathrm{P},H}\,, && \lambda \geq 0. \end{aligned}$$

In addition, $A_2(\,\cdot\,)$ is subadditive in the sense of

$$A_2(\lambda + \kappa) \;\leq\; A_2(\lambda) + A_2(\kappa)\,, \qquad \lambda, \kappa \geq 0.$$

Finally, if there exists a function $h : [0,1] \to [0,\infty)$ with $\lim_{\lambda \to 0^+} h(\lambda) = 0$ and $A_2(\lambda) \leq \lambda h(\lambda)$ for all $\lambda \in [0,1]$, then we have $A_2(\lambda) = 0$ for all $\lambda \geq 0$, and 0 minimizes $\mathcal{R}_{L,\mathrm{P}}$ in H.

Proof. The definition of the approximation error function immediately gives $A_2(0) = 0$. Moreover, $A_2(\,\cdot\,)$ is an infimum over a family of affine linear and increasing functions, and hence we see that A_2 is concave, continuous, and increasing by Lemma A.6.4. Furthermore, (5.31) follows from the concavity, namely

$$A_2(\lambda) = A_2\Big(\frac{\lambda}{\kappa}\kappa + \Big(1 - \frac{\lambda}{\kappa}\Big)0\Big) \geq \frac{\lambda}{\kappa} A_2(\kappa) + \Big(1 - \frac{\lambda}{\kappa}\Big) A_2(0) = \frac{\lambda}{\kappa} A_2(\kappa)\,.$$

In addition, for $\lambda \geq 0$, we obtain from the definition of A_2 that

$$A_2(\lambda) \leq \lambda \|0\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(0) - \mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}_{L,\mathrm{P}}(0) - \mathcal{R}^*_{L,\mathrm{P},H}\,.$$

In order to show the subadditivity, it suffices to consider $\lambda, \kappa > 0$. Without loss of generality, we may additionally assume $\lambda \leq \kappa$. Using the already proved (5.31) twice, we then obtain

$$A_2(\lambda + \kappa) \leq (\lambda + \kappa)\kappa^{-1} A_2(\kappa) = \lambda \kappa^{-1} A_2(\kappa) + A_2(\kappa) \leq A_2(\lambda) + A_2(\kappa)\,.$$

Finally, let us assume that we have $A_2(\lambda) \leq \lambda h(\lambda)$ for all $\lambda \in [0,1]$. Using our previous results, we then obtain

$$A_2(1) \;\leq\; \lambda^{-1} A_2(\lambda) \;\leq\; h(\lambda)\,, \qquad \lambda \in (0,1].$$

For $\lambda \to 0$, we then find $A_2(1) = 0$, and since A_2 is a non-negative and concave function with $A_2(0) = 0$, we then find $A_2(\lambda) = 0$ for all $\lambda \geq 0$. The last assertion is a trivial consequence of the latter. □

Now assume for a moment that the general SVM solution $f_{\mathrm{P},\lambda}$ exists for all $\lambda > 0$. The continuity of A_2 at 0 together with $A_2(0) = 0$ then shows both

$$\lim_{\lambda \to 0} \lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) = \mathcal{R}^*_{L,\mathrm{P},H} \tag{5.32}$$

and $\lim_{\lambda \to 0} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) = \mathcal{R}^*_{L,\mathrm{P},H}$. In other words, the (regularized) risk of the SVM solution $f_{\mathrm{P},\lambda}$ tends to the minimal risk in H for vanishing regularization term λ. Since this suggests that the behavior of A_2 becomes particularly interesting for $\lambda \to 0$, we will mainly focus on this limit behavior. To this end, we need the following preparatory lemma.

Lemma 5.16. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss, H be the RKHS of a measurable kernel on X, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Assume that there exists an $f_0^* \in H$ with $\mathcal{R}_{L,\mathrm{P}}(f_0^*) = \mathcal{R}^*_{L,\mathrm{P},H}$. Then there exists exactly one element $f^*_{L,\mathrm{P},H} \in H$ such that both*

$$\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P},H}) = \mathcal{R}^*_{L,\mathrm{P},H}$$

and $\|f^*_{L,\mathrm{P},H}\| \leq \|f^*\|$ for all $f^* \in H$ satisfying $\mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}^*_{L,\mathrm{P},H}$.

Proof. Let M denote the set of all elements f^* minimizing $\mathcal{R}_{L,\mathrm{P}}$ in H, i.e.

$$M := \left\{ f^* \in H : \mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}^*_{L,\mathrm{P},H} \right\}.$$

Uniqueness. Assume that we have two different elements $f_1^*, f_2^* \in M$ satisfying $\|f_1^*\| \leq \|f\|$ and $\|f_2^*\| \leq \|f\|$ for all $f \in M$. Then we obviously have $\|f_1^*\| = \|f_2^*\|$. Moreover, the convexity of $\mathcal{R}_{L,\mathrm{P}}$ implies $\frac{1}{2}(f_1^* + f_2^*) \in M$, and the last statement in Lemma A.5.9 shows

$$\left\| \frac{1}{2}(f_1^* + f_2^*) \right\|_H^2 < \frac{\|f_1^*\|_H^2 + \|f_2^*\|_H^2}{2} = \|f_1^*\|_H^2 . \tag{5.33}$$

Since this contradicts our assumptions on f_1^*, there is at most one $f^*_{L,\mathrm{P},H}$.

Existence. Since L is a convex loss, it is continuous, and consequently the risk functional $\mathcal{R}_{L,\mathrm{P}} : H \to [0, \infty]$ is lower semi-continuous by Lemma 2.15. This shows that the set M of minimizing elements is closed. Using Lemma A.2.8, we then see that the map $N : H \to [0, \infty]$ defined by

$$N(f) := \begin{cases} \|f\| & \text{if } f \in M \\ \infty & \text{otherwise} \end{cases}$$

is lower semi-continuous. In addition, M is convex by the convexity of L, and hence so is N. Now observe that the set $\{f \in H : N(f) \leq \|f_0^*\|\}$ is non-empty and bounded, and consequently N has a global minimum at some $f_0 \in H$ by Theorem A.6.9. Obviously, this f_0 is the desired $f^*_{L,\mathrm{P},H}$. □

Let us now assume for a moment that we are in the situation of Lemma 5.16. If $f_{\mathrm{P},\lambda}$ exists for some $\lambda > 0$, it necessarily satisfies

$$\|f_{\mathrm{P},\lambda}\| \;\leq\; \|f^*_{L,\mathrm{P},H}\| \tag{5.34}$$

since otherwise we would find a contradiction by

$$\lambda\|f^*_{L,\mathrm{P},H}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P},H}) \;<\; \lambda\|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})\,.$$

Moreover, observe that for $\lambda = 0$ a simple calculation shows

$$\lambda\|f^*_{L,\mathrm{P},H}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P},H}) = \inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f)\,,$$

and hence we write $f_{\mathrm{P},0} := f^*_{L,\mathrm{P},H}$. With this notation, we can now formulate our first main result describing the behavior of the function $\lambda \mapsto f_{\mathrm{P},\lambda}$.

Theorem 5.17 (Continuity in the regularization parameter). *Let* P *be a distribution on* $X \times Y$*,* $L : X \times Y \times \mathbb{R} \to [0,\infty)$ *be a convex* P*-integrable Nemitski loss, and* H *be the RKHS of a bounded measurable kernel on* X*. Furthermore, let* $(\lambda_n) \subset (0,\infty)$ *be a sequence that converges to a real number* $\lambda \in [0,\infty)$*. If the sequence* $(f_{\mathrm{P},\lambda_n})$ *is bounded, then* $f_{\mathrm{P},\lambda}$ *exists and we have*

$$\lim_{n\to\infty} \|f_{\mathrm{P},\lambda_n} - f_{\mathrm{P},\lambda}\|_H = 0\,.$$

Proof. Since $(f_{\mathrm{P},\lambda_n})$ is bounded, Theorem A.5.6 shows that there exist an $f^* \in H$ and a subsequence $(f_{\mathrm{P},\lambda_{n_i}})$ such that

$$f_{\mathrm{P},\lambda_{n_i}} \to f^* \qquad \text{with respect to the weak topology in } H.$$

Moreover, since this subsequence is bounded, we may additionally assume that there exists a $c \geq 0$ such that $\|f_{\mathrm{P},\lambda_{n_i}}\| \to c$. Now recall that the Dirac functionals are contained in the dual H', and therefore weak convergence in H implies pointwise convergence. Lemma 2.17 thus shows $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_{n_i}}) \to \mathcal{R}_{L,\mathrm{P}}(f^*)$. Furthermore, by (A.21), the weak convergence of $(f_{\mathrm{P},\lambda_{n_i}})$ implies

$$\|f^*\| \leq \liminf_{i\to\infty} \|f_{\mathrm{P},\lambda_{n_i}}\| = \lim_{i\to\infty} \|f_{\mathrm{P},\lambda_{n_i}}\| = c\,, \tag{5.35}$$

and hence the continuity of $A_2(\,\cdot\,)$ established in Lemma 5.15 yields

$$\begin{aligned}
\lambda\|f^*\|^2 + \mathcal{R}_{L,\mathrm{P}}(f^*) - \mathcal{R}^*_{L,\mathrm{P},H} &\leq \lambda c^2 + \mathcal{R}_{L,\mathrm{P}}(f^*) - \mathcal{R}^*_{L,\mathrm{P},H} \\
&= \lim_{i\to\infty} \bigl(\lambda_{n_i}\|f_{\mathrm{P},\lambda_{n_i}}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_{n_i}}) - \mathcal{R}^*_{L,\mathrm{P},H}\bigr) \\
&= \lim_{i\to\infty} A_2(\lambda_{n_i}) \\
&= A_2(\lambda)\,.
\end{aligned} \tag{5.36}$$

Consequently, f^* minimizes the regularized L-risk. If $\lambda > 0$, this implies $f^* = f_{\mathrm{P},\lambda}$ by Lemma 5.1. Furthermore, if $\lambda = 0$, this means that f^* minimizes

$\mathcal{R}_{L,\mathrm{P}}$ in H, and by combining (5.35) with (5.34), we thus find $\|f^*\| \leq \|f^*_{L,\mathrm{P},H}\|$, i.e., we have $f^* = f^*_{L,\mathrm{P},H} = f_{\mathrm{P},0}$ by Lemma 5.16.

Now, using the equality $f^* = f_{\mathrm{P},\lambda}$ in (5.36), we obtain

$$\lambda\|f_{\mathrm{P},\lambda}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} \;=\; \lambda c^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H}\,,$$

i.e., we have $c = \|f_{\mathrm{P},\lambda}\|$. Therefore, we have both $f_{\mathrm{P},\lambda_{n_i}} \to f_{\mathrm{P},\lambda}$ weakly and $\|f_{\mathrm{P},\lambda_{n_i}}\| \to \|f_{\mathrm{P},\lambda}\|$. Together these convergences imply

$$\lim_{i\to\infty} \|f_{\mathrm{P},\lambda_{n_i}} - f_{\mathrm{P},\lambda}\|^2 = \lim_{i\to\infty} \bigl(\|f_{\mathrm{P},\lambda_{n_i}}\|^2 - 2\langle f_{\mathrm{P},\lambda_{n_i}}, f_{\mathrm{P},\lambda}\rangle + \|f_{\mathrm{P},\lambda}\|^2\bigr) = 0\,,$$

i.e., the subsequence $(f_{\mathrm{P},\lambda_{n_i}})$ converges to $f_{\mathrm{P},\lambda}$ with respect to $\|\cdot\|_H$. Finally, let us assume that (f_{P,λ_n}) does *not* converge to $f_{\mathrm{P},\lambda}$ in norm. Then there exists a $\delta > 0$ and a subsequence $(f_{\mathrm{P},\lambda_{n_j}})$ with $\|f_{\mathrm{P},\lambda_{n_j}} - f_{\mathrm{P},\lambda}\| > \delta$. However, applying the reasoning above to this subsequence gives a sub-subsequence converging to $f_{\mathrm{P},\lambda}$ and hence we find a contradiction. □

The preceding theorem has some interesting consequences on the behavior of both $\lambda \mapsto f_{\mathrm{P},\lambda}$ and $\lambda \mapsto A_2(\lambda)$. Let us begin with a result that characterizes the existence of $f^*_{L,\mathrm{P},H}$ in terms of $\lambda \mapsto \|f_{\mathrm{P},\lambda}\|$ and A_2.

Corollary 5.18. *Let* P *be a distribution on* $X \times Y$*,* $L : X \times Y \times \mathbb{R} \to [0,\infty)$ *be a convex* P*-integrable Nemitski loss, and* H *be the RKHS of a bounded measurable kernel on* X*. Then the following statements are equivalent:*

i) There exists an $f^* \in H$ *minimizing* $\mathcal{R}_{L,\mathrm{P}}$ *in* H*.*
ii) The function $\lambda \mapsto \|f_{\mathrm{P},\lambda}\|$ *is bounded on* $(0,\infty)$*.*
iii) There exists a constant $c > 0$ *with* $A_2(\lambda) \leq c\lambda$ *for all* $\lambda \geq 0$*.*

In addition, if one of the statements is satisfied, we have $A_2(\lambda) \leq \lambda\,\|f^*_{L,\mathrm{P},H}\|^2_H$ *for all* $\lambda \geq 0$ *and*

$$\lim_{\lambda\to 0^+} \|f_{\mathrm{P},\lambda} - f^*_{L,\mathrm{P},H}\|_H = 0\,.$$

Proof. ii) $\Rightarrow$ *i).* Let $(\lambda_n) \subset (0,\infty)$ be a sequence with $\lambda_n \to 0$. Our assumption then guarantees that the sequence $(f_{\mathrm{P},\lambda_n})$ is bounded, and hence Theorem 5.17 shows that $f_{\mathrm{P},0} = f^*_{L,\mathrm{P},H}$ exists and that we have $f_{\mathrm{P},\lambda_n} \to f^*_{L,\mathrm{P},H}$.

i) $\Rightarrow$ *iii).* By Lemma 5.16, we know that $f^*_{L,\mathrm{P},H}$ exists. Now the implication follows from the estimate

$$A_2(\lambda) \leq \lambda\|f^*_{L,\mathrm{P},H}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P},H}) - \mathcal{R}^*_{L,\mathrm{P},H} = \lambda\|f^*_{L,\mathrm{P},H}\|^2\,, \qquad \lambda \geq 0.$$

iii) $\Rightarrow$ *ii).* This implication follows from the estimate

$$\lambda\|f_{\mathrm{P},\lambda}\|^2 \leq \lambda\|f_{\mathrm{P},\lambda}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} = A_2(\lambda) \leq c\lambda\,. \qquad \square$$

Let us now assume for a moment that L is the hinge loss and that H is the RKHS of a universal (or just a strictly positive definite) kernel. For a training set D without contradicting samples, i.e., $x_i = x_j$ implies $y_i = y_j$, it is then

straightforward to see that there is an $f^* \in H$ with $\mathcal{R}_{L,\mathrm{D}}(f^*) = 0$. Obviously, this f^* minimizes $\mathcal{R}_{L,\mathrm{D}}$ in H and thus $f^*_{L,\mathrm{D},H} \in H$ does exist. Moreover, since $f^*_{L,\mathrm{D},H}$ has zero error on D and minimal norm among such minimizers, it coincides with the *hard margin* SVM solution. Consequently, Corollary 5.18 shows that the soft margin SVM solutions $f_{\mathrm{D},\lambda}$ converge to the hard margin solution if D is *fixed* and $\lambda \to 0$. Since the hard margin SVM solution does not make training errors, this convergence indicates that the soft margin SVM solutions may overfit if the regularization parameter is chosen too small.

Corollary 5.18 shows that a *linear* upper bound on A_2 of the form $A_2(\lambda) \leq c\lambda$ occurs if and only if $\mathcal{R}_{L,\mathrm{P}}$ has a global minimum in H. Now recall that Lemma 5.15 showed that such an upper bound on A_2 is optimal in the sense that every stronger upper bound of the form $A_2(\lambda) \leq \lambda h(\lambda)$, $\lambda \in [0,1]$, for some function $h : [0,1] \to [0,\infty)$ with $\lim_{\lambda \to 0^+} h(\lambda) = 0$, implies $A_2(\lambda) = 0$ for all $\lambda \geq 0$. Since the latter implies that 0 minimizes $\mathcal{R}_{L,\mathrm{P}}$ in H, we see that $f^*_{L,\mathrm{P},H}$ exists if and only if A_2 has an "optimal" upper bound.

Now assume that $f_{\mathrm{P},\lambda}$ exists for all $\lambda > 0$. By Lemma 5.15, we then find

$$\lambda \|f_{\mathrm{P},\lambda}\|_H^2 \leq A_2(\lambda) \leq \mathcal{R}_{L,\mathrm{P}}(0) - \mathcal{R}^*_{L,\mathrm{P},H}\,, \qquad \lambda > 0,$$

and hence we obtain $\lim_{\lambda \to \infty} f_{\mathrm{P},\lambda} = 0$. This justifies the notation $f_{\mathrm{P},\infty} := 0$. The next corollary shows the continuity of the (extended) function $\lambda \mapsto f_{\mathrm{P},\lambda}$.

Corollary 5.19. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss, P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss, and H be the RKHS of a bounded measurable kernel on X. Then*

$$\lambda \mapsto f_{\mathrm{P},\lambda}$$

is a continuous map from $(0,\infty]$ to H, and $\lambda \mapsto \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ is a continuous map from $(0,\infty]$ to $[0,\infty)$. Furthermore, if there exists an f^* minimizing $\mathcal{R}_{L,\mathrm{P}}$ in H, both maps are also defined and continuous at 0.

Proof. The first assertion is a direct consequence of Theorem 5.2 and Theorem 5.17, and the second assertion follows from combining the first assertion with Lemma 2.17. The last assertion follows from Corollary 5.18 and the notational convention $f_{\mathrm{P},0} := f^*_{L,\mathrm{P},H}$. □

Our next goal is to investigate whether the regularization term $\lambda \|f_{\mathrm{P},\lambda}\|_H^2$ and $A_2(\lambda)$ behave similarly for $\lambda \to 0$. To this end, we need some preparatory notions and results. Let us begin with the following definition, which introduces a differently regularized approximation error function.

Definition 5.20. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss, H be the RKHS of a measurable kernel on X, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Then the* ***$\infty$-approximation error function*** *$A_\infty : [0,\infty) \to [0,\infty)$ is defined by*

$$A_\infty(\lambda) := \inf_{f \in \lambda^{-1} B_H} \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H}\,, \qquad \lambda \geq 0.$$

Moreover, an element $f^{(\infty)}_{\mathrm{P},\lambda} \in \lambda^{-1} B_H$ *satisfying*

$$\mathcal{R}_{L,\mathrm{P}}(f^{(\infty)}_{\mathrm{P},\lambda}) = A_\infty(\lambda)$$

and $\|f^{(\infty)}_{\mathrm{P},\lambda}\| \leq \|f^*\|$ *for all* $f^* \in \lambda^{-1} B_H$ *with* $\mathcal{R}_{L,\mathrm{P}}(f^*) = A_\infty(\lambda)$ *is called a* ***minimal norm minimizer*** *of* $\mathcal{R}_{L,\mathrm{P}}$ *in* $\lambda^{-1} B_H$*.*

Note that, for RKHSs H satisfying $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$, the value $A_\infty(\lambda)$ is usually called the **approximation error** of the set $\lambda^{-1} B_H$.

We will see later that there is an intimate relation between the functions A_2 and A_∞ and their minimizers. Before we go into details, we first present two results establishing the existence and uniqueness of minimal norm minimizers.

Lemma 5.21 (Uniqueness of minimal norm minimizers). *Let* H *be the RKHS of a measurable kernel on* X*,* $L : X \times Y \times \mathbb{R} \to [0, \infty)$ *be a convex loss, and* P *be a distribution on* $X \times Y$ *with* $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$*. Then, for all* $\lambda > 0$*, there exists at most one minimal norm minimizer of* $\mathcal{R}_{L,\mathrm{P}}$ *in* $\lambda^{-1} B_H$*.*

Proof. Suppose that there are two different minimal norm minimizers $f_1^*, f_2^* \in \lambda^{-1} B_H$ of $\mathcal{R}_{L,\mathrm{P}}$ in $\lambda^{-1} B_H$. By the convexity of $\mathcal{R}_{L,\mathrm{P}}$, we then find that $\frac{1}{2}(f_1^* + f_2^*) \in \lambda^{-1} B_H$ also minimizes $\mathcal{R}_{L,\mathrm{P}}$. Moreover, we have (5.33), which contradicts our assumption that f^* is a minimal norm minimizer. □

The next lemma shows that the conditions ensuring the existence of general SVM solutions $f_{\mathrm{P},\lambda}$ also ensure the existence of minimal norm minimizers.

Lemma 5.22 (Existence of minimal norm minimizers). *Let* P *be a distribution on* $X \times Y$*,* $L : X \times Y \times \mathbb{R} \to [0, \infty)$ *be a convex* P*-integrable Nemitski loss, and* H *be the RKHS of a bounded measurable kernel on* X*. Then, for all* $\lambda > 0$*, there exists exactly one minimal norm minimizer* $f^{(\infty)}_{\mathrm{P},\lambda} \in H$*.*

Proof. We first show that there exists an $f^* \in \lambda^{-1} B_H$ that minimizes $\mathcal{R}_{L,\mathrm{P}}$ on $\lambda^{-1} B_H$. To this end, observe that $\mathcal{R}_{L,\mathrm{P}} : H \to [0, \infty)$ is continuous by Lemma 4.23 and Lemma 2.17. Therefore, the map $\mathcal{R} : H \to [0, \infty]$ defined by

$$\mathcal{R}(f) := \begin{cases} \mathcal{R}_{L,\mathrm{P}}(f) & \text{if } f \in \lambda^{-1} B_H \\ \infty & \text{otherwise} \end{cases}$$

is lower semi-continuous by Lemma A.2.8 and, in addition, $\mathcal{R}$ is also convex. Now recall that, since L is a P-integrable Nemitski loss, we have $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, and therefore every $f \in H$ with $\mathcal{R}(f) \leq \mathcal{R}_{L,\mathrm{P}}(0)$ must satisfy $\|f\| \leq \lambda^{-1}$. This shows that the set $\{f \in H : \mathcal{R}(f) \leq \mathcal{R}_{L,\mathrm{P}}(0)\}$ is non-empty and bounded, and consequently $\mathcal{R}$ has a global minimum at some $f^* \in \lambda^{-1} B_H$ by Theorem

A.6.9. Since $\mathcal{R}$ coincides with $\mathcal{R}_{L,\mathrm{P}}$ on $\lambda^{-1}B_H$, it is obvious that this f^* also minimizes $\mathcal{R}_{L,\mathrm{P}}$ on $\lambda^{-1}B_H$. In other words, the set

$$A := \left\{ f \in \lambda^{-1}B_H : \mathcal{R}_{L,\mathrm{P}}(f) \leq \mathcal{R}_{L,\mathrm{P}}(\tilde{f}) \text{ for all } \tilde{f} \in \lambda^{-1}B_H \right\}$$

is non-empty and, by the continuity of $\mathcal{R}_{L,\mathrm{P}} : H \to [0,\infty)$, we see that A is also closed. Moreover, $\|\cdot\|_H : H \to [0,\infty)$ is continuous. By applying Lemma A.2.8 and Theorem A.6.9 to the map $N : H \to [0,\infty]$ defined by

$$N(f) := \begin{cases} \|f\| & \text{if } f \in A \\ \infty & \text{otherwise,} \end{cases}$$

we hence see that $\|\cdot\|_H$ restricted to A has a minimum. □

Let us now compare the minimizers of A_2 and A_∞. We begin by showing that the existence of $f_{\mathrm{P},\lambda}$ implies that $f^{(\infty)}_{\mathrm{P},\gamma}$ exists for a suitable value γ.

Lemma 5.23. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, H be the RKHS of a measurable kernel on X, and P be a distribution on $X \times Y$. Furthermore, assume that, for some $\lambda > 0$, there exists a unique $f_{\mathrm{P},\lambda}$. If $f_{\mathrm{P},\lambda} \neq 0$, then $f^{(\infty)}_{\mathrm{P},\gamma}$ exists for $\gamma := \|f_{\mathrm{P},\lambda}\|^{-1}$ and we have $f^{(\infty)}_{\mathrm{P},\gamma} = f_{\mathrm{P},\lambda}$.*

Proof. Let us first show that $f_{\mathrm{P},\lambda}$ minimizes $\mathcal{R}_{L,\mathrm{P}}$ on $\gamma^{-1}B_H$. To this end, assume the converse, i.e., that there exists an $f \in \gamma^{-1}B_H$ with

$$\mathcal{R}_{L,\mathrm{P}}(f) < \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})\,.$$

The definition of γ then gives $\|f\| \leq \gamma^{-1} = \|f_{\mathrm{P},\lambda}\|$, and hence we find

$$\lambda\|f\|^2 + \mathcal{R}_{L,\mathrm{P}}(f) < \lambda\|f_{\mathrm{P},\lambda}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})\,. \tag{5.37}$$

Since the latter contradicts the definition of $f_{\mathrm{P},\lambda}$, we see that $f_{\mathrm{P},\lambda}$ minimizes $\mathcal{R}_{L,\mathrm{P}}$ on $\gamma^{-1}B_H$. Now assume that $f_{\mathrm{P},\lambda}$ is not a minimal norm minimizer. Then there exists an $f \in H$ with $\mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ and $\|f\| < \|f_{\mathrm{P},\lambda}\|$. Since this leads again to the false (5.37), we obtain the assertion. □

If L is a convex, integrable Nemitski loss, we have already seen that the corresponding general SVM solutions exist and depend continuously on the regularization parameter. This leads to the following corollary.

Corollary 5.24. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss, P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss, and H be the RKHS of a bounded measurable kernel on X with $\mathcal{R}^*_{L,\mathrm{P},H} < \mathcal{R}_{L,\mathrm{P}}(0)$. Then $\gamma : (0,\infty) \to (0,\infty)$ defined by $\gamma(\lambda) := \|f_{\mathrm{P},\lambda}\|^{-1}$, $\lambda > 0$, is a continuous map with*

$$f^{(\infty)}_{\mathrm{P},\gamma(\lambda)} = f_{\mathrm{P},\lambda}\,, \qquad \lambda > 0.$$

Consequently, $A_\infty : (0,\infty) \to (0,\infty)$ is an increasing and continuous map. Moreover, if there exists an f^* minimizing $\mathcal{R}_{L,\mathrm{P}}$ in H, we have $f^{(\infty)}_{\mathrm{P},\lambda} = f^*_{L,\mathrm{P},H}$ and $A_\infty(\lambda) = 0$ for all $0 \leq \lambda \leq \|f^*_{L,\mathrm{P},H}\|^{-1}$.

Proof. By Theorem 5.6 we know that $f_{\mathrm{P},\lambda} \neq 0$ for all $\lambda > 0$. The first assertion is hence a consequence of Corollary 5.19 and Lemma 5.23. The second assertion then follows from the first assertion and Lemma 2.17, and the third assertion is a consequence of the definition of A_∞. □

In order to appreciate the preceding corollary, let us assume that there does *not* exist an $f^* \in H$ minimizing $\mathcal{R}_{L,\mathrm{P}}$ in H. Then we know from Corollary 5.18 that $\lambda \mapsto \|f_{\mathrm{P},\lambda}\|$ is unbounded. Since $\lim_{\lambda\to\infty} f_{\mathrm{P},\lambda} = 0$, the intermediate value theorem then shows that for every $\gamma \in (0,\infty)$ there exists a $\lambda \in (0,\infty)$ with $\|f_{\mathrm{P},\lambda}\|^{-1} = \gamma$. Consequently, we obtain

$$\bigl\{f_{\mathrm{P},\lambda} : \lambda \in (0,\infty)\bigr\} = \bigl\{f^{(\infty)}_{\mathrm{P},\lambda} : \lambda \in (0,\infty)\bigr\}\,,$$

i.e., both approximation error functions produce the same set of minimizers, or **regularization path**, although they use a different form of regularization.

Let us finally compare the growth rates of the approximation error functions.

Theorem 5.25 (Quantitative comparison of A_2 and A_∞). *Let H be the RKHS of a measurable kernel on X, $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Then, for all $\lambda > 0$, the following statements are true:*

i) For every real number $\kappa \geq A_\infty(\lambda)$, we have

$$A_2\bigl(\lambda^2\kappa\bigr) \leq 2\kappa\,. \tag{5.38}$$

ii) For every real number $\gamma > A_2(\lambda)$, we have

$$A_\infty\bigl(\lambda^{1/2}\gamma^{-1/2}\bigr) \;\leq\; \gamma\,. \tag{5.39}$$

In addition, if $f_{\mathrm{P},\lambda}$ exists, then (5.39) also holds for $\gamma := A_2(\lambda)$.

Proof. i). For $\varepsilon > 0$ there exists an $f_\varepsilon \in \lambda^{-1}B_H$ with $\mathcal{R}_{L,\mathrm{P}}(f_\varepsilon) - \mathcal{R}^*_{L,\mathrm{P},H} \leq \kappa + \varepsilon$. If $f_\varepsilon \neq 0$, we have $\lambda \leq \|f_\varepsilon\|^{-1}$, and hence we obtain $\lambda^2\kappa \leq \|f_\varepsilon\|^{-2}\kappa =: \lambda_\varepsilon$. By the monotonicity of $A_2(\,\cdot\,)$, the latter yields

$$A_2\bigl(\lambda^2\kappa\bigr) \;\leq\; A_2(\lambda_\varepsilon) \;\leq\; \lambda_\varepsilon\|f_\varepsilon\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_\varepsilon) - \mathcal{R}^*_{L,\mathrm{P},H} \;\leq\; 2\kappa + \varepsilon\,.$$

Letting $\varepsilon \to 0$, we then obtain the assertion. The case $f_\varepsilon = 0$ can be shown analogously by setting $\lambda_\varepsilon := \lambda^2\kappa$.

ii). Since $\gamma > A_2(\lambda)$, there exists an $f \in H$ with

$$\lambda\|f\|^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H} \leq \gamma,$$

and since $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H} \geq 0$, this f satisfies $\lambda\|f\|^2 \leq \gamma$. The latter gives $\|f\| \leq (\gamma/\lambda)^{1/2} =: \kappa^{-1}$, and hence we find

$$A_\infty(\kappa) \;\leq\; \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H} \;\leq\; \lambda\|f\|^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H} \;\leq\; \gamma\,.$$

Finally, if $f_{\mathrm{P},\lambda}$ exists, we have $\lambda\|f_{\mathrm{P},\lambda}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} = A_2(\lambda)$, and hence repeating the reasoning above gives the second assertion. □

If $\lambda \mapsto A_\infty(\lambda)$ behaves polynomially, then Theorem 5.25 can be used to show that the asymptotic behavior of $\lambda \mapsto A_2(\lambda)$ is polynomial and completely determined by that of $\lambda \mapsto A_\infty(\lambda)$. We refer to Exercise 5.11 for details. For non-polynomial behavior, however, this is no longer true. Indeed, if there exists a function minimizing $\mathcal{R}_{L,\mathrm{P}}$ in H, then Lemma 5.15 and Corollary 5.18 showed that

$$\lambda A_2(1) \le A_2(\lambda) \le \lambda \|f^*_{L,\mathrm{P},H}\|_H^2\,, \qquad \lambda \in [0,1],$$

whereas Corollary 5.24 showed that $A_\infty(\lambda) = 0$ for all $0 \le \lambda \le \|f^*_{L,\mathrm{P},H}\|_H^{-1}$.

Let us finally discuss how the behavior of $A_2(\lambda)$ for $\lambda \to 0$ influences the behavior of $\|f_{\mathrm{P},\lambda}\|$ for $\lambda \to 0$. To this end, let us recall that we found

$$\|f_{\mathrm{P},\lambda}\|_H \le \sqrt{\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda}} \tag{5.40}$$

at the beginning of Section 5.1. Unfortunately, this bound is *always* suboptimal for $\lambda \to 0$. Indeed, we have $\lambda\|f_{\mathrm{P},\lambda}\|^2 \le A_2(\lambda)$, and hence we obtain

$$\|f_{\mathrm{P},\lambda}\| \le \sqrt{\frac{A_2(\lambda)}{\lambda}}\,, \qquad \lambda > 0. \tag{5.41}$$

Since $A_2(\lambda) \to 0$ for $\lambda \to 0$, the latter bound is strictly sharper than (5.40). This observation will be of great importance when investigating the stochastic properties of empirical SVM solutions in Chapter 7.

5.5 Approximation Error of RKHSs

We saw in (5.32) that, for vanishing regularization parameter, the risk of the general SVM solution tends to the minimal risk $\mathcal{R}^*_{L,\mathrm{P},H}$ of the used RKHS H. However, our ultimate interest is to investigate whether the risk of the empirical SVM tends to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$. Following the intuition that the risk of the empirical SVM solution $f_{\mathrm{D},\lambda}$ is close to that of the corresponding infinite-sample SVM solution $f_{\mathrm{P},\lambda}$, we can conjecture that $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})$ is close to $\mathcal{R}^*_{L,\mathrm{P},H}$. Moreover, in Section 6.4 we will show that this conjecture is indeed correct, and consequently we need to know that

$$\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$$

in order to show that $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})$ is close to $\mathcal{R}^*_{L,\mathrm{P}}$. The goal of this section is to establish necessary and sufficient conditions on H for this equality to hold.

Let us begin with a rather technical observation. In Chapter 3, we often used *complete* measurable spaces X in order to ensure the measurability of many considered functions such as Bayes decision functions. On the other hand, we considered continuous kernels with "large" RKHSs over (compact) metric spaces X in Section 4.6. However, in general, the Borel σ-algebra $\mathcal{B}(X)$

is *not* complete, and hence one may ask which σ-algebra we should use for computing the Bayes risk. Fortunately, it turns out by (A.8) that we have

$$\mathcal{R}^*_{L,\mathrm{P}} = \mathcal{R}^*_{L,\hat{\mathrm{P}}}\,,$$

where $\hat{\mathrm{P}}$ is the extension of P to the P-completion $\mathcal{B}_{\mathrm{P}}(X)$ of $\mathcal{B}(X)$. In other words, the Bayes risk does not change when using these different σ-algebras, and hence we can always choose the σ-algebra that best fits our needs.

In the following, we will often use intermediate approximation results where the space H in $\mathcal{R}^*_{L,\mathrm{P},H}$ is replaced by another set of measurable functions. Let us formalize this idea by the following definition.

Definition 5.26. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, P be a distribution on $X \times Y$, and $F \subset \mathcal{L}_0(X)$ be a set of measurable functions. Then the* ***minimal L-risk*** *over F is*

$$\mathcal{R}^*_{L,\mathrm{P},F} := \inf\bigl\{\mathcal{R}_{L,\mathrm{P}}(f) : f \in F\bigr\}\,.$$

Now our first result shows that, for integrable Nemitski losses, the Bayes risk can be approximated by *bounded* measurable functions.

Proposition 5.27. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss and P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss. Then we have*

$$\mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)} = \mathcal{R}^*_{L,\mathrm{P}}\,.$$

Proof. Let us fix a measurable $f : X \to \mathbb{R}$ with $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$. Then the functions $f_n := \mathbf{1}_{\{|f|\le n\}} f$, $n \ge 1$, are bounded, and an easy calculation yields

$$\begin{aligned}\bigl|\mathcal{R}_{L,\mathrm{P}}(f_n) - \mathcal{R}_{L,\mathrm{P}}(f)\bigr| &\le \int_{\{|f|>n\}\times Y} \bigl|L(x,y,0) - L\bigl(x,y,f(x)\bigr)\bigr|\,d\mathrm{P}(x,y)\\ &\le \int_{\{|f|>n\}\times Y} b(x,y) + h(0) + L\bigl(x,y,f(x)\bigr)\,d\mathrm{P}(x,y)\end{aligned}$$

for all $n \ge 1$. In addition, the integrand in the last integral is integrable since $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ and $b \in \mathcal{L}_1(\mathrm{P})$, and consequently Lebesgue's theorem yields $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}_{L,\mathrm{P}}(f)$ for $n \to \infty$. From this, we easily get the assertion. □

Using Proposition 5.27, we can now establish our first condition on F, which ensures $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$.

Theorem 5.28 (Approximation by pointwise dense sets). *Let P be a distribution on $X \times Y$, $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous P-integrable Nemitski loss, and $F \subset L_\infty(\mathrm{P}_X)$. Assume that for all $g \in L_\infty(\mathrm{P}_X)$ there exists a sequence $(f_n) \subset F$ such that $\sup_{n\ge1}\|f_n\|_\infty < \infty$ and*

$$\lim_{n\to\infty} f_n(x) \;=\; g(x)$$

for P_X-almost all $x \in X$. Then we have $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$.

Proof. By Proposition 5.27, we know that $\mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)} = \mathcal{R}^*_{L,\mathrm{P}}$, and since $F \subset L_\infty(\mathrm{P}_X)$ we also have $\mathcal{R}^*_{L,\mathrm{P},F} \geq \mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)}$. In order to show the converse inequality, we fix a $g \in L_\infty(\mathrm{P}_X)$. Let $(f_n) \subset H$ be a sequence of functions according to the assumptions of the theorem. Lemma 2.17 then yields $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}_{L,\mathrm{P}}(g)$, and hence we find $\mathcal{R}^*_{L,\mathrm{P},F} \leq \mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)}$. □

With the help of Theorem 5.28, we now show that the RKHSs of universal kernels approximate the Bayes risk of continuous, integrable Nemitski losses.

Corollary 5.29. *Let X be a compact metric space, H be the RKHS of a universal kernel on X, P be a distribution on $X \times Y$, and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous P-integrable Nemitski loss. Then we have*

$$\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}\,.$$

Proof. Let us fix a $g \in L_\infty(\mathrm{P}_X)$. By Theorem A.5.25, there then exists a sequence $(g_n) \subset C(X)$ with $\|g_n - g\|_1 \to 0$ for $n \to \infty$. By clipping g_n at $\pm\|g\|_\infty$ and considering a suitable subsequence, we see that we can assume without loss of generality that $\|g_n\|_\infty \leq \|g\|_\infty$ for all $n \geq 1$ and $g_n(x) \to g(x)$ for P_X-almost all $x \in X$. Moreover, the universality of H gives a sequence $(f_n) \subset H$ with $\|f_n - g_n\|_\infty \leq 1/n$ for all $n \geq 1$. Since this yields both $\|f_n\|_\infty \leq 1 + \|g\|_\infty$ for all $n \geq 1$ and $f_n(x) \to g(x)$ for P_X-almost all $x \in X$, we obtain the assertion by Theorem 5.28. □

The next theorem, which is a consequence of Theorem 4.61, provides a result similar to Corollary 5.29 for discrete input spaces.

Theorem 5.30. *Let X be a countable set and k be a bounded kernel on X that satisfies $k(\,\cdot\,,x) \in c_0(X)$ for all $x \in X$ and*

$$\sum_{x,x' \in X} k(x,x')f(x)f(x') > 0 \tag{5.42}$$

for all $f \in \ell_1(X)$ with $f \neq 0$. Then the RKHS H of k satisfies

$$\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$$

for all closed $Y \subset \mathbb{R}$, all distributions P on $X \times Y$, and all continuous, P-integrable Nemitski losses $L : X \times Y \times \mathbb{R} \to [0,\infty)$.

Proof. We have already seen in Theorem 4.61 that H is dense in $c_0(X)$. By Lemma 2.17, we hence find $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P},c_0(X)}$. Therefore it remains to show

$$\mathcal{R}^*_{L,\mathrm{P},c_0(X)} = \mathcal{R}^*_{L,\mathrm{P}}\,. \tag{5.43}$$

To this end, let ν be the counting measure on X and $h : X \to [0,1]$ be the map that satisfies $\mathrm{P}_X = h\nu$. In addition, recall that we have $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$

since L is a P-integrable Nemitski loss. Given an $\varepsilon > 0$, there hence exists a *finite* set $A \subset X$ with

$$\sum_{x \in X \setminus A} h(x) \int_Y L(x, y, 0) d\mathrm{P}(y|x) \leq \varepsilon \,.$$

In addition, there exists a $g : X \to \mathbb{R}$ with $\mathcal{R}_{L,\mathrm{P}}(g) \leq \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon$. Let us define $f := \mathbf{1}_A g$. Then we have $f \in c_0(X)$ and

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f) &= \sum_{x \in A} h(x) \int_Y L\big(x, y, g(x)\big) d\mathrm{P}(y|x) + \sum_{x \in X \setminus A} h(x) \int_Y L(x, y, 0) d\mathrm{P}(y|x) \\ &\leq \mathcal{R}_{L,\mathrm{P}}(g) + \varepsilon \,.\end{aligned}$$

From this we easily infer (5.43). □

If L is actually a continuous, P-integrable Nemitski loss of some order $p \in [1, \infty)$, we have already seen in Lemma 2.17 that the risk functional $\mathcal{R}_{L,\mathrm{P}}$ is even continuous on $L_p(\mathrm{P}_X)$. This leads to the following result.

Theorem 5.31 (Approximation by p-integrable functions). *Let* P *be a distribution on $X \times Y$ and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a continuous* P*-integrable Nemitski loss of order $p \in [1, \infty)$. Then, for every dense $F \subset L_p(\mathrm{P}_X)$, we have*

$$\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}} \,.$$

Proof. Since $L_\infty(\mathrm{P}_X) \subset L_p(\mathrm{P}_X)$, we have $\mathcal{R}^*_{L,\mathrm{P},L_p(\mathrm{P}_X)} = \mathcal{R}^*_{L,\mathrm{P}}$ by Proposition 5.27. Now the assertion easily follows from the denseness of F in $L_p(\mathrm{P}_X)$ and the continuity of $\mathcal{R}_{L,\mathrm{P}} : L_p(\mathrm{P}_X) \to [0, \infty)$ established in Lemma 2.17. □

Recall that we have shown in Theorem 4.63 that the RKHSs H_γ of the Gaussian RBF kernels are dense in $L_p(\mu)$ for all $p \in [1, \infty)$ and all distributions μ on $\mathbb{R}^d$. With the help of the preceding theorem, it is then easy to establish $\mathcal{R}^*_{L,\mathrm{P},H_\gamma} = \mathcal{R}^*_{L,\mathrm{P}}$ for almost all of the margin-based and distance-based loss functions considered in Sections 2.3 and 2.4. The details are left to the reader as an additional exercise.

Let us now derive some *necessary* conditions for $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$. To this end, we need the following lemma.

Lemma 5.32. *Let X be a measurable space and μ be a probability measure on X. Assume that we have a subspace $F \subset L_\infty(\mu)$ such that for all measurable $A \subset X$ there exists a sequence $(f_n) \subset F$ with*

$$\lim_{n \to \infty} f_n(x) = \mathbf{1}_A(x)$$

for μ-almost all $x \in X$. Then, for all $g \in L_\infty(\mu)$, there exists a sequence $(f_n) \subset F$ with $\lim_{n \to \infty} f_n(x) = g(x)$ for μ-almost all $x \in X$.

Proof. Observe that, for measurable step functions $g \in L_\infty(\mu)$, the assertion immediately follows from the fact that F is a vector space. Let us now fix an arbitrary $g \in L_\infty(\mu)$. For $n \geq 1$, there then exists a measurable step function $g_n \in L_\infty(\mu)$ with $\|g_n - g\|_\infty \leq 1/n$. For this g_n, there then exists a sequence $(f_{m,n})_{m\geq 1} \subset F$ with $\lim_{m\to\infty} f_{m,n}(x) = g_n(x)$ for μ-almost all $x \in X$. By Egorov's Theorem A.3.8, we find a measurable $A_n \subset X$ with $\mu(X\backslash A_n) \leq 1/n$ and

$$\lim_{m\to\infty} \|(f_{m,n} - g_n)_{|A_n}\|_\infty = 0\,.$$

Consequently, there is an index $m_n \geq 1$ with $\|(f_{m_n,n} - g_n)_{|A_n}\|_\infty \leq 1/n$. By putting all estimates together, we now obtain

$$\mu\biggl(\Bigl\{x \in X : \bigl|f_{m_n,n}(x) - g(x)\bigr| \leq \frac{2}{n}\Bigr\}\biggr) \;\geq\; 1 - \frac{1}{n}\,, \qquad\qquad n \geq 1.$$

Therefore the sequence $(f_{m_n,n})_{n\geq 1}$ converges to g in probability μ, and hence there exists a subsequence of it that converges to g almost surely. □

With the help of the preceding lemma, we can now formulate a necessary condition for *convex* supervised loss functions. For its formulation, we need to recall from Section 3.1 that $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ denotes the set of exact minimizers of the inner L-risk of Q. Moreover, recall Definition 3.5, where we said that a distribution P on $X \times Y$ is of type $\mathcal{Q}$ if $\mathcal{Q}$ is a set of distributions on Y and $\mathrm{P}(\,\cdot\,|x) \in \mathcal{Q}$ for P_X-almost all $x \in X$.

Theorem 5.33 (Pointwise denseness is necessary). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex supervised loss for which there exist two distributions Q_1 and Q_2 on Y and $t_1^*,\ t_2^* \in \mathbb{R}$ such that $t_1^* \neq t_2^*$, $\mathcal{M}_{L,\mathrm{Q}_1}(0^+) = \{t_1^*\}$, and $\mathcal{M}_{L,\mathrm{Q}_2}(0^+) = \{t_2^*\}$. Furthermore, let X be a measurable space and μ be a distribution on X. Assume that $F \subset L_\infty(\mu)$ is a subspace with*

$$\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}} \tag{5.44}$$

for all $\{\mathrm{Q}_1,\mathrm{Q}_2\}$-type distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$. Then, for all $g \in L_\infty(\mu)$, there exists a sequence $(f_n) \subset F$ such that $\lim_{n\to\infty} f_n(x) = g(x)$ for μ-almost all $x \in X$.

Proof. Let $\mathcal{A}$ be the σ-algebra of the measurable space X. We fix an $A_1 \in \mathcal{A}$ and write $A_2 := \emptyset$. Let us define two distributions P_1 and P_2 on $X \times Y$ by

$$\mathrm{P}_i(\,\cdot\,|x) := \begin{cases} \mathrm{Q}_1 & \text{if } x \in A_i \\ \mathrm{Q}_2 & \text{if } x \in X\backslash A_i \end{cases}$$

and $(\mathrm{P}_i)_X := \mu$ for $i = 1, 2$. Our assumptions on Q_1 and Q_2 guarantee $\mathcal{C}^*_{L,\mathrm{Q}_1} < \infty$ and $\mathcal{C}^*_{L,\mathrm{Q}_2} < \infty$ by Lemma 3.10, and hence we find $\mathcal{R}^*_{L,\mathrm{P}_i} < \infty$ for $i = 1, 2$. Moreover, every Bayes decision function of $\mathcal{R}_{L,\mathrm{P}_i}$, $i = 1, 2$, has μ-almost surely the form

$$f^*_{L,\mathrm{P}_i} := t^*_1 \mathbf{1}_{A_i} + t^*_2 \mathbf{1}_{X\setminus A_i}\,, \qquad i = 1, 2.$$

Now, (5.44) yields $\mathcal{R}^*_{L,\mathrm{P}_i,F} = \mathcal{R}^*_{L,\mathrm{P}_i}$, $i = 1, 2$, and hence there are sequences $(f^{(1)}_n) \subset F$ and $(f^{(2)}_n) \subset F$ with

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}_i}(f^{(i)}_n) = \mathcal{R}^*_{L,\mathrm{P}_i}\,, \qquad i = 1, 2. \tag{5.45}$$

Recalling that convex supervised losses are self-calibrated, Corollary 3.62 then shows for $i = 1, 2$ that

$$\lim_{n\to\infty} f^{(i)}_n = f^*_{L,\mathrm{P}_i} \tag{5.46}$$

in probability $\hat{\mu}$, where $\hat{\mu}$ is the extension of μ to the μ-completed σ-algebra $\mathcal{A}_\mu$, see Lemma A.3.3. Now observe that all functions in (5.46) are $\mathcal{A}$-measurable, and hence (5.46) actually holds in probability μ. Consequently, there exist subsequences $(f^{(1)}_{n_j})$ and $(f^{(2)}_{n_j})$ for which (5.46) holds μ-almost surely. For

$$f_j := \frac{1}{t^*_1 - t^*_2}\bigl(f^{(1)}_{n_j} - f^{(2)}_{n_j}\bigr)\,, \qquad j \geq 1,$$

we then have $f_j \in F$, and in addition our construction yields

$$\lim_{j\to\infty} f_j = \frac{1}{t^*_1 - t^*_2}\bigl(f^*_{L,\mathrm{P}_1} - f^*_{L,\mathrm{P}_2}\bigr) = \frac{1}{t^*_1 - t^*_2}\bigl(t^*_1 \mathbf{1}_{A_1} + t^*_2 \mathbf{1}_{X\setminus A_1} - t^*_2 \mathbf{1}_X\bigr) = \mathbf{1}_{A_1}$$

μ-almost surely. By Lemma 5.32, we thus obtain the assertion. □

For RKHSs, Theorem 5.33 has an interesting implication, which is presented in the following corollary.

Corollary 5.34 (Kernels should be strictly positive definite). *Let L be a loss function satisfying the assumptions of Theorem 5.33. Furthermore, let X be a measurable space and k be a measurable kernel on X whose RKHS H satisfies $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$ for all $\{\mathrm{Q}_1, \mathrm{Q}_2\}$-type distributions P on $X \times Y$. Then k is strictly positive definite.*

Proof. Let $x_1, \ldots, x_n \in X$ be mutually different points and μ be the associated empirical distribution. Obviously, it suffices to show that the kernel matrix $K := (k(x_i, x_j))$ has full rank. Let us assume the converse, i.e., we assume that there exists an $y \in \mathbb{R}^n$ with $K\alpha \neq y$ for all $\alpha \in \mathbb{R}^n$. Since $K\mathbb{R}^n$ is closed, there then exists an $\varepsilon > 0$ with $\|K\alpha - y\|_\infty \geq \varepsilon$ for all $\alpha \in \mathbb{R}^n$. Now note that every $\bar{f} \in H$ that is orthogonal to $\operatorname{span}\{k(\,\cdot\,, x_i) : i = 1, \ldots, n\}$ satisfies

$$\bar{f}(x_j) = \langle \bar{f}, k(\,\cdot\,, x_j)\rangle_H = 0\,, \qquad j = 1, \ldots, n.$$

By decomposing H into $\operatorname{span}\{k(\,\cdot\,, x_i) : i = 1, \ldots, n\}$ and its orthogonal complement, we consequently see that for every $f \in H$ there is an $\alpha \in \mathbb{R}^n$ with

$$f(x_j) = \sum_{i=1}^n \alpha_i k(x_j, x_i)\,, \qquad j = 1, \ldots, n,$$

and hence for all $f \in H$ there is an index $j \in \{1, \ldots, n\}$ with $|f(x_j) - y_j| > \varepsilon$. On the other hand, k is bounded on $\{1, \ldots, n\}$, and hence Theorem 5.33 gives a sequence $(f_n) \subset H$ with $f_n(x_i) \to y_i$ for all $i \in \{1, \ldots, n\}$. From this we easily find a contradiction. □

In order to illustrate the previous results, let us assume for a moment that we have a distribution μ on X and a convex loss $L : Y \times \mathbb{R} \to [0, \infty)$ that satisfy the assumptions of both Theorem 5.28 and Theorem 5.33. In addition, let $F \subset L_\infty(\mu)$ be a subspace. Now assume that we are interested in the question of whether

$$\mathcal{R}^*_{L,\mathrm{P},F} \;=\; \mathcal{R}^*_{L,\mathrm{P}} \tag{5.47}$$

holds for a reasonably large class of distributions P with $\mathrm{P}_X = \mu$. Theorem 5.33 then shows that a *necessary* condition for (5.47) to hold is that F be "dense" in $L_\infty(\mu)$ with respect to the μ-almost sure convergence[2], i.e., every $g \in L_\infty(\mu)$ is the μ-almost sure limit of a suitable sequence $(f_n) \subset F$. However, the *sufficient* condition of Theorem 5.28 for (5.47) to hold requires that there actually be such a sequence (f_n) that is *uniformly* bounded. In other words, there is a gap between the necessary and the sufficient conditions. Now recall that in Theorem 5.31 we have already weakened the sufficient condition for a more restricted class of loss functions. In the following, we will present a stronger necessary condition by making additional assumptions on the distributions Q_1 and Q_2 of Theorem 5.33, so in the end we obtain characterizations on F for (5.47) to hold for a variety of important loss functions. Let us begin with the following simple lemma.

Lemma 5.35. *Let X be a measurable space, μ be a probability measure on X, and $q > 0$. Assume that we have a subspace $F \subset L_q(\mu)$ such that for all measurable $A \subset X$ there exists a sequence $(f_n) \subset F$ with*

$$\lim_{n\to\infty} \|f_n - \mathbf{1}_A\|_{L_q(\mu)} = 0\,. \tag{5.48}$$

Then F is dense in $L_q(\mu)$.

Proof. If $g \in L_q(\mu)$ is a measurable step function, there obviously exists a sequence $(f_n) \subset F$ with $\lim_{n\to\infty} \|f_n - g\|_q = 0$. Moreover, if $g \in L_q(\mu)$ is bounded and n is an integer, there exists a measurable step function g_n with $\|g_n - g\|_\infty \le 1/n$. In addition, we have just seen that there exists an $f_n \in F$ with $\|f_n - g_n\|_q \le 1/n$, and hence we find $\lim_{n\to\infty} \|f_n - g\|_q = 0$. Finally, for general $g \in L_q(\mu)$, we find an approximating sequence by first approximating g with the bounded measurable functions $g_n := \mathbf{1}_{|g|\le n} g$, $n \ge 1$, and then approximating these g_n with suitable functions $f_n \in F$. □

[2] Note that in general there is no topology generating the almost sure convergence, and thus "dense" is not really defined. However, the almost sure convergence in Theorem 5.33 and Theorem 5.28 can be replaced by the convergence in probability (see Exercise 5.12), and since this convergence originates from a metric, we then have a precise meaning of "dense".

With the help of the preceding lemma, we can now establish a stronger necessary condition on F for $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$ to hold. For its formulation, we need to recall the self-calibration function defined in (3.67).

Theorem 5.36 (Denseness in $L_q(\mu)$ is necessary). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex supervised loss such that there exist two distributions Q_1 and Q_2 on Y and $t_1^*, t_2^* \in \mathbb{R}$ with $t_1^* \neq t_2^*$, $\mathcal{M}_{L,\mathrm{Q}_1}(0^+) = \{t_1^*\}$ and $\mathcal{M}_{L,\mathrm{Q}_2}(0^+) = \{t_2^*\}$. In addition, assume that their self-calibration functions satisfy*

$$\delta_{\max,\check{L},L}(\varepsilon, \mathrm{Q}_i) \geq B\varepsilon^q\,, \qquad \varepsilon > 0,\ i = 1,2,$$

for some constants $B > 0$ and $q > 0$. Furthermore, let X be a measurable space, μ be a distribution on X, and $F \subset L_q(\mu)$ be a subspace with

$$\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$$

for all $\{\mathrm{Q}_1, \mathrm{Q}_2\}$-type distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$. Then F is dense in $L_q(\mu)$.

Proof. Following the argument used in the proof of Theorem 5.33, we may assume without loss of generality that X is a complete measurable space. Let us now fix a measurable $A_1 \subset X$ and write $A_2 := \emptyset$. Furthermore, we define the distributions P_i, their Bayes decision functions f^*_{L,P_i}, and the approximating sequences $(f_n^{(i)}) \subset F$, $i = 1,2$, as in the proof of Theorem 5.33. Then (5.45) together with (3.69) for $p := \infty$ yields

$$\lim_{n\to\infty} \|f_n^{(i)} - f^*_{L,\mathrm{P}_i}\|_{L_q(\mu)} = 0\,.$$

For $f_n := \frac{1}{t_1^* - t_2^*}\bigl(f_n^{(1)} - f_n^{(2)}\bigr), n \geq 1$, we then obtain $\lim_{n\to\infty} \|f_n - \mathbf{1}_{A_1}\|_{L_q(\mu)} = 0$, and hence we obtain the assertion by Lemma 5.35. □

By combining Theorem 5.31 with Theorem 5.36, we now obtain the following characterization of subspaces F satisfying $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$.

Corollary 5.37 (Characterization). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex supervised Nemitski loss of order $p \in [1,\infty)$, i.e.*

$$L(y,t) \leq b(y) + c|t|^p\,, \qquad (y,t) \in Y \times \mathbb{R}, \tag{5.49}$$

for a suitable constant $c > 0$ and a measurable function $b : Y \to [0,\infty)$. Furthermore, let Q_1 and Q_2 be distributions on Y with $b \in L_1(\mathrm{Q}_1) \cap L_1(\mathrm{Q}_2)$ and $t_1^*, t_2^* \in \mathbb{R}$ with $t_1^* \neq t_2^*$, $\mathcal{M}_{L,\mathrm{Q}_1}(0^+) = \{t_1^*\}$ and $\mathcal{M}_{L,\mathrm{Q}_2}(0^+) = \{t_2^*\}$. In addition, assume that their self-calibration functions satisfy

$$\delta_{\max,\check{L},L}(\varepsilon, \mathrm{Q}_i) \geq B\,\varepsilon^p\,, \qquad \varepsilon > 0,\ i = 1,2,$$

for some constant $B > 0$. Furthermore, let X be a measurable space, μ be a distribution on X, and $F \subset L_p(\mu)$ be a subspace. Then the following statements are equivalent:

i) F is dense in $L_p(\mu)$.
ii) For all distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$ for which L is a P-integrable Nemitski loss of order $p \in [1, \infty)$, we have $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$.
iii) For all $\{\mathrm{Q}_1, \mathrm{Q}_2\}$-type distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$, we have $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$.

Proof. *i)* $\Rightarrow$ *ii)*. Use Theorem 5.31.

ii) $\Rightarrow$ *iii)*. By (5.49) and $b \in L_1(\mathrm{Q}_1) \cap L_1(\mathrm{Q}_2)$, we see that L is a P-integrable Nemitski loss of order $p \in [1, \infty)$ for all $\{\mathrm{Q}_1, \mathrm{Q}_2\}$-type distributions P on $X \times Y$.

iii) $\Rightarrow$ *i)*. Use Theorem 5.36. □

Let us now illustrate that many important losses satisfy the assumptions of Corollary 5.37. For some more examples, we refer to Exercise 5.13.

Example 5.38. For $p \geq 1$, let L be the p-th **power absolute distance loss** defined in Example 2.39. Moreover, let $\mathrm{Q}_1 := \delta_{\{y_1\}}$ and $\mathrm{Q}_1 := \delta_{\{y_2\}}$ be two Dirac distributions on $\mathbb{R}$ with $y_1 \neq y_2$. Then L, Q_1, and Q_2 satisfy the assumptions of Corollary 5.37.

In order to see this, recall that L is a Nemitski loss of order p by Example 2.39 and Lemma 2.36. Furthermore, for the Dirac measure $\mathrm{Q} := \delta_{\{y\}}$ at an arbitrary $y \in \mathbb{R}$, we have $\mathcal{C}_{L,\mathrm{Q}}(t) = |t - y|^p$, $t \in \mathbb{R}$. Consequently, we have $\mathcal{C}^*_{L,\mathrm{Q}} = 0$ and $\mathcal{M}_{L,\mathrm{Q}}(0^+) = \{y\}$. With these equalities, it is easy to check that the self-calibration function of L is

$$\delta_{\max,\check{L},L}(\varepsilon, \mathrm{Q}) = \varepsilon^p\,, \qquad \varepsilon \geq 0.$$

Since $y_1 \neq y_2$, we then see that the assumptions of Corollary 5.37 are satisfied, and hence we have $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$ if and only if F is dense in $L_p(\mathrm{P}_X)$. Moreover, it also shows that restricting the class of distributions to noise-free distributions P, i.e., to distributions with $\mathrm{P}(\,\cdot\,|x) = \delta_{\{g(x)\}}$ for measurable $g : X \to \mathbb{R}$, does not change this characterization. ◁

Example 5.39. Let $\epsilon > 0$ and L be the **ϵ-insensitive loss** defined in Example 2.42. Moreover, for $y_1 \neq y_2$, we define $\mathrm{Q}_i := \frac{1}{2}\delta_{\{y_i - \epsilon\}} + \frac{1}{2}\delta_{\{y_i + \epsilon\}}$, $i = 1, 2$. Then L, Q_1, and Q_2 satisfy the assumptions of Corollary 5.37 for $p = 1$.

In order to see this, we first recall with Example 2.42 that L is a Nemitski loss of order 1. Let us define $\psi(r) := \max\{0, |r| - \epsilon\}$, $r \in \mathbb{R}$. Then we have

$$2\mathcal{C}_{L,\mathrm{Q}_i}(t) = \psi(y_i - \epsilon - t) + \psi(y_i + \epsilon - t)\,, \qquad t \in \mathbb{R}\,,\, i = 1, 2,$$

and thus we have $\mathcal{C}_{L,\mathrm{Q}_i}(y_i) = 0 \leq \mathcal{C}_{L,\mathrm{Q}_i}(t)$ for all $t \in \mathbb{R}$. For $t \geq 0$, this yields

$$\mathcal{C}_{L,\mathrm{Q}_i}(y_i \pm t) - \mathcal{C}^*_{L,\mathrm{Q}_i} = \frac{1}{2}\psi(\epsilon + t) + \frac{1}{2}\psi(\epsilon - t) \geq \frac{1}{2}\psi(\epsilon + t) = \frac{t}{2}\,,$$

and hence we find $\mathcal{M}_{L,\mathrm{Q}_i}(0^+) = \{y_i\}$ and $\delta_{\max,\check{L},L}(\varepsilon, \mathrm{Q}) \geq \varepsilon/2$ for $\varepsilon \geq 0$. ◁

The following example shows a similar result for the hinge loss, which is used as a surrogate for the classification loss.

Example 5.40. Let L be the **hinge loss** defined in Example 2.27. Furthermore, let Q_1, Q_2 be distributions on $Y := \{-1, 1\}$ with $\eta_1 := \mathrm{Q}_1(\{1\}) \in (0, 1/2)$ and $\eta_2 := \mathrm{Q}_2(\{1\}) \in (1/2, 1)$. Then L, Q_1, and Q_2 satisfy the assumptions of Corollary 5.37 for $p = 1$.

In order to see this, recall that L_{hinge} is a Lipschitz continuous loss, and hence it is a Nemitski loss of order 1 by (2.11). Moreover, Example 3.7 shows that $\mathcal{M}_{L_{\text{hinge}},\eta}(0^+) = \{\operatorname{sign}(2\eta - 1)\}$ for $\eta \neq 0, \frac{1}{2}, 1$, and Exercise 3.15 shows that

$$\delta_{\max,L_{\text{hinge}}}(\varepsilon, \eta) = \varepsilon \, \min\{\eta, 1 - \eta, 2\eta - 1\}\,, \qquad \varepsilon \geq 0. \qquad \triangleleft$$

Note that, unlike the distributions in Example 5.38, the distributions in Example 5.40 are noisy. This is due to the fact that only noise makes the hinge loss minimizer unique (see Example 3.7 and Figure 3.1 on page 54). Moreover, note that using, e.g., the least squares loss for a classification problem requires $L_2(\mu)$-denseness, which in general is a strictly stronger condition than the $L_1(\mu)$-denseness, which is sufficient for the hinge loss and the logistic loss for classification. This is somewhat remarkable since the target functions for the former two losses are bounded, whereas in general the target function for the logistic loss for classification is not even integrable.

We saw in Theorem 5.30 that "strict positive definiteness of k for infinite sequences" is *sufficient* for $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$. On the other hand, "strict positive definiteness for finite sequences" is *necessary* for $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$ by Corollary 5.34, and hence it seems natural to ask whether the latter, weaker condition is also sufficient. However, with the developed theory, it is easy to check that in general this is not the case. Indeed, recall that we saw in Theorem 4.62 that there exists a strictly positive definite kernel whose RKHS H is not dense in the spaces $L_1(\mu)$. Consequently, this RKHS cannot satisfy $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$ for the loss functions considered in Examples 5.38 to 5.40.

Let us finally present an example of a set of bounded continuous functions that drastically fails to have good universal approximation properties.

Proposition 5.41. *Let $Y := \{-1, 1\}$ and $\mathcal{P}^d_n$ be the set of all polynomials on $X := [0, 1]^d$ whose degree is less than $n + 1$. Then, for all $\varepsilon > 0$, there exists a distribution P on $X \times Y$ with $\mathcal{R}^*_{L_{\text{class}},\mathrm{P}} = 0$ and*

$$\mathcal{R}^*_{L_{\text{class}},\mathrm{P},\mathcal{P}^d_n} \;\geq\; \frac{1}{2} - \varepsilon \;.$$

Moreover, P can be chosen such that the "classes" $\{x \in X : \mathrm{P}(y = 1|x) = 0\}$ and $\{x \in X : \mathrm{P}(y = 1|x) = 1\}$ have strictly positive distance.

Proof. We first treat the case $d = 1$. To this end, we fix an integer $m \geq (3n + 2)/\varepsilon$ and write

$$I_i := [(i+\varepsilon)/m, (i+1-\varepsilon)/m]\,, \qquad i = 0, \ldots, m-1.$$

Moreover, let μ_{I_i} be the Lebesgue measure on I_i and let P be defined by

$$\mathrm{P}_X := (1-2\varepsilon)^{-1} \sum_{i=0}^{m-1} \mu_{I_i}\,,$$

$$\mathrm{P}\bigl(y=1\big|x\bigr) := \begin{cases} 1 & \text{if } x \in I_i \text{ for some even } i \in \{0,\ldots,m-1\} \\ 0 & \text{otherwise.} \end{cases}$$

For a fixed polynomial $f \in \mathcal{P}_n^1$, we now write $x_1 < \ldots < x_k$, $k \le n$ for its mutually different and ordered zeros in $(0,1)$. In addition, we write $x_0 := 0$ and $x_{k+1} := 1$. Finally, we define $a_j := |\{i : I_i \subset [x_j, x_{j+1}]\}|$ for $j = 0, \ldots, k$. Obviously, there are at most k intervals I_i that do not lie between two consecutive zeros, and hence we get

$$\sum_{j=0}^{k} a_j \;\ge\; m-k \;\ge\; m-n\,.$$

Moreover, at most $\lfloor (a_j+1)/2 \rfloor$ intervals I_i are correctly classified on $[x_j, x_{j+1}]$ by the function $\operatorname{sign} \circ f$, and consequently at least

$$\sum_{j=0}^{k} \Bigl\lfloor \frac{a_j}{2} \Bigr\rfloor \ge \sum_{j=0}^{k} \Bigl(\frac{a_j}{2} - 1\Bigr) \ge \frac{m-n}{2} - (k+1) \ge \frac{m-3n-2}{2}$$

intervals I_i are not correctly classified on $[0,1]$ by $\operatorname{sign} \circ f$. Since $\mathrm{P}_X(I_i) = 1/m$, we thus obtain

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) \;\ge\; \frac{1}{m} \sum_{j=0}^{k} \Bigl\lfloor \frac{a_j}{2} \Bigr\rfloor \;\ge\; \frac{1}{2} - \frac{3n+2}{m} \;\ge\; \frac{1}{2} - \varepsilon\,.$$

Finally, for the case $d > 1$, let $I : [0,1] \to [0,1]^d$ be the embedding defined by $t \mapsto (t, 0, \ldots, 0)$, $t \in \mathbb{R}$. Moreover, consider the above distribution P embedded into $[0,1]^d$ via I. Then observe that, given an $f \in \mathcal{P}_n^d$, its restriction $f_{|I([0,1])}$ can be interpreted as a polynomial in $\mathcal{P}_n^1$, and from this it is straightforward to prove the assertion. □

5.6 Further Reading and Advanced Topics

The first representer theorem for empirical distributions was established for some specific losses, including the least squares loss, by Kimeldorf and Wahba (1971). The version we presented in Theorem 5.5 is a simplified version of a more general representer theorem for empirical distributions proved by

Schölkopf *et al.* (2001b). For a brief discussion on some related results, we refer to these authors as well as to the book by Schölkopf and Smola (2002). Finally, Dinuzzo *et al.* (2007) recently established a refinement of Theorem 5.8 for empirical distributions.

The first considerations on general SVM solutions were made by Zhang (2001). In particular, he showed that the general SVM solutions exist for the hinge loss and that they converge to the hard margin SVM solution for $\lambda \to 0$ (see Exercise 5.10 for a precise statement). Moreover, he also mentions the general representer theorem for the hinge loss, though he does not prove it. The existence of general SVM solutions for a wide class of L_{class}-surrogates was then established by Steinwart (2005), and a corresponding representer theorem was established by Steinwart (2003). The existence of general SVM solutions for convex integrable Nemitski losses of some type $p \geq 1$ as well as the corresponding general representer theorem was then shown by De Vito *et al.* (2004). The results presented in Section 5.2 are closely related to their findings, although their representer theorem is stated for supervised losses and for SVM optimization problems having an additional "offset".

The Lipschitz continuity of general SVM solutions we presented in Section 5.3 was again found by Zhang (2001) for differentiable losses. Independently, Bousquet and Elisseeff (2002) established a similar result for empirical SVM solutions (see Exercise 5.6). The presentation given in Section 5.3 mainly follows that of Christmann and Steinwart (2007).

The relation between the approximation error functions A_2 and A_∞ was investigated by Steinwart and Scovel (2005a). The presentation of the corresponding part of Section 5.4 closely follows their work, though it is fair to say that some of the results, such as Lemma 5.23, are to some extent folklore. Furthermore, it is interesting to note that the functions A_2 and A_∞ are often closely related to concepts from approximation theory. To be more precise, let E and F be Banach spaces such that $E \subset F$ and $\mathrm{id} : E \to F$ is continuous. Then the **K-functional** is defined by

$$K(y,t) := \inf_{x \in E} t\|x\|_E + \|y - x\|_F \,, \qquad y \in F,\, t > 0.$$

Moreover, for $r \in (0,1)$, the **interpolation space** $(E,F)_r$ is the set of elements $y \in F$ for which

$$\|y\|_{(E,F)_r} := \sup_{t>0} K(y,t)t^{-r} < \infty.$$

One can show that $((E,F)_r, \|\cdot\|_{(E,F)_r})$ is a Banach space, and for many classical spaces E and F this interpolation space can actually be described in closed form. We refer to the books by Bergh and Löfström (1976), Triebel (1978), and Bennett and Sharpley (1988). In particular, for a Euclidean ball $X \subset \mathbb{R}^d$, Theorem 7.31 of Adams and Fournier (2003) shows that the Sobolev space $W^k(X)$ is continuously embedded into $(W^m(X), L_2(X))_{k/m}$ for all $0 < k < m$ and that this embedding is almost sharp. Besides the K-functional, one can also consider the functional

$$A(y,t) := \inf_{x \in tB_E} \|x - y\|_F \,, \qquad y \in F,\, t > 0.$$

Not surprisingly, both functionals are closely related. Indeed, with techniques similar to those of Theorem 5.25, Smale and Zhou (2003) showed that

$$y \in (E,F)_r \qquad \Longleftrightarrow \qquad c := \sup_{t>0} A(y,t) t^{\frac{r}{1-r}} < \infty \,,$$

and in this case we actually have $c^{1-r} \leq \|y\|_{(E,F)_r} \leq 2c^{1-r}$. Let us now illustrate how these abstract results relate to the approximation error functions. To this end, let us first assume that L is the least squares loss. Moreover, we fix an RKHS H over X with bounded measurable kernel and a distribution P on $X \times \mathbb{R}$ with $|\mathrm{P}|_2 < \infty$. Then we have $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} = \|f - f^*_{L,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)}$, and by setting $E := H$ and $F := L_2(\mathrm{P}_X)$, we thus find $A_\infty(t) = A^2(f^*_{L,\mathrm{P}}, t^{-1})$ for all $t > 0$. Using the relation between A_∞ and A_2, we conclude that

$$f^*_{L,\mathrm{P}} \in \bigl(H, L_2(\mathrm{P}_X)\bigr)_r \qquad \Longleftrightarrow \qquad \exists\, c \geq 0 \;\forall \lambda > 0 : A_2(\lambda) \leq c\lambda^r \,,$$

and if $f^*_{L,\mathrm{P}} \in (H, L_2(\mathrm{P}_X))_r$ we may actually choose $c := \|f^*_{L,\mathrm{P}}\|^2_{(H,L_2(\mathrm{P}_X))_r}$. In particular, if $H = W^m(X)$, $f^*_{L,\mathrm{P}} \in W^k(X)$ for some $k < m$, and P_X is the uniform distribution on X, then there exists a constant $c > 0$ that is independent of $f^*_{L,\mathrm{P}}$ such that

$$A_2(\lambda) \leq c\, \|f^*_{L,\mathrm{P}}\|^2_{W^k(X)} \lambda^{k/m} \,, \qquad \lambda > 0 \,.$$

In this direction, it is also interesting to note that Smale and Zhou (2003) showed that, given a *fixed* width $\gamma > 0$, the approximation error function of the corresponding Gaussian RBF kernel can only satisfy a non-trivial bound of the form $A_2(\lambda) \leq c\lambda^\beta$ if $f^*_{L,\mathrm{P}}$ is C^∞. Furthermore, for the least squares loss, the approximation error function is also related to the integral operator of the kernel of H. To explain this, let X be a compact metric space and $T_k : L_2(\mathrm{P}_X) \to L_2(\mathrm{P}_X)$ be the integral operator associated to the kernel k and the measure P_X. Then one can show (see, e.g., Theorem 4.1 in Cucker and Zhou, 2007) that

$$f^*_{L,\mathrm{P}} \in T_k^{r/2}(L_2(\mathrm{P}_X)) \qquad \Longrightarrow \qquad \exists\, c \geq 0 \;\forall \lambda > 0 : A_2(\lambda) \leq c\lambda^r \,.$$

In addition, this theorem also shows that if the latter estimate on A_2 holds and $\operatorname{supp} \mathrm{P}_X = X$, then we have $f^*_{L,\mathrm{P}} \in T_k^{(r-\varepsilon)/2}(L_2(\mathrm{P}_X))$ for all $\varepsilon > 0$. Finally, let us assume that L is a Lipschitz continuous loss for which a Bayes decision function $f^*_{L,\mathrm{P}}$ exists. Then we have $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \leq \|f - f^*_{L,\mathrm{P}}\|_{L_1(\mathrm{P}_X)}$, and by setting $E := H$ and $F := L_1(\mathrm{P}_X)$ we thus find $A_\infty(t) \leq A(f^*_{L,\mathrm{P}}, t^{-1})$ for all $t > 0$. From this we conclude that

$$f^*_{L,\mathrm{P}} \in \bigl(H, L_1(\mathrm{P}_X)\bigr)_r \qquad \Longrightarrow \qquad \forall \lambda > 0 : A_2(\lambda) \leq \|f^*_{L,\mathrm{P}}\|^{\frac{2}{2-r}}_{(H,L_1(\mathrm{P}_X))_r} \lambda^{\frac{r}{2-r}} \,.$$

Qualitative approximation properties of function classes in terms of excess risks have been investigated for quite a while. For example, certain neural network architectures have been known for more than 15 years to be universal approximators. For some information in this regard, we refer to p. 518f of the book by Devroye *et al.* (1996). Moreover, the characterization for the least squares loss given in Example 5.38 is, of course, trivial if we recall the formula for the excess least squares risk given in Example 2.6. Moreover, Example 5.38 can also be shown without using the self-calibration function. Furthermore, the fact that $L_\infty(\mu)$-denseness is sufficient for most commonly used loss functions is also to some extent folklore. The more sophisticated sufficient condition given in Theorem 5.31 together with the general necessary conditions found in Section 5.5 are taken from Steinwart *et al.* (2006b). Finally, Corollary 5.29 together with the universality of certain kernels was first shown by Steinwart (2001, 2005) and independently by Hammer and Gersmann (2003).

5.7 Summary

In this chapter, we investigated general SVM solutions and their properties. To this end, we showed in the first section that for a large class of convex losses these solutions exist and are unique. Being motivated by the representer theorem for empirical solutions, we then established a general representer theorem, which additionally describes the form of the representing function, in the second section. In the third section, we used this general representer theorem to show that the general SVM solutions depend on the underlying distribution in a Lipschitz continuous fashion.

In the fourth section, we investigated the behavior of the general SVM solution and its associated (regularized) risk for vanishing regularization parameters. In particular, we showed that this risk tends to the best possible risk obtainable in the RKHS. In addition, we compared the regularization scheme of SVMs with a more classical notion of approximation error.

In the last section, we investigated under which assumptions the RKHS is rich enough to achieve the Bayes risk. Here we first derived a sufficient condition for universal kernels. We then obtained a characterization for a large class of loss functions and illustrated these findings with some examples.

5.8 Exercises

5.1. Existence of general SVM solutions for pinball loss ($\star$)
Formulate a condition on P that ensures the existence of a general SVM solution when using the pinball loss. Does this condition change for different values of τ?

5.2. A representer theorem for the logistic loss ($\star\star$)
Consider a strictly positive definite kernel and the logistic loss for classification. Show that all coefficients in (5.6) do not vanish.

5.3. A representer theorem for some distance-based losses ($\star\star$)
Formulate a generalized representer theorem for the loss functions $L_{p\text{-dist}}$, $p \geq 1$, and $L_{\text{r-logist}}$. Compare the different norm bounds for the representing function h. What does (5.10) mean for the least squares loss?

5.4. Generalized representer theorem for margin-based losses ($\star\star$)
Formulate a generalized representer theorem for convex, margin-based loss functions. How can (5.19) and (5.21) be simplified?

5.5. Generalized representer theorem for the hinge loss ($\star\star\star$)
Formulate a generalized representer theorem for the hinge loss. What is the form of (5.19) and (5.21)? Then assume that P is an empirical distribution with respect to a data set $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$. Investigate the form of the coefficients in (5.6) and describe when a specific coefficient vanishes. Finally, find a condition on the kernel that ensures that the coefficients are uniquely determined.

5.6. Stability of empirical SVM solutions ($\star\star$)
Let $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ be a sample set and $\bar{D} := ((x_1, y_1), \ldots, (x_{n-1}, y_{n-1})) \in (X \times Y)^{n-1}$ be the sample set we obtain from D by removing the last sample. Furthermore, let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss function, H be a RKHS over X with canonical feature map $\Phi : X \to H$, and $\lambda > 0$. Show that

$$\left\| f_{D,\lambda} - f_{\bar{D},\lambda} \right\|_H \;\leq\; \frac{2\|k\|_\infty}{\lambda n} \cdot |L|_{B_\lambda,1} \,,$$

where B_λ is defined by (5.29).

5.7. Existence of general SVM solutions for non-convex losses ($\star\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a continuous loss function and P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss. Furthermore, let H be the RKHS of a bounded measurable kernel over X. Show that for all $\lambda > 0$ there exists a general SVM solution $f_{P,\lambda}$.

Hint: Adapt the technique used in the proof of Theorem 5.17.

5.8. Approximation error function without RKHSs ($\star\star\star$)
Investigate which results from Section 5.4 still hold if the RKHS H in the definition of the approximation error functions is replaced by a general normed space consisting of measurable functions $f : X \to \mathbb{R}$.

5.9. Approximation error function with general exponent ($\star\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function, H be the RKHS of a measurable

kernel over X, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. For $p \geq 1$, define the *p-approximation error function* $A_p : [0,\infty) \to [0,\infty)$ by

$$A_p(\lambda) := \inf_{f \in H} \lambda \|f\|_H^p + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H}\,, \qquad \lambda \geq 0.$$

Establish a (suitably modified) version of Lemma 5.15. Furthermore, show the uniqueness and existence of a minimizer of $A_p(\lambda)$ under the assumptions of Lemma 5.1 and Theorem 5.2. Then establish (suitably modified) versions of Theorem 5.17 and its Corollaries 5.18, 5.19, and 5.24. Finally, discuss the consequences of the latter with respect to variable exponent p.

5.10. SVM behavior for classes with strictly positive distances ($\star\star$)
Let X be a compact metric space, $Y := \{-1, 1\}$, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}} = 0$. Let us write $\eta(x) := \mathrm{P}(y = 1|x)$, $x \in X$. Assume that the "classes"

$$\{x \in X : \eta = 0\} \cap \operatorname{supp} \mathrm{P}_X$$

and

$$\{x \in X : \eta = 1\} \cap \operatorname{supp} \mathrm{P}_X$$

have strictly positive distance and that H is the RKHS of a universal kernel on X. Show that $f^*_{L_{\mathrm{hinge}},\mathrm{P},H}$ exists and that the general SVM solutions with respect to the hinge loss satisfy $\lim_{\lambda \to 0^+} f_{L_{\mathrm{hinge}},\mathrm{P},H,\lambda} = f^*_{L_{\mathrm{hinge}},\mathrm{P},H}$.

5.11. Same behavior of different approximation functions ($\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss function, H be the RKHS of a measurable kernel over X, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Moreover, for $p \geq 1$, define the *p-approximation error function* $A_p : [0,\infty) \to [0,\infty)$ as in Exercise 5.9. Finally, let $c > 0$, $\lambda_0 > 0$, and $\alpha > 0$ be fixed constants. Show the following statements:

i) $A_p(\lambda) \leq c^{\frac{p}{\alpha+p}} \lambda^{\frac{\alpha}{\alpha+p}}$ for all $\lambda > 0$ implies $A_\infty(\lambda) \leq c\lambda^\alpha$ for all $\lambda > 0$.
ii) $A_\infty(\lambda) \leq c\lambda^\alpha$ for all $\lambda > 0$ implies $A_p(\lambda) \leq 2c^{\frac{p}{\alpha+p}} \lambda^{\frac{\alpha}{\alpha+p}}$ for all $\lambda > 0$.
iii) $A_p(\lambda) \geq 2c^{\frac{p}{\alpha+p}} \lambda^{\frac{\alpha}{\alpha+p}}$ for all $\lambda \in [0, \lambda_0]$ implies $A_\infty(\lambda) \geq c\lambda^\alpha$ for all $\lambda > 0$ with $\lambda^p A_\infty(\lambda) \leq \lambda_0$.
iv) $A_\infty(\lambda) \geq c\lambda^\alpha$ for all $\lambda \in [0, \lambda_0]$ implies $A_p(\lambda) \geq c^{\frac{p}{\alpha+p}} \lambda^{\frac{\alpha}{\alpha+p}}$ for all $\lambda > 0$ with $\frac{\lambda}{A_p(\lambda)} \leq \lambda_0^p$.

With the help of these estimates, discuss the optimality of (5.41) for $p = 2$.
Hint: First prove an analogue to Theorem 5.25.

5.12. Some other conditions for universal approximators ($\star\star$)
Show that Theorem 5.28 still holds true if we only assume that for all $g \in L_\infty(\mathrm{P}_X)$ there exists a sequence $(f_n) \subset F$ with $\sup_{n \geq 1} \|f_n\|_\infty < \infty$ and $\lim_{n \to \infty} f_n = g$ in probability P_X.

5.13. Universal approximators for some margin-based losses ($\star\star$)
Discuss the squared hinge loss and the least squares loss in the sense of Example 5.40.

6

Basic Statistical Analysis of SVMs

Overview. *So far we have not considered the fact that SVMs typically deal with observations from a random process. This chapter addresses this issue by so-called oracle inequalities that relate the risk of an empirical SVM solution to that of the corresponding infinite-sample SVM. In particular, we will see that the analysis of the learning ability of SVMs can be split into a statistical part described by the oracle inequalities and a deterministic part based on the approximation error function investigated in the previous chapter.*

Prerequisites. *The first three sections require only basic knowledge of probability theory as well as some notions from the introduction in Chapter 1 and Sections 2.1 and 2.2 on loss functions. In the last two sections, we additionally need Sections 4.2, 4.3, and 4.6 on kernels and Chapter 5 on infinite-sample SVMs.*

Usage. *The oracle inequalities of this chapter are necessary for Chapter 8 on classification. In addition, they are helpful in Chapter 11, where practical strategies for selecting hyper parameters are discussed.*

Let us recall from the introduction that the goal of learning from a training set D is to find a decision function f_D such that $\mathcal{R}_{L,\mathrm{P}}(f_D)$ is close to the minimal risk $\mathcal{R}^*_{L,\mathrm{P}}$. Since we typically assume that the empirical data set D consists of i.i.d. observations from an unknown distribution P, the decision function f_D and its associated risk $\mathcal{R}_{L,\mathrm{P}}(f_D)$ become random variables. Informally, the "learning ability" of a learning method $D \mapsto f_D$ can hence be described by an answer to the following question:

$$\textit{What is the probability that } \mathcal{R}_{L,\mathrm{P}}(f_D) \textit{ is close to } \mathcal{R}^*_{L,\mathrm{P}}\textit{?}$$

The main goal of this chapter is to present basic concepts and techniques for addressing this question for SVMs. To this end, we introduce two key notions of statistical learning in Section 6.1, namely *consistency* and *learning rates*, that formalize possible answers to the question above. While consistency is of purely asymptotic nature, learning rates provide a framework that is more closely related to practical needs. On the other hand, we will see in Section 6.1 that consistency can often be ensured *without* assumptions on P, while learning with guaranteed rates *almost always* requires assumptions on the unknown

distribution P. In the second section, we then establish some basic concentration inequalities that will be the key tools for investigating the statistical properties of SVMs in this chapter. Subsequently we will illustrate their use in Section 6.3, where we investigate empirical risk minimization. The last two sections are devoted to the actual statistical analysis of SVMs. In Section 6.4, we establish two *oracle inequalities* that, for a fixed regularization parameter, relate the risk of empirical SVM decision functions to the approximation error function. These oracle inequalities will then be used to establish basic forms of both consistency and learning rates for SVMs using *a priori* defined regularization parameters. Thereby it turns out that the fastest learning rates our analysis provides require some knowledge about the distribution P. Unfortunately, however, the required type of knowledge on P is rarely available in practice, and hence these rates are in general not achievable with *a priori* defined regularization parameters. Finally, in Section 6.5 we present and analyze a simple method for determining the regularization parameter for SVMs in a *data-dependent* way. Here it will turn out that this method is *adaptive* in the sense that it achieves the fastest learning rates of our previous analysis *without* knowing any characteristics of P.

6.1 Notions of Statistical Learning

In this section, we introduce some basic notions that describe "learning" in a more formal sense. Let us begin by defining learning methods.

Definition 6.1. *Let X be a set and $Y \subset \mathbb{R}$. A* ***learning method*** *$\mathcal{L}$ on $X \times Y$ maps every data set $D \in (X \times Y)^n$, $n \geq 1$, to a function $f_D : X \to \mathbb{R}$.*

By definition, *any* method that assigns to every training set D of arbitrary but finite length a function f_D is a learning method. In particular, the meaning of "learning" is not specified in this definition. However, before we can define what we actually mean by "learning", we have to introduce a rather technical assumption dealing with the measurability of learning methods. Fortunately, we will see later that this measurability is usually fulfilled for SVMs and related learning methods. Therefore, readers not interested in these technical, yet mathematically important, details may jump directly to Definition 6.4.

Before we introduce the required measurability notion for learning methods, let us first recall (see Lemma A.3.3) that the P-completion $\mathcal{A}_{\mathrm{P}}$ of a σ-algebra $\mathcal{A}$ is the smallest σ-algebra that contains $\mathcal{A}$ and all subsets of P-zero sets in $\mathcal{A}$. Moreover, the universal completion of $\mathcal{A}$ is defined as the intersection of all completions $\mathcal{A}_{\mathrm{P}}$, where P runs through the set of all probability measures defined on $\mathcal{A}$. In order to avoid notational overload, we *always* assume in this chapter that $(X \times Y)^n$ is equipped with the universal completion of the product σ-algebra on $(X \times Y)^n$, where the latter is usually defined from the context. Moreover, the canonical extension of a product measure P^n to this completion will also be denoted by P^n if no confusion can arise.

Definition 6.2. *Let $X \neq \emptyset$ be a set equipped with some σ-algebra and $Y \subset \mathbb{R}$ be a closed non-empty subset equipped with the Borel σ-algebra. We say that the* ***learning method*** *$\mathcal{L}$ on $X \times Y$ is* ***measurable*** *if for all $n \geq 1$ the map*

$$\begin{aligned}(X \times Y)^n \times X &\to \mathbb{R}\\ (D, x) &\mapsto f_D(x)\end{aligned}$$

is measurable with respect to the universal completion of the product σ-algebra on $(X \times Y)^n \times X$, where f_D denotes the decision function produced by $\mathcal{L}$.

In the following sections, we will see that both ERM and SVMs are measurable learning methods under rather natural assumptions, and therefore we omit presenting examples of measurable learning methods in this section.

Now note that for measurable learning methods the maps $x \mapsto f_D(x)$ are measurable for all fixed $D \in (X \times Y)^n$. Consequently, the risks $\mathcal{R}_{L,\mathrm{P}}(f_D)$ are defined for all $D \in (X \times Y)^n$ and all $n \geq 1$. The following lemma ensures that for measurable learning methods the "probability" for sets of the form $\{D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) \leq \varepsilon\}$, $\varepsilon \geq 0$, is defined.

Lemma 6.3. *Let $\mathcal{L}$ be a measurable learning method on $X \times Y$ producing the decision functions f_D. Then, for all loss functions $L : X \times Y \times \mathbb{R} \to [0, \infty)$, all probability measures P on $X \times Y$, and all $n \geq 1$, the maps*

$$\begin{aligned}(X \times Y)^n &\to [0, \infty]\\ D &\mapsto \mathcal{R}_{L,\mathrm{P}}(f_D)\end{aligned}$$

are measurable with respect to the universal completion of the product σ-algebra on $(X \times Y)^n$.

Proof. By the measurability of $\mathcal{L}$ and L, we obtain the measurability of the map $(D, x, y) \mapsto L(x, y, f_D(x))$. Now the assertion follows from the measurability statement in Tonelli's Theorem A.3.10. □

With these preparations, we can now introduce our first notion of learning.

Definition 6.4. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, P be a distribution on $X \times Y$, and $\mathcal{L}$ be a measurable learning method on $X \times Y$. Then $\mathcal{L}$ is said to be* ***L-risk consistent*** *for P if, for all $\varepsilon > 0$, we have*

$$\lim_{n \to \infty} \mathrm{P}^n \Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) \leq \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon \Big) = 1 \,. \tag{6.1}$$

Moreover, $\mathcal{L}$ is called ***universally L-risk consistent*** *if it is L-risk consistent for all distributions P on $X \times Y$.*

When the training set is sufficiently large, consistent learning methods produce nearly optimal decision functions with high probability. In other words,

in the long run, a consistent method is able to "learn" with high probability decision functions that achieve nearly optimally the learning goal defined by the loss function L. Moreover, universally consistent learning methods accomplish this *without* knowing any specifics of the data-generating distribution P. From an orthodox machine learning point of view, which prohibits assumptions on P, universal consistency is thus a minimal requirement for any reasonable learning method. However, one might wonder whether this point of view, though mathematically compelling, is, at least sometimes, too unrealistic. To illustrate this suspicion, let us consider binary classification on $X := [0,1]$. Then every Bayes decision function is of the form $\mathbf{1}_{X_1} - \mathbf{1}_{X_{-1}}$, where $X_{-1}, X_1 \subset [0,1]$ are suitable sets. In addition, let us restrict our discussion to nontrivial distributions P on $X \times \{-1,1\}$, i.e., to distributions satisfying both $\mathrm{P}_X(X_1) > 0$ and $\mathrm{P}_X(X_{-1}) > 0$. For "elementary" classification problems, where X_1 and X_{-1} are *finite* unions of intervals, consistency then seems to be a natural minimal requirement for any reasonable learning methods. On the other hand, "monster" distributions P, such as the one where X_1 is the Cantor set, X_{-1} is its complement, and P_X is a mixture of the Hausdorff measure on X_1 and the Lebesgue measure on $[-1,1]$, seem to be less realistic for practical applications, and hence it may be harder to argue that learning methods should be able to learn for such P. In many situations, however, we cannot *a priori* exclude elementary distributions that are disturbed by some small yet not vanishing amount attributed to a malign or even monster distribution. Therefore, universal consistency can also be viewed as a notion of robustness that prevents a learning method from asymptotically failing in the presence of deviations from (implicitly) assumed features of P.

One of the drawbacks of the notion of (universal) consistency is that it does not specify the *speed* of convergence in (6.1). In other words, consistency is a truly asymptotic notion in the sense that it does not give us any confidence about how well the method has learned for a given data set D of fixed length n. Therefore, our next goal is to introduce a notion of learning that has a less asymptotic nature. We begin by reformulating consistency.

Lemma 6.5 (Learning rate). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, P be a distribution on $X \times Y$, and $\mathcal{L}$ be a measurable learning method satisfying*

$$\sup_{D \in (X \times Y)^n} \mathcal{R}_{L,\mathrm{P}}(f_D) < \infty\,, \qquad\qquad n \geq 1,$$

for its decision functions f_D. Then the following statements are equivalent:

i) $\mathcal{L}$ is L-risk consistent for P.

ii) There exist a constant $c_{\mathrm{P}} > 0$ and a decreasing sequence $(\varepsilon_n) \subset (0,1]$ that converges to 0 such that for all $\tau \in (0,1]$ there exists a constant $c_\tau \in [1,\infty)$ only depending on τ such that, for all $n \geq 1$ and all $\tau \in (0,1]$, we have

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) \leq \mathcal{R}^*_{L,\mathrm{P}} + c_{\mathrm{P}}\, c_\tau\, \varepsilon_n\Big) \;\geq\; 1 - \tau\,. \qquad (6.2)$$

In this case, $\mathcal{L}$ is said to ***learn with rate*** (ε_n) ***and confidence*** $(c_\tau)_{\tau \in (0,1]}$.

Proof. The fact that (6.2) implies L-risk consistency is trivial, and hence it remains to show the converse implication. To this end, we define the function $F : (0,\infty) \times \mathbb{N} \to [0,\infty)$ by

$$F(\varepsilon, n) := \mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) > \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon\Big), \qquad \varepsilon \in (0,\infty),\, n \geq 1.$$

The L-risk consistency then shows $\lim_{n\to\infty} F(\varepsilon, n) = 0$ for all $\varepsilon > 0$, and hence Lemma A.1.4 yields a decreasing sequence $(\varepsilon_n) \subset (0,1]$ converging to 0 such that $\lim_{n\to\infty} F(\varepsilon_n, n) = 0$. For a fixed $\tau \in (0,1]$, there consequently exists an $n_\tau \geq 1$ such that

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) > \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon_n\Big) \leq \tau\,, \qquad n > n_\tau. \tag{6.3}$$

For $n \geq 1$, we write $b_n := \sup_{D \in (X\times Y)^n} \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P}}$. Then the boundedness of $\mathcal{L}$ shows $b_n < \infty$ for all $n \geq 1$ and, by the definition of b_n, we also have

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) > \mathcal{R}^*_{L,\mathrm{P}} + b_n\Big) \leq \tau\,, \qquad n = 1, \ldots, n_\tau. \tag{6.4}$$

Let us define $c_\tau := \varepsilon_{n_\tau}^{-1} \max\{1, b_1, \ldots, b_{n_\tau}\}$. Then we have $b_n \leq c_\tau \varepsilon_n$ for all $n = 1, \ldots, n_\tau$ and, since $c_\tau \geq 1$, we also have $\varepsilon_n \leq c_\tau \varepsilon_n$ for all $n > n_\tau$. Using these estimates in (6.4) and (6.3), respectively, yields (6.2). □

Note that in (6.2) the constant c_P depends on the distribution P. Therefore, if we do not know P, then in general we do not know c_P. In other words, even if we know that $\mathcal{L}$ learns with rate (ε_n) for all distributions P, this knowledge does not give us any confidence about how well the method has learned in a specific application. Unfortunately, however, the following results show that the situation is even worse in the sense that in general there exists no method that learns with a fixed rate and confidence for all distributions P. Before we state these results, we remind the reader that Lebesgue absolutely continuous distributions on subsets of $\mathbb{R}^d$ are atom-free. A precise definition of atom-free measures is given in Definition A.3.12.

Theorem 6.6 (No-free-lunch theorem). *Let $(a_n) \subset (0, 1/16]$ be a decreasing sequence that converges to 0. Moreover, let $(X, \mathcal{A}, \mu)$ be an atom-free probability space, $Y := \{-1, 1\}$, and L_{class} be the binary classification loss. Then, for every measurable learning method $\mathcal{L}$ on $X \times Y$, there exists a distribution* P *on $X \times Y$ with $\mathrm{P}_X = \mu$ such that $\mathcal{R}^*_{L_{\text{class}},\mathrm{P}} = 0$ and*

$$\mathbb{E}_{D\sim\mathrm{P}^n} \mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_D) \geq a_n\,, \qquad n \geq 1.$$

Since the proof of this theorem is out of the scope of this book, we refer the interested reader to Theorem 7.2 in the book by Devroye *et al.* (1996). In this regard, we also note that Lyapunov's Theorem A.3.13 can be easily utilized to generalize their proof to a fixed atom-free distribution. The details are discussed in Exercise 6.4.

Informally speaking, the no-free-lunch theorem states that, for sufficiently malign distributions, the *average* risk of any classification method may tend arbitrarily slowly to zero. Our next goal is to use this theorem to show that in general no learning method enjoys a uniform learning rate. The first result in this direction deals with the classification loss.

Corollary 6.7 (No uniform rate for classification). *Let $(X, \mathcal{A}, \mu)$ be an atom-free probability space, $Y := \{-1, 1\}$, and $\mathcal{L}$ be a measurable learning method on $X \times Y$. Then, for all decreasing sequences $(\varepsilon_n) \subset (0, 1]$ that converge to 0 and all families $(c_\tau)_{t \in (0,1]} \subset [1, \infty)$, there exists a distribution P on $X \times Y$ satisfying $\mathrm{P}_X = \mu$ and $\mathcal{R}^*_{L_{\text{class}},\mathrm{P}} = 0$ such that $\mathcal{L}$ does not learn with rate (ε_n) and confidence $(c_\tau)_{\tau \in (0,1]}$.*

Proof. For brevity's sake, we write $L := L_{\text{class}}$. Let us assume that the assertion is false, i.e., that there exist a decreasing sequence $(\varepsilon_n) \subset (0, 1]$ that converges to 0 and constants $c_\tau \in [1, \infty)$, $\tau \in (0, 1]$, such that, for all distributions P on $X \times Y$ satisfying $\mathrm{P}_X = \mu$ and $\mathcal{R}^*_{L,\mathrm{P}} = 0$, the method $\mathcal{L}$ learns with rate (ε_n) and confidence $(c_\tau)_{\tau \in (0,1]}$. In other words, we assume

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) > c_{\mathrm{P}}\, c_\tau\, \varepsilon_n\Big) \;\le\; \tau \tag{6.5}$$

for all $n \ge 1$ and $\tau \in (0, 1]$, where c_{P} is a constant independent of n and τ. Let us define $F : (0, 1] \times \mathbb{N} \to [0, \infty)$ by $F(\tau, n) := \tau^{-1} c_\tau \varepsilon_n$. Then we have $\lim_{n\to\infty} F(\tau, n) = 0$ for all $\tau \in (0, 1]$, and consequently an obvious modification of Lemma A.1.4 yields a decreasing sequence $(\tau_n) \subset (0, 1]$ converging to 0 such that $\lim_{n\to\infty} F(\tau_n, n) = 0$. We define $a_n := 1/16$ if $\tau_n \ge 1/32$ and $a_n := 2\tau_n$ otherwise. By the no-free-lunch theorem, there then exists a distribution P on $X \times Y$ such that $\mathrm{P}_X = \mu$, $\mathcal{R}^*_{L,\mathrm{P}} = 0$, and

$$\begin{aligned} a_n &\le \mathbb{E}_{D\sim\mathrm{P}^n} \mathcal{R}_{L,\mathrm{P}}(f_D) \\ &= \int\limits_{\mathcal{R}_{L,\mathrm{P}}(f_D) \le c_{\mathrm{P}}\, c_\tau\, \varepsilon_n} \mathcal{R}_{L,\mathrm{P}}(f_D)\, d\mathrm{P}^n(D) + \int\limits_{\mathcal{R}_{L,\mathrm{P}}(f_D) > c_{\mathrm{P}}\, c_\tau\, \varepsilon_n} \mathcal{R}_{L,\mathrm{P}}(f_D)\, d\mathrm{P}^n(D) \\ &\le c_{\mathrm{P}}\, c_\tau\, \varepsilon_n + \tau \end{aligned}$$

for all $n \ge 1$ and $\tau \in (0, 1]$, where in the last estimate we used (6.5) together with $\mathcal{R}_{L,\mathrm{P}}(f_D) \le 1$. Consequently, we find

$$\frac{a_n - \tau_n}{c_{\tau_n} \varepsilon_n} \;\le\; c_{\mathrm{P}}\,, \qquad\qquad n \ge 1.$$

On the other hand, our construction yields

$$\lim_{n\to\infty} \frac{a_n - \tau_n}{c_{\tau_n} \varepsilon_n} = \lim_{n\to\infty} \frac{\tau_n}{c_{\tau_n} \varepsilon_n} = \lim_{n\to\infty} \frac{1}{F(\tau_n, n)} = \infty\,,$$

and hence we have found a contradiction. □

In the following, we show that the result of Corollary 6.7 is true not only for the classification loss but for basically all loss functions. The idea of this generalization is that whenever we have a loss function L that describes a learning goal where at least two different labels have to be distinguished, then this learning problem is in some sense harder than binary classification and hence cannot be learned with a uniform rate. Since the precise statement of this idea is rather cumbersome and requires notations from Section 3.1, we suggest that the first-time reader skips this part and simply remembers the informal result described above. Moreover, for *convex* losses L, the conditions below can be substantially simplified. We refer the reader to Exercise 6.5 for a precise statement and examples.

Corollary 6.8 (No uniform learning rate). *Let $(X, \mathcal{A}, \mu)$ be an atom-free probability space, $Y \subset \mathbb{R}$ be a closed subset, and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss. Assume that there exist two distributions Q_1 and Q_2 on Y such that $\mathcal{M}_{L,\mathrm{Q}_1,x}(0^+) \neq \emptyset$, $\mathcal{M}_{L,\mathrm{Q}_2,x}(0^+) \neq \emptyset$, and*

$$\overline{\mathcal{M}_{L,\mathrm{Q}_1,x}(0^+)} \cap \mathcal{M}_{L,\mathrm{Q}_2,x}(0^+) = \emptyset$$

for all $x \in X$. For $x \in X$, we define

$$\begin{aligned}
\mathcal{M}_{1,x} &:= \bigl\{t \in \mathbb{R} : \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_2,x}(0^+)) \geq \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_1,x}(0^+))\bigr\}\,,\\
\mathcal{M}_{2,x} &:= \bigl\{t \in \mathbb{R} : \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_2,x}(0^+)) < \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_1,x}(0^+))\bigr\}\,.
\end{aligned}$$

If there exists a measurable $h : X \to (0, 1]$ such that for all $x \in X$ we have

$$\begin{aligned}
\inf_{t \in \mathcal{M}_{1,x}} \mathcal{C}_{L,\mathrm{Q}_2,x}(t) - \mathcal{C}^*_{L,\mathrm{Q}_2,x} &\geq h(x)\,,\\
\inf_{t \in \mathcal{M}_{2,x}} \mathcal{C}_{L,\mathrm{Q}_1,x}(t) - \mathcal{C}^*_{L,\mathrm{Q}_1,x} &\geq h(x)\,,
\end{aligned}$$

then the conclusion of Corollary 6.7 remains true if we replace L_{class} by L.

Proof. Clearly, the distribution $\bar{\mu} := \|h\|^{-1}_{L_1(\mu)} h\mu$ on X is atom-free. Moreover, for a distribution P on $X \times Y$ that is of type $\{\mathrm{Q}_1, \mathrm{Q}_2\}$ and satisfies $\mathrm{P}_X = \mu$, we associate the distribution $\bar{\mathrm{P}}$ on $X \times \{-1, 1\}$ that is defined by $\bar{\mathrm{P}}_X = \bar{\mu}$ and

$$\bar{\mathrm{P}}(y = 1|x) := \begin{cases} 0 & \text{if } \mathrm{P}(\,\cdot\,|x) = \mathrm{Q}_1 \\ 1 & \text{if } \mathrm{P}(\,\cdot\,|x) = \mathrm{Q}_2\,. \end{cases}$$

In other words, $\bar{\mathrm{P}}(\,\cdot\,|x)$ produces almost surely a negative label if $\mathrm{P}(\,\cdot\,|x) = \mathrm{Q}_1$ and almost surely a positive label if $\mathrm{P}(\,\cdot\,|x) = \mathrm{Q}_2$. From this it becomes obvious that $\mathcal{R}^*_{L_{\text{class}},\bar{\mathrm{P}}} = 0$. For $x \in X$ and $t \in \mathbb{R}$, we further define

$$\pi(t, x) := \begin{cases} -1 & \text{if } t \in \mathcal{M}_{1,x} \\ 1 & \text{if } t \in \mathcal{M}_{2,x}\,. \end{cases}$$

In other words, $\pi(t,x)$ becomes negative if t is closer to the minimizing set $\mathcal{M}_{L,\mathrm{Q}_1,x}(0^+)$ of the distribution Q_1 than to that of Q_2. Here the idea behind this construction is that Q_1 is identified with negative classification labels in the definition of $\bar{\mathrm{P}}$. Moreover, note that this definition is *independent* of P. In addition, since $t \mapsto \operatorname{dist}(t, A)$ is continuous for arbitrary $A \subset \mathbb{R}$ and $x \mapsto \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_i,x}(0^+))$, $i \in \{1,2\}$, is measurable by Aumann's measurable selection principle stated in Lemma A.3.18, we see by Lemma A.3.17 that $\pi : \mathbb{R} \times X \to \mathbb{R}$ is measurable. Now a simple calculation shows

$$h(x)\mathcal{C}_{L_{\text{class}},\bar{\mathrm{P}}(\,\cdot\,|x)}(\pi(t,x)) \;\leq\; \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}\,, \qquad x \in X,\, t \in \mathbb{R},$$

and thus we find $\|h\|_{L_1(\mu)}\mathcal{R}_{L_{\text{class}},\bar{\mathrm{P}}}(\pi \circ f) \leq \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}$ for all measurable $f : X \to \mathbb{R}$, where $\pi \circ f(x) := \pi(f(x), x)$, $x \in X$. Now let $\mathcal{L}$ be a measurable learning method producing decision functions f_D. Then $\pi \circ \mathcal{L}$ defined by $D \mapsto \pi \circ f_D$ is also a measurable learning method since π is measurable. Moreover, $\pi \circ \mathcal{L}$ is independent of P. Assume that $\mathcal{L}$ learns all distributions P of type $\{\mathrm{Q}_1, \mathrm{Q}_2\}$ that satisfy $\mathrm{P}_X = \mu$ with a uniform rate. Then our considerations above show that $\pi \circ \mathcal{L}$ learns all associated classification problems $\bar{\mathrm{P}}$ with the same uniform rate, but by Corollary 6.7 this is impossible. □

The results above show that in general we cannot *a priori* guarantee with a fixed confidence that a learning method finds a nearly optimal decision function. This is a fundamental limitation for statistical learning methods that we *cannot* elude by, e.g., cleverly combining different learning methods since such a procedure itself constitutes a learning method. In other words, the only way to resolve this issue is to make *a priori* assumptions on the data generating distribution P. However, since in almost no case will we be able to *rigorously check* whether P actually satisfies the imposed assumptions, such an approach has only very limited utility for *a priori guaranteeing* good generalization performance. On the other hand, by establishing learning rates for different types of assumptions on P, we can *understand* for which kind of distributions the learning method considered learns easily and for which it does not. In turn, such knowledge can then be used in practice where one *has to* decide which learning methods are likely to be appropriate for a specific application.

6.2 Basic Concentration Inequalities

We will see in the following sections that our statistical analysis of both ERM and SVMs relies heavily on bounds on the probabilities

$$\mathrm{P}^n\big(\{D \in (X \times Y)^n : |\mathcal{R}_{L,\mathrm{D}}(f) - \mathcal{R}_{L,\mathrm{P}}(f)| > \varepsilon\}\big)\,.$$

In this section, we thus establish some basic bounds on such probabilities.

Let us begin with an elementary yet powerful inequality that will be the key ingredient for *all* the more advanced results that follow (see also Exercise 6.2 for a slightly refined estimate).

Theorem 6.9 (Markov's inequality). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space. Then, for all measurable functions $f : \Omega \to \mathbb{R}$ and all $t > 0$, we have*

$$\mathrm{P}\big(\{\omega \in \Omega : f(\omega) \geq t\}\big) \;\leq\; \frac{\mathbb{E}_{\mathrm{P}}|f|}{t}\,.$$

Proof. For $A_t := \{\omega \in \Omega : f(\omega) \geq t\}$, we obviously have $t\,\mathbf{1}_{A_t} \leq f\mathbf{1}_{A_t} \leq |f|$, and hence we obtain $t\,\mathrm{P}(A_t) = \mathbb{E}_{\mathrm{P}}\, t\,\mathbf{1}_{A_t} \leq \mathbb{E}_{\mathrm{P}}|f|$. □

From Markov's inequality, it is straightforward to derive Chebyshev's inequality $\mathrm{P}(\{\omega \in \Omega : |f(\omega)| \geq t\}) \leq t^{-2}\mathbb{E}_{\mathrm{P}}|f|^2$. The following result also follows from Markov's inequality.

Theorem 6.10 (Hoeffding's inequality). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, $a < b$ be two real numbers, $n \geq 1$ be an integer, and $\xi_1, \ldots, \xi_n : \Omega \to [a, b]$ be independent random variables. Then, for all $\tau > 0$, we have*

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}\big(\xi_i - \mathbb{E}_{\mathrm{P}}\xi_i\big) \geq (b-a)\sqrt{\frac{\tau}{2n}}\right) \leq e^{-\tau}\,.$$

Proof. We begin with a preliminary consideration. To this end, let $\tilde{a} < \tilde{b}$ be two real numbers and $\xi : \Omega \to [\tilde{a}, \tilde{b}]$ be a random variable with $\mathbb{E}_{\mathrm{P}}\xi = 0$. Note that from this assumptions we can immediately conclude that $\tilde{a} \leq 0$ and $\tilde{b} \geq 0$. Moreover, observe that for $x \in [\tilde{a}, \tilde{b}]$ we have

$$x = \frac{\tilde{b} - x}{\tilde{b} - \tilde{a}}\tilde{a} + \frac{x - \tilde{a}}{\tilde{b} - \tilde{a}}\tilde{b}\,,$$

and hence the convexity of the exponential function implies

$$e^{tx} \leq \frac{\tilde{b} - x}{\tilde{b} - \tilde{a}}e^{t\tilde{a}} + \frac{x - \tilde{a}}{\tilde{b} - \tilde{a}}e^{t\tilde{b}}\,, \qquad t > 0.$$

Since $\mathbb{E}_{\mathrm{P}}\xi = 0$, we then obtain

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}e^{t\xi} \leq \mathbb{E}_{\mathrm{P}}\left(\frac{\tilde{b} - \xi}{\tilde{b} - \tilde{a}}e^{t\tilde{a}} + \frac{\xi - \tilde{a}}{\tilde{b} - \tilde{a}}e^{t\tilde{b}}\right) &= \frac{\tilde{b}}{\tilde{b} - \tilde{a}}e^{t\tilde{a}} - \frac{\tilde{a}}{\tilde{b} - \tilde{a}}e^{t\tilde{b}} \\ &= e^{t\tilde{a}}\left(1 + \frac{\tilde{a}}{\tilde{b} - \tilde{a}} - \frac{\tilde{a}}{\tilde{b} - \tilde{a}}e^{t(\tilde{b}-\tilde{a})}\right)\end{aligned}$$

for all $t > 0$. Let us now write $p := -\tilde{a}\,(\tilde{b} - \tilde{a})^{-1}$. Then we observe that $\tilde{a} \leq 0$ implies $p \geq 0$, and $\tilde{b} \geq 0$ implies $p \leq 1$. For $s \in \mathbb{R}$, we hence find $e^s > 0 \geq 1 - 1/p$, from which we conclude that $1 - p + pe^s > 0$. Consequently, $\phi_p(s) := \ln(1 - p + pe^s) - ps$ is defined for all $s \in \mathbb{R}$. Moreover, these definitions together with our previous estimate yield

$$\mathbb{E}_{\mathrm{P}}e^{t\xi} \leq e^{-tp(\tilde{b}-\tilde{a})}\left(1 - p + pe^{t(\tilde{b}-\tilde{a})}\right) = e^{\phi_p(t(\tilde{b}-\tilde{a}))}\,.$$

Now observe that we have $\phi_p(0) = 0$ and $\phi_p'(s) = \frac{pe^s}{1-p+pe^s} - p$. From the latter, we conclude that $\phi_p'(0) = 0$ and

$$\phi_p''(s) = \frac{(1-p+pe^s)pe^s - pe^s pe^s}{(1-p+pe^s)^2} = \frac{(1-p)pe^s}{(1-p+pe^s)^2} \leq \frac{(1-p)pe^s}{4(1-p)pe^s} = \frac{1}{4}$$

for all $s \in \mathbb{R}$. By Taylor's formula with Lagrangian remainder, we hence find

$$\phi_p(s) = \phi_p(0) + \phi_p'(0)s + \frac{1}{2}\phi_p''(s')s^2 \leq \frac{s^2}{8}\,, \qquad s > 0,$$

where $s' \in [0, s]$ is a suitable real number. Consequently, we obtain

$$\mathbb{E}_{\mathrm{P}} e^{t\xi} \leq e^{\phi_p(t(\tilde{b}-\tilde{a}))} \leq \exp\left(\frac{t^2(\tilde{b}-\tilde{a})^2}{8}\right), \qquad t > 0\,.$$

Applying this estimate to the random variables $\xi_i - \mathbb{E}\xi_i : \Omega \to [a - \mathbb{E}\xi_i, b - \mathbb{E}\xi_i]$, where $\mathbb{E} := \mathbb{E}_{\mathrm{P}}$, we now find

$$\mathbb{E}_{\mathrm{P}} e^{t(\xi_i - \mathbb{E}\xi_i)} \leq \exp\left(\frac{t^2(b-a)^2}{8}\right), \qquad t > 0,\ i = 1, \ldots, n.$$

Using this estimate together with Markov's inequality and the independence assumption, we hence obtain with $\mathbb{E} := \mathbb{E}_{\mathrm{P}}$ that

$$\begin{aligned} \mathrm{P}\left(\sum_{i=1}^n (\xi_i - \mathbb{E}\xi_i) \geq \varepsilon n\right) &\leq e^{-t\varepsilon n}\, \mathbb{E} \exp\left(t \sum_{i=1}^n (\xi_i - \mathbb{E}\xi_i)\right) \leq e^{-t\varepsilon n} \prod_{i=1}^n \mathbb{E} e^{t(\xi_i - \mathbb{E}\xi_i)} \\ &\leq e^{-t\varepsilon n}\, e^{\frac{nt^2(b-a)^2}{8}} \end{aligned}$$

for all $\varepsilon > 0$ and $t > 0$. Now we obtain the assertion by considering $\varepsilon := (b-a)(\frac{\tau}{2n})^{1/2}$ and $t := \frac{4\varepsilon}{(b-a)^2}$. □

Our next goal is to present a concentration inequality that refines Hoeffding's inequality when we know not only the $\|\cdot\|_\infty$-norms of the random variables involved but also their variances, i.e., their $\|\cdot\|_2$-norms. To this end, we need the following technical lemma.

Lemma 6.11. *For all $x > -1$, we have $(1+x)\ln(1+x) - x \geq \frac{3}{2}\frac{x^2}{x+3}$.*

Proof. For $x > -1$, we define $f(x) := (1+x)\ln(1+x) - x$ and $g(x) := \frac{3}{2}\frac{x^2}{x+3}$. Then an easy calculation shows that, for all $x > -1$, we have

$$\begin{aligned} f'(x) &= \ln(1+x)\,, & f''(x) &= \frac{1}{1+x}\,, \\ g'(x) &= \frac{3x^2 + 18x}{2(x+3)^2}\,, & g''(x) &= \frac{27}{(x+3)^3}\,. \end{aligned}$$

Consequently, we have $f(0) = g(0) = 0$, $f'(0) = g'(0) = 0$, and $f''(x) \geq g''(x)$ for all $x > -1$. For $x \geq 0$, the fundamental theorem of calculus thus gives

$$f'(x) = \int_0^x f''(t)dt \geq \int_0^x g''(t)dt = g'(x)\,,$$

and by repeating this reasoning, we obtain the assertion for $x \geq 0$. For $x \in (-1, 0]$, we can show the assertion analogously. □

Now we can establish the announced refinement of Hoeffding's inequality.

Theorem 6.12 (Bernstein's inequality). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, $B > 0$ and $\sigma > 0$ be real numbers, and $n \geq 1$ be an integer. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to \mathbb{R}$ be independent random variables satisfying $\mathbb{E}_\mathrm{P}\xi_i = 0$, $\|\xi_i\|_\infty \leq B$, and $\mathbb{E}_\mathrm{P}\xi_i^2 \leq \sigma^2$ for all $i = 1, \ldots, n$. Then we have*

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^n \xi_i \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\right) \leq e^{-\tau}\,, \qquad \tau > 0.$$

Proof. By Markov's inequality and the independence of $\xi_1, \ldots, \xi_n$, we have

$$\mathrm{P}\left(\sum_{i=1}^n \xi_i \geq \varepsilon n\right) \leq e^{-t\varepsilon n}\,\mathbb{E}_\mathrm{P}\exp\left(t\sum_{i=1}^n \xi_i\right) \leq e^{-t\varepsilon n}\prod_{i=1}^n \mathbb{E}_\mathrm{P}\, e^{t\xi_i}$$

for all $t \geq 0$ and $\varepsilon > 0$. Furthermore, the properties of ξ_i imply

$$\mathbb{E}_\mathrm{P}\, e^{t\xi_i} = \sum_{k=0}^\infty \frac{t^k}{k!}\mathbb{E}_\mathrm{P}\xi_i^k \leq 1 + \sum_{k=2}^\infty \frac{t^k}{k!}\sigma^2 B^{k-2} = 1 + \frac{\sigma^2}{B^2}\big(e^{tB} - tB - 1\big)\,.$$

Using the simple estimate $1 + x \leq e^x$ for $x := \frac{\sigma^2}{B^2}(e^{tB} - tB - 1)$, we hence obtain

$$\begin{aligned}\mathrm{P}\left(\sum_{i=1}^n \xi_i \geq \varepsilon n\right) &\leq e^{-t\varepsilon n}\prod_{i=1}^n \Big(1 + \frac{\sigma^2}{B^2}\big(e^{tB} - tB - 1\big)\Big) \\ &\leq \exp\Big(-t\varepsilon n + \frac{\sigma^2 n}{B^2}\big(e^{tB} - tB - 1\big)\Big)\end{aligned}$$

for all $t \geq 0$. Now elementary calculus shows that the right-hand side of the inequality is minimized at

$$t^* := \frac{1}{B}\ln\Big(1 + \frac{\varepsilon B}{\sigma^2}\Big)\,.$$

Writing $y := \frac{\varepsilon B}{\sigma^2}$ and using Lemma 6.11, we furthermore obtain

$$\begin{aligned}-t^*\varepsilon n + \frac{\sigma^2 n}{B^2}\big(e^{t^*B} - t^*B - 1\big) = -\frac{n\sigma^2}{B^2}\big((1+y)\ln(1+y) - y\big) &\leq -\frac{3n\sigma^2}{2B^2}\frac{y^2}{y+3} \\ &= -\frac{3\varepsilon^2 n}{2\varepsilon B + 6\sigma^2}\,.\end{aligned}$$

Let us now define $\varepsilon := \sqrt{\frac{2\sigma^2\tau}{n} + \frac{B^2\tau^2}{9n^2}} + \frac{B\tau}{3n}$. Then we have $\tau = \frac{3\varepsilon^2 n}{2\varepsilon B + 6\sigma^2}$ and

$$\varepsilon \leq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n},$$

and thus we obtain the assertion. □

It is important to keep in mind that in situations where we know upper bounds σ^2 and B for the variances and the suprema, respectively, Bernstein's inequality is often sharper than Hoeffding's inequality. The details are discussed in Exercise 6.1.

Our next goal is to generalize Bernstein's inequality to Hilbert space valued random variables. To this end, we need the following more general result.

Theorem 6.13. *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, E be a separable Banach space, and $\xi_1, \ldots, \xi_n : \Omega \to E$ be independent E-valued P-integrable random variables. Then, for all $\varepsilon > 0$ and all $t \geq 0$, we have*

$$\mathrm{P}\left(\left\| \sum_{i=1}^n \xi_i \right\| \geq \varepsilon n\right) \leq \exp\left(-t\varepsilon n + t\mathbb{E}_{\mathrm{P}} \left\| \sum_{i=1}^n \xi_i \right\| + \sum_{i=1}^n \mathbb{E}_{\mathrm{P}}(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|)\right).$$

Proof. Let us consider the σ-algebras $\mathcal{F}_0 := \{\emptyset, \Omega\}$ and $\mathcal{F}_k := \sigma(\xi_1, \ldots, \xi_k)$, $k = 1, \ldots, n$. Furthermore, for $k = 1, \ldots, n$, we define

$$X_k := \mathbb{E}_{\mathrm{P}}\left(\left\| \sum_{i=1}^n \xi_i \right\| \,\middle|\, \mathcal{F}_k\right) - \mathbb{E}_{\mathrm{P}}\left(\left\| \sum_{i=1}^n \xi_i \right\| \,\middle|\, \mathcal{F}_{k-1}\right),$$

$$Y_k := \mathbb{E}_{\mathrm{P}}\left(\left\| \sum_{i=1}^n \xi_i \right\| - \left\| \sum_{i\neq k} \xi_i \right\| \,\middle|\, \mathcal{F}_k\right).$$

By a simple telescope sum argument, we then have

$$\sum_{i=1}^n X_i = \left\| \sum_{i=1}^n \xi_i \right\| - \mathbb{E}_{\mathrm{P}} \left\| \sum_{i=1}^n \xi_i \right\|. \tag{6.6}$$

Moreover, note that $\sum_{i\neq k} \xi_i$ is independent of ξ_k, and hence we obtain

$$\mathbb{E}_{\mathrm{P}}\left(\left\| \sum_{i\neq k} \xi_i \right\| \,\middle|\, \mathcal{F}_k\right) = \mathbb{E}_{\mathrm{P}}\left(\left\| \sum_{i\neq k} \xi_i \right\| \,\middle|\, \mathcal{F}_{k-1}, \xi_k\right) = \mathbb{E}_{\mathrm{P}}\left(\left\| \sum_{i\neq k} \xi_i \right\| \,\middle|\, \mathcal{F}_{k-1}\right).$$

Since $\mathcal{F}_{k-1} \subset \mathcal{F}_k$, we thus find

$$\begin{aligned} X_k &= \mathbb{E}_{\mathrm{P}}\left(\left\| \sum_{i=1}^n \xi_i \right\| \,\middle|\, \mathcal{F}_k\right) - \mathbb{E}_{\mathrm{P}}\left(\left\| \sum_{i=1}^n \xi_i \right\| \,\middle|\, \mathcal{F}_{k-1}\right) \\ &= \mathbb{E}_{\mathrm{P}}\left(\left\| \sum_{i=1}^n \xi_i \right\| - \left\| \sum_{i\neq k} \xi_i \right\| \,\middle|\, \mathcal{F}_k\right) - \mathbb{E}_{\mathrm{P}}\left(\left\| \sum_{i=1}^n \xi_i \right\| - \left\| \sum_{i\neq k} \xi_i \right\| \,\middle|\, \mathcal{F}_{k-1}\right) \\ &= Y_k - \mathbb{E}_{\mathrm{P}}(Y_k|\mathcal{F}_{k-1}) \end{aligned} \tag{6.7}$$

for all $k = 1, \dots, n$. Using $x \leq e^{x-1}$ for all $x \in \mathbb{R}$, we hence obtain

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}\big(e^{tX_k} \,|\, \mathcal{F}_{k-1}\big) &= e^{-t\mathbb{E}_{\mathrm{P}}(Y_k|\mathcal{F}_{k-1})}\mathbb{E}_{\mathrm{P}}\big(e^{tY_k} \,|\, \mathcal{F}_{k-1}\big) \\ &\leq e^{-t\mathbb{E}_{\mathrm{P}}(Y_k|\mathcal{F}_{k-1})}e^{\mathbb{E}_{\mathrm{P}}(e^{tY_k} \,|\, \mathcal{F}_{k-1})-1} \\ &= \exp\Big(\mathbb{E}_{\mathrm{P}}\big(e^{tY_k} - 1 - tY_k \,|\, \mathcal{F}_{k-1}\big)\Big)\,. \end{aligned} \tag{6.8}$$

Now, an easy calculation shows $e^x - e^{-x} \geq 2x$ for all $x \geq 0$, which in turn implies $e^{-x} - 1 - (-x) \leq e^x - 1 - x$ for all $x \geq 0$. From this we conclude that $e^x - 1 - x \leq e^{|x|} - 1 - |x|$ for all $x \in \mathbb{R}$. Moreover, it is straightforward to check that the function $x \mapsto e^x - 1 - x$ is increasing on $[0, \infty)$. In addition, the triangle inequality in E gives

$$|Y_k| \leq \mathbb{E}_{\mathrm{P}}\left(\Big|\,\|\sum_{i=1}^n \xi_i\,\| - \|\sum_{i\neq k}\xi_i\,\|\Big|\,\Big|\,\mathcal{F}_k\right) \leq \mathbb{E}_{\mathrm{P}}\left(\Big\|\sum_{i=1}^n \xi_i - \sum_{i\neq k}\xi_i\Big\|\,\Big|\,\mathcal{F}_k\right) = \|\xi_k\|\,,$$

where in the last step we used that $\|\xi_k\|$ is $\mathcal{F}_k$-measurable. Consequently, (6.8) implies

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}\big(e^{tX_k} \,|\, \mathcal{F}_{k-1}\big) &\leq \exp\Big(\mathbb{E}_{\mathrm{P}}\big(e^{t|Y_k|} - 1 - t|Y_k| \,\big|\, \mathcal{F}_{k-1}\big)\Big) \\ &\leq \exp\Big(\mathbb{E}_{\mathrm{P}}\big(e^{t\|\xi_k\|} - 1 - t\|\xi_k\| \,\big|\, \mathcal{F}_{k-1}\big)\Big) \\ &= \exp\Big(\mathbb{E}_{\mathrm{P}}\big(e^{t\|\xi_k\|} - 1 - t\|\xi_k\|\big)\Big)\,, \end{aligned} \tag{6.9}$$

where in the last step we used that ξ_k is independent of $\mathcal{F}_{k-1}$. Moreover, $\sum_{i=1}^{k-1} X_i$ is $\mathcal{F}_{k-1}$-measurable, and writing $\mathbb{E} := \mathbb{E}_{\mathrm{P}}$ we hence have

$$\mathbb{E}\Big(e^{t\sum_{i=1}^{k-1} X_i}\mathbb{E}\big(e^{tX_k}\,\big|\mathcal{F}_{k-1}\big)\Big) = \mathbb{E}\Big(\mathbb{E}\big(e^{t\sum_{i=1}^{k-1} X_i}e^{tX_k}\,\big|\mathcal{F}_{k-1}\big)\Big) = \mathbb{E}e^{t\sum_{i=1}^{k} X_i}\,.$$

Combining this last equation with (6.9) now yields

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}e^{t\sum_{i=1}^{k} X_i} &= \mathbb{E}_{\mathrm{P}}\Big(e^{t\sum_{i=1}^{k-1} X_i}\,\mathbb{E}_{\mathrm{P}}\big(e^{tX_k}\,\big|\mathcal{F}_{k-1}\big)\Big) \\ &\leq \mathbb{E}_{\mathrm{P}}\Big(e^{t\sum_{i=1}^{k-1} X_i}\Big)\cdot\exp\Big(\mathbb{E}_{\mathrm{P}}\big(e^{t\|\xi_k\|} - 1 - t\|\xi_k\|\big)\Big)\,, \end{aligned}$$

and by successively applying this inequality we hence obtain

$$\mathbb{E}_{\mathrm{P}}e^{t\sum_{i=1}^{n} X_i} \leq \prod_{i=1}^n \exp\Big(\mathbb{E}_{\mathrm{P}}\big(e^{t\|\xi_i\|} - 1 - t\|\xi_k\|\big)\Big) = \exp\left(\sum_{i=1}^n \mathbb{E}_{\mathrm{P}}\big(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|\big)\right)$$

for all $t \geq 0$. By Markov's inequality and (6.6), we thus find

$$\begin{aligned}
\mathrm{P}\Big(\Big\|\sum_{i=1}^n \xi_i\Big\| \geq \varepsilon n\Big) &\leq e^{-t\varepsilon n}\,\mathbb{E}_{\mathrm{P}} \exp\Big(t\Big\|\sum_{i=1}^n \xi_i\Big\|\Big) \\
&= e^{-t\varepsilon n}\,\mathbb{E}_{\mathrm{P}} \exp\Big(t\sum_{i=1}^n X_i + t\mathbb{E}_{\mathrm{P}}\Big\|\sum_{i=1}^n \xi_i\Big\|\Big) \\
&\leq \exp\Big(-t\varepsilon n + t\mathbb{E}_{\mathrm{P}}\Big\|\sum_{i=1}^n \xi_i\Big\| + \sum_{i=1}^n \mathbb{E}_{\mathrm{P}}\big(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|\big)\Big)
\end{aligned}$$

for all $\varepsilon > 0$ and all $t \geq 0$. □

With the help of the previous theorem we can now establish the following Hilbert space version of Bernstein's inequality.

Theorem 6.14 (Bernstein's inequality in Hilbert spaces). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, H be a separable Hilbert space, $B > 0$, and $\sigma > 0$. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to H$ be independent random variables satisfying $\mathbb{E}_{\mathrm{P}}\xi_i = 0$, $\|\xi_i\|_\infty \leq B$, and $\mathbb{E}_{\mathrm{P}}\|\xi_i\|_H^2 \leq \sigma^2$ for all $i = 1, \ldots, n$. Then we have*

$$\mathrm{P}\left(\Big\|\frac{1}{n}\sum_{i=1}^n \xi_i\Big\|_H \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \sqrt{\frac{\sigma^2}{n}} + \frac{2B\tau}{3n}\right) \;\leq\; e^{-\tau}\,, \qquad\qquad \tau > 0.$$

Proof. We will prove the assertion by applying Theorem 6.13. To this end, we first observe that the independence of $\xi_1, \ldots, \xi_n$ yields $\mathbb{E}\langle \xi_i, \xi_j\rangle_H = \langle \mathbb{E}\xi_i, \mathbb{E}\xi_j\rangle_H = 0$ for all $i \neq j$, where $\mathbb{E} := \mathbb{E}_{\mathrm{P}}$. Consequently, we obtain

$$\mathbb{E}\Big\|\sum_{i=1}^n \xi_i\Big\|_H \leq \left(\mathbb{E}\Big\|\sum_{i=1}^n \xi_i\Big\|_H^2\right)^{1/2} = \left(\sum_{i=1}^n \mathbb{E}\|\xi_i\|_H^2\right)^{1/2} \leq \sqrt{n\sigma^2}\,. \tag{6.10}$$

In addition, the series expansion of the exponential function yields

$$\sum_{i=1}^n \mathbb{E}\big(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|\big) = \sum_{i=1}^n \sum_{j=2}^\infty \frac{t^j}{j!}\mathbb{E}\|\xi_i\|_H^j \leq \sum_{i=1}^n \sum_{j=2}^\infty \frac{t^j}{j!} B^{j-2}\mathbb{E}\|\xi_i\|_H^2$$

for all $t \geq 0$, and therefore we find

$$\sum_{i=1}^n \mathbb{E}\big(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|\big) \leq \frac{\sigma^2}{B^2}\sum_{i=1}^n \sum_{j=2}^\infty \frac{t^j}{j!} B^j = \frac{n\sigma^2}{B^2}\big(e^{tB} - 1 - tB\big)$$

for all $t \geq 0$. By Theorem 6.13, we hence obtain

$$\begin{aligned}
\mathrm{P}\Big(\Big\|\sum_{i=1}^n \xi_i\Big\|_H \geq \varepsilon n\Big) &\leq \exp\Big(-t\varepsilon n + t\mathbb{E}\Big\|\sum_{i=1}^n \xi_i\Big\| + \sum_{i=1}^n \mathbb{E}(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|)\Big) \\
&\leq \exp\Big(-t\varepsilon n + t\sqrt{n\sigma^2} + \frac{n\sigma^2}{B^2}\big(e^{tB} - 1 - tB\big)\Big) \qquad (6.11)
\end{aligned}$$

for all $t \geq 0$. Let us now restrict our considerations to $\varepsilon \geq \sigma n^{-1/2}$. Then we have $y := \frac{\varepsilon B}{\sigma^2} - \frac{B}{\sqrt{n\sigma^2}} > 0$, and consequently it is easy to see that the function on the right-hand side of (6.11) is minimized at

$$t^* := \frac{1}{B} \ln\left(1 + \frac{\varepsilon B}{\sigma^2} - \frac{B}{\sqrt{n\sigma^2}}\right).$$

Moreover, Lemma 6.11 yields

$$\begin{aligned} -t^*\varepsilon n + t^*\sqrt{n\sigma^2} + \frac{n\sigma^2}{B^2}\bigl(e^{t^*B} - 1 - t^*B\bigr) &= -\frac{n\sigma^2}{B^2}\Bigl((1+y)\ln(1+y) - y\Bigr) \\ &\leq -\frac{3n\sigma^2}{2B^2}\frac{y^2}{y+3}. \end{aligned}$$

By combining this estimate with (6.11), we then find

$$\mathrm{P}\left(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\|_H \geq \varepsilon n\right) \;\leq\; \exp\left(-\frac{3n\sigma^2}{2B^2}\frac{y^2}{y+3}\right).$$

Let us now define $\varepsilon := \sqrt{\frac{2\sigma^2\tau}{n} + \frac{B^2\tau^2}{9n^2}} + \frac{B\tau}{3n} + \sqrt{\frac{\sigma^2}{n}}$. Then, an easy calculation shows

$$y = \frac{\varepsilon B}{\sigma^2} - \frac{B}{\sqrt{n\sigma^2}} = \sqrt{\frac{2B^2\tau}{n\sigma^2} + \frac{B^4\tau^2}{9n^2\sigma^4}} + \frac{B^2\tau}{3n\sigma^2},$$

and hence we find $\tau = -\frac{3n\sigma^2}{2B^2}\frac{y^2}{y+3}$. Now the assertion follows from

$$\varepsilon = \sqrt{\frac{2\sigma^2\tau}{n} + \frac{B^2\tau^2}{9n^2}} + \frac{B\tau}{3n} + \sqrt{\frac{\sigma^2}{n}} \leq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n} + \sqrt{\frac{\sigma^2}{n}}. \qquad \square$$

The following Hilbert space version of Hoeffding's inequality is an immediate consequence of Theorem 6.14. We will use it in Section 6.4 to derive an oracle inequality for SVMs.

Corollary 6.15 (Hoeffding's inequality in Hilbert spaces). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, H be a separable Hilbert space H, and $B > 0$. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to H$ be independent H-valued random variables satisfying $\|\xi_i\|_\infty \leq B$ for all $i = 1, \ldots, n$. Then, for all $\tau > 0$, we have*

$$\mathrm{P}\left(\Bigl\|\frac{1}{n}\sum_{i=1}^n (\xi_i - \mathbb{E}_\mathrm{P}\xi_i)\Bigr\|_H \geq B\sqrt{\frac{2\tau}{n}} + B\sqrt{\frac{1}{n}} + \frac{4B\tau}{3n}\right) \leq e^{-\tau}.$$

Proof. Let us define $\eta_i := \xi_i - \mathbb{E}_\mathrm{P}\xi_i$, $i = 1, \ldots, n$. Then we have $\mathbb{E}_\mathrm{P}\eta_i = 0$, $\|\eta_i\|_\infty \leq 2B$, and

$$\mathbb{E}_\mathrm{P}\|\eta_i\|_H^2 = \mathbb{E}_\mathrm{P}\langle\xi_i, \xi_i\rangle - 2\mathbb{E}_\mathrm{P}\langle\xi_i, \mathbb{E}_\mathrm{P}\xi_i\rangle + \langle\mathbb{E}_\mathrm{P}\xi_i, \mathbb{E}_\mathrm{P}\xi_i\rangle \leq \mathbb{E}_\mathrm{P}\langle\xi_i, \xi_i\rangle \leq B^2$$

for all $i = 1, \ldots, n$. Applying Theorem 6.14 to $\eta_1, \ldots, \eta_n$ then yields the assertion. $\square$

6.3 Statistical Analysis of Empirical Risk Minimization

In this section, we investigate statistical properties for empirical risk minimization. Although this learning method is not our primary object of interest, there are two reasons why we consider it before investigating SVMs. First, the method is so elementary that the basic ideas of its analysis are not hidden by technical considerations. This will give us good preparation for the more involved analysis of SVMs in Section 6.4. Second, the results we establish will be utilized in Section 6.5, where we investigate how the regularization parameter of SVMs can be chosen in a data-dependent and adaptive way.

Let us begin by formally introducing empirical risk minimization.

Definition 6.16. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss and $\mathcal{F} \subset \mathcal{L}_0(X)$ be a non-empty set. A learning method whose decision functions f_D satisfy*

$$\mathcal{R}_{L,\mathrm{D}}(f_D) = \inf_{f \in \mathcal{F}} \mathcal{R}_{L,\mathrm{D}}(f) \tag{6.12}$$

for all $n \geq 1$ and $D \in (X \times Y)^n$ is called ***empirical risk minimization (ERM)*** *with respect to L and $\mathcal{F}$.*

By definition, empirical risk minimization produces decision functions that minimize the empirical risk over $\mathcal{F}$. The *motivation* for this approach is based on the law of large numbers which says that for *fixed* $f \in \mathcal{F}$ we have

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{D}}(f) = \mathcal{R}_{L,\mathrm{P}}(f)$$

if the training sets D of length n are identically and independently distributed according to some probability measure P on $X \times Y$. This limit relation suggests that in order to find a minimizer of the true risk $\mathcal{R}_{L,\mathrm{P}}$, it suffices to find a minimizer of its empirical approximation $\mathcal{R}_{L,\mathrm{D}}$. Unfortunately, however, minimizing $\mathcal{R}_{L,\mathrm{D}}$ over $\mathcal{F} := \mathcal{L}_0(X)$ or $\mathcal{F} := \mathcal{L}_\infty(X)$ can lead to "overfitted" decision functions, as discussed in Exercise 6.7, and hence ERM typically minimizes over a smaller set $\mathcal{F}$ of functions. Moreover, note that for general losses L and sets of functions $\mathcal{F}$ there does not necessarily exist a function f_D satisfying (6.12). In addition, there are also situations in which multiple minimizers exist, and consequently one should always be aware that ERM usually is not a uniquely determined learning method. Let us now show that there usually exists a measurable ERM if there exists an ERM.[1]

Lemma 6.17 (Measurability of ERM). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss and $\mathcal{F} \subset \mathcal{L}_0(X)$ be a subset that is equipped with a complete and separable metric dominating the pointwise convergence. Then, if there exists an ERM, there also exists a measurable ERM.*

[1] Since this is little more than a technical requirement for the following results, the first-time reader may skip this lemma.

Before we present the proof of this lemma, let us first note that *finite* subsets $\mathcal{F} \subset \mathcal{L}_0(X)$ equipped with the discrete metric satisfy the assumptions above. In addition, the existence of an ERM is automatically guaranteed in this case, and hence there always exists a measurable ERM for finite $\mathcal{F}$. Similarly, for *closed, separable* $\mathcal{F} \subset \mathcal{L}_\infty(X)$, there exists a measurable ERM whenever there exists an ERM.

Proof. Lemma 2.11 shows that the map $(x, y, f) \mapsto L(x, y, f(x))$ defined on $X \times Y \times \mathcal{F}$ is measurable. From this it is easy to conclude that the map $\varphi : (X \times Y)^n \times \mathcal{F} \to [0, \infty)$ defined by

$$\varphi(D, f) := \mathcal{R}_{L,\mathrm{D}}(f)\,, \qquad\qquad D \in (X \times Y)^n,\ f \in \mathcal{F},$$

is measurable with respect to the product topology of $(X \times Y)^n \times \mathcal{F}$. By taking $F(D) := \mathcal{F}$, $D \in (X \times Y)^n$, in Aumann's measurable selection principle (see Lemma A.3.18), we thus see that there exists an ERM such that $D \mapsto f_D$ is measurable with respect to the universal completion of the product σ-algebra of $(X \times Y)^n$. Consequently, the map $(X \times Y)^n \times X \to \mathcal{F} \times X$ defined by $(D, x) \mapsto (f_D, x)$ is measurable with respect to the universal completion of the product σ-algebra of $(X \times Y)^n \times X$. In addition, Lemma 2.11 shows that the map $\mathcal{F} \times X \to \mathbb{R}$ defined by $(f, x) \mapsto f(x)$ is measurable. Combining both maps, we then obtain the measurability of $(D, x) \mapsto f_D(x)$. □

Let us now analyze the statistical properties of ERM. To this end, let us assume that $\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} < \infty$. Moreover, let us fix a $\delta > 0$ and a function $f_\delta \in \mathcal{F}$ such that $\mathcal{R}_{L,\mathrm{P}}(f_\delta) \leq \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} + \delta$. Then a simple calculation shows

$$\begin{aligned}
\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} &\leq \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}_{L,\mathrm{D}}(f_D) + \mathcal{R}_{L,\mathrm{D}}(f_D) - \mathcal{R}_{L,\mathrm{P}}(f_\delta) + \delta \\
&\leq \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}_{L,\mathrm{D}}(f_D) + \mathcal{R}_{L,\mathrm{D}}(f_\delta) - \mathcal{R}_{L,\mathrm{P}}(f_\delta) + \delta \\
&\leq 2 \sup_{f \in \mathcal{F}} \bigl| \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f) \bigr| + \delta\,,
\end{aligned}$$

and by letting $\delta \to 0$ we thus find

$$\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} \;\leq\; 2 \sup_{f \in \mathcal{F}} \bigl| \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f) \bigr|\,. \tag{6.13}$$

Let us now assume that $\mathcal{F}$ is a finite set with cardinality $|\mathcal{F}|$ and that $B > 0$ is a real number such that

$$L(x, y, f(x)) \leq B\,, \qquad\qquad (x, y) \in X \times Y,\ f \in \mathcal{F}. \tag{6.14}$$

Note that the latter assumption ensures the earlier imposed $\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} < \infty$. For a measurable ERM, (6.13) together with Hoeffding's inequality then yields

$$
\begin{aligned}
&\mathrm{P}^n\left(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} \geq B\sqrt{\frac{2\tau}{n}}\right) \\
&\leq \mathrm{P}^n\left(D \in (X \times Y)^n : \sup_{f\in\mathcal{F}} \left|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\right| \geq B\sqrt{\frac{\tau}{2n}}\right) \\
&\leq \sum_{f\in\mathcal{F}} \mathrm{P}^n\left(D \in (X \times Y)^n : \left|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\right| \geq B\sqrt{\frac{\tau}{2n}}\right) \\
&\leq 2\,|\mathcal{F}|e^{-\tau}\,. \qquad (6.15)
\end{aligned}
$$

By elementary algebraic transformations, we thus find the following result.

Proposition 6.18 (Oracle inequality for ERM). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, $\mathcal{F} \subset \mathcal{L}_0(X)$ be a non-empty finite set, and $B > 0$ be a constant such that (6.14) holds. Then, for all measurable ERMs, all distributions* P *on $X \times Y$, and all $\tau > 0$, $n \geq 1$, we have*

$$
\mathrm{P}^n\left(D \in (X\times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) < \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2\,|\mathcal{F}|)}{n}}\right) \;\geq\; 1 - e^{-\tau}\,.
$$

Inequalities like the one above are called **oracle inequalities** since they compare the empirically obtained decision function with the one an omniscient oracle, having an infinite amount of observation, would obtain when pursuing the same goal, which in the case above is minimizing the L-risk over $\mathcal{F}$.

Proposition 6.18 shows that with high probability the function f_D approximately minimizes the risk $\mathcal{R}_{L,\mathrm{P}}$ in $\mathcal{F}$. In other words, the heuristic of replacing the unknown risk $\mathcal{R}_{L,\mathrm{P}}$ by the empirical risk $\mathcal{R}_{L,\mathrm{D}}$ is justified for *finite* sets $\mathcal{F}$. However, the assumption that $\mathcal{F}$ is finite is quite restrictive, and hence our next goal is to remove it. To this end we first observe that we cannot use a simple limit argument for $|\mathcal{F}| \to \infty$ in Proposition 6.18 since the term $B\sqrt{2\tau + 2\ln(2\,|\mathcal{F}|)}\,n^{-1/2}$ is unbounded in $|\mathcal{F}|$. To resolve this problem we introduce the following fundamental concept, which will enable us to approximate infinite $\mathcal{F}$ by *finite* subsets.

Definition 6.19. *Let (T,d) be a metric space and $\varepsilon > 0$. We call $S \subset T$ an ε-**net** of T if for all $t \in T$ there exists an $s \in S$ with $d(s,t) \leq \varepsilon$. Moreover, the ε-**covering number** of T is defined by*

$$
\mathcal{N}(T,d,\varepsilon) \;:=\; \inf\left\{n \geq 1 : \exists\, s_1,\ldots,s_n \in T \text{ such that } T \subset \bigcup_{i=1}^n B_d(s_i,\varepsilon)\right\},
$$

where $\inf\emptyset := \infty$ and $B_d(s,\varepsilon) := \{t \in T : d(t,s) \leq \varepsilon\}$ denotes the closed ball with center $s \in T$ and radius ε.

Moreover, if (T,d) is a subspace of a normed space $(E, \|\cdot\|)$ and the metric is given by $d(x,x') = \|x - x'\|$, $x, x' \in T$, we write $\mathcal{N}(T, \|\cdot\|, \varepsilon) := \mathcal{N}(T,d,\varepsilon)$.

In simple words, an ε-net approximates T up to ε. Moreover, the covering number $\mathcal{N}(T,d,\varepsilon)$ is the size of the smallest possible ε-net, i.e., it is the smallest number of points that are needed to approximate the set T up to ε. Note that if T is compact, then all covering numbers are finite, i.e., $\mathcal{N}(T,d,\varepsilon) < \infty$ for all $\varepsilon > 0$. Moreover, $\mathcal{N}(T,d,\varepsilon)$ is a decreasing function in ε and $\sup_{\varepsilon>0} \mathcal{N}(T,d,\varepsilon) < \infty$ if and only if T is finite.

Besides covering numbers, we will also need the following "inverse" concept.

Definition 6.20. *Let (T,d) be a metric space and $n \geq 1$ be an integer. Then the n-th **(dyadic) entropy number** of (T,d) is defined by*

$$e_n(T,d) \;:=\; \inf\left\{\varepsilon > 0 : \exists\, s_1,\ldots,s_{2^{n-1}} \in T \text{ such that } T \subset \bigcup_{i=1}^{2^{n-1}} B_d(s_i,\varepsilon)\right\}.$$

Moreover, if (T,d) is a subspace of a normed space $(E,\|\cdot\|)$ and the metric d is given by $d(x,x') = \|x-x'\|$, $x,x' \in T$, we write

$$e_n(T,\|\cdot\|) := e_n(T,E) := e_n(T,d)\,.$$

Finally, if $S : E \to F$ is a bounded, linear operator between the normed spaces E and F, we write $e_n(S) := e_n(SB_E, \|\cdot\|_F)$.

Note that the (dyadic) entropy numbers consider ε-nets of cardinality 2^{n-1} instead of ε-nets of cardinality n. The reason for this is that this choice ensures that the entropy numbers share some basic properties with other s-numbers such as the singular numbers introduced in Section A.5.2. Basic properties of entropy numbers and their relation to singular numbers together with some bounds for important function classes can be found in Section A.5.6.

The following lemma shows that bounds on entropy numbers imply bounds on covering numbers (see Exercise 6.8 for the inverse implication).

Lemma 6.21 (Equivalence of covering and entropy numbers). *Let (T,d) be a metric space and $a > 0$ and $q > 0$ be constants such that*

$$e_n(T,d) \;\leq\; a\, n^{-1/q}\,, \qquad\qquad n \geq 1.$$

Then, for all $\varepsilon > 0$, we have

$$\ln \mathcal{N}(T,d,\varepsilon) \;\leq\; \ln(4)\cdot\left(\frac{a}{\varepsilon}\right)^q.$$

Proof. Let us fix a $\delta > 0$ and an $\varepsilon \in (0,a]$. Then there exists an integer $n \geq 1$ such that

$$a(1+\delta)(n+1)^{-1/q} \;\leq\; \varepsilon \;\leq\; a(1+\delta)n^{-1/q}\,. \tag{6.16}$$

Since $e_n(T,d) < a(1+\delta)n^{-1/q}$, there then exists an $a(1+\delta)n^{-1/q}$-net S of T with $|S| \leq 2^{n-1}$, i.e., we have

$$\mathcal{N}\bigl(T, d, a(1+\delta)n^{-1/q}\bigr) \;\leq\; 2^{n-1}\,.$$

Moreover, (6.16) implies $2^{1/q}\varepsilon \geq 2^{1/q}a(1+\delta)(n+1)^{-1/q} \geq a(1+\delta)n^{-1/q}$ and $n \leq \bigl(\frac{(1+\delta)a}{\varepsilon}\bigr)^q$. Consequently, we obtain

$$\ln \mathcal{N}(T, d, 2^{\frac{1}{q}}\varepsilon) \leq \ln\bigl(2\,\mathcal{N}(T, d, a(1+\delta)n^{-\frac{1}{q}})\bigr) \leq n \ln 2 \leq \ln(2) \cdot \left(\frac{(1+\delta)a}{\varepsilon}\right)^q .$$

Since $\ln \mathcal{N}(T, d, \varepsilon) = 0$ for all $\varepsilon > a$, we then find the assertion. □

With the help of covering numbers, we can now investigate the statistical properties of ERM over certain *infinite* sets $\mathcal{F}$.

Proposition 6.22 (Oracle inequality for ERM). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a locally Lipschitz continuous loss and P be a distribution on $X \times Y$. Moreover, let $\mathcal{F} \subset \mathcal{L}_\infty(X)$ be non-empty and compact, and $B > 0$ and $M > 0$ be constants satisfying (6.14) and $\|f\|_\infty \leq M$, $f \in \mathcal{F}$, respectively. Then, for all measurable ERMs and all $\varepsilon > 0$, $\tau > 0$, and $n \geq 1$, we have*

$$\mathrm{P}^n\left(\mathcal{R}_{L,\mathrm{P}}(f_D) \geq \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2\,\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon))}{n}} + 4\varepsilon|L|_{M,1}\right) \leq e^{-\tau}\,.$$

Before we prove this proposition, let us first note that the compactness of $\mathcal{F}$ together with the continuity of $\mathcal{R}_{L,\mathrm{D}} : \mathcal{L}_\infty(X) \to [0,\infty)$ ensures the existence of an empirical risk minimizer. Moreover, the compactness of $\mathcal{F}$ implies that $\mathcal{F}$ is a closed and separable subset of $\mathcal{L}_\infty(X)$. The remarks after Lemma 6.17 then show that there *exists* a measurable ERM. In addition, Proposition 6.22 remains true if one replaces the compactness assumption by $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) < \infty$ for all $\varepsilon > 0$. However, in this case the *existence* of an ERM is no longer "automatically" guaranteed. Finally, if L is not locally Lipschitz continuous, variants of Proposition 6.22 still hold if the covering numbers are replaced by other notions measuring the "size" of $\mathcal{F}$. For the classification loss, corresponding results are briefly mentioned in Section 6.6.

Proof. For a fixed $\varepsilon > 0$, the compactness of $\mathcal{F}$ shows that there exists an ε-net $\mathcal{F}_\varepsilon$ of $\mathcal{F}$ with $|\mathcal{F}_\varepsilon| = \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) < \infty$. For $f \in \mathcal{F}$, there thus exists a $g \in \mathcal{F}_\varepsilon$ with $\|f - g\|_\infty \leq \varepsilon$, and hence we find

$$\begin{aligned}
&\bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr| \\
&\leq \bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| + \bigl|\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr| + \bigl|\mathcal{R}_{L,\mathrm{D}}(g) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr| \\
&\leq 2\,\varepsilon\,|L|_{M,1} + \bigl|\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr|\,,
\end{aligned}$$

where in the last step we used the local Lipschitz continuity of the L-risks established in Lemma 2.19. By taking suprema on the right- and left-hand sides we thus obtain

$$\sup_{f\in\mathcal{F}}\bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr| \;\leq\; 2\,\varepsilon\,|L|_{M,1} + \sup_{g\in\mathcal{F}_\varepsilon}\bigl|\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr|\,,$$

and combining this estimate with (6.13), the latter inequality leads to

$$\begin{aligned}
&\mathrm{P}^n\left(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} \geq B\sqrt{\frac{2\tau}{n}} + 4\varepsilon|L|_{M,1}\right) \\
&\leq \mathrm{P}^n\left(D \in (X \times Y)^n : \sup_{g \in \mathcal{F}_\varepsilon} \left|\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\right| \geq B\sqrt{\frac{\tau}{2n}}\right) \\
&\leq 2\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)\, e^{-\tau} \qquad (6.17)
\end{aligned}$$

for all $\varepsilon, \tau > 0$. Some algebraic transformations then yield the assertion. □

6.4 Basic Oracle Inequalities for SVMs

In Section 6.3, we introduced some basic techniques to analyze the statistical properties of empirical risk minimization. Since the only difference between ERM and SVMs is the additional regularization term $\lambda\|\cdot\|_H^2$, it seems plausible that these techniques can be adapted to the analysis of SVMs. This will be the idea of the second oracle inequality for SVMs we establish in this section. Moreover, we will also provide some bounds on the covering numbers for certain RKHSs. First, however, we will present another technique for establishing oracle inequalities for SVMs. This technique, which requires fewer assumptions on the kernel and the input space, combines a stability argument with the Hilbert space valued version of Hoeffding's inequality proved in Section 6.2. Finally, we illustrate how the established oracle inequalities can be used to establish both consistency and learning rates for SVMs.

Before we present the first oracle inequality, we have to ensure that SVMs are measurable learning methods. This is done in the following lemma.

Lemma 6.23 (Measurability of SVMs). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss and H be a separable RKHS over X with measurable kernel k. Then, for all $\lambda > 0$, the corresponding SVM that produces the decision functions $f_{\mathrm{D},\lambda}$ for $D \in (X \times Y)^n$ and $n \geq 1$ is a measurable learning method, and the maps $D \mapsto f_{\mathrm{D},\lambda}$ mapping $(X \times Y)^n$ to H are measurable.*

Proof. Obviously, H is a separable metric space and Lemma 4.24 ensures $H \subset \mathcal{L}_0(X)$. Moreover, the Dirac functionals are continuous on H by the definition of RKHSs, and hence the metric of H dominates the pointwise convergence. Finally, the norm $\|\cdot\|_H : H \to \mathbb{R}$ is continuous and hence measurable. Analogously to the proof of Lemma 6.17, we hence conclude that $\varphi : (X \times Y)^n \times H \to [0,\infty)$ defined by

$$\varphi(D, f) := \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)\,, \qquad D \in (X \times Y)^n,\ f \in H,$$

is measurable. In addition, Lemma 5.1 shows that $f_{\mathrm{D},\lambda}$ is the only element in H satisfying

$$\varphi(D, f_{\mathrm{D},\lambda}) = \inf_{f \in H} \varphi(D, f)\,, \qquad D \in (X \times Y)^n,$$

and consequently the measurability of $D \mapsto f_{\mathrm{D},\lambda}$ with respect to the universal completion of the product σ-algebra of $(X \times Y)^n$ follows from Aumann's measurable selection principle (see Lemma A.3.18). As in the proof of Lemma 6.17, we then obtain the first assertion. □

Let us recall that in this chapter we always assume that $(X \times Y)^n$ is equipped with the universal completion of the product σ-algebra of $(X \times Y)^n$. In addition, given a distribution P on $X \times Y$, we always write P^n for the canonical extension of the n-fold product measure of P to this completion. Note that these conventions together with Lemmas 6.23 and 6.3 make it possible to ignore measurability questions for SVMs.

Let us now establish a first oracle inequality for SVMs.

Theorem 6.24 (Oracle inequality for SVMs). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex, locally Lipschitz continuous loss satisfying $L(x, y, 0) \le 1$ for all $(x, y) \in X \times Y$, H be a separable RKHS over X with measurable kernel k satisfying $\|k\|_\infty \le 1$, and P be a distribution on $X \times Y$. For fixed $\lambda > 0$, $n \ge 1$, and $\tau > 0$, we then have with probability P^n not less than $1 - e^{-\tau}$ that*

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} < A_2(\lambda) + \lambda^{-1} |L|^2_{\lambda^{-\frac{1}{2}},1} \left(\sqrt{\frac{8\tau}{n}} + \sqrt{\frac{4}{n}} + \frac{8\tau}{3n} \right),$$

where $A_2(\,\cdot\,)$ denotes the corresponding approximation error function.

Before we prove Theorem 6.24, we note that the condition $L(x, y, 0) \le 1$ is satisfied for all margin-based losses $L(y, t) = \varphi(yt)$ for which we have $\varphi(0) \le 1$. In particular, all examples considered in Section 2.3, namely the (truncated) least squares loss, the hinge loss, and the logistic loss for classification, fall into this category. Furthermore, "restricted" distance-based losses i.e., losses $L : [-1, 1] \times \mathbb{R} \to [0, \infty)$ of the form $L(y, t) = \psi(y - t)$, $y \in [-1, 1]$, $t \in \mathbb{R}$, satisfy $L(y, 0) \le 1$, $y \in [-1, 1]$, if and only if $\psi(r) \le 1$ for all $r \in [-1, 1]$. Note that the least squares loss, the logistic loss for regression, Huber's loss for $\alpha \le \sqrt{2}$, the ϵ-insensitive loss, and the pinball loss satisfy this assumption.

Proof. Let $\Phi : X \to H$ denote the canonical feature map of k. By Corollary 5.10, there exists a bounded measurable function $h : X \times Y \to \mathbb{R}$ such that

$$\|h\|_\infty \le |L|_{\lambda^{-1/2},1}\,,$$
$$\big\| f_{\mathrm{P},\lambda} - f_{\mathrm{D},\lambda} \big\|_H \le \frac{1}{\lambda} \big\| \mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\mathrm{D}} h\Phi \big\|_H\,,$$

for all $D \in (X \times Y)^n$. Moreover, since $\|h(x, y)\Phi(x)\|_H \le \|h\|_\infty \le |L|_{\lambda^{-1/2},1}$ for all $(x, y) \in X \times Y$, we find by Corollary 6.15 that

$$\mathrm{P}^n \left(D \in (X \times Y)^n : \big\| \mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\mathrm{D}} h\Phi \big\|_H \ge |L|_{\lambda^{-1/2},1} \left(\sqrt{\frac{2\tau}{n}} + \sqrt{\frac{1}{n}} + \frac{4\tau}{3n} \right) \right) \le e^{-\tau}.$$

Combining these estimates yields

$$\mathrm{P}^n\left(D\in(X\times Y)^n : \|f_{\mathrm{D},\lambda}-f_{\mathrm{P},\lambda}\|_H \geq \lambda^{-1}|L|_{\lambda^{-\frac{1}{2}},1}\left(\sqrt{\frac{2\tau}{n}}+\sqrt{\frac{1}{n}}+\frac{4\tau}{3n}\right)\right) \leq e^{-\tau}.$$

Furthermore, $\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) \leq \lambda\|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{P},\lambda})$ implies

$$\begin{aligned}
&\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} - A_2(\lambda)\\
&= \lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \lambda\|f_{\mathrm{P},\lambda}\|_H^2 - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})\\
&= \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda})\\
&\qquad +\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) - \lambda\|f_{\mathrm{P},\lambda}\|_H^2 - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})\\
&\leq \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{P},\lambda}) - \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda})\,. \qquad (6.18)
\end{aligned}$$

Moreover, $\|f_{\mathrm{Q},\lambda}\|_\infty \leq \|f_{\mathrm{Q},\lambda}\|_H \leq \lambda^{-1/2}$ holds for all distributions Q on $X\times Y$ by Lemma 4.23, (5.4), and $\mathcal{R}_{L,\mathrm{Q}}(0) \leq 1$. Consequently, for every distribution Q on $X\times Y$, we have

$$\mathcal{R}_{L,\mathrm{Q}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{Q}}(f_{\mathrm{P},\lambda}) \leq |L|_{\lambda^{-1/2},1}\|f_{\mathrm{D},\lambda} - f_{\mathrm{P},\lambda}\|_H$$

by Lemma 2.19. Applying this estimate to (6.18) twice yields

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} - A_2(\lambda) \leq 2|L|_{\lambda^{-1/2},1}\|f_{\mathrm{D},\lambda} - f_{\mathrm{P},\lambda}\|_H\,,$$

and by combining this inequality with the above concentration inequality we obtain the assertion. □

We will later see that a key feature of the oracle inequality above is the fact that it holds under somewhat minimal assumptions. In addition, the technique used in its proof is very flexible, as we will see, e.g., in Chapter 9 when dealing with regression problems having *unbounded* noise. On the downside, however, the oracle inequality above often leads to suboptimal learning rates. In order to illustrate this, we first need the following oracle inequality.

Theorem 6.25 (Oracle inequality for SVMs using benign kernels). *Let X be a compact metric space and $L : X\times Y\times\mathbb{R}\to[0,\infty)$ be a convex, locally Lipschitz continuous loss satisfying $L(x,y,0)\leq 1$ for all $(x,y)\in X\times Y$. Moreover, let H be the RKHS of a continuous kernel k on X with $\|k\|_\infty\leq 1$ and P be a probability measure on $X\times Y$. Then, for fixed $\lambda>0$, $n\geq 1$, $\varepsilon>0$, and $\tau>0$, we have with probability P^n not less than $1-e^{-\tau}$ that*

$$\begin{aligned}
&\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H}\\
&< A_2(\lambda) + 4\varepsilon|L|_{\lambda^{-\frac{1}{2}},1} + \big(|L|_{\lambda^{-\frac{1}{2}},1}\lambda^{-\frac{1}{2}}+1\big)\sqrt{\frac{2\tau+2\ln\big(2\mathcal{N}(B_H,\|\cdot\|_\infty,\lambda^{\frac{1}{2}}\varepsilon)\big)}{n}}.
\end{aligned}$$

Proof. By Corollary 4.31, $\mathrm{id} : H \to C(X)$ is compact, i.e., the $\|\cdot\|_\infty$-closure $\overline{B_H}$ of the unit ball B_H is a compact subset of $C(X)$. From this we conclude that $\mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) < \infty$ for all $\varepsilon > 0$. In addition, the compactness of X implies that X is separable, and hence Lemma 4.33 shows that H is separable. Consequently, the SVM is measurable. Moreover, from (6.18) and $\|f_{\mathrm{Q},\lambda}\|_\infty \leq \|f_{\mathrm{Q},\lambda}\|_H \leq \lambda^{-1/2}$ for all distributions Q on $X \times Y$, we conclude that

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} - A_2(\lambda) \leq 2 \sup_{\|f\|_H \leq \lambda^{-1/2}} |\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)| \,.$$

In addition, for $f \in \lambda^{-1/2} B_H$ and $B := |L|_{\lambda^{-1/2},1} \lambda^{-1/2} + 1$, we have

$$\big|L(x,y,f(x))\big| \leq \big|L(x,y,f(x)) - L(x,y,0)\big| + L(x,y,0) \leq B$$

for all $(x,y) \in X \times Y$. Now let $\mathcal{F}_\varepsilon$ be an ε-net of $\lambda^{-1/2} B_H$ with cardinality

$$|\mathcal{F}_\varepsilon| = \mathcal{N}\big(\lambda^{-1/2} B_H, \|\cdot\|_\infty, \varepsilon\big) = \mathcal{N}\big(B_H, \|\cdot\|_\infty, \lambda^{1/2}\varepsilon\big) \,.$$

As in (6.17), we then conclude that for $\tau > 0$ we have

$$\begin{aligned} &\mathrm{P}^n\left(\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} \geq A_2(\lambda) + B\sqrt{\frac{2\tau}{n}} + 4\varepsilon |L|_{\lambda^{-1/2},1}\right) \\ &\leq 2\,\mathcal{N}(B_H, \|\cdot\|_\infty, \lambda^{1/2}\varepsilon)\, e^{-\tau} \,. \end{aligned}$$

By simple algebraic calculations, we then obtain the assertion. □

The right-hand side of the oracle inequality of Theorem 6.25 involves $\|\cdot\|_\infty$-covering numbers of the unit ball B_H of the RKHS H. By Corollary 4.31, these covering numbers are finite, and hence the right-hand side is non-trivial for certain values of ε, λ, and n. In order to derive consistency and learning rates from Theorem 6.25, however, we need quantitative statements on the covering numbers. This is the goal of the following two results, which for later purposes are stated in terms of entropy numbers reviewed in Section A.5.6.

Theorem 6.26 (Entropy numbers for smooth kernels). *Let $\Omega \subset \mathbb{R}^d$ be an open subset, $m \geq 1$, and k be an m-times continuously differentiable kernel on Ω. Moreover, let $X \subset \Omega$ be a closed Euclidean ball and let $H_{|X}$ denote the RKHS of the restricted kernel $k_{|X\times X}$. Assume that we have an $r_0 \in (1,\infty]$ such that $rX \subset \Omega$ for all $r \in [1, r_0)$. Then there exists a constant $c_{m,d,k}(X) > 0$ such that*

$$e_i\big(\mathrm{id} : H_{|rX} \to \ell_\infty(rX)\big) \;\leq\; c_{m,d,k}(X)\, r^m\, i^{-m/d} \,, \qquad i \geq 1, r \in [1, r_0).$$

Proof. By the definition of the space $C^0(\overline{r\mathring{X}})$ given in Section A.5.5, there exists a (unique) norm-preserving extension operator $\hat{} : C^0(\overline{r\mathring{X}}) \to C(rX)$, i.e., we have $\hat{f}_{|r\mathring{X}} = f$ and $\|f\|_\infty = \|\hat{f}\|_\infty$ for all $f \in C^0(\overline{r\mathring{X}})$. Moreover, recall

that Corollary 4.36 showed that the RKHS H of k is embedded into $C^m(\Omega)$, and using the compactness of X together with (4.24) we then conclude that the restriction operator $\cdot_{|r\mathring{X}} : H_{|rX} \to C^m(\overline{r\mathring{X}})$ is continuous. Since $H_{|rX}$ consists of continuous functions, we thus obtain the commutative diagram

$$\begin{array}{ccc} H_{|rX} & \xrightarrow{\text{id}} & C(rX) \\ \downarrow{\scriptstyle \cdot_{|r\mathring{X}}} & & \uparrow{\scriptstyle \hat{}} \\ C^m(\overline{r\mathring{X}}) & \xrightarrow[\text{id}]{} & C^0(\overline{r\mathring{X}}) \end{array}$$

Now the multiplicity (A.38) together with (A.46), (A.47), (A.40), and the fact that $C(rX)$ is isometrically embedded into $\ell_\infty(rX)$ yields the assertion. □

Let us briefly translate the result above into the language of covering numbers. To this end, we assume that X and k satisfy the assumptions of Theorem 6.26. Lemma 6.21 then shows that

$$\ln \mathcal{N}(B_{H_{|rX}}, \|\cdot\|_\infty, \varepsilon) \;\le\; a\,\varepsilon^{-2p}, \qquad \varepsilon > 0. \tag{6.19}$$

for $2p := d/m$ and $a := \ln(4) \cdot (c_{m,d}(X))^{d/m}\, r^d$. Now recall that Taylor and Gaussian RBF kernels are infinitely often differentiable and hence (6.19) holds for arbitrarily small $p > 0$. For Gaussian RBF kernels, however, the parameter γ is usually *not* fixed (see Section 8.2), and hence it is important to know how the constant a depends on γ. This is the goal of the next theorem.

Theorem 6.27 (Entropy numbers for Gaussian kernels). *Let $X \subset \mathbb{R}^d$ be a closed Euclidean ball and $m \ge 1$ be an integer. Then there exists a constant $c_{m,d}(X) > 0$ such that, for all $0 < \gamma \le r$ and all $i \ge 1$, we have*

$$e_i\big(\text{id} : H_\gamma(rX) \to \ell_\infty(rX)\big) \;\le\; c_{m,d}(X)\, r^m\, \gamma^{-m} i^{-\frac{m}{d}}\,.$$

Proof. For $x \in r\gamma^{-1}X$ and $f \in H_\gamma(rX)$, we write $\tau_\gamma f(x) := f(\gamma x)$. Proposition 4.37 applied to the dilation factor γ, the kernel parameter 1, and the set $r\gamma^{-1}X$ then shows that $\tau_\gamma : H_\gamma(rX) \to H_1(r\gamma^{-1}X)$ is an isometric isomorphism. Moreover, the dilation $\tau_{1/\gamma} : \ell_\infty(r\gamma^{-1}X) \to \ell_\infty(rX)$ is clearly an isometric isomorphism, too. In addition, we have the commutative diagram

$$\begin{array}{ccc} H_\gamma(rX) & \xrightarrow{\text{id}} & \ell_\infty(rX) \\ \downarrow{\scriptstyle \tau_\gamma} & & \uparrow{\scriptstyle \tau_{1/\gamma}} \\ H_1(r\gamma^{-1}X) & \xrightarrow[\text{id}]{} & \ell_\infty(r\gamma^{-1}X) \end{array}$$

and hence we obtain the assertion by Theorem 6.26 and (A.38). □

Our last goal in this section is to illustrate how the above oracle inequalities can be used to establish both consistency and learning rates for SVMs. For conceptional simplicity, we thereby restrict our considerations to *Lipschitz continuous* losses L with $|L|_1 \leq 1$, but similar results can be easily derived for locally Lipschitz continuous losses, too. Now recall that the Lipschitz continuity together with $L(x, y, 0) \leq 1$ yields $L(x, y, t) \leq 1 + |t|$, and hence L is a P-integrable Nemitski loss of order 1 for all distributions P on $X \times Y$. In the following we further assume for simplicity that we use a *fixed* RKHS H that in addition is assumed to be dense in $L_1(\mu)$ for all distributions μ on X. Here we recall that we have intensively investigated such RKHSs in Section 4.6. Moreover, Theorem 5.31 showed for such H that $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$, i.e., the Bayes risk can be approximated by functions from H.

In the following, we only consider the situation of Theorem 6.25 since for Theorem 6.24 the results are similar (see Exercise 6.9 for precise statements and a comparison of the resulting learning rates). Since Theorem 6.25 involves covering numbers, we assume for simplicity that there exist constants $a \geq 1$ and $p > 0$ such that

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \;\leq\; a\,\varepsilon^{-2p}\,, \qquad\qquad \varepsilon > 0. \tag{6.20}$$

By Theorem 6.26 and Lemma 6.21, we see that both Taylor and Gaussian kernels satisfy this assumption for all $p > 0$. Moreover, we saw in Section 4.6 that *(a)* Taylor kernels often have RKHSs that are dense in $L_1(\mu)$ and *(b)* Gaussian kernels always satisfy this denseness assumption. Consequently, these kernels are ideal candidates for our discussion.

In order to illustrate the utility of the oracle inequalities obtained let us now fix a $\lambda \in (0, 1]$ and a $\tau \geq 1$. For

$$\varepsilon := \left(\frac{p}{2}\right)^{1/(1+p)} \left(\frac{2a}{n}\right)^{1/(2+2p)} \lambda^{-1/2}\,.$$

Theorem 6.25 together with Lemma A.1.5 and $(p+1)(2/p)^{p/(1+p)} \leq 3$ then shows that

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} < A_2(\lambda) + \frac{3}{\lambda^{1/2}} \left(2\Big(\frac{2a}{n}\Big)^{\frac{1}{2+2p}} + \Big(\frac{2\tau}{n}\Big)^{\frac{1}{2}}\right) \tag{6.21}$$

holds with probability P^n not less than $1 - e^{-\tau}$.

Let us now assume that for sample size n we choose a $\lambda_n \in (0, 1]$ such that $\lim_{n\to\infty} \lambda_n = 0$ and

$$\lim_{n\to\infty} \lambda_n^{1+p} n = \infty\,. \tag{6.22}$$

Lemma 5.15, see also (5.32), then shows that the right-hand side of (6.21) converges to 0, and hence we have $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}$ in probability. In other words, we have shown that, for RKHSs satisfying both the denseness assumption above and (6.20), the SVM is universally L-risk consistent whenever the regularization sequence tends to zero in a controlled way described by (6.22).

In order to establish learning rates, let us additionally assume that there exist constants $c > 0$ and $\beta \in (0, 1]$ such that

$$A_2(\lambda) \le c\lambda^{\beta}, \qquad \lambda \ge 0. \tag{6.23}$$

Then a straightforward calculation shows that the asymptotically best choice for λ_n in (6.21) is a sequence that behaves like $n^{-\frac{1}{(1+p)(2\beta+1)}}$ and that the resulting learning rate is given by

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}} \le C\sqrt{\tau}\, n^{-\frac{\beta}{(2\beta+1)(1+p)}}\Big) \ge 1 - e^{-\tau},$$

where C is a constant independent of τ and n. It is important to note that the regularization sequence (λ_n) that achieves this rate *depends* on β. Unfortunately, however, we will almost never know the value of β, and hence we cannot choose the "optimal" regularization sequence suggested by Theorem 6.25. In the following section, we will therefore investigate how this problem can be addressed by choosing λ in a data-dependent way.

6.5 Data-Dependent Parameter Selection for SVMs

In this section, we first present a simple method for choosing the regularization parameter λ in a *data-dependent* way. We will then show that this method is *adaptive* in the sense that it does not need to know characteristics of the distribution such as (6.23) to achieve the learning rates we obtained in the previous section by knowing these characteristics.

Let us begin by describing this parameter selection method, which in some sense is a simplification of cross-validation considered in Section 11.3.

Definition 6.28. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss that can be clipped at 1, H be an RKHS over X, and $\Lambda := (\Lambda_n)$ be a sequence of finite subsets $\Lambda_n \subset (0, 1]$. Given a $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$, we define*

$$D_1 := ((x_1, y_1), \ldots, (x_m, y_m)),$$
$$D_2 := ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n)),$$

where $m := \lfloor n/2 \rfloor + 1$ and $n \ge 3$. Then use D_1 as a training set by computing the SVM decision functions

$$f_{\mathrm{D}_1,\lambda} := \arg\min_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}_1}(f), \qquad \lambda \in \Lambda_n, \tag{6.24}$$

and use D_2 to determine λ by choosing a $\lambda_{D_2} \in \Lambda_n$ such that

$$\mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) = \min_{\lambda \in \Lambda_n} \mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda}), \tag{6.25}$$

where $\widehat{f}_{\mathrm{D}_1,\lambda}$ denotes the clipped version of $f_{\mathrm{D}_1,\lambda}$. Every learning method that produces the resulting decision functions $f_{\mathrm{D}_1,\lambda_{D_2}}$ is called a ***training validation support vector machine (TV-SVM)*** *with respect to Λ.*

Informally speaking, the idea of TV-SVMs[2] is to use the *training set* D_1 to build a couple of SVM decision functions and then use the decision function that best performs on the independent *validation set* D_2. Here we note that Theorem 5.5 ensures that the SVM solutions $f_{D_1,\lambda}$, $\lambda \in \Lambda_n$, found in the training step (6.24) exist, and hence there exists a TV-SVM. However, note that in general the validation step (6.25) does *not* provide a unique regularization parameter λ_{D_2}, and hence the TV-SVM, like ERM, is not a uniquely defined learning method. The following lemma shows that for all interesting cases there exists a measurable TV-SVM.

Lemma 6.29 (Measurability of TV-SVMs). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss that can be clipped at 1, and let H be a separable RKHS over X having a measurable kernel. Then there exists a measurable TV-SVM.*

Proof. Lemma 6.23 showed that $(D,x) \mapsto f_{D_1,\lambda}(x)$ is measurable, and hence $\varphi : (X \times Y)^n \times \Lambda_n \to [0,\infty)$ defined by

$$\varphi(D,\lambda) := \mathcal{R}_{L,D_2}(\widehat{f}_{D_1,\lambda}), \qquad D \in (X \times Y)^n,\ \lambda \in \Lambda_n,$$

is measurable. The rest of the proof is analogous to the proofs of Lemmas 6.17 and 6.23. □

Our next goal is to establish oracle inequalities for TV-SVMs. To this end, we need the following lemma that describes how the term on the right-hand side of our oracle inequalities for SVMs can be approximately minimized.

Lemma 6.30. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, H be the RKHS of a measurable kernel over X, P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,P,H} < \infty$, and $A_2 : [0,\infty) \to [0,\infty)$ be the corresponding approximation error function. We fix a bounded interval $I \subset (0,\infty)$. In addition, let $\alpha, c \in (0,\infty)$ be two constants and Λ be a finite ε-net of I for some fixed $\varepsilon > 0$. Then we have*

$$\min_{\lambda\in\Lambda}\bigl(A_2(\lambda) + c\lambda^{-\alpha}\bigr) \;\le\; A_2(2\varepsilon) + \inf_{\lambda\in I}\bigl(A_2(\lambda) + c\lambda^{-\alpha}\bigr).$$

Proof. Let us assume that Λ is of the form $\Lambda = \{\lambda_1,\ldots,\lambda_m\}$ with $\lambda_{i-1} < \lambda_i$ for all $i = 2,\ldots,m$. We write $\lambda_0 := \inf I$. Our first goal is to show that

$$\lambda_i - \lambda_{i-1} \;\le\; 2\varepsilon, \qquad i = 1,\ldots,m. \tag{6.26}$$

To this end, we fix an $i \in \{1,\ldots,m\}$ and write $\bar\lambda := (\lambda_i + \lambda_{i-1})/2 \in I \cup \{\lambda_0\}$. Since $\Lambda \cup \{\lambda_0\}$ is an ε-net of $I \cup \{\lambda_0\}$, we then have $\lambda_i - \bar\lambda \le \varepsilon$ or $\bar\lambda - \lambda_{i-1} \le \varepsilon$. Simple algebra shows that in both cases we find (6.26). For $\delta > 0$, we now fix a $\lambda^* \in I$ such that

$$A_2(\lambda^*) + c(\lambda^*)^{-\alpha} \le \inf_{\lambda\in I}\bigl(A_2(\lambda) + c\lambda^{-\alpha}\bigr) + \delta. \tag{6.27}$$

[2] For simplicity, we only consider (almost) equally sized data sets D_1 and D_2, but the following results and their proofs remain almost identical for different splits.

Then there exists an index $i \in \{1, \ldots, m\}$ such that $\lambda_{i-1} \leq \lambda^* \leq \lambda_i$, and by (6.26) we conclude that $\lambda^* \leq \lambda_i \leq \lambda^* + 2\varepsilon$. By the monotonicity and subadditivity of $A_2(\,\cdot\,)$ established in Lemma 5.15, we thus find

$$\begin{aligned}\min_{\lambda\in\Lambda}\bigl(A_2(\lambda) + c\lambda^{-\alpha}\bigr) \leq A_2(\lambda_i) + c\lambda_i^{-\alpha} &\leq A_2(\lambda^* + 2\varepsilon) + c(\lambda^*)^{-\alpha} \\ &\leq A_2(\lambda^*) + c(\lambda^*)^{-\alpha} + A_2(2\varepsilon)\,.\end{aligned}$$

Combining this estimate with (6.27) then yields the assertion. □

With the help of the Lemma 6.30, we can now establish our first oracle inequality for TV-SVMs. For simplicity, it only considers the situation investigated at the end of Section 6.4, but generalizations are easy to establish.

Theorem 6.31 (Oracle inequality for TV-SVMs and benign kernels). *Let X be a compact metric space and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex, Lipschitz continuous loss with $|L|_1 \leq 1$. Assume that L can be clipped at 1 and that it satisfies $L(x, y, 0) \leq 1$ for all $(x, y) \in X \times Y$. Furthermore, let H be the RKHS of a continuous kernel k on X satisfying $\|k\|_\infty \leq 1$ and*

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \;\leq\; a\varepsilon^{-2p}\,, \qquad\qquad \varepsilon > 0, \tag{6.28}$$

where $a \geq 1$ and $p > 0$ are constants. Moreover, for $n \geq 4$ and $\varepsilon > 0$, let $\Lambda_n \subset (0, 1]$ be a finite ε-net of $(0, 1]$ of cardinality $|\Lambda_n|$. For fixed $\tau > 0$ and $\tau_n := 2 + \tau + \ln|\Lambda_n|$, we then have with probability P^n not less than $1 - e^{-\tau}$ that

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P},H} < \inf_{\lambda\in(0,1]}\left(A_2(\lambda) + \frac{13}{\sqrt{\lambda}}\left(\left(\frac{a}{n}\right)^{\frac{1}{2+2p}} + \sqrt{\frac{\tau_n}{n}}\right)\right) + A_2(2\varepsilon)\,.$$

Consequently, if we use ε_n-nets Λ_n with $\varepsilon_n \to 0$ and $n^{-1}\ln|\Lambda_n| \to 0$, then the resulting TV-SVM is consistent for all P satisfying $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$. Finally, if $\varepsilon_n \leq 1/n$ and $|\Lambda_n|$ grows polynomially in n, then the TV-SVM learns with rate

$$n^{-\frac{\beta}{(2\beta+1)(1+p)}} \tag{6.29}$$

for all distributions P that satisfy $A_2(\lambda) \leq c\lambda^\beta$ for some constants $c > 0$ and $\beta \in (0, 1]$ and all $\lambda \geq 0$.

Proof. Let us define $m := \lfloor n/2 \rfloor + 1$. Since $m \geq n/2$, we obtain similarly to (6.21) that with probability P^m not less than $1 - |\Lambda_n|e^{-\tau}$ we have

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_1,\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} < A_2(\lambda) + \frac{3}{\lambda^{1/2}}\left(2\Bigl(\frac{4a}{n}\Bigr)^{\frac{1}{2+2p}} + \Bigl(\frac{2\tau+2}{n}\Bigr)^{\frac{1}{2}}\right)$$

for all $\lambda \in \Lambda_n$ *simultaneously*. In addition, we have $L(x, y, \mathcal{T}) \leq |L|_1 + L(x, y, 0) \leq 2 =: B$, and hence Proposition 6.18 yields

$$\mathrm{P}^{n-m}\left(D_2 : \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) < \inf_{\lambda\in\Lambda_n} \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda}) + 4\sqrt{\frac{2\tau+2\ln(2|\Lambda_n|)}{n}}\right) \geq 1-e^{-\tau},$$

where we used $n-m \geq n/2-1 \geq n/4$. Since $\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda}) \leq \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_1,\lambda})$, we conclude that with probability P^n not less than $1-(|\Lambda_n|+1)e^{-\tau}$ we have

$$\begin{aligned}
\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P},H} &< \inf_{\lambda\in\Lambda_n}\left(A_2(\lambda) + \frac{3}{\lambda^{1/2}}\left(2\Big(\frac{4a}{n}\Big)^{\frac{1}{2+2p}} + \Big(\frac{2\tau+2}{n}\Big)^{\frac{1}{2}}\right)\right) \\
&\quad + 4\sqrt{\frac{2\tau+2\ln(2|\Lambda_n|)}{n}} \\
&\leq \inf_{\lambda\in(0,1]}\left(A_2(\lambda) + \frac{3}{\lambda^{1/2}}\left(2\Big(\frac{4a}{n}\Big)^{\frac{1}{2+2p}} + \Big(\frac{2\tau+2}{n}\Big)^{\frac{1}{2}}\right)\right) \\
&\quad + A_2(2\varepsilon) + 4\sqrt{\frac{2\tau+2\ln(2|\Lambda_n|)}{n}}\,,
\end{aligned}$$

where in the last step we used Lemma 6.30. From this we easily obtain the first assertion. The second and third assertions then follow by the arguments used at the end of Section 6.4. □

Note that the preceding proof heavily relied on the assumption that L can be clipped. Indeed, without this assumption, Proposition 6.18 only shows that

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_1,\lambda_{D_2}}) < \inf_{\lambda\in\Lambda_n} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_1,\lambda}) + 4\sup_{\lambda\in\Lambda_n} \lambda^{-1/2}\sqrt{\frac{2\tau+2\ln(2|\Lambda_n|)}{n}}$$

holds with probability not less than $1-e^{-\tau}$. Since for n^{-1}-nets Λ_n of $(0,1]$ we have $\sup_{\lambda\in\Lambda_n}\lambda^{-1/2} \geq n^{1/2}$, it becomes obvious that the preceding proof does not provide consistency or the rates (6.29) if L cannot be clipped. In other words, the fact that L is clippable ensures that the error of the parameter selection step does *not* dominate the error of the SVM training step.

Let us now recall the end of Section 6.4, where we saw that SVMs satisfying the covering number assumption (6.28) and the approximation error assumption $A_2(\lambda) \leq c\lambda^\beta$ can learn with rate (6.29). Unfortunately, however, this rate required a regularization sequence $\lambda_n := n^{-\frac{1}{(1+p)(2\beta+1)}}$, i.e., the rate was only achievable if we had knowledge on the distribution P, the RKHS H, and their interplay. Of course, we almost never know the exponent β that bounds the approximation error function, and hence it remained unclear whether the learning rate (6.29) was actually realizable. Theorem 6.31 now shows that the TV-SVM does achieve this learning rate *without* knowing the exponent β. Moreover, the theorem also shows that we do not even have to know the exponent p in the covering number assumption (6.28) to achieve this rate. Of course, this p is independent of P and hence in principle *a priori* known. In practice, however, covering number bounds are often extremely difficult to establish for new RKHSs, and hence the independence of the TV-SVM from

this exponent is an important feature. Furthermore, the following oracle inequality for TV-SVMs shows that this learning method achieves non-trivial learning rates even if there is no exponent p satisfying (6.28).

Theorem 6.32 (Oracle inequality for TV-SVMs). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex Lipschitz continuous loss that can be clipped at 1 and that satisfies $|L|_1 \leq 1$ and $L(x,y,0) \leq 1$ for all $(x,y) \in X \times Y$. Furthermore, let H be a separable RKHS with measurable kernel k over X satisfying $\|k\|_\infty \leq 1$. Moreover, for $n \geq 4$ and $\varepsilon > 0$, let $\Lambda_n \subset (0,1]$ be a finite ε-net of $(0,1]$. For fixed $\tau > 0$, we then have with probability P^n not less than $1 - e^{-\tau}$ that*

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P},H} < \inf_{\lambda\in(0,1]} \left(A_2(\lambda) + \frac{14}{\lambda}\left(\sqrt{\frac{\tau+\ln(2|\Lambda_n|)}{n}} + \frac{\tau+\ln(2|\Lambda_n|)}{n} \right) \right) + A_2(2\varepsilon)\,.$$

In particular, if we use ε_n-nets Λ_n with $\varepsilon_n \to 0$ and $n^{-1}\ln|\Lambda_n| \to 0$, then the resulting TV-SVM is consistent for all P with $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$. Moreover, for $\varepsilon_n \leq n^{-1/2}$ and $|\Lambda_n|$ growing polynomially in n, the TV-SVM learns with rate

$$\left(\frac{\ln(n+1)}{n} \right)^{\frac{\beta}{2\beta+2}} \tag{6.30}$$

for all distributions P that satisfy $A_2(\lambda) \leq c\lambda^\beta$ for some constants $c > 0$ and $\beta \in (0,1]$ and all $\lambda \geq 0$.

Proof. Repeat the proof of Theorem 6.31, but use Theorem 6.24 instead of Theorem 6.25. □

Theorem 6.32 shows that the TV-SVM learns with a specific rate if an approximation error assumption is satisfied. Moreover, this rate equals the "optimal" rate we can derive from Theorem 6.24 up to a logarithmic factor (see Exercise 6.9), i.e., the TV-SVM is again adaptive with respect to the unknown exponent β bounding the approximation error function. Moreover, by combining the two oracle inequalities for the TV-SVM, we see that the TV-SVM is in some sense also adaptive to the size of the input domain. To illustrate this, let us consider the space $X := \mathbb{R}^d$. Moreover, assume that we have an RKHS H over X such that the covering number bound (6.28) is satisfied for some exponent $p(X')$ whenever we consider the restriction of H to some compact subset $X' \subset \mathbb{R}^d$. By combining Theorem 6.31 with Theorem 6.32, we then see that the TV-SVM learns with rate (6.30) if the support of P_X is not compact and with rate

$$\min\left\{ \left(\frac{\ln(n+1)}{n} \right)^{\frac{\beta}{2\beta+2}}, n^{-\frac{\beta}{(1+p(X'))(2\beta+1)}} \right\}$$

if $X' := \operatorname{supp}(\mathrm{P}_X)$ is compact. In this sense, the TV-SVM is adaptive not only to the approximation error assumption (6.23) but also to the input domain

of the data. Finally, note that these considerations can be refined using the more advanced techniques of the next chapter. We refer to Section 8.3, where this is worked out in detail for binary classification.

6.6 Further Reading and Advanced Topics

The first learning method that was shown to be universally consistent (see Stone, 1977) was the so-called nearest-neighbor method. Since then, universal consistency has been established for a variety of different methods. Many examples of such methods for classification and regression can be found in the books by Devroye *et al.* (1996) and Györfi *et al.* (2002), respectively. Moreover, besides the no-free-lunch theorem, which was proved by Devroye (1982), Devroye *et al.* (1996) also present some other fundamental limitations in statistical learning theory. These limitations include the non-existence of an overall best-performing classification method, the no-free-lunch theorem under certain additional assumptions on P, and the non-existence of a method that estimates the Bayes risk with a uniform rate. Moreover, learning rates (and their optimality) for certain regression methods are presented in great detail by Györfi *et al.* (2002).

The classical concentration inequalities presented in Section 6.2 were proven by Hoeffding (1963) and Bernstein (1946). Sharper versions of Bernstein's inequality were found by Bennett (1962) and Hoeffding (1963). For a more detailed discussion on these inequalities, we refer to Hoeffding (1963) and Bousquet (2003a). Finally, Theorem 6.13 and the Hilbert space valued versions of Bernstein's and Hoeffding's inequalities were taken from Chapter 3 of Yurinsky (1995). Note that the crucial step in deriving these Hilbert space valued versions is the estimate (6.10), which by symmetrization holds (up to some constant) in every Banach space of type 2. Moreover, weaker versions of (6.10) can actually be established whenever the Banach space has some non-trivial type. For more information on the type concept for Banach spaces, we refer to Chapter 11 of Diestel *et al.* (1995).

The discussion in Section 6.3 is nowadays folklore in the machine learning literature. The idea of estimating the excess risk of an empirical risk minimizer by a supremum (6.13) goes back to Vapnik and Chervonenkis (1974). Generalizations of this bound to infinite sets $\mathcal{F}$ require bounds on the "size" or "complexity" of $\mathcal{F}$. Probably the most classical such complexity measure is the so-called Vapnik-Chervonenkis (VC) dimension, which can be applied if, e.g., L is the binary classification loss. Furthermore, there are various extensions and generalizations of the VC dimension that make it possible to deal with other types of loss functions. We refer to the books by Vapnik (1998), Anthony and Bartlett (1999), and Vidyasagar (2002).

Using covering numbers as a complexity measure is another idea that frequently appears in the literature. Probably the easiest way to use these

numbers is presented in (6.17), but there also exist more sophisticated concentration inequalities, such as Lemma 3.4 of Alon *et al.* (1997) and Theorem 9.1 of Györfi *et al.* (2002), where the latter goes back to Pollard (1984). Covering numbers themselves were first investigated by Kolmogorov (1956) and Kolmogorov and Tikhomirov (1961). Since then various results for interesting function classes have been established. We refer to the books of Pinkus (1985), Carl and Stephani (1990), and Edmunds and Triebel (1996) for a detailed account and to Section A.5.6 for a brief overview.

Results similar to Theorem 6.25 were first established by Cucker and Smale (2002) and Steinwart (2005). Moreover, results in the spirit of Theorem 6.24 were found by Zhang (2001), Steinwart (2005), and in a different context by Bousquet and Elisseeff (2002). Universal consistency of SVMs for binary classification was first shown by Steinwart (2002), Zhang (2004b), and Steinwart (2005). Finally, consistency of SVMs for certain violations of the i.i.d. assumption was recently shown by Steinwart *et al.* (2008) and Steinwart and Anghel (2008) with techniques similar to the one used for Theorem 6.24.

In its simplistic form, the parameter selection method considered in Section 6.5 is little more than an illustration of how oracle inequalities can be used to analyze learning methods that *include* the parameter selection step. Nonetheless, the TV-SVM procedure is related to commonly used methods such as grid search and cross-validation, discussed in Section 11.3. A different approach for the parameter selection problem is considered by Lecué (2007b), who proposes to use the aggregated decision function

$$\sum_{\lambda \in \Lambda} w_\lambda \, \widehat{f}_{\mathrm{D}_1,\lambda} \,,$$

where the weights w_λ are computed in terms of $\mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda})$. More precisely, he considers weights of the form

$$w_\lambda := \frac{\exp(-|D_2|\,\mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda}))}{\sum_{\lambda' \in \Lambda} \exp(-|D_2|\,\mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda'}))}$$

and establishes, for example for the hinge loss, oracle inequalities for this approach. These oracle inequalities imply that this aggregation procedure is adaptive to characteristics of P considered in Chapter 8. Moreover, a similar weighting approach was taken by Bunea and Nobel (2005) for the least squares loss. For further methods and results, we refer to Bartlett (2008), Bunea *et al.* (2007), Dalalyan and Tsybakov (2007), Lecué (2007a), Tsybakov (2003), and the references therein.

6.7 Summary

In this chapter, we developed basic techniques for investigating the statistical properties of SVMs. To this end, we first introduced two notions of statistical

learning, namely the purely asymptotic notion of *consistency* and the more practically oriented notion of *learning rates*. We further presented the *no-free-lunch theorem*, which implied that uniform learning rates are impossible without assumptions on P.

In Section 6.2, we then established *concentration inequalities*, which described how close empirical averages of i.i.d. random variables are centered around their mean. The main results in this direction were *Hoeffding's inequality*, which gives an exponential tail for bounded real-valued random variables, and *Bernstein's inequality*, which improves this tail when the variance of the random variables is substantially smaller than their supremum norm. Finally, we generalized these inequalities to Hilbert space valued random variables.

In Section 6.3, we used these inequalities to analyze empirical risk minimization. We began by considering empirical risk minimizers over *finite* function classes and introduced *covering* and *entropy numbers* to generalize the basic idea to infinite function classes. The techniques developed for ERM were then modified in Section 6.4 to establish *oracle inequalities* for SVMs. There we also illustrated how these oracle inequalities can be used to establish both consistency and learning rates for SVMs whose regularization parameter only depends on the sample size. Unfortunately, however, the fastest learning rates we obtained required knowledge about certain characteristics of the data-generating distribution P. Since this knowledge is typically not available, we finally introduced and analyzed a data-dependent choice of the regularization parameter in Section 6.5. This selection method turned out to be consistent and, more important, we also saw that this method is *adaptive* to some unknown characteristics of P.

6.8 Exercises

6.1. Comparison of Hoeffding's and Bernstein's inequalities ($\star$)
Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, $B > 0$, and $\sigma > 0$. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to \mathbb{R}$ be independent and bounded random variables with $\|\xi_i\|_\infty \leq B$ and $\mathbb{E}\xi_i^2 \leq \sigma^2$ for all $i = 1, \ldots, n$. Finally, let $\tau > 0$ be a real number and $n \geq 1$ be an integer satisfying $n \geq \frac{8}{9}\tau$. Show that Bernstein's inequality is sharper than Hoeffding's inequality if and only if

$$\sigma < \left(1 - \sqrt{\frac{8\tau}{9n}}\right) B\,.$$

What happens if we additionally assume $\mathbb{E}\xi_i = 0$ for all $i = 1, \ldots, n$?

6.2. A variant of Markov's inequality ($\star\star$)
Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and $f : \Omega \to \mathbb{R}$ be a measurable function. Show that for all $t > 0$ the following inequalities hold:

$$\sum_{n=1}^{\infty} \mathrm{P}\big(\{\omega \in \Omega : |f(\omega)| \geq nt\}\big) \;\leq\; \frac{\mathbb{E}_\mathrm{P}|f|}{t} \;\leq\; 1 + \sum_{n=1}^{\infty} \mathrm{P}\big(\{\omega \in \Omega : |f(\omega)| \geq nt\}\big).$$

Hint: Apply Lemma A.3.11.

6.3. Chebyshev's inequality for sums of i.i.d. random variables ($\star\star$)
Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and $\xi_1, \ldots, \xi_n : \Omega \to \mathbb{R}$ be independent random variables for which there exists a constant $\sigma > 0$ such that $\mathbb{E}_{\mathrm{P}} \xi_i^2 \leq \sigma^2$ for all $i = 1, \ldots, n$.
i). Show the following inequality:

$$\mathrm{P}\left(\frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mathbb{E}\xi_i) \geq \sqrt{\frac{2\sigma^2 e^{\tau}}{n}} \right) \leq e^{-\tau}, \qquad \tau > 0.$$

ii). Compare this inequality with Hoeffding's and Bernstein's inequalities.
iii). Generalize the inequality above to Hilbert space valued random variables.

6.4. Proof of the no-free-lunch theorem($\star\star\star\star$)
Prove Theorem 6.6 using the proof of Theorem 7.2 by Devroye *et al.* (1996).

Hint: Fix an arbitrary decreasing sequence $(p_i) \subset (0, 1]$ with $\sum p_i = 1$. Using Lyapunov's Theorem A.3.13, which in particular states that $\{\mu(A) : A \in \mathcal{A}\} = [0, 1]$, construct a sequence (A_i) of mutually disjoint $A_i \in \mathcal{A}$ satisfying $\mu(A_i) = p_i$ for all $i \geq 1$. Use this to suitably modify the construction at the beginning of the proof of Theorem 7.2 by Devroye *et al.* (1996). Check that the rest of the proof can be kept unchanged.

6.5. No uniform rate for convex losses ($\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss function for which there exist two distributions Q_1 and Q_2 on Y that have mutually distinct L-risk minimizers, i.e., for all $x \in X$, we have $\mathcal{M}_{L,\mathrm{Q}_1,x}(0^+) \neq \emptyset$, $\mathcal{M}_{L,\mathrm{Q}_2,x}(0^+) \neq \emptyset$, and

$$\mathcal{M}_{L,\mathrm{Q}_1,x}(0^+) \cap \mathcal{M}_{L,\mathrm{Q}_2,x}(0^+) = \emptyset .$$

i). Show that L satisfies the assumptions of Corollary 6.8.
ii). Show that for margin-based and distance-based convex losses $L \neq 0$ there exist two distributions Q_1 and Q_2 on Y having mutually distinct L-risk minimizers.

Hint: For *i)* show that there exists a constant $c > 0$ such that for all $x \in X$ we have $\mathrm{dist}(t, \mathcal{M}_{L,\mathrm{Q}_2,x}(0^+)) \geq c$ if $t \in \mathcal{M}_{1,x}$ and $\mathrm{dist}(t, \mathcal{M}_{L,\mathrm{Q}_1,x}(0^+)) \geq c$ if $t \in \mathcal{M}_{2,x}$. Then repeat the argument used in the proof of Lemma 3.15.

6.6. Simple analysis of approximate empirical risk minimizers ($\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function, $B > 0$ be a real number, and $\mathcal{F} \subset \mathcal{L}_0(X)$ be a finite set of bounded measurable functions such that $L(x, y, f(x)) \leq B$ for all $(x, y) \in X \times Y$ and all $f \in \mathcal{F}$. In addition, assume that for some $\epsilon > 0$ we have a measurable learning algorithm that produces ϵ-*approximate* minimizers f_D of $\mathcal{R}_{L,\mathrm{D}}(\,\cdot\,)$, i.e.,

$$\mathcal{R}_{L,\mathrm{D}}(f_D) \leq \inf_{f \in \mathcal{F}} \mathcal{R}_{L,\mathrm{D}}(f) + \epsilon , \qquad D \in (X \times Y)^n .$$

Show that, for all $\tau > 0$ and all $n \geq 1$, the following inequality holds:

$$\mathrm{P}^n\left(D \in (X\times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) < \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2|\mathcal{F}|)}{n}} + \epsilon\right) \geq 1 - e^{-\tau}\,.$$

6.7. A simple example of overfitting ERM ($\star\star\star$)
Let $(X, \mathcal{A})$ be a measurable space such that $\{x\} \in \mathcal{A}$ for all $x \in X$. Furthermore, let $Y := \{-1, 1\}$, L_{class} be the binary classification loss, and $\mathcal{F} := \mathcal{L}_\infty(X)$. For $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$, define the function $f_D : X \to \mathbb{R}$ by

$$f_D := \frac{1}{n}\sum_{i=1}^n y_i \mathbf{1}_{\{x_i\}}\,.$$

i). Show that $D \mapsto f_D$ is a measurable empirical risk minimizer with respect to $\mathcal{F}$ and L_{class}.
ii). Let P be a distribution on $X \times Y$ such that $\mathrm{P}_X(\{x\}) = 0$ for all $x \in X$. Show that $\mathcal{R}_{L,\mathrm{P}}(f_D) = \mathcal{R}_{L,\mathrm{P}}(0)$.
iii). Find distributions P on $X \times Y$ such that $\mathcal{R}^*_{L,\mathrm{P}} = 0$ and $\mathcal{R}_{L,\mathrm{P}}(f_D) = 1/2$.

6.8. Entropy vs. covering numbers ($\star\star\star$)
Let (T, d) be a metric space and $a > 0$ and $q > 0$ be constants such that

$$\ln \mathcal{N}(T, d, \varepsilon) < \left(\frac{a}{\varepsilon}\right)^q, \qquad \varepsilon > 0.$$

Show that $e_n(T, d) \leq 3^{\frac{1}{q}} a\, n^{-\frac{1}{q}}$ for all $n \geq 1$.

6.9. Consistency and rates for SVMs using their stability ($\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex, Lipschitz continuous loss satisfying $L(x, y, 0) \leq 1$ for all $(x, y) \in X \times Y$, and $|L|_1 \leq 1$. Moreover, let H be a separable RKHS with measurable kernel k over X satisfying $\|k\|_\infty \leq 1$, and let P be a distribution on $X \times Y$ such that H is dense in $L_1(\mathrm{P}_X)$.
i). Show that with probability P^n not less than $1 - e^{-\tau}$ we have

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} < A_2(\lambda) + \lambda^{-1}\left(\sqrt{\frac{8\tau}{n}} + \sqrt{\frac{4}{n}} + \frac{8\tau}{3n}\right).$$

ii). Show that the SVM is consistent whenever we choose a sequence $(\lambda_n) \subset (0, 1]$ such that $\lim_{n\to\infty} \lambda_n = 0$ and $\lim_{n\to\infty} \lambda_n^2 n = \infty$.
iii). Assume that (6.23) holds, i.e., there exist constants $c > 0$ and $\beta \in (0, 1]$ such that $A_2(\lambda) \leq c\lambda^\beta$ for all $\lambda > 0$. Show that the asymptotically best choice for λ_n is a sequence that behaves like $n^{-\frac{1}{2\beta+2}}$ and that the resulting learning rate is given by

$$\mathrm{P}^n\left(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}} \leq \tilde{C}\tau n^{-\frac{\beta}{2\beta+2}}\right) \geq 1 - e^{-\tau}\,,$$

where $\tilde{C}$ is a constant independent of τ and n.
iv). Show that the learning rates established in *iii)* are faster than those of Theorem 6.25 if $p > 1/(2\beta + 1)$.

7

Advanced Statistical Analysis of SVMs (*)

Overview. *In the previous chapter, we established both consistency and learning rates using relatively simple oracle inequalities. The first goal of this chapter is to illustrate why these learning rates are too loose. We then establish refined oracle inequalities that lead to learning rates that are substantially faster than those of the previous chapter.*

Prerequisites. *The first two sections require only the statistical analysis from Chapter 6. The following two sections additionally need aspects of empirical process theory provided in Sections A.8 and A.9. The final section requires knowledge from Sections A.5.2 and A.5.6.*

Usage. *We will use the derived oracle inequalities in Chapters 8 and 9 when dealing with classification and regression, respectively.*

In the previous chapter, we presented two techniques to establish oracle inequalities for SVMs. We further showed how these oracle inequalities can be used to establish learning rates if assumptions on the approximation error function are made. The first goal of this chapter is to demonstrate in Section 7.1 that these learning rates are almost always suboptimal. We will then present a new technique to establish sharper oracle inequalities. We begin by considering ERM over finite sets of functions since this learning method is, as in Chapter 6, a suitable raw model for studying the basic principles of this technique. The resulting oracle inequality, which will later be used for parameter selection, is presented in Section 7.2. Unlike in the previous chapter, however, there is no simple yet effective way to extend this technique to infinite sets of functions. This forces us to introduce some heavy machinery from empirical process theory, which is summarized in Section A.8. We also need a new concentration inequality, known as Talagrand's inequality, which, unlike the concentration inequalities of the previous chapter, deals with suprema of functions directly. Since the highly non-trivial proof of Talagrand's inequality is out of the scope of this chapter, it is deferred to Section A.9. In Section 7.3, we will carefully introduce these tools in the process of establishing an oracle inequality for ERM over *infinite* sets of functions, so that the reader immediately gets an idea, of how the different tools work together. We then adapt this approach to SVMs and related modifications of ERM in Section 7.4. Finally, we revisit entropy numbers for RKHSs in Section 7.5.

7.1 Why Do We Need a Refined Analysis?

In this section, we show that the oracle inequalities established in the previous chapter almost always lead to suboptimal learning rates. This will give us the motivation to look for more advanced techniques in the following sections.

Let us begin by recalling the situation of Theorem 6.25. To this end, we assume for simplicity that the loss L is Lipschitz continuous with $|L|_1 \leq 1$. In addition, we again assume that there are constants $a \geq 1$ and $p > 0$ such that

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \;\leq\; a\varepsilon^{-2p}\,, \qquad\qquad \varepsilon > 0.$$

Theorem 6.25 then implied the oracle inequality (6.21), i.e., for fixed $n \geq 1$, $\lambda \in (0,1]$, and $\tau \geq 1$, we have[1]

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} < A_2(\lambda) + \frac{3}{\lambda^{1/2}} \left(2\Big(\frac{2a}{n}\Big)^{\frac{1}{2+2p}} + \Big(\frac{2\tau}{n}\Big)^{\frac{1}{2}} \right) \tag{7.1}$$

with probability P^n not less than $1 - e^{-\tau}$. The positive aspect of this oracle inequality is that it holds for *all* distributions P on $X \times Y$. From the machine learning perspective, this distribution independence is highly desirable since one of its basic assumptions is that the data-generating distribution P is unknown. However, this distribution independence is also the weakness of the oracle inequality above since for most distributions it is overly pessimistic in the sense that the resulting learning rates are suboptimal. To explain this, let us assume for simplicity that for sample size n we have chosen a regularization parameter $\lambda_n \in (0,1]$. Now recall that in the proof of Theorem 6.25 we used the trivial estimate

$$\|f_{\mathrm{D},\lambda_n}\|_H \leq \lambda_n^{-1/2} \tag{7.2}$$

to determine the function class over which the SVM actually minimizes. However, (7.1) then shows that with high probability we have $\|f_{\mathrm{D},\lambda_n}\|_H \leq (\varepsilon_n/\lambda_n)^{1/2}$, where ε_n is a shorthand for the right-hand side of (7.1). Assuming that we have chosen λ_n such that $\varepsilon_n \to 0$ for $n \to \infty$, we hence see that with high probability we have an estimate on $\|f_{\mathrm{D},\lambda_n}\|_H$ that is sharper than (7.2) for large n. To refine the analysis of Theorem 6.25, we could now exclude in the proof of Theorem 6.25 the set of samples D where this sharper estimate is not satisfied. As a consequence, we would work with a smaller function class and with a smaller bound B on the suprema. Now recall that both the size of the function class and B have a significant impact on the oracle inequality of Theorem 6.25 and hence on (7.1). To illustrate this, assume that we have chosen $\lambda_n := n^{-\gamma}$ for some $0 < \gamma < 1/(1+p)$ and all $n \geq 1$. Moreover,

[1] Here, as in the rest of this chapter, we assume that $(X \times Y)^n$ is equipped with the universal completion of the product σ-algebra of $(X \times Y)^n$. In addition, P^n denotes the canonical extension of the n-fold product measure of P to this completion. Recall that these conventions together with Lemmas 6.23 and 6.3 make it possible to ignore measurability questions for SVMs.

assume that $A_2(\lambda) \leq c\lambda^\beta$ for some constants $c > 0$, $\beta \in (0, 1]$, and all $\lambda > 0$. By following the path above, we would then obtain an improvement of (7.1) by a factor of the form $n^{-\alpha}$ for some $\alpha > 0$. This discussion shows that the original estimate (7.1) indeed leads to suboptimal learning rates. Moreover, it shows another dilemma: the improved version of (7.1) leads directly to a further improvement of (7.2), which in turn yields a further improvement of (7.1), and so on. On the other hand, it is not hard to see that such an iteration would increase the arising constants since we have to exclude more and more sets of small probability, and hence it seems likely that an analysis following this path would be rather technical. In addition, it would require choosing λ_n *a priori*, and hence we would not obtain an oracle inequality that can be used for the analysis of parameter selection procedures such as the TV-SVM.

Interestingly, the phenomenon above is not the only source for the general suboptimality of Theorem 6.25. However, the second source is more involved, and therefore we only illustrate it for ERM.[2] To this end, let us fix a finite set $\mathcal{F} \subset \mathcal{L}_\infty(X)$ of functions and a loss function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ such that $L(x, y, f(x)) \leq B$ for some constant $B > 0$ and all $(x, y) \in X \times Y$ and $f \in \mathcal{F}$. In addition, we assume that P is a distribution for which there exists an $f^* \in \mathcal{F}$ with $\mathcal{R}_{L,\mathrm{P}}(f^*) = 0$. This implies $L(x, y, f^*(x)) = 0$ for P-almost all $(x, y) \in X \times Y$, and hence we have $\mathcal{R}_{L,\mathrm{D}}(f^*) = 0$ for P^n-almost all $D \in (X \times Y)^n$. From this we conclude that $\mathcal{R}_{L,\mathrm{D}}(f_D) = 0$ almost surely. For $f \in \mathcal{F}$, we now define $h_f \in \mathcal{L}_\infty(X \times Y)$ by

$$h_f(x, y) := L(x, y, f(x)) \,, \qquad (x, y) \in X \times Y \,.$$

Since L is non-negative, this definition immediately yields

$$\mathbb{E}_\mathrm{P} h_f^2 \;\leq\; B\mathbb{E}_\mathrm{P} h_f \,. \tag{7.3}$$

Furthermore, for $r > 0$ and $f \in \mathcal{F}$, we define

$$g_{f,r} := \frac{\mathbb{E}_\mathrm{P} h_f - h_f}{\mathbb{E}_\mathrm{P} h_f + r} \,.$$

For $f \in \mathcal{F}$ with $\mathbb{E}_\mathrm{P} h_f = 0$, we then have $\mathbb{E}_\mathrm{P} h_f^2 = 0$ by (7.3) and hence we obtain $\mathbb{E}_\mathrm{P} g_{f,r}^2 = 0 \leq \frac{B}{2r}$. Moreover, for $f \in \mathcal{F}$ with $\mathbb{E}_\mathrm{P} h_f \neq 0$, we find

$$\mathbb{E}_\mathrm{P} g_{f,r}^2 \leq \frac{\mathbb{E}_\mathrm{P} h_f^2}{(\mathbb{E}_\mathrm{P} h_f + r)^2} \leq \frac{\mathbb{E}_\mathrm{P} h_f^2}{2r\mathbb{E}_\mathrm{P} h_f} \leq \frac{B}{2r} \,,$$

where we used (7.3) and the trivial estimate $2ab \leq (a + b)^2$ for $a, b \geq 0$. In addition, we have

$$\|g_{f,r}\|_\infty = \sup_{(x,y)\in X\times Y} \left| \frac{\mathbb{E}_\mathrm{P} h_f - h_f(x, y)}{\mathbb{E}_\mathrm{P} h_f + r} \right| = \frac{\|\mathbb{E}_\mathrm{P} h_f - h_f\|_\infty}{\mathbb{E}_\mathrm{P} h_f + r} \leq \frac{B}{r}$$

[2] At the end of Section 7.4, we see that this phenomenon indeed occurs for SVMs.

and $\mathrm{E}_{\mathrm{P}} g_{f,r} = 0$. Applying Bernstein's inequality and the union bound, we hence obtain

$$\mathrm{P}^n \left(D \in (X \times Y)^n : \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathrm{D}} g_{f,r} \geq \sqrt{\frac{B\tau}{nr}} + \frac{2B\tau}{3nr} \right) \leq |\mathcal{F}| e^{-\tau} \,,$$

and since $f_D \in \mathcal{F}$, the definition of $g_{f_D,r}$ thus yields

$$\mathrm{P}^n \left(D \in (X \times Y)^n : \mathbb{E}_{\mathrm{P}} h_{f_D} - \mathbb{E}_{\mathrm{D}} h_{f_D} \geq \left(\mathbb{E}_{\mathrm{P}} h_{f_D} + r \right) \left(\sqrt{\frac{B\tau}{nr}} + \frac{2B\tau}{3nr} \right) \right) \leq |\mathcal{F}| e^{-\tau}.$$

Because $\mathbb{E}_{\mathrm{D}} h_{f_D} = \mathcal{R}_{L,\mathrm{D}}(f_D) = 0$ almost surely, we hence conclude that

$$\mathrm{P}^n \left(D \in (X \times Y)^n : \left(1 - \sqrt{\frac{B\tau}{nr}} - \frac{2B\tau}{3nr} \right) \mathbb{E}_{\mathrm{P}} h_{f_D} \geq \sqrt{\frac{rB\tau}{n}} + \frac{2B\tau}{3n} \right) \leq |\mathcal{F}| e^{-\tau}.$$

For $r := \frac{4B\tau}{n}$, the relation $\mathbb{E}_{\mathrm{P}} h_{f_D} = \mathcal{R}_{L,\mathrm{P}}(f_D)$ thus yields

$$\mathrm{P}^n \left(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) \geq \frac{8B\tau}{n} \right) \leq |\mathcal{F}| e^{-\tau}. \tag{7.4}$$

Compared with (6.15), the latter estimate replaces the square root term $B(\frac{2\tau}{n})^{1/2}$ by the substantially faster decaying linear term $\frac{8B\tau}{n}$. In other words, for the specific type of distribution considered above, our analysis of ERM in Chapter 6 is too loose by a factor of $n^{-1/2}$.

The reason for this improvement is the *variance bound* (7.3), which guarantees a small variance whenever we have a small error $\mathcal{R}_{L,\mathrm{P}}(f) = \mathbb{E}_{\mathrm{P}} h_f$. Interestingly, such a variance bound can also easily be used in a brute-force analysis that does not start with an initial small error. Indeed, let us assume that we would begin our analysis by using Bernstein's inequality together with the trivial variance bound $\mathbb{E}_{\mathrm{P}} h_{f_D}^2 \leq B^2$. We would then obtain a small upper bound on $\mathbb{E}_{\mathrm{P}} h_{f_D}$ that holds with high probability. Using (7.3), this would give us a smaller bound on $\mathbb{E}_{\mathrm{P}} h_{f_D}^2$, which in turn would improve our first bound on $\mathbb{E}_{\mathrm{P}} h_{f_D}$ by another application of Bernstein's inequality. Obviously, this brute-force analysis would be very similar to the iterative proof procedure we discussed around (7.2). This observation suggests that Theorem 6.25 is also suboptimal if P guarantees a variance bound in the sense of (7.3), and we will see in Section 7.4 that this is indeed the case. Remarkably, however, the argument that led to (7.4) avoided this iterative argument by considering the function $g_{f,r}$ in Bernstein's inequality. This trick, in a refined form, will be *the* central idea for the refined analysis of this chapter.

7.2 A Refined Oracle Inequality for ERM

The goal of this section is to generalize the technique of using a variance bound in conjunction with Bernstein's inequality to derive oracle inequalities

for ERM. Although these generalizations still will not be powerful enough to deal with SVMs, they will already illustrate the key ideas. In addition, the derived oracle inequality for ERM will later be used for investigating the adaptivity of data-dependent parameter selection strategies such as the one of TV-SVMs.

Let us begin with the following elementary and widely known lemma.

Lemma 7.1. *For $q \in (1,\infty)$, define $q' \in (1,\infty)$ by $\frac{1}{q} + \frac{1}{q'} = 1$. Then we have*

$$ab \le \frac{a^q}{q} + \frac{b^{q'}}{q'}$$

and $(qa)^{2/q}(q'b)^{2/q'} \le (a+b)^2$ for all $a, b \ge 0$.

Proof. For $a = 0$, the first assertion is trivial, and hence it suffices to consider the case $a > 0$. We define $h_a(b) := a^q/q + b^{q'}/q' - ab$, $b \ge 0$. Obviously, the derivative of this function is $h_a'(b) = b^{q'-1} - a$, and hence h_a has a unique global minimum at $b^* := a^{1/(q'-1)}$. Using $q'/(q'-1) = q$, we find $h_a(b^*) = 0$, which then gives the desired first inequality. The second inequality then follows from the first by a simple variable transformation and $a^2 + b^2 \le (a+b)^2$. □

Before we present the improved oracle inequality for ERM, let us introduce the shorthand $L \circ f$ for the function $(x,y) \mapsto L(x,y,f(x))$, where $f : X \to \mathbb{R}$ is an arbitrary function and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is a loss.

Theorem 7.2 (Improved oracle inequality for ERM). *Consider a measurable ERM with respect to the loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ and the finite set $\mathcal{F} \subset \mathcal{L}_0(X)$. Moreover, let P be a distribution on $X \times Y$ that has a Bayes decision function $f^*_{L,\mathrm{P}}$. Assume that there exist constants $B > 0$, $\vartheta \in [0,1]$, and $V \ge B^{2-\vartheta}$ such that for all $f \in \mathcal{F}$ we have*

$$\|L \circ f - L \circ f^*_{L,\mathrm{P}}\|_\infty \le B\,, \tag{7.5}$$

$$\mathbb{E}_\mathrm{P}\big(L \circ f - L \circ f^*_{L,\mathrm{P}}\big)^2 \le V \cdot \big(\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}})\big)^\vartheta\,. \tag{7.6}$$

Then, for all fixed $\tau > 0$ and $n \ge 1$, we have with probability P^n not less than $1 - e^{-\tau}$ that

$$\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P}} \;<\; 6\big(\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} - \mathcal{R}^*_{L,\mathrm{P}}\big) + 4\left(\frac{8V\big(\tau + \ln(1+|\mathcal{F}|)\big)}{n}\right)^{\frac{1}{2-\vartheta}}.$$

Proof. We first note that since $\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P}} \le B$ and $V \ge B^{2-\vartheta}$, it suffices to consider the case $n \ge 8\tau$. For $f \in \mathcal{F}$ we define $h_f := L \circ f - L \circ f^*_{L,\mathrm{P}}$, and in addition we fix an $f_0 \in \mathcal{F}$. Since $\mathcal{R}_{L,\mathrm{D}}(f_D) \le \mathcal{R}_{L,\mathrm{D}}(f_0)$, we then have $\mathbb{E}_\mathrm{D} h_{f_D} \le \mathbb{E}_\mathrm{D} h_{f_0}$, and consequently we obtain

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}_{L,\mathrm{P}}(f_0) &= \mathbb{E}_\mathrm{P} h_{f_D} - \mathbb{E}_\mathrm{P} h_{f_0} \\ &\le \mathbb{E}_\mathrm{P} h_{f_D} - \mathbb{E}_\mathrm{D} h_{f_D} + \mathbb{E}_\mathrm{D} h_{f_0} - \mathbb{E}_\mathrm{P} h_{f_0} \end{aligned} \tag{7.7}$$

for all $D \in (X \times Y)^n$. Let us first estimate $\mathbb{E}_D h_{f_0} - \mathbb{E}_P h_{f_0}$ for the case $\vartheta > 0$. To this end, we observe that $\|h_{f_0} - \mathbb{E}_P h_{f_0}\|_\infty \leq 2B$ and

$$\mathbb{E}_P(h_{f_0} - \mathbb{E}_P h_{f_0})^2 \leq \mathbb{E}_P h_{f_0}^2 \leq V(\mathbb{E}_P h_{f_0})^\vartheta .$$

In addition, for $q := \frac{2}{2-\vartheta}$, $q' := \frac{2}{\vartheta}$, $a := \left(\frac{2^{1-\vartheta}\vartheta^\vartheta V\tau}{n}\right)^{1/2}$, and $b := \left(\frac{2\mathbb{E}_P h_{f_0}}{\vartheta}\right)^{\vartheta/2}$, Lemma 7.1 shows that

$$\sqrt{\frac{2\tau V(\mathbb{E}_P h_{f_0})^\vartheta}{n}} \leq \left(1-\frac{\vartheta}{2}\right)\left(\frac{2^{1-\vartheta}\vartheta^\vartheta V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \mathbb{E}_P h_{f_0} \leq \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \mathbb{E}_P h_{f_0},$$

and hence Bernstein's inequality shows that we have

$$\mathbb{E}_D h_{f_0} - \mathbb{E}_P h_{f_0} < \mathbb{E}_P h_{f_0} + \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{4B\tau}{3n} \tag{7.8}$$

with probability P^n not less than $1 - e^{-\tau}$. Furthermore, note that, for $\vartheta = 0$, the same inequality holds by Hoeffding's inequality and $\|h_{f_0}\|_\infty \leq B \leq \sqrt{V}$.

To estimate the remaining term $\mathbb{E}_P h_{f_D} - \mathbb{E}_D h_{f_D}$, we define the functions

$$g_{f,r} := \frac{\mathbb{E}_P h_f - h_f}{\mathbb{E}_P h_f + r}, \qquad f \in \mathcal{F},\, r > 0 .$$

Obviously, we have $\|g_{f,r}\|_\infty \leq 2Br^{-1}$. Moreover, for $\vartheta > 0$, $b := \mathbb{E}_P h_f \neq 0$, $q := \frac{2}{2-\vartheta}$, $q' := \frac{2}{\vartheta}$, and $a := r$, the second inequality of Lemma 7.1 yields

$$\mathbb{E}_P g_{f,r}^2 \leq \frac{\mathbb{E}_P h_f^2}{(\mathbb{E}_P h_f + r)^2} \leq \frac{(2-\vartheta)^{2-\vartheta}\vartheta^\vartheta\, \mathbb{E}_P h_f^2}{4r^{2-\vartheta}(\mathbb{E}_P h_f)^\vartheta} \leq V r^{\vartheta-2} .$$

Furthermore, for $\vartheta > 0$ and $\mathbb{E}_P h_f = 0$, we have $\mathbb{E}_P h_f^2 = 0$ by (7.6), which in turn implies $\mathbb{E}_P g_{f,r}^2 \leq V r^{\vartheta-2}$. Finally, in the case $\vartheta = 0$, we easily obtain $\mathbb{E}_P g_{f,r}^2 \leq \mathbb{E}_P h_f^2\, r^{-2} \leq V r^{\vartheta-2}$, and hence we have $\mathbb{E}_P g_{f,r}^2 \leq V r^{\vartheta-2}$ in all cases. Consequently, Bernstein's inequality yields

$$P^n\left(D \in (X \times Y)^n : \sup_{f\in\mathcal{F}} \mathbb{E}_D g_{f,r} < \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr}\right) \;\geq\; 1 - |\mathcal{F}|e^{-\tau} \tag{7.9}$$

for all $r > 0$. Now observe that for $D \in (X \times Y)^n$ satisfying

$$\sup_{f\in\mathcal{F}} \mathbb{E}_D g_{f,r} < \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr},$$

the definition of $g_{f_D,r}$ and $f_D \in \mathcal{F}$ imply

$$\mathbb{E}_P h_{f_D} - \mathbb{E}_D h_{f_D} < \mathbb{E}_P h_{f_D}\left(\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr}\right) + \sqrt{\frac{2V\tau r^\vartheta}{n}} + \frac{4B\tau}{3n} .$$

By combining this estimate with (7.7), (7.8), and (7.9), we thus see that

$$\mathbb{E}_{\mathrm{P}} h_{f_D} < 2\mathbb{E}_{\mathrm{P}} h_{f_0} + \mathbb{E}_{\mathrm{P}} h_{f_D} \left(\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr} \right) + \sqrt{\frac{2V\tau r^{\vartheta}}{n}} + \left(\frac{2V\tau}{n} \right)^{\frac{1}{2-\vartheta}} + \frac{8B\tau}{3n}$$

holds with probability P^n not less than $1 - (1 + |\mathcal{F}|)e^{-\tau}$. Let us now define $r := \left(\frac{8V\tau}{n}\right)^{1/(2-\vartheta)}$. Then we obviously have

$$\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} = \frac{1}{2} \qquad \text{and} \qquad \sqrt{\frac{2V\tau r^{\vartheta}}{n}} = \frac{r}{2}\,,$$

and by using $V \geq B^{2-\vartheta}$ and $n \geq 8\tau$ we also find

$$\frac{4B\tau}{3nr} \;=\; \frac{1}{6} \cdot \frac{8\tau}{n} \cdot \frac{B}{r} \;\leq\; \frac{1}{6} \cdot \left(\frac{8\tau}{n}\right)^{\frac{1}{2-\vartheta}} \cdot \frac{V^{\frac{1}{2-\vartheta}}}{r} \;=\; \frac{1}{6}$$

and $\frac{8B\tau}{3n} \leq \frac{r}{3}$. In addition, $2 \leq 4^{\frac{1}{2-\vartheta}}$ and the definition of r yield $(\frac{2V\tau}{n})^{\frac{1}{2-\vartheta}} \leq \frac{r}{2}$, and hence we have with probability P^n not less than $1 - (1 + |\mathcal{F}|)e^{-\tau}$ that

$$\mathbb{E}_{\mathrm{P}} h_{f_D} < 2\mathbb{E}_{\mathrm{P}} h_{f_0} + \frac{2}{3}\mathbb{E}_{\mathrm{P}} h_{f_D} + \frac{4}{3}r\,.$$

We now obtain the assertion by some simple algebraic transformations and taking a function $f_0 \in \mathcal{F}$ such that $\mathcal{R}_{L,\mathrm{P}}(f_0) = \min\{\mathcal{R}_{L,\mathrm{P}}(f) : f \in \mathcal{F}\}$. □

Note that in the proof of Theorem 7.2 we did not strive to obtain the smallest possible constants. Instead we tried to keep both the proof and the oracle inequality as simple as possible.

Obviously, the crucial assumption of Theorem 7.2 is the variance bound (7.6). Unfortunately, for many loss functions it is a non-trivial task to establish such a bound, as we will see in Sections 8.3 and 9.5, where we consider this issue for the hinge loss and the pinball loss, respectively. On the other hand, the following example shows that for the least squares loss and *bounded* Y, a variance bound always holds.

Example 7.3. Let $M > 0$ and $Y \subset [-M, M]$ be a closed subset. Moreover, let L be the **least squares loss**, X be a non-empty set equipped with some σ-algebra, and P be a distribution on $X \times Y$. By the non-negativity of L and $f^*_{L,\mathrm{P}}(x) = \mathbb{E}_{\mathrm{P}}(Y|x) \in [-M, M]$, $x \in X$, we first observe that

$$\left|L(y, f(x)) - L(y, f^*_{L,\mathrm{P}}(x))\right| \leq \sup_{y', t \in [-M,M]} (y' - t)^2 = 4M^2$$

for all measurable $f : X \to [-M, M]$ and all $(x, y) \in X \times Y$. Consequently, (7.5) holds for $B := 4M^2$. Moreover, we also have

$$\begin{aligned}\left(L(y, f(x)) - L(y, f^*_{L,\mathrm{P}}(x))\right)^2 &= \left((f(x) + f^*_{L,\mathrm{P}}(x) - 2y)(f(x) - f^*_{L,\mathrm{P}}(x))\right)^2 \\ &\leq 16M^2 \left(f(x) - f^*_{L,\mathrm{P}}(x)\right)^2,\end{aligned}$$

and hence we find

$$\mathbb{E}_{\mathrm{P}}(L \circ f - L \circ f^*_{L,\mathrm{P}})^2 \leq 16M^2\, \mathbb{E}_{\mathrm{P}}(f - f^*_{L,\mathrm{P}})^2 \tag{7.10}$$

$$= 16M^2\, \mathbb{E}_{\mathrm{P}}(L \circ f - L \circ f^*_{L,\mathrm{P}}) \tag{7.11}$$

In other words, (7.6) holds for $V := 16M^2$ and $\vartheta = 1$. ◁

Note that the variance bound established in the preceding example holds for the optimal exponent $\vartheta = 1$, which leads to a $1/n$ behavior of the oracle inequality presented in Theorem 7.2. Besides this, however, the preceding example also provides us with a "template approach" for establishing variance bounds. Indeed, (7.10) only uses the local Lipschitz continuity of the least squares loss when restricted to label domain Y, while (7.11) is a very special case of self-calibration. Interestingly, all later established variance bounds will essentially follow this pattern of combining Lipschitz continuity with self-calibration. Unlike for the least squares loss, which is nicely self-calibrated for *all* distributions having bounded label space Y, for most other interesting losses, establishing non-trivial self-calibration properties requires identifying suitable distributions. Unfortunately, in many cases this is a non-trivial task.

7.3 Some Advanced Machinery

One of the main ideas in the proof of Theorem 7.2 was to apply Bernstein's inequality to functions of the form

$$g_{f,r} := \frac{\mathbb{E}_{\mathrm{P}} h_f - h_f}{\mathbb{E}_{\mathrm{P}} h_f + r}\,, \qquad f \in \mathcal{F},\, r > 0\,, \tag{7.12}$$

where $h_f := L \circ f - L \circ f^*_{L,\mathrm{P}}$ and $\mathcal{F}$ was a *finite* set of functions. However, we have already seen in Chapter 6 that the statistical analysis of SVMs requires *infinite* sets of functions, namely balls of the RKHS used. Now a straightforward generalization of Theorem 7.2 to such $\mathcal{F}$ would assume $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) < \infty$ for all $\varepsilon > 0$ (see Exercise 7.3 for the details of such an extension). However, we will see later in Section 7.4 (see also Exercise 7.5) that the covering numbers with respect to the supremum norm often are too large. This suboptimality motivates this section, whose goal is to present more sophisticated tools for generalizing Theorem 7.2 to infinite function classes. However, these tools require much more involved proofs than the ones seen so far, and hence some of the more complicated results and their proofs are presented in Sections A.8 and A.9. By postponing these parts, we hope to give the reader a better understanding of how these tools work together.

Since most of the following results involve expectations of suprema over *uncountable* sets, we first clarify measurability issues. To this end, we introduce the following notion.

Definition 7.4. *Let (T,d) be a metric space and $(Z,\mathcal{A})$ be a measurable space. A family of maps $(g_t)_{t\in T}\subset\mathcal{L}_0(Z)$ is called a* ***Carathéodory family*** *if $t\mapsto g_t(z)$ is continuous for all $z\in Z$. Moreover, if T is separable or complete, we say that $(g_t)_{t\in T}$ is* ***separable*** *or* ***complete****, respectively.*

In the following we call a subset $\mathcal{G}\subset\mathcal{L}_0(Z)$ a **(separable or complete) Carathéodory set** if there exists a (separable or complete) metric space (T,d) and a Carathéodory family $(g_t)_{t\in T}\subset\mathcal{L}_0(Z)$ such that $\mathcal{G}=\{g_t : t\in T\}$. Note that, by the continuity of $t\mapsto g_t(z)$, Carathéodory sets satisfy

$$\sup_{g\in\mathcal{G}} g(z) = \sup_{t\in T} g_t(z) = \sup_{t\in S} g_t(z)\,, \qquad z\in Z, \tag{7.13}$$

for all dense $S\subset T$. In particular, for separable Carathéodory sets $\mathcal{G}$, there exists a *countable* and dense $S\subset T$, and hence the map $z\mapsto\sup_{t\in T} g_t(z)$ is measurable for such $\mathcal{G}$. Finally, recall Lemma A.3.17, which shows that the map $(z,t)\mapsto g_t(z)$ is measurable if T is separable *and* complete.

Now our first result generalizes Bernstein's inequality to suprema of separable Carathéodory sets.

Theorem 7.5 (Simplified Talagrand's inequality). *Let $(Z,\mathcal{A},\mathrm{P})$ be a probability space and $\mathcal{G}\subset\mathcal{L}_0(Z)$ be a separable Carathéodory set. Furthermore, let $B\geq 0$ and $\sigma\geq 0$ be constants such that $\mathbb{E}_{\mathrm{P}}g=0$, $\mathbb{E}_{\mathrm{P}}g^2\leq\sigma^2$, and $\|g\|_\infty\leq B$ for all $g\in\mathcal{G}$. For $n\geq 1$, we define $G:Z^n\to\mathbb{R}$ by*

$$G(z_1,\ldots,z_n) := \sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^n g(z_j)\right|, \qquad z=(z_1,\ldots,z_n)\in Z^n.$$

Then, for all $\tau>0$ and all $\gamma>0$, we have

$$\mathrm{P}^n\left(\left\{z\in Z^n : G(z)\geq(1+\gamma)\mathbb{E}_{\mathrm{P}^n}G+\sqrt{\frac{2\tau\sigma^2}{n}}+\left(\frac{2}{3}+\frac{1}{\gamma}\right)\frac{\tau B}{n}\right\}\right)\;\leq\; e^{-\tau}\,.$$

Proof. By (7.13), we may assume without loss of generality that $\mathcal{G}$ is countable. For fixed $a,b>0$ we now define $h:(0,\infty)\to(0,\infty)$ by $h(\gamma):=\gamma a+\gamma^{-1}b$, $\gamma>0$. Then elementary calculus shows that h has a global minimum at $\gamma^*:=\sqrt{b/a}$, and consequently we have $2\sqrt{ab}=h(\gamma^*)\leq h(\gamma)=\gamma a+\gamma^{-1}b$ for all $\gamma>0$. From this we conclude that

$$\sqrt{\frac{2\tau(\sigma^2+2B\mathbb{E}_{\mathrm{P}^n}G)}{n}}\leq\sqrt{\frac{2\tau\sigma^2}{n}}+2\sqrt{\frac{\tau B\mathbb{E}_{\mathrm{P}^n}G}{n}}\leq\sqrt{\frac{2\tau\sigma^2}{n}}+\gamma\mathbb{E}_{\mathrm{P}^n}G+\frac{\tau B}{\gamma n}\,,$$

and hence we obtain the assertion by applying Talagrand's inequality stated in Theorem A.9.1. □

Following the idea of the proof of Theorem 7.2, our first goal is to apply the preceding theorem to the family of maps $(g_{f,r})_{f\in\mathcal{F}}$, where $g_{f,r}$ is defined by (7.12). To this end, we first show that this is a separable Carathéodory family.

Lemma 7.6. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous loss, P be a distribution on $X \times Y$, and $f^*_{L,\mathrm{P}}$ be a Bayes decision function. Moreover, let $\mathcal{F} \subset \mathcal{L}_0(X)$ be equipped with a separable metric that dominates the pointwise convergence in the sense of (2.8). If there exists a constant $B > 0$ such that $\|L \circ f - L \circ f^*_{L,\mathrm{P}}\|_\infty \leq B$ for all $f \in \mathcal{F}$, then $(g_{f,r})_{f\in\mathcal{F}}$, where $g_{f,r}$ is defined by (7.12) and $r > 0$ is fixed, is a separable Carathéodory family.*

Proof. Since the metric of $\mathcal{F}$ dominates the pointwise convergence we see that for fixed $(x,y) \in X \times Y$ the $\mathbb{R}$-valued map $f \mapsto f(x)$ defined on $\mathcal{F}$ is continuous. Using the continuity of L, it is then easy to conclude that $(h_f)_{f\in\mathcal{F}}$, where $h_f := L\circ f - L\circ f^*_{L,\mathrm{P}}$ is a Carathéodory family. Moreover, we have $\|h_f\|_\infty \leq B$ for all $f \in \mathcal{F}$, and hence Lebesgue's dominated convergence theorem shows that $f \mapsto \mathbb{E}_\mathrm{P} h_f$ is continuous. From this we obtain the assertion. □

To ensure that $(g_{f,r})_{f\in\mathcal{F}}$ is a separable Carathéodory family, we assume in the rest of this section that the assumptions of Lemma 7.6 are satisfied.

Let us now consider the other assumptions of Theorem 7.5. To this end, recall that we have already seen in the proof of Theorem 7.2 that bounds of the form (7.5) and (7.6) lead to the estimates $\|g_{f,r}\|_\infty \leq Br^{-1}$ and $\mathbb{E}_\mathrm{P} g^2_{f,r} \leq Vr^{\vartheta-2}$, respectively. Moreover, we obviously have $\mathbb{E}_\mathrm{P} g_{f,r} = 0$. Applying Theorem 7.5 for $\gamma = 1$, we hence see that for all $\tau > 0$ we have

$$\sup_{f\in\mathcal{F}} \frac{\mathbb{E}_\mathrm{P} h_f - \mathbb{E}_\mathrm{D} h_f}{\mathbb{E}_\mathrm{P} h_f + r} < 2\mathbb{E}_{D\sim\mathrm{P}^n} \sup_{f\in\mathcal{F}} \frac{\mathbb{E}_\mathrm{P} h_f - \mathbb{E}_\mathrm{D} h_f}{\mathbb{E}_\mathrm{P} h_f + r} + \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{5B\tau}{3nr} \tag{7.14}$$

with probability P^n not less than $1 - e^{-\tau}$. In other words, besides the expectation on the right-hand side, we have basically obtained the key estimate (7.9) of the proof of Theorem 7.2 for general, *not necessarily finite* sets $\mathcal{F}$. The rest of this section is thus devoted to techniques for bounding the additional expectation. We begin with a method that removes the denominator.

Theorem 7.7 (Peeling). *Let $(Z, \mathcal{A}, \mathrm{P})$ be a probability space, (T,d) be a separable metric space, $h : T \to [0,\infty)$ be a continuous function, and $(g_t)_{t\in T} \subset \mathcal{L}_0(Z)$ be a Carathéodory family. We define $r^* := \inf\{h(t) : t \in T\}$. Moreover, let $\varphi : (r^*,\infty) \to [0,\infty)$ be a function such that $\varphi(4r) \leq 2\varphi(r)$ and*

$$\mathbb{E}_{z\sim\mathrm{P}} \sup_{\substack{t\in T\\ h(t)\leq r}} |g_t(z)| \;\leq\; \varphi(r)$$

for all $r > r^*$. Then, for all $r > r^*$, we have

$$\mathbb{E}_{z\sim\mathrm{P}} \sup_{t\in T} \frac{g_t(z)}{h(t) + r} \;\leq\; \frac{4\varphi(r)}{r}\,.$$

Proof. For $z \in Z$ and $r > r^*$, we have

$$\sup_{t\in T} \frac{g_t(z)}{h(t)+r} \leq \sup_{\substack{t\in T\\ h(t)\leq r}} \frac{|g_t(z)|}{r} + \sum_{i=0}^{\infty} \sup_{\substack{t\in T\\ h(t)\in[r4^i, r4^{i+1}]}} \frac{|g_t(z)|}{r4^i + r}\,,$$

where we used the convention $\sup \emptyset := 0$. Consequently, we obtain

$$\mathbb{E}_{z\sim \mathrm{P}} \sup_{t\in T} \frac{g_t(z)}{h(t)+r} \le \frac{\varphi(r)}{r} + \frac{1}{r}\sum_{i=0}^{\infty} \frac{1}{4^i+1}\, \mathbb{E}_{z\sim \mathrm{P}} \sup_{\substack{t\in T \\ h(t)\le r4^{i+1}}} |g_t(z)|$$
$$\le \frac{1}{r}\left(\varphi(r) + \sum_{i=0}^{\infty} \frac{\varphi(r4^{i+1})}{4^i+1}\right).$$

Moreover, induction yields $\varphi(r4^{i+1}) \le 2^{i+1}\varphi(r)$, $i \ge 0$, and hence we obtain

$$\mathbb{E}_{z\sim \mathrm{P}} \sup_{t\in T} \frac{g_t(z)}{h(t)+r} \le \frac{\varphi(r)}{r}\left(1 + \sum_{i=0}^{\infty} \frac{2^{i+1}}{4^i+1}\right) \le \frac{\varphi(r)}{r}\left(1 + \frac{2}{1+1} + \sum_{i=1}^{\infty} 2^{-i+1}\right).$$

From this estimate, we easily obtain the assertion. □

Note that the preceding proof actually yields a constant slightly smaller than 4. Indeed, numerical evaluation suggests a value of approximately 3.77, but since the goal of this section is to present the general picture and not a path that leads to the smallest possible constants, we will work with the slightly larger but simpler constant.

Let us now return to the concentration inequality (7.14) we derived from Talagrand's inequality. To apply the peeling argument, we write

$$\mathcal{H}_r := \left\{h_f : f \in \mathcal{F} \text{ and } \mathbb{E}_{\mathrm{P}} h_f \le r\right\}, \qquad r > 0. \tag{7.15}$$

Furthermore, for $r^* := \inf\{r > 0 : \mathcal{H}_r \neq \emptyset\} = \inf\{\mathbb{E}_{\mathrm{P}} h_f : f \in \mathcal{F}\}$, we assume that we have a function $\varphi_n : (r^*, \infty) \to [0, \infty)$ satisfying $\varphi_n(4r) \le 2\varphi_n(r)$ and

$$\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{h\in\mathcal{H}_r} \left|\mathbb{E}_{\mathrm{P}} h - \mathbb{E}_{\mathrm{D}} h\right| \;\le\; \varphi_n(r) \tag{7.16}$$

for all $r > r^*$. By Theorem 7.7 and (7.14), we then see that we have

$$\sup_{f\in\mathcal{F}} \frac{\mathbb{E}_{\mathrm{P}} h_f - \mathbb{E}_{\mathrm{D}} h_f}{\mathbb{E}_{\mathrm{P}} h_f + r} < \frac{8\varphi_n(r)}{r} + \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{5B\tau}{3nr} \tag{7.17}$$

with probability P^n not less than $1-e^{-\tau}$. Consequently, our next goal is to find functions φ_n satisfying (7.16). To do this, we need the following definitions.

Definition 7.8. *Let $(\Theta, \mathcal{C}, \nu)$ be a probability space and $\varepsilon_i : \Theta \to \{-1, 1\}$, $i = 1, \ldots, n$, be independent random variables with $\nu(\varepsilon_i = 1) = \nu(\varepsilon_i = -1) = 1/2$ for all $i = 1, \ldots, n$. Then $\varepsilon_1, \ldots, \varepsilon_n$ is called a **Rademacher sequence** with respect to ν.*

Statistically speaking, Rademacher sequences model repeated coin flips. Besides this obvious interpretation, they are, however, also an important tool in several branches of pure mathematics.

The following definition uses Rademacher sequences to introduce a new type of expectation of suprema. This new type will then be used to find functions φ_n satisfying (7.16).

Definition 7.9. *Let $\mathcal{H} \subset \mathcal{L}_0(Z)$ be a non-empty set and $\varepsilon_1, \ldots, \varepsilon_n$ be a Rademacher sequence with respect to some distribution ν. Then, for $D := (z_1, \ldots, z_n) \in Z^n$, the n-th* ***empirical Rademacher average of*** *$\mathcal{H}$ is defined by*

$$\mathrm{Rad}_D(\mathcal{H}, n) := \mathbb{E}_\nu \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(z_i) \right| . \tag{7.18}$$

Note that in the preceding definition we do not need to ensure that the supremum on the right-hand side of (7.18) is measurable since the expectation $\mathbb{E}_\nu$ reduces to a finite sum over 2^n summands.

Some simple structural properties of empirical Rademacher averages can be found in Exercise 7.2, whereas more advanced properties are collected in Section A.8. For now we need the following result, which follows directly from Corollary A.8.2.

Proposition 7.10 (Symmetrization). *Let $\mathcal{H} \subset \mathcal{L}_\infty(Z)$ be a separable Carathéodory set with $\sup_{h \in \mathcal{H}} \|h\|_\infty < \infty$ and P be a distribution on Z. Then for all $n \geq 1$ we have*

$$\mathbb{E}_{D \sim \mathrm{P}^n} \sup_{h \in \mathcal{H}} \left| \mathbb{E}_\mathrm{P} h - \mathbb{E}_\mathrm{D} h \right| \;\leq\; 2\, \mathbb{E}_{D \sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}, n) . \tag{7.19}$$

Combining the preceding proposition with (7.16), we see that it suffices to bound $\mathbb{E}_{D \sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r, n)$. To achieve this, we will first focus on bounding empirical Rademacher averages. We begin with the following lemma, which can be used to bound Rademacher averages of *finite* function classes.

Lemma 7.11. *Let $(\Theta, \mathcal{C}, \nu)$ be a probability space and $g_1, \ldots, g_m \in \mathcal{L}_0(\Theta)$ be functions for which there exists a constant $K > 0$ satisfying $\mathbb{E}_\nu e^{\lambda g_i} \;\leq\; e^{\lambda^2 K^2}$ for all $\lambda \in \mathbb{R}$ and $i = 1, \ldots, m$. Then we have*

$$\mathbb{E}_\nu \max_{1 \leq i \leq m} |g_i| \;\leq\; 2K \sqrt{\ln(2m)} .$$

Proof. Writing $\max_{1 \leq j \leq m} \pm \lambda g_i := \max\{\lambda g_1, \ldots, \lambda g_m, -\lambda g_1, \ldots, -\lambda g_m\}$ for $\lambda > 0$, we obtain

$$\mathbb{E}_\nu \max_{1 \leq i \leq m} |g_i| \;=\; \lambda^{-1} \mathbb{E}_\nu \max_{1 \leq j \leq m} \pm \lambda g_i \;\leq\; \lambda^{-1} \ln\bigl(\mathbb{E}_\nu e^{\max_{1 \leq j \leq m} \pm \lambda g_i}\bigr)$$

by Jensen's inequality. Moreover, by the monotonicity of the exponential function, we have

$$\mathbb{E}_\nu e^{\max_{1 \leq j \leq m} \pm \lambda g_i} \leq \mathbb{E}_\nu \max_{1 \leq j \leq m} e^{\pm \lambda g_i} \leq \sum_{j=1}^{m} \mathbb{E}_\nu \bigl(e^{\lambda g_i} + e^{-\lambda g_i}\bigr) \leq 2m\, e^{\lambda^2 K^2} .$$

Combining both estimates, we thus find

$$\mathbb{E}_\nu \max_{1 \leq i \leq m} |g_i| \leq \lambda^{-1} \ln\bigl(2m\, e^{\lambda^2 K^2}\bigr) = \lambda^{-1} \ln(2m) + \lambda K^2$$

for all $\lambda > 0$. For $\lambda := K^{-1} \sqrt{\ln(2m)}$, we then obtain the assertion. □

The preceding lemma only provides a bound for suprema over *finitely* many functions. However, if we assume that we have a set of functions $\mathcal{H}$ that can be suitably approximated by a *finite* set of functions, then it seems natural to ask whether we can also bound the supremum over this possibly infinite set $\mathcal{H}$. The following theorem gives a positive answer to this question with the help of entropy numbers.

Theorem 7.12 (Dudley's chaining). *Let (T,d) be a separable metric space, $(\Theta,\mathcal{C},\nu)$ be a probability space, and $(g_t)_{t\in T}\subset\mathcal{L}_0(\Theta)$ be a Carathéodory family. Assume that there exist a $t_0\in T$ and a constant $K>0$ such that $g_{t_0}=0$ and*

$$\mathbb{E}_\nu\, e^{\lambda(g_s-g_t)} \;\le\; e^{\lambda^2K^2\cdot d^2(s,t)}\,, \qquad\qquad s,t\in T,\,\lambda\in\mathbb{R}. \tag{7.20}$$

Then we have

$$\mathbb{E}_\nu \sup_{t\in T}|g_t| \;\le\; 2\sqrt{\ln 4}\,K\biggl(\sum_{i=1}^\infty 2^{i/2}e_{2^i}(T,d)+\sup_{t\in T}d(t,t_0)\biggr).$$

Proof. Without loss of generality, we may assume that all entropy numbers are finite, i.e., $e_n(T,d)<\infty$ for all $n\ge 1$. We fix an $\varepsilon>0$. For $i\ge 1$, we define $s_{2^i}:=(1+\varepsilon)e_{2^i}(T,d)$, and in addition we write $s_1:=\sup_{t\in T}d(t,t_0)$. For $i\ge 1$, we further fix an s_{2^i}-net T_i of T such that $|T_i|\le 2^{2^i-1}$. In addition, we write $T_0:=\{t_0\}$. Then there exist maps $\pi_i:T\to T_i$, $i\ge 0$, such that $d(t,\pi_i(t))\le s_{2^i}$ for all $t\in T$ and $\pi_i(t)=t$ for all $t\in T_i$. Let us now fix a $j\ge 0$. We define maps $\gamma_i:T\to T_i$, $i=0,\ldots,j$, recursively by $\gamma_j:=\pi_j$ and $\gamma_{i-1}:=\pi_{i-1}\circ\gamma_i$, $i=j,\ldots,1$. Note that $T_0=\{t_0\}$ implies $\gamma_0(t)=t_0$ for all $t\in T$. For $k\ge j$ and $t\in T_j$, we hence obtain

$$|g_t|=|g_t-g_{t_0}|=\biggl|\sum_{i=1}^j\bigl(g_{\gamma_i(t)}-g_{\gamma_{i-1}(t)}\bigr)\biggr|\le\sum_{i=1}^k\max_{t\in T_i}\bigl|g_t-g_{\pi_{i-1}(t)}\bigr|\,,$$

which for $S_k:=T_0\cup\cdots\cup T_k$ implies

$$\max_{t\in S_k}|g_t|\le\sum_{i=1}^k\max_{t\in T_i}\bigl|g_t-g_{\pi_{i-1}(t)}\bigr|\,.$$

Moreover, (7.20) yields

$$\mathbb{E}_\nu e^{\lambda(g_t-g_{\pi_{i-1}(t)})}\le e^{\lambda^2K^2s_{2^{i-1}}^2}$$

for all $t\in T_i$, $i\ge 1$, and $\lambda\in\mathbb{R}$. Consequently, Lemma 7.11 implies

$$\mathbb{E}_\nu\max_{t\in S_k}|g_t|\le\sum_{i=1}^k\mathbb{E}_\nu\max_{t\in T_i}\bigl|g_t-g_{\pi_{i-1}(t)}\bigr|\le 2K\sum_{i=1}^k\sqrt{\ln(2\cdot 2^{2^i-1})}\,s_{2^{i-1}}\,.$$

Let us write $S:=\bigcup_{i=0}^\infty T_i$. Then we have $\max_{t\in S_k}|g_t|\nearrow\sup_{t\in S}|g_t|$ for $k\to\infty$, and by Beppo Levi's theorem we hence obtain

$$\mathbb{E}_\nu \sup_{t\in S} |g_t| \;\le\; 2\sqrt{\ln 4}\, K \sum_{i=0}^{\infty} 2^{i/2} s_{2^i}\,.$$

Since S is dense in T, we then obtain the assertion by (7.13), the definition of s_{2^i}, and taking the limit $\varepsilon \to 0$. □

With the help of Dudley's chaining, we can now establish our first bound on empirical Rademacher averages.

Theorem 7.13. *For every non-empty set $\mathcal{H} \subset \mathcal{L}_0(Z)$ and every finite sequence $D := (z_1, \ldots, z_n) \in Z^n$, we have*

$$\mathrm{Rad}_D(\mathcal{H}, n) \le \sqrt{\frac{\ln 16}{n}} \left(\sum_{i=1}^{\infty} 2^{i/2}\, e_{2^i}\big(\mathcal{H} \cup \{0\}, L_2(\mathrm{D})\big) + \sup_{h\in\mathcal{H}} \|h\|_{L_2(\mathrm{D})} \right).$$

Proof. We consider $T := \mathcal{H} \cup \{0\}$ as a subset of $L_2(\mathrm{D})$ and equip it with the metric d defined by $\|\cdot\|_{L_2(\mathrm{D})}$. Since $L_2(\mathrm{D})$ is finite-dimensional, (T, d) is a separable metric space. Let us now fix a Rademacher sequence $\varepsilon_1, \ldots, \varepsilon_n$ with respect to a distribution ν on some Θ. Moreover, for $h \in T$, we define the function $g_h : \Theta \to \mathbb{R}$ by $g_h(\theta) := \frac{1}{n}\sum_{i=1}^n \varepsilon_i(\theta) h(z_i)$, $\theta \in \Theta$. Obviously, this yields $g_0 = 0$ and $g_h \in \mathcal{L}_0(\Theta)$, $h \in T$. Moreover, the convergence with respect to $\|\cdot\|_{L_2(\mathrm{D})}$ implies pointwise convergence on $z_1, \ldots, z_n$, and hence, for each fixed $\theta \in \Theta$, the map $h \mapsto g_h(\theta)$ is continuous with respect to the metric d. In other words, $(g_h)_{h\in T}$ is a separable Carathéodory family. Let us now establish a bound of the form (7.20). To this end, we observe that for $a \in \mathbb{R}$ and $i = 1, \ldots, n$ we have $\mathbb{E}_\nu e^{a\varepsilon_i} = \frac{1}{2}(e^a + e^{-a}) \le e^{a^2/2}$. For $h \in L_2(\mathrm{D})$, the independence of $\varepsilon_1, \ldots, \varepsilon_n$ and $\|h\|^2_{L_2(\mathrm{D})} = \frac{1}{n}\sum_{i=1}^n h^2(z_i)$ thus yields

$$\mathbb{E}_\nu \exp\left(\frac{\lambda}{n}\sum_{i=1}^n \varepsilon_i h(z_i)\right) = \prod_{i=1}^n \mathbb{E}_\nu e^{\frac{\lambda h(z_i)}{n}\varepsilon_i} \le \prod_{i=1}^n e^{\frac{\lambda^2 h^2(z_i)}{2n^2}} = \exp\left(\frac{\lambda^2 \|h\|^2_{L_2(\mathrm{D})}}{2n}\right)$$

for all $\lambda \in \mathbb{R}$. From this we conclude that

$$\mathbb{E}_\nu \exp\big(\lambda(g_{h_1} - g_{h_2})\big) \le \exp\left(\frac{\lambda^2 \|h_1 - h_2\|^2_{L_2(\mathrm{D})}}{2n}\right)$$

for all $\lambda \in \mathbb{R}$ and all $h_1, h_2 \in T$. Consequently, (7.20) is satisfied for $K := (2n)^{-1/2}$, and hence we obtain the assertion by Theorem 7.12. □

Before we return to our main goal, i.e., estimating $\mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r, n)$ for $\mathcal{H}_r$ defined by (7.15), we need two simple lemmas. The first one compares the entropy numbers of $\mathcal{H} \cup \{0\}$ with those of $\mathcal{H}$.

Lemma 7.14. *For all $\mathcal{H} \subset \mathcal{L}_0(Z)$, all $D \in Z^n$, and all $i \ge 2$, we have*

$$\begin{aligned} e_1\big(\mathcal{H} \cup \{0\}, L_2(\mathrm{D})\big) &\le \sup_{h\in\mathcal{H}} \|h\|_{L_2(\mathrm{D})}\,, \\ e_i\big(\mathcal{H} \cup \{0\}, L_2(\mathrm{D})\big) &\le e_{i-1}\big(\mathcal{H}, L_2(\mathrm{D})\big)\,. \end{aligned}$$

Proof. The first inequality immediately follows from $0 \in \mathcal{H} \cup \{0\}$. To show the second inequality, we fix an ε-net $T \subset \mathcal{H}$ of $\mathcal{H}$ having cardinality $|T| \le 2^{i-2}$. Then $T \cup \{0\} \subset \mathcal{H} \cup \{0\}$ is an ε-net of $\mathcal{H} \cup \{0\}$ satisfying $|T \cup \{0\}| \le 2^{i-2} + 1$. Using $2^{i-2} + 1 \le 2^{i-1}$, $i \ge 2$, we then find the second assertion. □

The second technical lemma estimates sums of the form $\sum_{i=1}^{\infty} 2^{i/2} s_{2^i}$ for positive sequences (s_i) of known decay.

Lemma 7.15. *Let $(s_i) \subset [0, \infty)$ be a decreasing sequence for which there are $a > 0$ and $p \in (0, 1)$ such that $s_1 \le a\, 2^{-\frac{1}{2p}}$ and $s_i \le a\, i^{-\frac{1}{2p}}$ for all $i \ge 2$. Then we have*

$$\sum_{i=0}^{\infty} 2^{i/2} s_{2^i} \;\le\; \frac{\sqrt{2}\, C_p^p}{(\sqrt{2}-1)(1-p)}\, a^p\, s_1^{1-p}\,, \tag{7.21}$$

where

$$C_p := \frac{\sqrt{2}-1}{\sqrt{2} - 2^{\frac{2p-1}{2p}}} \cdot \frac{1-p}{p}\,. \tag{7.22}$$

Proof. Let us fix a $t \ge 0$ and a natural number $n \ge 1$ such that $n - 1 \le t < n$. Then a simple application of the geometric series yields

$$\sum_{i=0}^{n-1} 2^{i/2} s_{2^i} \le s_1 \sum_{i=0}^{n-1} 2^{i/2} = s_1\, \frac{2^{n/2} - 1}{\sqrt{2} - 1} \le \frac{\sqrt{2}\, s_1}{\sqrt{2} - 1}\, 2^{t/2}\,,$$

and a similar argument shows that

$$\sum_{i=n}^{\infty} 2^{i/2} s_{2^i} \le a \sum_{i=n}^{\infty} 2^{\frac{i}{2} - \frac{i}{2p}} = a\biggl(\sum_{i=0}^{\infty} 2^{\frac{i}{2} - \frac{i}{2p}} - \sum_{i=0}^{n-1} 2^{\frac{i}{2} - \frac{i}{2p}} \biggr) \le \frac{a\, 2^{\frac{t(p-1)}{2p}}}{1 - 2^{\frac{p-1}{2p}}}\,.$$

Combining both estimates, we then obtain that for all $t \ge 0$ we have

$$\sum_{i=0}^{\infty} 2^{i/2} s_{2^i} \le c_1 e^{\alpha_1 t} + c_2 e^{\alpha_2 t}\,, \tag{7.23}$$

where $c_1 := \frac{\sqrt{2}}{\sqrt{2}-1} s_1$, $\alpha_1 := \frac{\ln 2}{2}$, $c_2 := a\bigl(1 - 2^{\frac{p-1}{2p}}\bigr)^{-1}$, and $\alpha_2 := \frac{(p-1)\ln 2}{2p}$. Now it is easy to check that the function $t \mapsto c_1 e^{\alpha_1 t} + c_2 e^{\alpha_2 t}$ is minimized at

$$t^* := \frac{1}{\alpha_2 - \alpha_1} \ln\biggl(-\frac{\alpha_1 c_1}{\alpha_2 c_2} \biggr) = \frac{2p}{\ln 2} \ln\biggl(\frac{a C_p}{s_1} \biggr) \ge \frac{2p}{\ln 2} \ln\bigl(2^{\frac{1}{2p}} C_p \bigr)\,.$$

In order to show $t^* \ge 0$, we write $f(x) := (\sqrt{2} - 1) 2^x x$ and $g(x) = 2^{x/2} - 1$ for $x \ge 0$. Since for $x_p := \frac{1}{p} - 1$ we have

$$2^{\frac{1}{2p}} C_p = (\sqrt{2} - 1) \cdot \frac{2^{(\frac{1}{p} - 1)\frac{1}{2}}}{1 - 2^{-(\frac{1}{p} - 1)\frac{1}{2}}} \cdot \biggl(\frac{1}{p} - 1 \biggr) = \frac{f(x_p)}{g(x_p)}\,,$$

it then suffices to show $f(x) \geq g(x)$ for all $x > 0$. However, the latter follows from $f(0) = g(0) = 0$,

$$f'(x) = (\sqrt{2}-1)2^x(x\ln 2 + 1) \geq 2^x(\sqrt{2}-1) > 2^{x/2}\ln\sqrt{2} = g'(x)\,, \qquad x \geq 0,$$

and the fundamental theorem of calculus. Plugging t^* into (7.23) together with some simple but tedious calculations then yields the assertion. □

With the help of the preceding lemmas, we can now establish an upper bound for expectations of empirical Rademacher averages.

Theorem 7.16. *Let $\mathcal{H} \subset \mathcal{L}_0(Z)$ be a separable Carathéodory set and* P *be a distribution on Z. Suppose that there exist constants $B \geq 0$ and $\sigma \geq 0$ such that $\|h\|_\infty \leq B$ and $\mathbb{E}_{\mathrm{P}} h^2 \leq \sigma^2$ for all $h \in \mathcal{H}$. Furthermore, assume that for a fixed $n \geq 1$ there exist constants $p \in (0,1)$ and $a \geq B$ such that*

$$\mathbb{E}_{D\sim \mathrm{P}^n} e_i(\mathcal{H}, L_2(\mathrm{D})) \;\leq\; a\, i^{-\frac{1}{2p}}\,, \qquad\qquad i \geq 1. \tag{7.24}$$

Then there exist constants $C_1(p) > 0$ and $C_2(p) > 0$ depending only on p such that

$$\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H}, n) \leq \max\Bigl\{C_1(p)\, a^p \sigma^{1-p} n^{-\frac{1}{2}},\, C_2(p)\, a^{\frac{2p}{1+p}} B^{\frac{1-p}{1+p}} n^{-\frac{1}{1+p}}\Bigr\}\,.$$

Proof. For $s_i := \mathbb{E}_{D\sim \mathrm{P}^n} e_i(\mathcal{H}\cup\{0\}, L_2(\mathrm{D}))$, $i \geq 2$, and $\tilde{a} := 2^{\frac{1}{2p}} a$, Lemma 7.14 yields

$$s_i \leq \mathbb{E}_{D\sim \mathrm{P}^n} e_{i-1}(\mathcal{H}, L_2(\mathrm{D})) \leq a(i-1)^{-\frac{1}{2p}} \leq \tilde{a} i^{-\frac{1}{2p}}\,. \tag{7.25}$$

Furthermore, for

$$\delta_D := \sup_{h\in\mathcal{H}} \|h\|_{L_2(\mathrm{D})}\,, \qquad\qquad D \in Z^n\,,$$

and $s_1 := \mathbb{E}_{D\sim \mathrm{P}^n}\delta_D$, we have $s_1 \leq B \leq a = \tilde{a}2^{-\frac{1}{2p}}$. Moreover, the monotonicity of the entropy numbers together with Lemma 7.14 shows that $s_2 \leq \mathbb{E}_{D\sim \mathrm{P}^n} e_1(\mathcal{H}\cup\{0\}, L_2(\mathrm{D})) \leq s_1$, and hence it is easy to conclude that (s_i) is decreasing. By Theorem 7.13 and Lemma 7.15, we hence find

$$\begin{aligned}\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H}, n) &\leq \sqrt{\frac{\ln 16}{n}} \sum_{i=0}^{\infty} 2^{i/2}\, s_{2^i} \\ &\leq \frac{1}{\sqrt{n}} \cdot \frac{2\sqrt{\ln 16}}{(\sqrt{2}-1)(1-p)} C_p^p \left(\mathbb{E}_{D\sim \mathrm{P}^n}\delta_D\right)^{1-p} a^p\,.\end{aligned}$$

Moreover, Corollary A.8.5 yields

$$\mathbb{E}_{D\sim \mathrm{P}^n}\delta_D \leq (\mathbb{E}_{D\sim \mathrm{P}^n}\delta_D^2)^{1/2} \leq \left(\sigma^2 + 8B\,\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H}, n)\right)^{1/2}\,,$$

and hence we obtain

$$\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n) \leq \frac{2\sqrt{\ln 16}\,C_p^p}{(\sqrt{2}-1)(1-p)}\cdot\frac{a^p}{\sqrt{n}}\,\big(\sigma^2+8B\,\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n)\big)^{\frac{1-p}{2}}\,.$$

In the case $\sigma^2 \geq 8B\,\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n)$, we conclude that

$$\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n) \leq \frac{2\sqrt{\ln 256}\,C_p^p}{(\sqrt{2}-1)(1-p)2^{p/2}}\cdot\frac{a^p}{\sqrt{n}}\cdot\sigma^{1-p}\,,$$

and in the case $\sigma^2 < 8B\,\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n)$ we have

$$\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n) \leq \frac{8\sqrt{\ln 16}\,C_p^p}{(\sqrt{2}-1)(1-p)4^p}\cdot\frac{a^p}{\sqrt{n}}\cdot\big(B\,\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n)\big)^{\frac{1-p}{2}}\,.$$

Simple algebraic transformations then yield the assertion. □

Numerical calculations show that the constants obtained in the proof of Theorem 7.16 satisfy $(1-p)C_1(p) \in [6.8, 12.25]$ and $(1-p)C_2(p) \in [5.68, 6.8]$. Further calculations for $C_1(p)$ yield $\lim_{p\to 0} C_1(p) \leq 11.38$ and $\lim_{p\to 1}(1-p)$ $C_1(p) \approx 6.8$. Finally, similar calculations for $C_2(p)$ give $\lim_{p\to 0} C_2(p) \approx 5.68$ and $\lim_{p\to 1}(1-p)C_2(p) \leq 6.8$. However, it is obvious from the proof above that we can, e.g., obtain smaller values for $C_2(p)$ for the price of larger values for $C_1(p)$, and hence the calculations above only illustrate how $C_1(p)$ and $C_2(p)$ can be estimated.

Before we use these estimates on the Rademacher averages to bound the term

$$\mathbb{E}_{D\sim \mathrm{P}^n}\sup_{h\in\mathcal{H}_r}\big|\mathbb{E}_{\mathrm{P}}h - \mathbb{E}_{\mathrm{D}}h\big|$$

on the left-hand side of (7.16), we present a simple lemma that estimates the entropy numbers of a set $\mathcal{H} := \{L\circ f - L\circ f^*_{L,\mathrm{P}} : f \in \mathcal{F}\}$ by the entropy numbers of the "base" set $\mathcal{F}$.

Lemma 7.17. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a locally Lipschitz continuous loss, P be a distribution on $X \times Y$, and $q \in [1,\infty]$ and $M > 0$ be constants. Assume that we have $\mathbb{E}_{(x,y)\sim \mathrm{P}}L^q(x,y,0) < \infty$. Furthermore, let $\mathcal{F} \subset \mathcal{L}_q(\mathrm{P}_X)$ be a non-empty set and $f_0 \in \mathcal{L}_0(\mathrm{P}_X)$ be an arbitrary function. We write*

$$\mathcal{H} := \big\{L\circ \widehat{f} - L\circ \widehat{f_0} : f \in \mathcal{F}\big\}\,,$$

where $\widehat{\cdot}$ is the clipping operation at M defined in (2.14). Then we have

$$e_n\big(\mathcal{H}, \|\cdot\|_{L_q(\mathrm{P})}\big) \;\leq\; |L|_{M,1}\cdot e_n\big(\mathcal{F}, \|\cdot\|_{L_q(\mathrm{P}_X)}\big)\,, \qquad n \geq 1.$$

Proof. For $L\circ\widehat{\mathcal{F}} := \{L\circ\widehat{f} : f \in \mathcal{F}\}$, we have $L\circ\widehat{\mathcal{F}} \subset \mathcal{L}_q(\mathrm{P})$ by (2.11) and $\mathbb{E}_{(x,y)\sim \mathrm{P}}L^q(x,y,0) < \infty$. In addition, we obviously have

$$e_n\big(\mathcal{H}, \|\cdot\|_{L_q(\mathrm{P})}\big) \;\leq\; e_n\big(L\circ\widehat{\mathcal{F}}, \|\cdot\|_{L_q(\mathrm{P})}\big)\,, \qquad n \geq 1.$$

Now let $f_1, \ldots, f_{2^{n-1}}$ be an ε-net of $\mathcal{F}$ with respect to $\|\cdot\|_{L_q(\mathrm{P}_X)}$. For $f \in \mathcal{F}$, there then exists an $i \in \{1, \ldots, 2^{n-1}\}$ such that $\|f - f_i\|_{L_q(\mathrm{P}_X)} \leq \varepsilon$. Moreover, the clipping operation obviously satisfies $|\widehat{s} - \widehat{t}| \leq |s - t|$ for all $s, t \in \mathbb{R}$. For $q < \infty$, we thus obtain

$$\begin{aligned}
\|L \circ \widehat{f} - L \circ \widehat{f}_i\|^q_{L_q(\mathrm{P})} &= \int_{X \times Y} \bigl|L(x, y, \widehat{f}(x)) - L(x, y, \widehat{f}_i(x))\bigr|^q \, d\mathrm{P}(x, y) \\
&\leq |L|^q_{M,1} \int_X \bigl|\widehat{f}(x) - \widehat{f}_i(x)\bigr|^q \, d\mathrm{P}_X(x) \\
&\leq |L|^q_{M,1} \cdot \varepsilon^q \,,
\end{aligned}$$

and consequently, $L \circ \widehat{f}_1, \ldots, L \circ \widehat{f}_{2^{n-1}}$ is an $|L|_{M,1} \cdot \varepsilon$-net of $L \circ \widehat{\mathcal{F}}$ with respect to $\|\cdot\|_{L_q(\mathrm{P})}$. From this we easily find the assertion. The case $q = \infty$ can be shown analogously. □

Let us now return to our example at the beginning of this section, where we applied Talagrand's inequality to the set $\mathcal{G}_r := \{g_{f,r} : f \in \mathcal{F}\}$ defined by (7.12). To this end, let us assume that L is a locally Lipschitz continuous loss function[3] and $\mathcal{F} \subset \mathcal{L}_0(X)$ is a non-empty subset. Assume that $\mathcal{F}$ is equipped with a complete, separable metric dominating the pointwise convergence and that all $f \in \mathcal{F}$ satisfy $\|f\|_\infty \leq M$ for a suitable constant $M > 0$. Moreover, we assume that there exists a $B > 0$ such that

$$L(x, y, t) \leq B \,, \qquad (x, y) \in X \times Y, \, t \in [-M, M].$$

In addition, let P be a distribution on $X \times Y$ and $f^*_{L,\mathrm{P}} : X \to [-M, M]$ be a Bayes decision function. Note that combining these assumptions, we see that the supremum bound (7.5) of Theorem 7.2 holds for all $f \in \mathcal{F}$. Furthermore, these assumption match those of Proposition 6.22, which established our first oracle inequality for ERM over infinite function classes. Assume further that there exist constants $\vartheta \in [0, 1]$ and $V \geq B^{2-\vartheta}$ such that for all $f \in \mathcal{F}$ we have

$$\mathbb{E}_\mathrm{P}\bigl(L \circ f - L \circ f^*_{L,\mathrm{P}}\bigr)^2 \leq V \cdot \bigl(\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}})\bigr)^\vartheta \,. \tag{7.26}$$

Let us now define $\mathcal{H}_r$ by (7.15), i.e., $\mathcal{H}_r := \{h_f : f \in \mathcal{F} \text{ and } \mathbb{E}_\mathrm{P} h_f \leq r\}$, where $h_f := L \circ f - L \circ f^*_{L,\mathrm{P}}$. Finally, let us assume that there exist constants $a > 0$ and $p \in (0, 1)$ such that

$$\mathbb{E}_{D \sim \mathrm{P}^n} e_i(\mathcal{F}, L_2(\mathrm{D})) \;\leq\; a \, i^{-\frac{1}{2p}} \,, \qquad i, n \geq 1. \tag{7.27}$$

Now note that $\|f\|_\infty \leq M$ implies $\widehat{f} = f$ for all $f \in \mathcal{F}$, where $\widehat{\cdot}$ denotes the clipping operation at M. Applying Lemma 7.17 together with (A.36) therefore yields $\mathbb{E}_{D \sim \mathrm{P}^n} e_i(\mathcal{H}_r, L_2(\mathrm{D})) \leq 2a|L|_{M,1} i^{-\frac{1}{2p}}$ for all $i, n \geq 1$ and all

[3] Loss functions that can be clipped will be considered in the following section. Here we assume for the sake of simplicity that clipping is superfluous.

$r > r^* := \inf\{\mathbb{E}_P h_f : f \in \mathcal{F}\}$. In addition, (7.26) implies $\mathbb{E}_P h^2 \le V r^\vartheta$ for all $h \in \mathcal{H}_r$, and thus we obtain

$$\mathbb{E}_{D\sim P^n} \mathrm{Rad}_D(\mathcal{H}_r, n) \le C \max\left\{ r^{\frac{\vartheta(1-p)}{2}} n^{-\frac{1}{2}}, n^{-\frac{1}{1+p}} \right\}$$

by Theorem 7.16, where the constant C depends on a, p, M, B, and V. By (7.17) and Proposition 7.10, we thus find that

$$\sup_{f\in\mathcal{F}} \frac{\mathbb{E}_P h_f - \mathbb{E}_D h_f}{\mathbb{E}_P h_f + r} < \frac{16C}{\sqrt{n r^{2-\vartheta(1-p)}}} + \frac{16C}{n^{1/(1+p)} r} + \sqrt{\frac{2V\tau}{n r^{2-\vartheta}}} + \frac{5B\tau}{3nr}$$

holds with probability P^n not less than $1 - e^{-\tau}$. Let us now assume that we consider a measurable ERM over the set $\mathcal{F}$. By replacing (7.9) with the inequality above, the proof of Theorem 7.2 then yields after some basic yet exhausting calculations[4] that for fixed $\tau \ge 1$ and $n \ge 1$ we have

$$\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \;\le\; 6\big(\mathcal{R}^*_{L,P,\mathcal{F}} - \mathcal{R}^*_{L,P}\big) + c\sqrt{\tau} n^{-\frac{1}{2-\vartheta+\vartheta p}} \tag{7.28}$$

with probability P^n not less than $1 - e^{-\tau}$, where c is a constant independent of τ and n. To appreciate this oracle inequality, we briefly compare it with our previous results for ERM. Our first approach using covering numbers led to Proposition 6.22, which for classes $\mathcal{F} \subset \mathcal{L}_\infty(X)$ satisfying the assumption

$$e_i(\mathcal{F}, \|\cdot\|_\infty) \;\le\; a\, i^{-\frac{1}{2p}}, \qquad\qquad i \ge 1, \tag{7.29}$$

showed that for fixed $\tau \ge 1$ and $n \ge 1$ we have

$$\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \;\le\; (\mathcal{R}^*_{L,P,\mathcal{F}} - \mathcal{R}^*_{L,P}) + \tilde{c}\sqrt{\tau} n^{-\frac{1}{2+2p}} \tag{7.30}$$

with probability P^n not less than $1 - e^{-\tau}$, where $\tilde{c}$ is a constant independent of τ and n. Now note that, for $p \in (0,1)$, condition (7.29) implies (7.27) and hence the oracle inequality (7.28) derived by the more involved techniques also holds. Moreover, we have $2 + 2p > 2 - \vartheta + \vartheta p$ by at least $2p$, and consequently (7.30) has worse behavior in the sample size n than (7.28). In this regard, it is also interesting to note that a brute-force approach (see Exercise 7.3) for generalizing Theorem 7.2 with the help of covering numbers does provide an improvement compared with (7.30). However, this improved oracle inequality is still not as sharp as (7.28).

Another difference between the oracle inequalities above is that (7.30) estimates against $\mathcal{R}^*_{L,P,\mathcal{F}} - \mathcal{R}^*_{L,P}$, which at first glance seems to be more interesting than $6(\mathcal{R}^*_{L,P,\mathcal{F}} - \mathcal{R}^*_{L,P})$ considered in (7.28). However, in the following, we are mainly interested in situations where the sets $\mathcal{F}$ become larger with the sample size, and hence this effect will be negligible for our purposes.

Finally, to compare (7.28) with the oracle inequality of Theorem 7.2, we assume that $\mathcal{F}$ is a *finite* set. Then it is easy to see that the entropy assumption (7.27) is satisfied for all $p \in (0,1)$, and hence (7.28) essentially recovers, i.e., modulo an arbitrarily small change in the exponent of n, Theorem 7.2.

[4] Since we will establish a more general inequality in the next section we omit the details of the estimation as well as an explicit upper bound on the constant c.

7.4 Refined Oracle Inequalities for SVMs

The goal of this section is to establish improved oracle inequalities for SVMs using the ideas of the previous section.

Let us begin by recalling Section 6.5, where we saw that oracle inequalities for SVMs can be used to investigate data-dependent parameter selection strategies if the loss can be *clipped*. In the following, we will therefore focus solely on such loss functions. Moreover, we will first consider the following modification of regularized empirical risk minimization.

Definition 7.18. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss that can be clipped, $\mathcal{F} \subset \mathcal{L}_0(X)$ be a subset, $\Upsilon : \mathcal{F} \to [0,\infty)$ be a function, and $\epsilon \geq 0$. A learning method whose decision functions f_D satisfy*

$$\Upsilon(f_D) + \mathcal{R}_{L,\mathrm{D}}(\widehat{f}_D) \;\leq\; \inf_{f\in\mathcal{F}} \Upsilon(f) + \mathcal{R}_{L,\mathrm{D}}(f) + \epsilon \tag{7.31}$$

for all $n \geq 1$ and $D \in (X\times Y)^n$ is called ***ϵ-approximate clipped regularized empirical risk minimization (ϵ-CR-ERM)*** *with respect to L, $\mathcal{F}$, and Υ.*

Moreover, in the case $\epsilon = 0$, we simply speak of ***clipped regularized empirical risk minimization (CR-ERM)****.*

Note that on the right-hand side of (7.31) the unclipped loss is considered, and hence CR-ERM does *not* necessarily minimize the regularized clipped empirical risk $\Upsilon(\,\cdot\,) + \mathcal{R}_{L,\mathrm{D}}(\widehat{\,\cdot\,})$. Moreover, in general CR-ERMs do *not* minimize the regularized risk $\Upsilon(\,\cdot\,) + \mathcal{R}_{L,\mathrm{D}}(\,\cdot\,)$ either, because on the left-hand side of (7.31) the clipped function is considered. However, if we have a minimizer of the *un*clipped regularized risk, then it automatically satisfies (7.31). In particular, SVM decision functions satisfy (7.31) for the regularizer $\Upsilon := \lambda\|\cdot\|_H^2$ and $\epsilon := 0$. In other words, SVMs are CR-ERMs.

Before we establish an oracle inequality for CR-ERMs, let us first ensure that there exist measurable versions. This is done in the following lemma.

Lemma 7.19 (Measurability of CR-ERMs). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss that can be clipped and $\mathcal{F} \subset \mathcal{L}_0(X)$ be a subset that is equipped with a complete, separable metric dominating the pointwise convergence. Moreover, let $\Upsilon : \mathcal{F} \to [0,\infty)$ be a function that is measurable with respect to the corresponding Borel σ-algebra on $\mathcal{F}$. Then, for all $\epsilon > 0$, there exists a measurable ϵ-CR-ERM with respect to L, $\mathcal{F}$, and Υ. Moreover, if there exists a CR-ERM, then there also exists a measurable CR-ERM. Finally, in both cases, the map $D \mapsto \Upsilon(f_D)$ mapping $(X \times Y)^n$ to $[0,\infty)$ is measurable for all $n \geq 1$.*

Proof. The maps $(D,f) \mapsto \Upsilon(f) + \mathcal{R}_{L,\mathrm{D}}(\widehat{f})$ and $(D,f) \mapsto \Upsilon(f) + \mathcal{R}_{L,\mathrm{D}}(f)$ are measurable by Lemma 2.11, where for the measurability of the first map we consider the clipped version of L, i.e.,

$$\widehat{L}(x,y,t) := L(x,y,\widehat{t}\,), \qquad (x,y) \in X \times Y,\, t \in \mathbb{R}.$$

Consequently, the maps above are also measurable with respect to the universal completion of the σ-algebra on $(X \times Y)^n$. By part *iii)* of Lemma A.3.18, the map $h : (X \times Y)^n \times \mathcal{F} \to \mathbb{R}$ defined by

$$h(D, f) := \Upsilon(f) + \mathcal{R}_{L,\mathrm{D}}(\widehat{f}\,) - \inf_{f' \in \mathcal{F}} \bigl(\Upsilon(f') + \mathcal{R}_{L,\mathrm{D}}(f')\bigr)\,,$$

$D \in (X \times Y)^n$, $f \in \mathcal{F}$, is therefore measurable with respect to the universal completion of the σ-algebra on $(X \times Y)^n$. For $A := (-\infty, \epsilon]$, we now consider the map $F : (X \times Y)^n \to 2^{\mathcal{F}}$ defined by

$$F(D) := \bigl\{f \in \mathcal{F} : h(D, f) \in A\bigr\}\,, \qquad\qquad D \in (X \times Y)^n.$$

Obviously, $F(D)$ contains exactly the functions f satisfying (7.31). Moreover, in the case $\epsilon > 0$, we obviously have $F(D) \neq \emptyset$ for all $D \in (X \times Y)^n$, while in the case $\epsilon = 0$ this follows from the existence of a CR-ERM. Therefore, part *ii)* of Lemma A.3.18 shows that there exists a measurable map $D \mapsto f_D$ such that $f_D \in F(D)$ for all $D \in (X \times Y)^n$. The first two assertions can then be shown by a literal repetition of the proof of Lemma 6.17. The last assertion follows from the measurability of Υ. □

Before we present the first main result of this section, which establishes an oracle inequality for general ϵ-CR-ERMs, we first need to introduce a few more notations. To this end, we assume that L, $\mathcal{F}$, and Υ satisfy the assumptions of Lemma 7.19. Moreover, let P be a distribution on $X \times Y$ such that there exists a Bayes decision function $f^*_{L,\mathrm{P}} : X \to [-M, M]$, where $M > 0$ is the constant at which L can be clipped. For

$$r^* := \inf_{f \in \mathcal{F}} \Upsilon(f) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) - \mathcal{R}^*_{L,\mathrm{P}} \tag{7.32}$$

and $r > r^*$, we write

$$\mathcal{F}_r := \{f \in \mathcal{F} \,:\, \Upsilon(f) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) - \mathcal{R}^*_{L,\mathrm{P}} \leq r\}\,, \tag{7.33}$$

$$\mathcal{H}_r := \{L \circ \widehat{f} \, - L \circ f^*_{L,\mathrm{P}} \,:\, f \in \mathcal{F}_r\}\,, \tag{7.34}$$

where, as usual, $L \circ g$ denotes the function $(x, y) \mapsto L(x, y, g(x))$. Furthermore, assume that there exist constants $B > 0$, $\vartheta \in [0, 1]$, and $V \geq B^{2-\vartheta}$ such that

$$L(x, y, t) \leq B\,, \tag{7.35}$$

$$\mathbb{E}_{\mathrm{P}}\bigl(L \circ \widehat{f} \, - L \circ f^*_{L,\mathrm{P}}\bigr)^2 \leq V \cdot \bigl(\mathbb{E}_{\mathrm{P}}(L \circ \widehat{f} \, - L \circ f^*_{L,\mathrm{P}})\bigr)^{\vartheta}\,, \tag{7.36}$$

for all $(x, y) \in X \times Y$, $t \in [-M, M]$, and $f \in \mathcal{F}$. In the following, (7.35) is called a **supremum bound** and (7.36) is called a **variance bound**.

With the help of these notions, we can now formulate the first main result of this section.

Theorem 7.20 (Oracle inequality for CR-ERMs). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous loss that can be clipped at $M > 0$ and that satisfies (7.35) for a constant $B > 0$. Moreover, let $\mathcal{F} \subset \mathcal{L}_0(X)$ be a subset that is equipped with a complete, separable metric dominating the pointwise convergence, and let $\Upsilon : \mathcal{F} \to [0,\infty)$ be a continuous function. Given a distribution P on $X \times Y$ that satisfies (7.36), we define r^* and $\mathcal{H}_r$ by (7.32) and (7.34), respectively. Assume that for fixed $n \geq 1$ there exists a $\varphi_n : [0,\infty) \to [0,\infty)$ such that $\varphi_n(4r) \leq 2\varphi_n(r)$ and*

$$\mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r, n) \;\leq\; \varphi_n(r) \tag{7.37}$$

for all $r > r^*$. Finally, fix an $f_0 \in \mathcal{F}$ and a $B_0 \geq B$ such that $\|L \circ f_0\|_\infty \leq B_0$. Then, for all fixed $\epsilon \geq 0$, $\tau > 0$, and $r > 0$ satisfying

$$r > \max\left\{30\varphi_n(r), \left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}}, \frac{5B_0\tau}{n}, r^*\right\}, \tag{7.38}$$

every measurable ϵ-CR-ERM satisfies

$$\Upsilon(f_D) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_D) - \mathcal{R}^*_{L,\mathrm{P}} \;\leq\; 6\big(\Upsilon(f_0) + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}\big) + 3r + 3\epsilon$$

with probability P^n not less than $1 - 3e^{-\tau}$.

Note that the value of r in Theorem 7.20 is only *implicitly* determined by the relation $r > 30\varphi_n(r)$ appearing in (7.38). Of course, for particular types of functions φ_n, this implicit definition can be made explicit, and we will see in the following that for SVMs this is even a relatively easy task.

Proof. As usual, we define $h_f := L \circ f - L \circ f^*_{L,\mathrm{P}}$ for all $f \in \mathcal{L}_0(X)$. By the definition of f_D, we then have

$$\Upsilon(f_D) + \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D} \leq \Upsilon(f_0) + \mathbb{E}_{\mathrm{D}} h_{f_0} + \epsilon\,,$$

and consequently we obtain

$$\begin{aligned}
&\Upsilon(f_D) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_D) - \mathcal{R}^*_{L,\mathrm{P}} \\
&= \Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} \\
&\leq \Upsilon(f_0) + \mathbb{E}_{\mathrm{D}} h_{f_0} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D} + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} + \epsilon \\
&= (\Upsilon(f_0) + \mathbb{E}_{\mathrm{P}} h_{f_0}) + (\mathbb{E}_{\mathrm{D}} h_{f_0} - \mathbb{E}_{\mathrm{P}} h_{f_0}) + (\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D}) + \epsilon
\end{aligned} \tag{7.39}$$

for all $D \in (X \times Y)^n$. Let us first bound the term $\mathbb{E}_{\mathrm{D}} h_{f_0} - \mathbb{E}_{\mathrm{P}} h_{f_0}$. To this end, we further split this difference into

$$\mathbb{E}_{\mathrm{D}} h_{f_0} - \mathbb{E}_{\mathrm{P}} h_{f_0} = \big(\mathbb{E}_{\mathrm{D}}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})\big) + \big(\mathbb{E}_{\mathrm{D}} h_{\widehat{f}_0} - \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_0}\big)\,. \tag{7.40}$$

Now observe that $L \circ f_0 - L \circ \widehat{f}_0 \geq 0$ implies $h_{f_0} - h_{\widehat{f}_0} = L \circ f_0 - L \circ \widehat{f}_0 \in [0, B_0]$, and hence we obtain

$$\mathbb{E}_{\mathrm{P}}\big((h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})\big)^2 \leq \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})^2 \leq B_0\, \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})\,.$$

Consequently, Bernstein's inequality stated in Theorem 6.12 and applied to the function $h := (h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})$ shows that

$$\mathbb{E}_{\mathrm{D}}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0}) < \sqrt{\frac{2\tau B_0\, \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})}{n}} + \frac{2B_0\tau}{3n}$$

holds with probability P^n not less than $1 - e^{-\tau}$. Moreover, using $\sqrt{ab} \leq \frac{a}{2} + \frac{b}{2}$, we find

$$\sqrt{\frac{2\tau B_0\, \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})}{n}} \leq \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0}) + \frac{B_0\tau}{2n}\,,$$

and consequently we have

$$\mathbb{E}_{\mathrm{D}}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0}) < \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0}) + \frac{7B_0\tau}{6n} \tag{7.41}$$

with probability P^n not less than $1 - e^{-\tau}$.

In order to bound the remaining term in (7.40), we now recall the proof of Theorem 7.2, where we note that (7.35) implies (7.5). Consequently, (7.8) shows that with probability P^n not less than $1 - e^{-\tau}$ we have

$$\mathbb{E}_{\mathrm{D}} h_{\widehat{f}_0} - \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_0} < \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_0} + \Big(\frac{2V\tau}{n}\Big)^{\frac{1}{2-\vartheta}} + \frac{4B\tau}{3n}\,.$$

By combining this estimate with (7.41) and (7.40), we now obtain that with probability P^n not less than $1 - 2e^{-\tau}$ we have

$$\mathbb{E}_{\mathrm{D}} h_{f_0} - \mathbb{E}_{\mathrm{P}} h_{f_0} < \mathbb{E}_{\mathrm{P}} h_{f_0} + \Big(\frac{2V\tau}{n}\Big)^{\frac{1}{2-\vartheta}} + \frac{4B\tau}{3n} + \frac{7B_0\tau}{6n}\,, \tag{7.42}$$

i.e., we have established a bound on the second term in (7.39).

Let us now consider the case $n < 72\tau$. Then the assumption $B^{2-\vartheta} \leq V$ implies $B < r$. Combining (7.42) with (7.39) and using both $B \leq B_0$ and $\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D} \leq 2B$ we hence find

$$\begin{aligned}
&\Upsilon(f_D) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_D) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq \Upsilon(f_0) + 2\mathbb{E}_{\mathrm{P}} h_{f_0} + \Big(\frac{2V\tau}{n}\Big)^{\frac{1}{2-\vartheta}} + \frac{5B_0\tau}{2n} + (\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D}) + \epsilon \\
&\leq 6\big(\Upsilon(f_0) + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}\big) + 3r + \varepsilon
\end{aligned}$$

with probability P^n not less than $1 - 2e^{-\tau}$.

Consequently, it remains to consider the case $n \geq 72\tau$. In order to establish a non-trivial bound on the term $\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D}$ in (7.39), we define functions

$$g_{f,r} := \frac{\mathbb{E}_{\mathrm{P}} h_{\widehat{f}} - h_{\widehat{f}}}{\Upsilon(f) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}} + r}, \qquad f \in \mathcal{F},\ r > r^*.$$

Obviously, for $f \in \mathcal{F}$, we then have $\|g_{f,r}\|_\infty \le 2Br^{-1}$, and for $\vartheta > 0$, $q := \frac{2}{2-\vartheta}$, $q' := \frac{2}{\vartheta}$, $a := r$, and $b := \mathbb{E}_{\mathrm{P}} h_{\widehat{f}} \neq 0$, the second inequality of Lemma 7.1 yields

$$\mathbb{E}_{\mathrm{P}} g_{f,r}^2 \le \frac{\mathbb{E}_{\mathrm{P}} h_{\widehat{f}}^2}{(\mathbb{E}_{\mathrm{P}} h_{\widehat{f}} + r)^2} \le \frac{(2-\vartheta)^{2-\vartheta}\vartheta^\vartheta\, \mathbb{E}_{\mathrm{P}} h_{\widehat{f}}^2}{4r^{2-\vartheta}(\mathbb{E}_{\mathrm{P}} h_{\widehat{f}})^\vartheta} \le V r^{\vartheta-2}. \tag{7.43}$$

Moreover, for $\vartheta > 0$ and $\mathbb{E}_{\mathrm{P}} h_{\widehat{f}} = 0$, we have $\mathbb{E}_{\mathrm{P}} h_{\widehat{f}}^2 = 0$ by the variance bound (7.36), which in turn implies $\mathbb{E}_{\mathrm{P}} g_{f,r}^2 \le V r^{\vartheta-2}$. Finally, it is not hard to see that $\mathbb{E}_{\mathrm{P}} g_{f,r}^2 \le V r^{\vartheta-2}$ also holds for $\vartheta = 0$. By simple modifications of Lemma 7.6 and its proof, we further see that all families of maps considered below are Carathéodory families. Symmetrization by Proposition 7.10 thus yields

$$\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{f\in\mathcal{F}_r} \bigl|\mathbb{E}_{\mathrm{D}}(\mathbb{E}_{\mathrm{P}} h_{\widehat{f}} - h_{\widehat{f}})\bigr| \le 2\mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r, n) \le 2\varphi_n(r).$$

Peeling by Theorem 7.7 together with $\mathcal{F}_r = \{f \in \mathcal{F} : \Upsilon(f) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}} \le r\}$ hence gives

$$\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{f\in\mathcal{F}} \bigl|\mathbb{E}_{\mathrm{D}} g_{f,r}\bigr| \ \le\ \frac{8\varphi_n(r)}{r}.$$

By Talagrand's inequality in the form of Theorem 7.5 applied to $\gamma := 1/4$, we therefore obtain

$$\mathrm{P}^n\left(D \in (X\times Y)^n : \sup_{f\in\mathcal{F}} \mathbb{E}_{\mathrm{D}} g_{f,r} < \frac{10\varphi_n(r)}{r} + \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{28B\tau}{3nr}\right) \ \ge\ 1 - e^{-\tau}$$

for all $r > r^*$. Using the definition of $g_{f_D,r}$, we thus have with probability P^n not less than $1 - e^{-\tau}$ that

$$\begin{aligned}\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D} &< \bigl(\Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D}\bigr)\left(\frac{10\varphi_n(r)}{r} + \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{28B\tau}{3nr}\right)\\ &\quad + 10\varphi_n(r) + \sqrt{\frac{2V\tau r^\vartheta}{n}} + \frac{28B\tau}{3n}.\end{aligned}$$

By combining this estimate with (7.39) and (7.42), we then obtain that

$$\begin{aligned}\Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} &< \Upsilon(f_0) + 2\mathbb{E}_{\mathrm{P}} h_{f_0} + \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{7B_0\tau}{6n} + \epsilon\\ &\quad + \bigl(\Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D}\bigr)\left(\frac{10\varphi_n(r)}{r} + \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{28B\tau}{3nr}\right)\\ &\quad + 10\varphi_n(r) + \sqrt{\frac{2V\tau r^\vartheta}{n}} + \frac{32B\tau}{3n}\end{aligned} \tag{7.44}$$

holds with probability P^n not less than $1 - 3e^{-\tau}$. Consequently, it remains to bound the various terms. To this end, we first observe that $r \geq 30\varphi_n(r)$ implies $10\varphi_n(r)r^{-1} \leq 1/3$ and $10\varphi_n(r) \leq r/3$. Moreover, $r \geq \left(\frac{72V\tau}{n}\right)^{1/(2-\vartheta)}$ yields

$$\left(\frac{2V\tau}{nr^{2-\vartheta}}\right)^{1/2} \leq \frac{1}{6} \qquad \text{and} \qquad \left(\frac{2V\tau r^{\vartheta}}{n}\right)^{1/2} \leq \frac{r}{6}\,.$$

In addition, $n \geq 72\tau$, $V \geq B^{2-\vartheta}$, and $r \geq \left(\frac{72V\tau}{n}\right)^{1/(2-\vartheta)}$ imply

$$\frac{28B\tau}{3nr} = \frac{7}{54} \cdot \frac{72\tau}{n} \cdot \frac{B}{r} \leq \frac{7}{54} \cdot \left(\frac{72\tau}{n}\right)^{\frac{1}{2-\vartheta}} \cdot \frac{V^{\frac{1}{2-\vartheta}}}{r} \leq \frac{7}{54}$$

and $\frac{32B\tau}{3n} \leq \frac{4r}{27}$. Finally, using $6 \leq 36^{1/(2-\vartheta)}$, we obtain $\left(\frac{2V\tau}{n}\right)^{1/(2-\vartheta)} \leq \frac{r}{6}$. Using these elementary estimates in (7.44), we see that

$$\Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} < \Upsilon(f_0) + 2\mathbb{E}_{\mathrm{P}} h_{f_0} + \frac{7B_0\tau}{6n} + \epsilon + \frac{17}{27}\big(\Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D}\big) + \frac{22r}{27}$$

holds with probability P^n not less than $1 - 3e^{-\tau}$. We now obtain the assertion by some simple algebraic transformations together with $r > \frac{5B_0\tau}{n}$. □

Starting from Theorem 7.20, it is straightforward to recover the ERM oracle inequality (7.28) obtained in the previous section. The details are left to the interested reader.

Before we apply the general oracle inequality above to SVMs, we need the following lemma, which estimates entropy numbers for RKHSs.

Lemma 7.21 (A general entropy bound for RKHSs). *Let k be a kernel on X with RKHS H. Moreover, let $n \geq 2$, $D := (x_1, \ldots, x_n) \in X^n$, and* D *be the associated empirical measure. Then there exists a constant $K \geq 1$ independent of X, H, D, and n such that*

$$\sum_{i=1}^{\infty} 2^{i/2} e_{2^i}(\mathrm{id} : H \to L_2(\mathrm{D})) \leq K\sqrt{\ln n}\, \|k\|_{L_2(\mathrm{D})}\,.$$

Proof. We have $\operatorname{rank}(\mathrm{id} : H \to L_2(\mathrm{D})) \leq \dim L_2(\mathrm{D}) \leq n$, and hence the definition (A.29) of the approximation numbers yields $a_i(\mathrm{id} : H \to L_2(\mathrm{D})) = 0$ for all $i > n$. By Carl's inequality in the form of (A.43), we thus find that

$$\begin{aligned}\sum_{i=0}^{\infty} 2^{i/2} e_{2^i}(\mathrm{id} : H \to L_2(\mathrm{D})) &\leq c_{2,1} \sum_{i=0}^{\log_2 n} 2^{i/2} a_{2^i}(\mathrm{id} : H \to L_2(\mathrm{D})) \\ &\leq c_{2,1}\sqrt{1+\log_2 n}\left(\sum_{i=0}^{\log_2 n} 2^i a_{2^i}^2(\mathrm{id} : H \to L_2(\mathrm{D}))\right)^{\frac{1}{2}},\end{aligned}$$

where in the last step we used Hölder's inequality. Moreover, the sequence (a_i) defined by $a_i := a_i(\mathrm{id} : H \to L_2(\mathrm{D}))$, $i \geq 1$, is decreasing, and hence we have

$$\sum_{i=0}^{\log_2 n} 2^i a_{2^i}^2 \leq 2 \sum_{i=0}^{\infty} 2^{i-1} a_{2^i}^2 \leq 2 \sum_{i=1}^{\infty} \sum_{j=2^{i-1}}^{2^i-1} a_j^2 = 2 \sum_{i=1}^{\infty} a_i^2 \,.$$

Now recall that we have seen in Theorems 4.26 and 4.27 that $\mathrm{id} : H \to L_2(\mathrm{D})$ is the adjoint S_k^* of the Hilbert-Schmidt operator $S_k : L_2(\mathrm{D}) \to H$ defined by (4.17), and hence S_k^* is Hilbert-Schmidt and in particular compact. Moreover, recall (A.29) and the subsequent paragraph, where we have seen that the approximation numbers equal the singular numbers for compact operators acting between Hilbert spaces. With this information, (A.27), and the estimates above, we now find that for $K := 3c_{2,1}$ we have

$$\begin{aligned}
\sum_{i=0}^{\infty} 2^{i/2} e_{2^i}(\mathrm{id} : H \to L_2(\mathrm{D})) &\leq K \sqrt{\log_2 n} \biggl(\sum_{i=1}^{\infty} a_i^2(\mathrm{id} : H \to L_2(\mathrm{D})) \biggr)^{1/2} \\
&= K \sqrt{\log_2 n} \biggl(\sum_{i=1}^{\infty} s_i^2(S_k^* : H \to L_2(\mathrm{D})) \biggr)^{1/2} \\
&= K \sqrt{\log_2 n} \, \|S_k^*\|_{\mathsf{HS}} \,.
\end{aligned}$$

Finally, recall that an operator shares its Hilbert-Schmidt norm with its adjoint, and hence we find $\|S_k^*\|_{\mathsf{HS}} = \|S_k\|_{\mathsf{HS}} = \|k\|_{L_2(\mathrm{D})}$ by Theorem 4.27. □

Let us now formulate our first improved oracle inequality for SVMs, which holds under somewhat minimal assumptions.

Theorem 7.22 (Oracle inequality for SVMs). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a Lipschitz continuous loss with $|L|_1 \leq 1$ that can be clipped at $M > 0$ and that satisfies the supremum bound (7.35) for a constant $B \geq 1$. Furthermore, let P be a distribution on $X \times Y$ and H be a separable RKHS of a bounded measurable kernel k on X with $\|k\|_\infty \leq 1$. Then there exists a constant $K > 0$ such that for all fixed $\tau \geq 1$, $n \geq 2$, and $\lambda > 0$ the SVM associated with L and H satisfies*

$$\lambda \|f_{\mathrm{D},\lambda}\|^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{P}}^* \leq 9A_2(\lambda) + K \frac{\ln n}{\lambda n} + \frac{15\tau}{n} \sqrt{\frac{A_2(\lambda)}{\lambda}} + \frac{300B\tau}{\sqrt{n}}$$

with probability P^n not less than $1 - 3e^{-\tau}$, where $A_2(\,\cdot\,)$ denotes the approximation error function associated with L and H.

Proof. We will use Theorem 7.20 for the regularizer $\Upsilon : H \to [0, \infty)$ defined by $\Upsilon(f) := \lambda \|f\|_H^2$. To this end, we write

$$\mathcal{F}_r := \{f \in H : \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) - \mathcal{R}_{L,\mathrm{P}}^* \leq r\} \,, \tag{7.45}$$

$$\mathcal{H}_r := \{L \circ \widehat{f} - L \circ f_{L,\mathrm{P}}^* : f \in \mathcal{F}_r\} \,. \tag{7.46}$$

From $\lambda \|f\|_H^2 \leq \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) - \mathcal{R}_{L,\mathrm{P}}^*$, we conclude $\mathcal{F}_r \subset (r/\lambda)^{1/2} B_H$, and hence we have

$$e_i(\mathcal{F}_r, L_2(\mathrm{D}_X)) \leq 2(r/\lambda)^{1/2} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X))$$

by (A.36). Consequently, Lemmas 7.17 and 7.21 yield

$$\sum_{i=0}^{\infty} 2^{\frac{i}{2}} e_{2^i}(\mathcal{H}_r, L_2(\mathrm{D})) \leq \sum_{i=0}^{\infty} 2^{\frac{i}{2}} e_{2^i}(\mathcal{F}_r, L_2(\mathrm{D}_X)) \leq K_1 \Big(\frac{r}{\lambda}\Big)^{1/2} \sqrt{\ln n}\,,$$

where $K_1 > 0$ is a universal constant. Therefore, Theorem 7.13 together with Lemma 7.14 shows that

$$\begin{aligned}
\mathrm{Rad}_D(\mathcal{H}_r, n) &\leq \sqrt{\frac{\ln 16}{n}} \Big(\sum_{i=1}^{\infty} 2^{\frac{i}{2}}\, e_{2^i}\big(\mathcal{H}_r \cup \{0\}, L_2(\mathrm{D})\big) + \sup_{h \in \mathcal{H}_r} \|h\|_{L_2(\mathrm{D})} \Big) \\
&\leq \sqrt{\frac{\ln 16}{n}} \Big(2 \sum_{i=0}^{\infty} 2^{\frac{i}{2}}\, e_{2^i}(\mathcal{H}_r, L_2(\mathrm{D})) + B \Big) \\
&\leq K_2 \Big(\frac{r}{\lambda n}\Big)^{1/2} \sqrt{\ln n} + B \sqrt{\frac{\ln 16}{n}}\,, \qquad (7.47)
\end{aligned}$$

where $K_2 > 0$ is another universal constant. Let us denote the right-hand side of (7.47) by $\varphi_n(r)$. Some elementary calculations then show that the condition $r \geq 30\varphi_n(r)$ is fulfilled for

$$r \geq \max\left\{ K_3 \frac{\ln n}{\lambda n},\ \frac{100B}{\sqrt{n}} \right\},$$

where $K_3 > 0$ is another universal constant. Moreover, the variance bound (7.36) is satisfied for $\vartheta := 0$ and $V := B^2$. In other words, the requirement $r \geq (\frac{72V\tau}{n})^{1/(2-\vartheta)}$ is fullfilled for $r \geq 9B\sqrt{\tau/n}$. Consequently, it remains to fix an $f_0 \in H$ and estimate the corresponding B_0. Let us choose $f_0 := f_{\mathrm{P},\lambda}$. Then we have

$$L(x, y, f_{\mathrm{P},\lambda}(x)) \leq L(x, y, 0) + \big| L(x, y, f_{\mathrm{P},\lambda}(x)) - L(x, y, 0) \big| \leq B + \|f_{\mathrm{P},\lambda}\|_\infty$$

for all $(x, y) \in (X \times Y)$, and combining this estimate with $\|f_{\mathrm{P},\lambda}\|_\infty \leq \|f_{\mathrm{P},\lambda}\|_H$ and

$$\lambda \|f_{\mathrm{P},\lambda}\|_H^2 \leq \lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} = A_2(\lambda)$$

yields $\|L \circ f_{\mathrm{P},\lambda}\|_\infty \leq B + \sqrt{\lambda^{-1} A_2(\lambda)} =: B_0$. For

$$r := K_3 \frac{\ln n}{\lambda n} + \frac{100B\tau}{\sqrt{n}} + \frac{5\tau}{n} \sqrt{\frac{A_2(\lambda)}{\lambda}} + r^*\,,$$

where r^* is defined by (7.32), it is then easy to check that (7.38) is satisfied. Applying Theorem 7.20 together with $r^* \leq A_2(\lambda)$ and some elementary calculations then yields the assertion. □

It is interesting to note that Theorem 7.22 as well as the following oracle inequality also hold (in a suitably modified form) for "ϵ-approximate clipped SVMs". We refer to Exercise 7.4 for a precise statement.

Let us now show how we can improve Theorem 7.22 when additional knowledge on the RKHS in terms of entropy numbers is available.

Theorem 7.23 (Oracle inequality for SVMs using benign kernels). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a locally Lipschitz continuous loss that can be clipped at $M > 0$ and satisfies the supremum bound (7.35) for a $B > 0$. Moreover, let H be a separable RKHS of a measurable kernel over X and P be a distribution on $X \times Y$ such that the variance bound (7.36) is satisfied for constants $\vartheta \in [0,1]$, $V \geq B^{2-\vartheta}$, and all $f \in H$. Assume that for fixed $n \geq 1$ there exist constants $p \in (0,1)$ and $a \geq B$ such that*

$$\mathbb{E}_{D_X \sim \mathrm{P}_X^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \;\leq\; a\, i^{-\frac{1}{2p}}\,, \qquad\qquad i \geq 1. \tag{7.48}$$

Finally, fix an $f_0 \in H$ and a constant $B_0 \geq B$ such that $\|L \circ f_0\|_\infty \leq B_0$. Then, for all fixed $\tau > 0$ and $\lambda > 0$, the SVM using H and L satisfies

$$\begin{aligned}\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} &\leq 9\bigl(\lambda\|f_0\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}\bigr) \\ &\quad + K\Bigl(\frac{a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\Bigl(\frac{72V\tau}{n}\Bigr)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}\end{aligned}$$

with probability P^n not less than $1 - 3e^{-\tau}$, where $K \geq 1$ is a constant only depending on p, M, B, ϑ, and V.

Proof. We first note that it suffices to consider the case $a^{2p} \leq \lambda^p n$. Indeed, for $a^{2p} > n\lambda^p$, we have

$$\begin{aligned}\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} &\leq \lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) + B \\ &\leq \mathcal{R}_{L,\mathrm{D}}(0) + B \\ &\leq 2B\Bigl(\frac{a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}}\,.\end{aligned}$$

In other words, the assertion is trivially satisfied for $a^{2p} > n\lambda^p$ whenever $K \geq 2B$. Let us now define r^* by (7.32), and for $r > r^*$ we define $\mathcal{F}_r$ and $\mathcal{H}_r$ by (7.45) and (7.46), respectively. Since $\mathcal{F}_r \subset (r/\lambda)^{1/2} B_H$, Lemma 7.17 together with (7.48) and (A.36) then yields

$$\begin{aligned}\mathbb{E}_{D\sim\mathrm{P}^n} e_i(\mathcal{H}_r, L_2(\mathrm{D})) &\leq |L|_{M,1}\mathbb{E}_{D_X\sim\mathrm{P}_X^n} e_i(\mathcal{F}_r, L_2(\mathrm{D}_X)) \\ &\leq 2|L|_{M,1}\Bigl(\frac{r}{\lambda}\Bigr)^{1/2} a\, i^{-\frac{1}{2p}}\,.\end{aligned}$$

Moreover, for $f \in \mathcal{F}_r$, we have $\mathbb{E}_\mathrm{P}(L\circ\widehat{f} - L\circ f^*_{L,\mathrm{P}})^2 \leq Vr^\vartheta$, and consequently Theorem 7.16 applied to $\mathcal{H} := \mathcal{H}_r$ shows that (7.37) is satisfied for

$$\varphi_n(r) := \max\Bigg\{ C_1(p) 2^p |L|_{M,1}^p a^p \Big(\frac{r}{\lambda}\Big)^{\frac{p}{2}} (V r^{\vartheta})^{\frac{1-p}{2}} n^{-\frac{1}{2}}, \\ C_2(p) \big(2^p |L|_{M,1}^p a^p\big)^{\frac{2}{1+p}} \Big(\frac{r}{\lambda}\Big)^{\frac{p}{1+p}} B^{\frac{1-p}{1+p}} n^{-\frac{1}{1+p}} \Bigg\},$$

where $C_1(p)$ and $C_2(p)$ are the constants appearing in Theorem 7.16. Furthermore, some elementary calculations using $2 - p - \vartheta + \vartheta p \geq 1$ and $a^{2p} \leq \lambda^p n$ show that the condition $r \geq 30\varphi_n(r)$ is satisfied if

$$r \geq \tilde{K} \Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\vartheta+\vartheta p}},$$

where

$$\tilde{K} := \max\Big\{ \big(30 \cdot 2^p C_1(p) |L|_{M,1}^p V^{\frac{1-p}{2}}\big)^{\frac{2}{2-p-\vartheta+\vartheta p}}, 30 \cdot 120^p C_2^{1+p}(p) |L|_{M,1}^{2p} B^{1-p} \Big\}.$$

Using $r^* \leq A_2(\lambda) \leq \lambda \|f_0\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}_{L,\mathrm{P}}^*$, the assertion thus follows from Theorem 7.20 for $K := \max\{3\tilde{K}, 2B\}$. □

Let us now compare the oracle inequalities for SVMs obtained in this section with the simple oracle inequalities derived in Theorems 6.25 and 6.24. To this end, let us briefly recall the assumptions made at the end of Section 6.4, where we established the first consistency results and learning rates for SVMs: in the following, L is a Lipschitz continuous loss with $L(x, y, 0) \leq 1$ for all $(x, y) \in X \times Y$, and $|L|_1 \leq 1$. Moreover, H is a fixed separable RKHS over X having a bounded measurable kernel with $\|k\|_\infty \leq 1$. We assume that H is dense in $L_1(\mu)$ for all distributions μ on X, so that by Theorem 5.31 we have $\mathcal{R}_{L,\mathrm{P},H}^* = \mathcal{R}_{L,\mathrm{P}}^*$ for all probability measures P on $X \times Y$. We will also need the entropy number assumption

$$e_i(\mathrm{id} : H \to \ell_\infty(X)) \;\leq\; a\, i^{-\frac{1}{2p}}, \qquad\qquad i \geq 1, \tag{7.49}$$

where $a \geq 1$ and $p \in (0, 1)$ are fixed constants. Recall that by Lemma 6.21 and Exercise 6.8 this assumption is essentially equivalent to the covering number assumption (6.20). Finally, we assume in the following comparison that (λ_n) is an *a priori*[5] chosen sequence of strictly positive numbers converging to zero.

Our first goal is to consider conditions on (λ_n) ensuring universal L-risk consistency. To this end, let us first recall (see Exercise 6.9) that the oracle inequality of Theorem 6.24 ensured universal consistency if $\lambda_n^2 n \to \infty$, whereas the inequality of Theorem 6.25 ensured this for the weaker condition $\lambda_n^{1+p} n \to \infty$, where p is the exponent appearing in (7.49). Remarkably, Theorem 7.22 guarantees universal consistency for the even milder condition $\lambda_n n / \ln n \to \infty$ *without* using an entropy number assumption. Finally, if such an assumption in the form of (7.49) is satisfied, then Theorem 7.23 ensures universal consistency

[5] Data-dependent choices for λ will be discussed after the comparison.

whenever $\lambda_n^{\max\{1/2,p\}} n \to \infty$. Obviously, this is the weakest condition of the four listed. In addition, Theorem 7.23 actually allows us to replace the entropy number condition (7.49) by (7.48), which in some cases is substantially weaker, as we will see in Section 7.5.

Let us now compare the learning rates we can derive from the four oracle inequalities. To this end, we assume, as usual, that there exist constants $c > 0$ and $\beta \in (0, 1]$ such that $A_2(\lambda) \leq c\lambda^\beta$ for all $\lambda \geq 0$. Under this assumption, we have seen in Exercise 6.9 that Theorem 6.24 leads to the learning rate

$$n^{-\frac{\beta}{2\beta+2}} . \tag{7.50}$$

Moreover, at the end of Section 6.4, we derived the learning rate

$$n^{-\frac{\beta}{(2\beta+1)(1+p)}} \tag{7.51}$$

from Theorem 6.25, where p is the exponent from (7.49). On the other hand, optimizing the oracle inequality of Theorem 7.22 with the help of Lemma A.1.7 yields the learning rate

$$n^{-\frac{\beta}{\beta+1}} \ln n , \tag{7.52}$$

which is better than (7.50) by a factor of two in the exponent. Moreover, it is even better than (7.51), which was derived under more restrictive assumptions than (7.52). Finally, by assuming (7.49), or the potentially weaker condition (7.48), Theorem 7.23 applied with $f_0 := f_{\mathrm{P},\lambda}$ and $B_0 := B + \sqrt{\lambda^{-1} A_2(\lambda)}$ together with Lemma A.1.7 provides the learning rate

$$n^{-\min\{\frac{2\beta}{\beta+1}, \frac{\beta}{\beta(2-p-\vartheta+\vartheta p)+p}\}} , \tag{7.53}$$

which reduces to $n^{-\min\{\frac{2\beta}{\beta+1}, \frac{\beta}{\beta(2-p)+p}\}}$ if no variance bound assumption, i.e., $\vartheta = 0$ is made. It is simple to check that the latter is even faster than (7.52).

So far, we have only considered Lipschitz continuous losses; however, Theorem 7.23 also holds for losses that are only *locally* Lipschitz continuous. Let us now briefly describe how the rates above change when considering such losses. For the sake of simplicity we only consider the least squares loss, $Y \subset [-1, 1]$, and the choice $f_0 := f_{\mathrm{P},\lambda}$. Then, because of the growth behavior of the least squares loss, we can only choose $B_0 := 2 + 2\lambda^{-1} A_2(\lambda)$, which, by Theorem 7.23 and Example 7.3, leads to the rate

$$n^{-\min\{\beta, \frac{\beta}{\beta+p}\}} . \tag{7.54}$$

It is easy to see that the learning rates (7.52) and (7.53) are achieved for regularization sequences of the form $\lambda_n := n^{-\rho/\beta}$, where ρ is the exponent in the corresponding learning rate. To achieve these learning rates with an *a priori* chosen sequence, we therefore need to know the value of β and for (7.53) also the values of ϑ and p. Since this is unrealistic, we now demonstrate that

the TV-SVM defined in Definition 6.28 is adaptive to these parameters by choosing the regularization parameter in a data-dependent way. For brevity's sake, we focus on the situation of Theorem 7.23; however, it is straightforward to show a similar result for the more general situation of Theorem 7.22.

Theorem 7.24 (Oracle inequality for TV-SVMs and benign kernels). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a Lipschitz continuous loss with $|L|_1 \leq 1$ that can be clipped at $M > 0$ and satisfies the supremum bound (7.35) for a $B \geq 1$. Moreover, let H be a separable RKHS of a measurable kernel over X with $\|k\|_\infty \leq 1$ and let P be a distribution on $X \times Y$ such that the variance bound (7.36) is satisfied for constants $\vartheta \in [0,1]$, $V \geq B^{2-\vartheta}$, and all $f \in H$. We further assume that H is dense in $L_1(\mathrm{P}_X)$. In addition, let $p \in (0,1)$ and $a \geq B$ be constants such that*

$$\mathbb{E}_{D_X \sim \mathrm{P}_X^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \;\leq\; a\, i^{-\frac{1}{2p}}\,, \qquad n \geq 1,\, i \geq 1.$$

Moreover, for $n \geq 4$ and $\varepsilon > 0$, let $\Lambda_n \subset (0,1]$ be a finite ε-net of $(0,1]$ with cardinality $|\Lambda_n|$. For fixed $\tau \geq 1$, we then have for every corresponding, measurable TV-SVM with probability P^n not less than, $1 - e^{-\tau}$ that

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P}} < 6 \inf_{\lambda \in (0,1]} \biggl(9A_2(\lambda) + K\Bigl(\frac{2a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{30\tau_n}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}} \biggr) \\ + 6A_2(2\varepsilon) + 45\Bigl(\frac{64V\tau_n}{n}\Bigr)^{\frac{1}{2-\vartheta}} + \frac{180B\tau_n}{n},\end{aligned}$$

where D_1 and D_2 are the training and validation set built from $D \in (X \times Y)^n$, K is the constant appearing in Theorem 7.23, and $\tau_n := \tau + \ln(1 + 3|\Lambda_n|)$.

Consequently, if we use ε_n-nets Λ_n with $\varepsilon_n \to 0$ and $n^{-1}\ln(|\Lambda_n|) \to 0$, the resulting TV-SVM is consistent. Moreover, for $\varepsilon_n \leq 1/n^2$ and $|\Lambda_n|$ growing polynomially in n, the TV-SVM learns with rate (7.53) if there exist constants $c > 0$ and $\beta \in (0,1]$ such that $A_2(\lambda) \leq c\lambda^\beta$ for all $\lambda \geq 0$.

Note that it is easy to derive modifications of Theorem 7.24 that consider locally Lipschitz continuous loss functions. For example, for the least squares loss we only have to replace the term $\sqrt{\lambda^{-1}A_2(\lambda)}$ by $2\lambda^{-1}A_2(\lambda)$ in the oracle inequality. Moreover, for this loss the conditions ensuring consistency remain unchanged and the resulting learning rate becomes (7.54).

Proof. For $f_0 := f_{\mathrm{P},\lambda}$, we have already seen in the proof of Theorem 7.22 that $\|L \circ f_0\|_\infty \leq B + \sqrt{\lambda^{-1}A_2(\lambda)}$. Since $m := \lfloor n/2 \rfloor + 1$ implies $m \geq n/2$, we hence see by Theorem 7.23 that with probability P^m not less than $1 - 3|\Lambda_n|e^{-\tau}$ we have

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} \leq 9A_2(\lambda) + K\Bigl(\frac{2a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{30\tau}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}} \\ + 3\Bigl(\frac{144V\tau}{n}\Bigr)^{\frac{1}{2-\vartheta}} + \frac{30B\tau}{n}\end{aligned}$$

for all $\lambda \in \Lambda_n$ *simultaneously*. Moreover, $n \geq 4$ implies $n - m \geq n/2 - 1 \geq n/4$, and therefore Theorem 7.2 shows that, for *fixed* $D_1 \in (X \times Y)^m$ and $\tilde{\tau}_n := \tau + \ln(1 + |\Lambda_n|)$, the probability P^{n-m} of having a $D_2 \in (X \times Y)^{n-m}$ such that

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P}} \leq 6 \inf_{\lambda \in \Lambda_n} \bigl(\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda}) - \mathcal{R}^*_{L,\mathrm{P}}\bigr) + 4\biggl(\frac{32V\tilde{\tau}_n}{n}\biggr)^{\frac{1}{2-\vartheta}}$$

is not less than $1 - e^{-\tau}$. Combining both estimates, we conclude that with probability P^n not less than $(1 - 3|\Lambda_n|e^{-\tau})(1 - e^{-\tau})$ we have

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P}} &\leq 6 \inf_{\lambda \in \Lambda_n} \biggl(9A_2(\lambda) + K\Bigl(\frac{2a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{30\tau}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}}\biggr) \\ &\quad + 41\Bigl(\frac{64V\tau}{n}\Bigr)^{\frac{1}{2-\vartheta}} + \frac{180B\tau}{n} + 4\biggl(\frac{32V\tilde{\tau}_n}{n}\biggr)^{\frac{1}{2-\vartheta}}.\end{aligned}$$

Now recall that $\lambda \mapsto \lambda^{-1}A_2(\lambda)$ is decreasing by Lemma 5.15, and hence an almost literal repetition of the proof of Lemma 6.30 yields

$$\begin{aligned}&\inf_{\lambda \in \Lambda_n} \biggl(9A_2(\lambda) + K\Bigl(\frac{2a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{30\tau}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}}\biggr) \\ &\leq A_2(2\varepsilon) + \inf_{\lambda \in (0,1]} \biggl(9A_2(\lambda) + K\Bigl(\frac{2a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{30\tau}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}}\biggr).\end{aligned}$$

Using $(1 - 3|\Lambda_n|e^{-\tau})(1 - e^{-\tau}) \geq 1 - (1 + 3|\Lambda_n|)e^{-\tau}$ together with a variable transformation that adjusts the probability P^n to be not less than $1 - e^{-\tau}$ and some simple estimates then yields the asserted oracle inequality. The other assertions are direct consequences of this oracle inequality. □

7.5 Some Bounds on Average Entropy Numbers

The goal of this section is to illustrate that the entropy assumption (7.48) used in Theorems 7.23 and 7.24 is weaker than the entropy bound (7.49) used in the simple analysis of Section 6.4. To this end, we first relate the average entropy numbers in (7.48) to the eigenvalues of the integral operator defined by the kernel of the RKHS used.

Let us begin by establishing some preparatory lemmas. To this end, we fix a Hilbert space H. For $v, w \in H$, we define $v \otimes w : H \to H$ by $v \otimes w(u) := \langle u, v\rangle w$, $u \in H$. Obviously, $v \otimes w$ is bounded and linear with $\operatorname{rank} v \otimes w \leq 1$, and hence it is a Hilbert-Schmidt operator, i.e., $v \otimes w \in \mathsf{HS}(H)$. Since in the following we need various facts on Hilbert-Schmidt operators, we encourage the reader to review the corresponding material presented at the end of Section A.5.2.

Lemma 7.25 (Feature maps of squared kernels). *Let H be an RKHS over X with kernel k and canonical feature map $\Phi : X \to H$. Then $\Psi : X \to \mathsf{HS}(H)$ defined by $\Psi(x) := \Phi(x) \otimes \Phi(x)$, $x \in X$, is a feature map of the squared kernel $k^2 : X \times X \to \mathbb{R}$, and we have $\|\Psi(x)\|_{\mathsf{HS}} = \|\Phi(x)\|_H^2$ for all $x \in X$.*

Proof. For $f \in H$ and $x \in X$, the reproducing property yields

$$\Psi(x)f := \Psi(x)(f) = \langle f, \Phi(x)\rangle_H \, \Phi(x) = f(x)\Phi(x)\,. \tag{7.55}$$

Let $(e_i)_{i\in I}$ be an ONB of H. By the definition (A.28) of the inner product of $\mathsf{HS}(H)$ and (4.9), we then obtain

$$\begin{aligned}\langle \Psi(x), \Psi(x')\rangle_{\mathsf{HS}(H)} = \sum_{i\in I} \langle \Psi(x)e_i, \Psi(x')e_i\rangle_H &= \sum_{i\in I} \langle e_i(x)\Phi(x), e_i(x')\Phi(x')\rangle_H \\ &= \langle \Phi(x), \Phi(x')\rangle_H \sum_{i\in I} e_i(x)e_i(x') \\ &= k^2(x, x') \end{aligned} \tag{7.56}$$

for all $x, x' \in X$, i.e., we have shown the first assertion. The second assertion follows from (7.56) and $\|\Psi(x)\|_{\mathsf{HS}} = \sqrt{k^2(x,x)} = k(x,x) = \|\Phi(x)\|_H^2$. □

Let us now assume that k is a measurable kernel on X with separable RKHS H and μ is a probability measure on X with $\|k\|_{L_2(\mu)} < \infty$. In the following, we write $S_{k,\mu} : L_2(\mu) \to H$ for the integral operator defined in (4.17) in order to emphasize that this operator depends not only on k but also on μ. Analogously, we write $T_{k,\mu} := S_{k,\mu}^* \circ S_{k,\mu} : L_2(\mu) \to L_2(\mu)$ for the integral operator considered in Theorem 4.27. Finally, we also need the **covariance operator** $C_{k,\mu} := S_{k,\mu} \circ S_{k,\mu}^* : H \to H$. The next lemma relates $C_{k,\mu}$ to the feature map Ψ of k^2.

Lemma 7.26. *Let k be a measurable kernel on X with separable RKHS H and ν be a probability measure on X such that $\|k\|_{L_2(\nu)} < \infty$. Then $\Psi : X \to \mathsf{HS}(H)$ defined in Lemma 7.25 is Bochner ν-integrable and $\mathbb{E}_\nu \Psi = C_{k,\nu}$.*

Proof. Obviously, Ψ is measurable, and since (7.56) implies

$$\mathbb{E}_{x\sim\nu} \|\Psi(x)\|_{\mathsf{HS}} = \mathbb{E}_{x\sim\nu} k(x,x) = \|k\|_{L_2(\nu)} < \infty\,,$$

we see that Ψ is Bochner integrable with respect to ν. Let us now fix an $f \in H$. Since by Theorem 4.26 the operator $S_{k,\nu}^*$ equals $\mathrm{id} : H \to L_2(\nu)$, we then have $C_{k,\nu} f = S_{k,\nu} \circ S_{k,\nu}^* f = \mathbb{E}_\nu f\Phi$. In addition, considering the bounded linear operator $\delta_f : \mathsf{HS}(H) \to H$ defined by $\delta_f(S) := Sf$, $S \in \mathsf{HS}(H)$, we find by (A.32) and (7.55) that

$$(\mathbb{E}_\nu \Psi)(f) = \delta_f\big(\mathbb{E}_{x\sim\nu}\Psi(x)\big) = \mathbb{E}_{x\sim\nu}\delta_f(\Psi(x)) = \mathbb{E}_{x\sim\nu}\big(\Psi(x)(f)\big) = \mathbb{E}_\nu f\Phi\,.$$

By combining our considerations, we then obtain $\mathbb{E}_\nu \Psi = C_{k,\nu}$. □

Before we can state the first main result of this section, we finally need the following two technical lemmas.

Lemma 7.27. *Let $(\alpha_i) \subset [0,1]$ be a sequence such that $\sum_{i=1}^{\infty} \alpha_i = m$ for some $m \in \mathbb{N}$ and $(\lambda_i) \subset [0,\infty)$ be a decreasing sequence. Then we have*

$$\sum_{i=1}^{\infty} \alpha_i \lambda_i \leq \sum_{i=1}^{m} \lambda_i \,.$$

Proof. Let us define $\gamma_i := \lambda_i - \lambda_m$, $i \geq 1$. Then we have $\gamma_i \geq 0$ if $i \leq m$ and $\gamma_i \leq 0$ if $i \geq m$, and consequently we obtain

$$\sum_{i=1}^{\infty} \alpha_i \gamma_i \leq \sum_{i=1}^{m} \alpha_i \gamma_i \leq \sum_{i=1}^{m} \gamma_i \,.$$

Moreover, we have

$$\sum_{i=1}^{\infty} \alpha_i \lambda_i - m\lambda_m = \sum_{i=1}^{\infty} \alpha_i \lambda_i - \sum_{i=1}^{\infty} \alpha_i \lambda_m = \sum_{i=1}^{\infty} \alpha_i \gamma_i$$

and

$$\sum_{i=1}^{m} \gamma_i = \sum_{i=1}^{m} \lambda_i - \sum_{i=1}^{m} \lambda_m = \sum_{i=1}^{m} \lambda_i - m\lambda_m \,,$$

and by combining all formulas, we then obtain the assertion. □

Given a compact and self-adjoint operator $T \in \mathcal{L}(H)$, we now investigate the set of non-zero eigenvalues $\{\lambda_i(T) : i \in I\}$ considered in the Spectral Theorem A.5.13. Following the notation of this theorem, we assume in the following that either $I = \{1, 2, \ldots, n\}$ or $I = \mathbb{N}$. In order to unify our considerations, we further need the **extended sequence of eigenvalues** of T, which in the case of finite I is the sequence $\lambda_1(T), \lambda_2(T), \ldots, \lambda_{|I|}(T), 0, \ldots$, and in the case of infinite I is the sequence $(\lambda_i(T))_{i \geq 1}$. With these notations, we can now establish the last preparatory lemma.

Lemma 7.28. *Let $T : H \to H$ be a compact, positive, and self-adjoint operator on a separable Hilbert space H and $(\lambda_i(T))_{i \geq 1}$ be its extended sequence of eigenvalues. Then, for all $m \geq 1$ and $\bar{m} := \min\{m, \dim H\}$, we have*

$$\sum_{i=1}^{m} \lambda_i(T) \;=\; \max_{\substack{V \text{subspace of } H \\ \dim V = \bar{m}}} \langle P_V, T \rangle_{\mathsf{HS}} \,.$$

Proof. Let $(e_i)_{i \in J}$ be an ONB of H with $I \subset J$ such that the subfamily $(e_i)_{i \in I}$ is the ONS provided by Theorem A.5.13. In addition, we assume for notational convenience that $J = \{1, \ldots, \dim H\}$ if $\dim H < \infty$ and $J = \mathbb{N}$ if $\dim H = \infty$ and $|I| < \infty$. Note that in both cases we have $\lambda_i(T) = 0$ for all

$i \in J\backslash I$ since we consider extended sequences of eigenvalues. Moreover, in the case $\dim H = |I| = \infty$, we define $\lambda_i(T) := 0$ for $i \in J \setminus I$. Let us now write $V_0 := \operatorname{span}\{e_1, \ldots, e_{\bar{m}}\}$, where we note that our notational assumptions ensure $\{1, \ldots, \bar{m}\} \subset J$. Obviously, the orthogonal projection $P_{V_0} : H \to H$ onto V_0 satisfies $P_{V_0} e_i = e_i$ if $i \in \{1, \ldots, \bar{m}\}$ and $P_{V_0} e_i = 0$ otherwise. Moreover, we have $Te_i = \lambda_i(T) e_i$ for all $i \in I$, and the spectral representation (A.23) shows $Te_i = 0$ for all $i \in J\backslash I$. Since our notational assumptions ensured $\lambda_i(T) = 0$ for all $i \in J \setminus I$, we thus find $Te_i = \lambda_i(T) e_i$ for all $i \in J$. The definition (A.28) of the inner product in $\mathsf{HS}(H)$ together with $\lambda_i(T) = 0$ for $i \in \mathbb{N}\backslash\{1, \ldots, \dim H\}$ hence yields

$$\sum_{i=1}^{m} \lambda_i(T) = \sum_{i=1}^{\bar{m}} \lambda_i(T) = \sum_{i \in J} \langle P_{V_0} e_i, Te_i \rangle_H = \langle P_{V_0}, T \rangle_{\mathsf{HS}} \leq \max_{\substack{V \text{subspace of } H \\ \dim V = \bar{m}}} \langle P_V, T \rangle_{\mathsf{HS}} \,.$$

Conversely, if V is an arbitrary subspace of H with $\dim V = \bar{m}$, the self-adjointness of the orthogonal projection P_V onto V shows that for $\alpha_i := \langle P_V e_i, e_i \rangle_H = \langle P_V^2 e_i, e_i \rangle_H$ we have

$$\alpha_i = \langle P_V^2 e_i, e_i \rangle_H = \langle P_V^* \circ P_V e_i, e_i \rangle_H = \langle P_V e_i, P_V e_i \rangle_H = \|P_V e_i\|_H^2 \in [0,1]$$

for all $i \in J$. Let us write $\alpha_i := 0$ if $\dim H < \infty$ and $i > \dim H$. Then we observe

$$\sum_{i=1}^{\infty} \alpha_i = \sum_{i \in J} \|P_V e_i\|_H^2 = \|P_V\|_{\mathsf{HS}}^2 = \bar{m} \,.$$

From this, Lemma 7.27, and $Te_i = \lambda_i(T) e_i$ for all $i \in J$, we conclude that

$$\sum_{i=1}^{m} \lambda_i(T) = \sum_{i=1}^{\bar{m}} \lambda_i(T) \geq \sum_{i=1}^{\infty} \alpha_i \lambda_i(T) = \sum_{i \in J} \langle P_V e_i, Te_i \rangle_H = \langle P_V, T \rangle_{\mathsf{HS}} \,,$$

and hence we have shown the assertion. □

Let us now formulate our first main result of this section, which relates the average eigenvalues of the empirical integral operators $T_{k,\mathrm{D}}$, $D \in X^n$, with the eigenvalues of the infinite-sample integral operator $T_{k,\mu}$.

Theorem 7.29 (Eigenvalues of empirical integral operators). *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then we have*

$$\mathbb{E}_{D \sim \mu^n} \sum_{i=1}^{\infty} \lambda_i(T_{k,\mathrm{D}}) = \sum_{i=1}^{\infty} \lambda_i(T_{k,\mu}) < \infty \tag{7.57}$$

for the extended sequences of eigenvalues. Furthermore, for all $m \geq 1$, these extended sequences of eigenvalues satisfy

$$\mathbb{E}_{D\sim\mu^n}\sum_{i=m}^{\infty}\lambda_i(T_{k,\mathrm{D}}) \le \sum_{i=m}^{\infty}\lambda_i(T_{k,\mu}) \tag{7.58}$$

and

$$\mathbb{E}_{D\sim\mu^n}\sum_{i=1}^{m}\lambda_i(T_{k,\mathrm{D}}) \ge \sum_{i=1}^{m}\lambda_i(T_{k,\mu})\,. \tag{7.59}$$

Proof. Let us begin with some preliminary considerations. To this end, let $(e_j)_{j\in J}$ be an ONB of H and ν be an arbitrary probability measure on X with $\|k\|_{L_2(\nu)} < \infty$. Recall that we have seen in front of the Spectral Theorem A.5.13 that the integral operator $T_{k,\nu} = S^*_{k,\nu}\circ S_{k,\nu}$ and the covariance operator $C_{k,\nu} = S_{k,\nu}\circ S^*_{k,\nu}$ have exactly the same non-zero eigenvalues with the same geometric multiplicities. Consequently, their extended sequences of eigenvalues coincide. Moreover, by (A.25), (A.27), (A.26), and Theorem 4.26, we find

$$\sum_{i=1}^{\infty}\lambda_i(T_{k,\nu}) = \sum_{i=1}^{\infty}s_i^2(S^*_{k,\nu}) = \|S^*_{k,\nu}\|^2_{\mathsf{HS}} = \sum_{j\in J}\|S^*_{k,\nu}e_j\|^2_{L_2(\nu)} = \sum_{j\in J}\mathbb{E}_\nu e_j^2\,,$$

where we note that $\|S^*_{k,\nu}\|_{\mathsf{HS}} = \|S_{k,\nu}\|_{\mathsf{HS}} < \infty$ by Theorem 4.27. Applying the equality above twice, we now obtain

$$\mathbb{E}_{D\sim\mu^n}\sum_{i=1}^{\infty}\lambda_i(T_{k,\mathrm{D}}) = \sum_{j\in J}\mathbb{E}_{D\sim\mu^n}\mathbb{E}_{\mathrm{D}}e_j^2 = \sum_{j\in J}\mathbb{E}_\mu e_j^2 = \sum_{i=1}^{\infty}\lambda_i(T_{k,\mu}),$$

i.e., we have shown (7.57). In order to prove (7.59), we first observe that Lemma 7.26 together with Lemma 7.28 and $\lambda_i(T_{k,\nu}) = \lambda_i(C_{k,\nu})$ for all $i \ge 1$ implies

$$\sum_{i=1}^{m}\lambda_i(T_{k,\nu}) = \max_{\substack{V\text{subspace of } H\\ \dim V=\bar m}}\langle P_V, C_{k,\nu}\rangle_{\mathsf{HS}} = \max_{\substack{V\text{subspace of } H\\ \dim V=\bar m}}\mathbb{E}_\nu\langle P_V, \Psi\rangle_{\mathsf{HS}}\,,$$

where again ν is an arbitrary probability measure on X with $\|k\|_{L_2(\nu)} < \infty$ and $\bar m := \min\{m, \dim H\}$. From this we conclude that

$$\begin{aligned}
\mathbb{E}_{D\sim\mu^n}\sum_{i=1}^{m}\lambda_i(T_{k,\mathrm{D}}) &= \mathbb{E}_{D\sim\mu^n}\max_{\substack{V\text{subspace of } H\\ \dim V=\bar m}}\mathbb{E}_{\mathrm{D}}\langle P_V, \Psi\rangle_{\mathsf{HS}}\\
&\ge \max_{\substack{V\text{subspace of } H\\ \dim V=\bar m}}\mathbb{E}_{D\sim\mu^n}\mathbb{E}_{\mathrm{D}}\langle P_V, \Psi\rangle_{\mathsf{HS}}\\
&= \max_{\substack{V\text{subspace of } H\\ \dim V=\bar m}}\mathbb{E}_{\mu}\langle P_V, \Psi\rangle_{\mathsf{HS}}\\
&= \sum_{i=1}^{m}\lambda_i(T_{k,\mu})\,.
\end{aligned}$$

Finally, (7.58) is a direct consequence of (7.57) and (7.59). □

Let us now recall that we have seen in Sections A.5.2 and A.5.6 that there is an intimate relationship between eigenvalues, singular numbers, approximation numbers, and entropy numbers. This relationship together with the preceding theorem leads to the second main result of this section.

Theorem 7.30 (Average entropy numbers of RKHSs). *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then for all $0 < p < 1$ there exists a constant $c_p \geq 1$ only depending on p such that for all $n \geq 1$ and $m \geq 1$ we have*

$$\mathbb{E}_{D\sim\mu^n} e_m(S^*_{k,\mathrm{D}}) \leq c_p\, m^{-1/p} \sum_{i=1}^{\min\{m,n\}} i^{1/p-1} \left(\frac{1}{i} \sum_{j=i}^{\infty} e_j^2(S^*_{k,\mu}) \right)^{1/2}.$$

Proof. For $q := 1$, Carl's inequality (A.42) shows that there exists a constant $c_p > 0$ such that for $m, n \geq 1$ and all $D \in X^n$ we have

$$\sum_{i=1}^{m} i^{1/p-1} e_i(S^*_{k,\mathrm{D}}) \leq c_p \sum_{i=1}^{m} i^{1/p-1} a_i(S^*_{k,\mathrm{D}}) = c_p \sum_{i=1}^{\min\{m,n\}} i^{1/p-1} a_i(S^*_{k,\mathrm{D}}),$$

where in the last step we used that $n \geq \operatorname{rank} S^*_{k,D}$ implies $a_i(S^*_{k,\mathrm{D}}) = 0$ for all $i > n$. Moreover, for $M := \min\{m, n\}$ and $\tilde{M} := \lfloor (M+1)/2 \rfloor$, we have

$$\begin{aligned}\sum_{i=1}^{M} i^{1/p-1} a_i(S^*_{k,\mathrm{D}}) &\leq \sum_{i=1}^{\tilde{M}} (2i-1)^{1/p-1} a_{2i-1}(S^*_{k,\mathrm{D}}) + \sum_{i=1}^{\tilde{M}} (2i)^{1/p-1} a_{2i}(S^*_{k,\mathrm{D}}) \\ &\leq 2^{1/p} \sum_{i=1}^{M} i^{1/p-1} a_{2i-1}(S^*_{k,\mathrm{D}}).\end{aligned}$$

Now recall from the end of Section A.5.2 that $a_i^2(S^*_{k,\mathrm{D}}) = s_i^2(S^*_{k,\mathrm{D}}) = s_i(S^*_{k,\mathrm{D}} S_{k,\mathrm{D}}) = \lambda_i(T_{k,\mathrm{D}})$ for all $i \geq 1$ and $D \in X^n$, and hence we obtain

$$\begin{aligned}\sum_{i=1}^{m} i^{1/p-1} \mathbb{E}_{D\sim\mu^n} e_i(S^*_{k,\mathrm{D}}) &\leq 2^{1/p} c_p \sum_{i=1}^{M} i^{1/p-1} \mathbb{E}_{D\sim\mu^n} a_{2i-1}(S^*_{k,\mathrm{D}}) \\ &\leq 2^{1/p} c_p \sum_{i=1}^{M} i^{1/p-1} \big(\mathbb{E}_{D\sim\mu^n} \lambda_{2i-1}(T_{k,\mathrm{D}}) \big)^{1/2}.\end{aligned}$$

Now for each $D \in X^n$ the sequence $(\lambda_i(T_{k,\mathrm{D}}))_{i\geq 1}$ is monotonically decreasing and hence so is $(\mathbb{E}_{D\sim\mu^n} \lambda_i(T_{k,\mathrm{D}}))_{i\geq 1}$. By Theorem 7.29, we hence find

$$i\, \mathbb{E}_{D\sim\mu^n} \lambda_{2i-1}(T_{k,\mathrm{D}}) \leq \sum_{j=i}^{2i-1} \mathbb{E}_{D\sim\mu^n} \lambda_j(T_{k,\mathrm{D}}) \leq \sum_{j=i}^{\infty} \lambda_j(T_{k,\mu})$$

for all $i \geq 1$, and consequently we obtain

$$\sum_{i=1}^{M} i^{1/p-1} \bigl(\mathbb{E}_{D\sim\mu^n} \lambda_{2i-1}(T_{k,\mathrm{D}})\bigr)^{1/2} \le \sum_{i=1}^{M} i^{1/p-1} \left(\frac{1}{i} \sum_{j=i}^{\infty} \lambda_j(T_{k,\mu})\right)^{1/2}.$$

Moreover, we have $\lambda_j(T_{k,\mu}) = s_i^2(S_{k,\mu}^*) = a_j^2(S_{k,\mu}^*) \le 4e_j^2(S_{k,\mu}^*)$, where in the last step we used (A.44). Combining the estimates above, we hence obtain

$$\sum_{i=1}^{m} i^{1/p-1} \mathbb{E}_{D\sim\mu^n} e_i(S_{k,\mathrm{D}}^*) \le 2^{1/p+1} c_p \sum_{i=1}^{M} i^{1/p-1} \left(\frac{1}{i} \sum_{j=i}^{\infty} e_j^2(S_{k,\mu}^*)\right)^{1/2}.$$

Finally, for $\tilde{m} := \lfloor (m+1)/2 \rfloor$, the monotonicity of the entropy numbers yields

$$\tilde{m}^{\frac{1}{p}} \, \mathbb{E}_{D\sim\mu^n} e_m(S_{k,\mathrm{D}}^*) \le \sum_{i=\tilde{m}}^{m} i^{\frac{1}{p}-1} \, \mathbb{E}_{D\sim\mu^n} e_i(S_{k,\mathrm{D}}^*) \le \sum_{i=1}^{m} i^{\frac{1}{p}-1} \, \mathbb{E}_{D\sim\mu^n} e_i(S_{k,\mathrm{D}}^*),$$

and since $m/2 \le \lfloor (m+1)/2 \rfloor = \tilde{m}$, we hence obtain the assertion. □

With the help of Theorem 7.30, we can now formulate the following condition, which ensures the average entropy number assumption (7.48).

Corollary 7.31. *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Assume that there exist constants $0 < p < 1$ and $a \ge 1$ such that*

$$e_i(\mathrm{id} : H \to L_2(\mu)) \le a \, i^{-\frac{1}{2p}}, \qquad i \ge 1. \tag{7.60}$$

Then there exists a constant $c_p > 0$ only depending on p such that

$$\mathbb{E}_{D\sim\mu^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D})) \le c_p \, a \, \bigl(\min\{i,n\}\bigr)^{\frac{1}{2p}} \, i^{-\frac{1}{p}}, \qquad i,n \ge 1.$$

Proof. By Theorem 4.26, the operator $S_{k,\nu}^*$ coincides with $\mathrm{id} : H \to L_2(\nu)$ for all distributions ν with $\|k\|_{L_2(\mu)} < \infty$. Since $0 < p < 1$, it is then easy to see that there exists a constant $\tilde{c}_p$ such that

$$\frac{1}{i} \sum_{j=i}^{\infty} e_j^2(S_{k,\mu}^*) \le a^2 \cdot \frac{1}{i} \sum_{j=i}^{\infty} j^{-\frac{1}{p}} \le \tilde{c}_p^2 \, a^2 \, i^{-\frac{1}{p}}, \qquad i \ge 1.$$

Using Theorem 7.30 and $\frac{1}{2p} - 1 > -1$, we hence find a constant $c_p' > 0$ such that

$$\mathbb{E}_{D\sim\mu^n} e_m(S_{k,\mathrm{D}}^*) \le c_p \tilde{c}_p \, a \, m^{-\frac{1}{p}} \sum_{i=1}^{\min\{m,n\}} i^{\frac{1}{2p}-1} \le c_p' \, a \, \bigl(\min\{m,n\}\bigr)^{\frac{1}{2p}} \, m^{-\frac{1}{p}}. \quad \square$$

Considering the proofs of Theorem 7.30 and Corollary 7.31, it is not hard to see that we can replace the entropy bound (7.60) by a bound on the *eigenvalues* of $T_{K,\mu}$. We refer to Exercise 7.7 for details.

Let us now return to the main goal of this section, which is to illustrate that the average entropy assumption (7.48) is weaker than the uniform entropy bound (7.49). To this end, we need the following definition.

Definition 7.32. *We say that a distribution μ on $\mathbb{R}^d$ has* ***tail exponent*** $\tau \in [0,\infty]$ *if*

$$\mu\bigl(\mathbb{R}^d \backslash r B_{\ell_2^d}\bigr) \;\le\; r^{-\tau}\,, \qquad\qquad r > 0. \tag{7.61}$$

Obviously, every distribution has tail exponent $\tau = 0$. On the other hand, a distribution has tail exponent $\tau = \infty$ if and only if the support of μ is contained in the closed unit ball $B_{\ell_2^d}$. Finally, the intermediate case $\tau \in (0,\infty)$ describes how much mass μ has *outside* the scaled Euclidean balls $rB_{\ell_2^d}$, $r > 1$.

Before we return to entropy numbers, we need another notation. To this end, let μ be a distribution on $\mathbb{R}^d$ and $X \subset \mathbb{R}^d$ be a measurable set with $\mu(X) > 0$. Then we define the distribution μ_X on $\mathbb{R}^d$ by

$$\mu_X(A) := \mu(A \cap X)/\mu(X)\,, \qquad\qquad A \subset \mathbb{R}^d \text{ measurable.}$$

With this notation, we can now formulate the following result.

Theorem 7.33 (Entropy numbers on unbounded domains). *Let H be an RKHS of a bounded kernel k on $\mathbb{R}^d$ with $\|k\|_\infty = 1$ and μ be a distribution on $\mathbb{R}^d$ that has tail exponent $\tau \in (0,\infty]$. We write $B := B_{\ell_2^d}$ and assume that there exist constants $a \ge 1$, $c \ge 1$, $\varsigma > 0$, and $p \in (0,1)$ such that*

$$e_i(\mathrm{id} : H \to L_2(\mu_{rB})) \;\le\; c\, a^{\varsigma}\, r^{\varsigma} i^{-\frac{1}{2p}}\,, \qquad\qquad i \ge 1,\, r \ge a^{-1}. \tag{7.62}$$

Then there exists a constant $c_{\varsigma,\tau} \ge 1$ only depending on ς and τ such that

$$e_i(\mathrm{id} : H \to L_2(\mu)) \;\le\; c\, c_{\varsigma,\tau}\, a^{\frac{\varsigma\tau}{2\varsigma+\tau}}\, i^{-\frac{2p\varsigma+\tau}{4p\varsigma+2p\tau}}\,, \qquad\qquad i \ge 1\,.$$

Proof. Obviously, for $\tau = \infty$, there is nothing to prove, and hence we assume $\tau < \infty$. Let us now fix an $\varepsilon > 0$, an $r \ge a^{-1}$, and an integer $i \ge 1$. Moreover, let $f_1, \ldots, f_{2^{i-1}}$ be an $(1+\varepsilon)e_i(B_H, L_2(\mu_{rB}))$-net of B_H and $f'_1, \ldots, f'_{2^{i-1}}$ be an $(1+\varepsilon)e_i(B_H, L_2(\mu_{\mathbb{R}^d\backslash rB}))$-net of B_H. For $j, l \in \{1, \ldots, 2^{i-1}\}$, we write

$$f_j \diamond f'_l := \mathbf{1}_{rB} f_j + \mathbf{1}_{\mathbb{R}^d\backslash rB} f'_l\,,$$

i.e., $f_j \diamond f'_l$ equals f_j on the scaled Euclidean ball rB, while it equals f'_l on the complement of this ball. Let us now investigate how well these functions approximate B_H in $L_2(\mu)$. To this end, we fix a $g \in B_H$. Then there exist two indexes $j, l \in \{1, \ldots, 2^{i-1}\}$ such that

$$\begin{aligned} \|g - f_j\|_{L_2(\mu_{rB})} &\le (1+\varepsilon) e_i(B_H, L_2(\mu_{rB}))\,, \\ \|g - f'_l\|_{L_2(\mu_{\mathbb{R}^d\backslash rB})} &\le (1+\varepsilon) e_i(B_H, L_2(\mu_{\mathbb{R}^d\backslash rB}))\,. \end{aligned}$$

With these estimates, we obtain

$$\begin{aligned} \|g - f_j \diamond f'_l\|^2_{L_2(\mu)} &= \mu(rB)\|g - f_j\|^2_{L_2(\mu_{rB})} + \mu(\mathbb{R}^d\backslash rB)\|g - f'_l\|^2_{L_2(\mu_{\mathbb{R}^d\backslash rB})} \\ &\le (1+\varepsilon)^2 \Bigl(e_i^2\bigl(B_H, L_2(\mu_{rB})\bigr) + r^{-\tau} e_i^2\bigl(B_H, L_2(\mu_{\mathbb{R}^d\backslash rB})\bigr)\Bigr)\,. \end{aligned}$$

Moreover, Carl's inequality (A.42) together with (A.27), (4.19), $\|k\|_\infty = 1$, and $a_m(S^*_{k,\nu}) = s_m(S^*_{k,\nu})$ for all $m \geq 1$ and all distributions ν on $\mathbb{R}^d$ yields a universal constant $K \geq 1$ independent of all other occurring terms such that

$$i e_i^2(B_H, L_2(\nu)) \leq \sum_{m=1}^{i} e_m^2(S^*_{k,\nu}) \leq K \|S^*_{k,\nu}\|^2_{\mathsf{HS}} \leq K\,. \tag{7.63}$$

Combining this estimate for $\nu := \mu_{\mathbb{R}^d \setminus rB}$ with $\sqrt{s+t} \leq \sqrt{s} + \sqrt{t}$ and the previous estimate, we then obtain

$$\|g - f_j \diamond f_l'\|_{L_2(\mu)} \leq (1+\varepsilon)\big(e_i(B_H, L_2(\mu_{rB})) + \sqrt{K} r^{-\tau/2} i^{-1/2}\big).$$

Since there are 2^{2i-2} functions $f_j \diamond f_l'$, the latter estimate together with (7.62) and $\varepsilon \to 0$ implies

$$e_{2i-1}(\mathrm{id} : H \to L_2(\mu)) \leq 2c\, a^\varsigma\, r^\varsigma i^{-\frac{1}{2p}} + 2\sqrt{K} r^{-\frac{\tau}{2}} i^{-\frac{1}{2}}\,,$$

where the factor 2 appears since in general we cannot guarantee $f_j \diamond f_l' \in H$ and hence (A.36) has to be applied. For $r := a^{-\frac{2\varsigma}{2\varsigma+\tau}} i^{\frac{1-p}{p(2\varsigma+\tau)}} \geq a^{-1}$, we thus arrive at

$$e_{2i-1}(\mathrm{id} : H \to L_2(\mu)) \leq 4c\sqrt{K}\, a^{\frac{\varsigma\tau}{2\varsigma+\tau}}\, i^{-\frac{2p\varsigma+\tau}{2p(2\varsigma+\tau)}}\,, \qquad i \geq 1,$$

and from this we easily obtain the assertion by the monotonicity of the entropy numbers. □

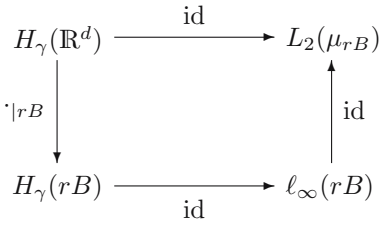
Let us now recall Example 4.32, where we saw that for all Gaussian RBF kernels the embedding $\mathrm{id} : H_\gamma(\mathbb{R}) \to C_b(\mathbb{R}^d)$ is *not* compact. Consequently, no entropy estimate of the form (7.49) is possible in this case. On the other hand, the following theorem together with Corollary 7.31 establishes an entropy assumption of the form (7.48), and hence the latter is indeed strictly weaker than (7.49).

Theorem 7.34 (Entropy numbers for Gaussian kernels). *Let μ be a distribution on $\mathbb{R}^d$ having tail exponent $\tau \in (0, \infty]$. Then, for all $\varepsilon > 0$ and $d/(d+\tau) < p < 1$, there exists a constant $c_{\varepsilon,p} \geq 1$ such that*

$$e_i\big(\mathrm{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mu)\big) \leq c_{\varepsilon,p}\, \gamma^{-\frac{(1-p)(1+\varepsilon)d}{2p}}\, i^{-\frac{1}{2p}}$$

for all $i \geq 1$ and $\gamma \in (0, 1]$.

Proof. Let B be the closed unit Euclidean ball centered at the origin and $r > 0$ be a strictly positive real number. Since $\mathbb{R}^d \setminus rB$ has measure zero with respect to the distribution μ_{rB}, we obtain the commutative diagram



where $\cdot|_{rB}$ denotes the restriction operator. The latter is a metric surjection (note that Corollary 4.43 implies that it is even an isometric isomorphism), and, in addition, we obviously have $\|\operatorname{id} : \ell_\infty(rB) \to L_2(\mu_{rB})\| \le 1$. For an integer $m \ge 1$, Theorem 6.27 then yields a constant $\tilde{c}_{m,d} \ge 1$ only depending on m and d such that

$$e_i\big(\operatorname{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mu_{rB})\big) \le e_i\big(\operatorname{id} : H_\gamma(rB) \to \ell_\infty(rB)\big) \le \tilde{c}_{m,d}\, r^m\, \gamma^{-m} i^{-\frac{m}{d}}$$

for all $0 < \gamma \le r$ and all $i \ge 1$. Let us now restrict our consideration to integers m with $m > d/2$. Applying Theorem 7.33 to the previous estimate, we then obtain

$$e_i\big(\operatorname{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mu)\big) \;\le\; c_{m,d,\tau}\gamma^{-\frac{m\tau}{2m+\tau}}\, i^{-\frac{md+m\tau}{2md+d\tau}}\,, \tag{7.64}$$

where $c_{m,d,\tau} \ge 1$ is a constant only depending on m, d, and τ. Moreover, using (7.63) with $\nu := \mu$, we see that there exists a universal constant $K \ge 1$ such that

$$e_i\big(\operatorname{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mu)\big) \;\le\; K\, i^{-1/2}\,, \qquad i \ge 1. \tag{7.65}$$

Interpolating (7.64) and (7.65) by Lemma A.1.3 with $t := 2$ and $r := \frac{2md+d\tau}{md+m\tau}$ shows that for all $s \in [r, 2]$ there exists a constant $c_{m,d,\tau,s} \ge 1$ such that

$$e_i\big(\operatorname{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mathrm{D})\big) \;\le\; c_{m,d,\tau,s}\, \gamma^{-\frac{(2-s)md}{s(2m-d)}}\, i^{-\frac{1}{s}} \tag{7.66}$$

for all $i \ge 1$ and $m \ge 1$. We now fix a p with $\frac{d}{d+\tau} < p < 1$ and define $s := 2p$. For $\varepsilon > 0$, there then exists an integer m such that both

$$r = \frac{2md + d\tau}{md + m\tau} < s \qquad \text{and} \qquad \frac{md}{2m-d} \le \frac{(1+\varepsilon)d}{2}\,.$$

Consequently, we can apply (7.66) to this s and m. □

7.6 Further Reading and Advanced Topics

The first argument, presented in Section 7.1, for the *suboptimality* of the approach of Chapter 6 was discovered by Steinwart and Scovel (2005a, 2007), who also used the described iterative scheme to establish some learning rates for SVMs. The second argument, based on the variance bound for distributions

having zero Bayes risk, is well-known. For the case of binary classification, we refer to Section 12.7 of Devroye *et al.* (1996) for a detailed account including certain infinite sets of functions and some historical remarks.

In the machine learning literature, the idea of using a variance bound to obtain improved oracle inequalities goes, to the best of our knowledge, back to Lee *et al.* (1998). This idea was later refined by, e.g., Mendelson (2001a, 2001b), Bartlett *et al.* (2005), and Bartlett *et al.* (2006). In addition, similar ideas were also developed in the statistical literature by, e.g., Mammen and Tsybakov (1999) and Massart (2000b). For a brief historical survey, we refer to the introduction of Bartlett *et al.* (2005). The *improved oracle inequality for ERM* established in Theorem 7.2 is rooted in the ideas of the articles above. Finally, oracle inequalities that do *not* have a constant in front of the approximation error term have recently been established for ERM and some aggregation procedures by Lecué (2007a).

The first version of *Talagrand's inequality* was proved by Talagrand (1996), who showed that there exist universal constants K, c_1, and c_2 such that

$$\mathrm{P}\bigl(\{z \in Z : g(z) - \mathbb{E}_{\mathrm{P}} g \geq \varepsilon\}\bigr) \leq K \exp\left(-\frac{\varepsilon^2}{2(c_1 v + c_2 B\varepsilon)}\right), \tag{7.67}$$

where $v := \mathbb{E}_{\mathrm{P}} \sup_{f\in\mathcal{F}} \sum_{j=1}^n f(z_j)^2$. Unfortunately, however, his constants are rather large and, in addition, the variance condition expressed in v is more complicated then the one in Theorem A.9.1. The first improvement of this inequality was achieved by Ledoux (1996), who showed by developing the entropy functional technique that (7.67) holds for $K = 2$, $c_1 = 42$, and $c_2 = 8$ if v is replaced by $v + \frac{4}{21} B \mathbb{E}_{\mathrm{P}} g$. The next improvement was established by Massart (2000a) by proving (7.67) for $K = 1$, $c_1 = 8$, and $c_2 = 2.5$. In addition, he showed a version of Theorem 7.5 in which $\sqrt{2\tau n\sigma^2} + (2/3 + \gamma^{-1})\tau B$ is replaced by $\sqrt{8\tau n\sigma^2} + (5/2 + 32\gamma^{-1})\tau B$. The last improvements were then achieved by Rio (2002) and Bousquet (2002a). We followed Bousquet's approach, which gives optimal constants. Finally, Klein and Rio (2005) recently established a version of Talagrand's inequality for independent, not necessarily identically distributed random variables and almost optimal constants. More applications of the entropy technique are presented by Chafaï (2004), and other applications of Talagrand's inequality are discussed by Boucheron *et al.* (2003).

The *peeling* argument of Theorem 7.7 for weighted empirical processes is a standard tool to estimate complexity, measures of complicated function classes by complexity measures of related "easier" function classes. Moreover, *symmetrization* is a standard tool in empirical process theory that goes back to Kahane (1968) and Hoffmann-Jørgensen (1974, 1977). In the proof of Theorem A.8.1, we followed Section 2.3 of van der Vaart and Wellner (1996). The *contraction principle* stated in Theorem A.8.4 was taken from p. 112 of Ledoux and Talagrand (1991), and its Corollary A.8.5 was first shown by Talagrand (1994). Finally, *Dudley's chaining* (see Dudley, 1967, and the

historical remarks on p. 269 of van der Vaart and Wellner, 1996) for deriving the *maximal inequality* of Theorem 7.12 was taken from Section 2.2 of van der Vaart and Wellner (1996) with the additional improvements reported by Bousquet (2002b).

Empirical *Rademacher averages* were first used in learning theory by Koltchinskii (2001) and Koltchinskii and Panchenko (2000, 2002) as a penalization method for model selection. Some empirical findings for this approach were first reported by Lozano (2000). This penalization method was later refined, for example, by Lugosi and Wegkamp (2004). Furthermore, Rademacher averages of localized function classes were used by Bartlett *et al.* (2002), Mendelson (2002), and Bartlett *et al.* (2005) in a way similar to that of Theorem 7.20 and in a refined way by Bartlett *et al.* (2004). An overview of the ideas was described by Bousquet (2003b). Relations of Rademacher averages to other complexity measures are described by Mendelson (2002, 2003a), and structural properties of Rademacher averages were investigated by Bartlett and Mendelson (2002). Moreover, Mendelson (2003b) estimated Rademacher averages of balls in RKHSs by the eigenvalues of the associated integral operator. Finally, using covering numbers to bound Rademacher averages also goes back to Mendelson (2002). In Section 7.3, we essentially followed his approach, though we decided to use entropy numbers instead of covering numbers since *a)* this yields slightly smaller constants and *b)* entropy numbers have a conceptionally easier relation to eigenvalues and approximation numbers.

To the best of our knowledge, the *clipping idea* was first used by Bartlett (1998) for the analysis of neural networks. In the context of SVMs, it first appeared in a paper by Bousquet and Elisseeff (2002). The approach for the oracle inequalities presented in Section 7.4 was taken from Steinwart *et al.* (2007), which in turn was inspired by the work of Wu *et al.* (2007). An oracle inequality for SVMs using losses that *cannot* be clipped was established by Steinwart *et al.* (2006c). Their proof is based on a rather complicated refinement of a technique developed by Bartlett *et al.* (2006). However, it is relatively easy to modify the proof of the oracle inequality for CR-ERM presented in Theorem 7.20 to deal with regularized *unclipped* ERM. Such a modification would then essentially reproduce the oracle inequality of Steinwart *et al.* (2006c). Finally, an oracle inequality for unclipped SVMs using the hinge loss was proved by Blanchard *et al.* (2008). It is interesting to compare their conjecture regarding the optimal exponent in the regularization term with Exercise 7.6.

Unfortunately, little is known about the sharpness of the derived oracle inequalities and their resulting best learning rates. In fact, to the best of our knowledge, the only known results consider (essentially) the case where L is the least squares loss, $H = W^m(X)$ for some Euclidean ball X, P_X is the uniform distribution, and $m > d/2$. If the regression function $f^*_{L,\mathrm{P}}$ is bounded and contained in $W^k(X)$ for some $0 < k \leq m$, it is then easy to check by the remarks made in Section 5.6, the entropy number bound (A.48), and Corollary

7.31 that the best learning rate Theorem 7.23 provides is $n^{-\min\{\frac{k}{m}, \frac{2k}{2k+d}\}}$. For $m - d/2 \leq k \leq m$, it is known that this rate is optimal in a minmax sense. We refer to Györfi *et al.* (2002) and the references therein for such optimal rates.

Theorem 7.29, which compares the average eigenvalues of empirical integral operators with the eigenvalues of the corresponding infinite-sample integral operator, was first shown by Shawe-Taylor *et al.* (2002, 2005) in the special case of continuous kernels over compact metric spaces. Zwald *et al.* (2004) generalized this result to bounded measurable kernels with separable RKHSs. We essentially followed their ideas for the proof of Theorem 7.29, while to the best of our knowledge the rest of Section 7.5 has not been published.

Exercise 7.7 allows us to replace the entropy number assumption (7.48) by a bound on the eigenvalues of the integral operator. For the hinge loss and unclipped SVM decision functions, a conceptionally similar oracle inequality was shown by Blanchard *et al.* (2008). A key step in their proof is an estimate in the spirit of Mendelson (2003b), i.e., a bound of the Rademacher averages of balls of RKHSs by the eigenvalues of the associated integral operator. In the polynomial regime (7.68), their resulting learning rates are, however, always worse than the ones we obtained in the discussion after Theorem 7.23. This is, of course, partly a result of the fact that these authors consider unclipped decision functions. On the other hand, Steinwart and Scovel (2005b) pointed out that the results of Blanchard *et al.* (2008) are suboptimal for SVMs under the uniform entropy assumption (7.49), and by modifying the oracle inequality of Steinwart *et al.* (2006c) we conjecture that this remains true under the eigenvalue assumption (7.68). Finally, an oracle inequality for SVMs using the least squares loss that involves the eigenvalues of the associated integral operator was established by Caponnetto and De Vito (2005).

7.7 Summary

In this chapter, we developed advanced techniques for establishing oracle inequalities for ERM and SVMs. The main reason for this development was the observation made in Section 7.1 that the oracle inequalities of Chapter 6 provide suboptimal learning rates. Here we identified two sources for the suboptimality, namely a supremum bound that is suboptimal in terms of the regularization parameter and a possibly existing variance bound that can be exploited using Bernstein's inequality rather than Hoeffding's inequality.

In Section 7.2, we then established an improved oracle inequality for ERM over finite sets of functions that involves a variance bound. This oracle inequality proved to be useful for parameter selection purposes in Section 7.4. Moreover, its proof presented the core idea for establishing similar oracle inequalities for SVMs. Unfortunately, however, it turned out that this core idea could not be directly generalized to infinite sets of functions. This forced us to introduce some advanced tools from empirical process theory such as

Talagrand's inequality, peeling, symmetrization, Rademacher averages, and Dudley's chaining in Section 7.3. At the end of this section, we then illustrated how to orchestrate these tools to derive an improved oracle inequality for ERM over infinite sets of functions.

In Section 7.4, we used the tools from the previous section to establish an oracle inequality for general, *clipped regularized empirical risk minimizers*. We then derived two oracle inequalities for clipped SVMs, one involving bounds on entropy numbers and one not. An extensive comparison to the oracle inequalities of Chapter 6 showed that the new oracle inequalities always provide faster learning rates. Furthermore, we combined the improved oracle inequality for ERM over finite sets of functions with the new oracle inequalities for clipped SVMs to derive some results for simple data-dependent choices of the regularization parameter. Finally, we showed in Section 7.5 that the entropy assumptions of the new oracle inequalities can be substantially weaker than the ones from Chapter 6.

7.8 Exercises

7.1. Optimality of peeling ($\star$)
Let $T \neq \emptyset$ be a set, $g, h : T \to [0, \infty)$ be functions, and $r^* := \inf\{h(t) : t \in T\}$. Furthermore, let $\varphi : (r^*, \infty) \to [0, \infty)$ be a function such that

$$\sup_{\substack{t \in T \\ h(t) \leq r}} g(t) \;\geq\; \varphi(r)$$

for all $r > r^*$. Show the inequality

$$\sup_{t \in T} \frac{g(t)}{h(t) + r} \;\geq\; \frac{\varphi(r)}{2r}\,, \qquad\qquad r > r^*.$$

Generalize this result to expectations over suprema.

7.2. Simple properties of empirical Rademacher averages ($\star$)
Let $\mathcal{F}, \mathcal{G} \subset \mathcal{L}_0(Z)$ be non-empty subsets, $\alpha \in \mathbb{R}$ be a real number, n be an integer, and $D := (z_1, \ldots, z_n) \in Z^n$ be a finite sequence. Show that the following relations hold:

$$\begin{aligned}
\mathrm{Rad}_D(\mathcal{F} \cup \mathcal{G}, n) &\leq \mathrm{Rad}_D(\mathcal{F}, n) + \mathrm{Rad}_D(\mathcal{G}, n)\,,\\
\mathrm{Rad}_D(\mathcal{F} \cup \{0\}, n) &= \mathrm{Rad}_D(\mathcal{F}, n)\,,\\
\mathrm{Rad}_D(\alpha\mathcal{F}, n) &= |\alpha| \mathrm{Rad}_D(\mathcal{F}, n)\,,\\
\mathrm{Rad}_D(\mathcal{F} + \mathcal{G}, n) &\leq \mathrm{Rad}_D(\mathcal{F}, n) + \mathrm{Rad}_D(\mathcal{G}, n)\,,\\
\mathrm{Rad}_D(\mathrm{co}\,\mathcal{F}, n) &= \mathrm{Rad}_D(\mathcal{F}, n)\,,
\end{aligned}$$

where in the last equation $\mathrm{co}\,\mathcal{F}$ denotes the convex hull of $\mathcal{F}$.

7.3. Another oracle inequality ($\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a Lipschitz continuous loss with $|L|_1 \leq 1$, $\mathcal{F} \subset \mathcal{L}_\infty(X)$ be a separable set satisfying $\|f\|_\infty \leq M$ for a suitable constant $M > 0$ and all $f \in \mathcal{F}$, and P be a distribution on $X \times Y$ that has a Bayes decision function $f^*_{L,\mathrm{P}}$ with $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) < \infty$. Assume that there exist constants $B > 0$, $\vartheta \in [0, 1]$, and $V \geq B^{2-\vartheta}$ such that for all $f \in \mathcal{F}$ we have

$$\begin{aligned} \|L \circ f - L \circ f^*_{L,\mathrm{P}}\|_\infty &\leq B\,, \\ \mathbb{E}_\mathrm{P}\bigl(L \circ f - L \circ f^*_{L,\mathrm{P}}\bigr)^2 &\leq V \cdot \bigl(\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}})\bigr)^\vartheta\,. \end{aligned}$$

For $f \in \mathcal{F}$, define $h_f := L \circ f - L \circ f^*_{L,\mathrm{P}}$. For a measurable ERM with respect to L and $\mathcal{F}$, show the following assertions:

i) $\|h_f - h_{f'}\|_\infty \leq \|f - f'\|_\infty$ for all $f, f' \in \mathcal{F}$.
ii) Let $\mathcal{C}$ be an ε-net of $\mathcal{F}$ with respect to $\|\cdot\|_\infty$. Then for all $D \in (X \times Y)^n$ there exists an $f \in \mathcal{C}$ such that $\mathbb{E}_\mathrm{P} h_{f_D} - \mathbb{E}_\mathrm{D} h_{f_D} \leq \mathbb{E}_\mathrm{P} h_f - \mathbb{E}_\mathrm{D} h_f + 2\varepsilon$.
iii) For all $\varepsilon > 0$, $r > 0$, and $\tau > 0$, we have

$$\mathbb{E}_\mathrm{P} h_{f_D} - \mathbb{E}_\mathrm{D} h_{f_D} < 2\varepsilon + \bigl(\mathbb{E}_\mathrm{P} h_{f_D} + \varepsilon\bigr)\left(\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr}\right) + \sqrt{\frac{2V\tau r^\vartheta}{n}} + \frac{4B\tau}{3n}$$

with probability P^n not less than $1 - \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)e^{-\tau}$.
iv) Given an $\varepsilon > 0$ and $\tau > 0$, we have

$$\begin{aligned} \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P}} \leq\, & 6\bigl(\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} - \mathcal{R}^*_{L,\mathrm{P}}\bigr) + 8\varepsilon \\ & + 4\left(\frac{8V\bigl(\tau + 1 + \ln \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)\bigr)}{n}\right)^{\frac{1}{2-\vartheta}} \end{aligned}$$

with probability P^n not less than $1 - e^{-\tau}$.
v) Let $a > 0$ and $p > 0$ be constants such that $\ln \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq a\varepsilon^{-2p}$ for all $\varepsilon > 0$. Then, for all $\tau \geq 1$, we have

$$\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P}} \leq 6\bigl(\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} - \mathcal{R}^*_{L,\mathrm{P}}\bigr) + 4\left(\frac{16V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + c_{p,\vartheta}\left(\frac{8Va}{n}\right)^{\frac{1}{2+2p-\vartheta}}$$

with probability P^n not less than $1 - e^{-\tau}$, where $c_{p,\vartheta} \in (0, 12]$ is a constant with $\lim_{p \to 0^+} c_{p,\vartheta} = 4$.
vi) Compare this oracle inequality with (7.28) and (7.30).

Hint: To prove *iii)*, fix an ε-net $\mathcal{C}$ of $\mathcal{F}$ with $|\mathcal{C}| = \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$. Then apply (7.9) to $\mathcal{C}$. Finally, follow the argument after (7.9) and apply *ii)*. To prove *iv)*, repeat the last steps of the proof of Theorem 7.2 using *iii)* for $r := (\frac{16V\tau}{n})^{1/(2-\vartheta)}$.

7.4. Oracle inequality for clipped approximate SVMs ($\star\star\star$)
Under the assumptions of Theorem 7.23, show that any clipped ϵ-approximate SVM satisfies

$$\mathrm{P}^n\Bigl(D : \lambda\|f_{\mathrm{D},\lambda}\|^2+\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda})-\mathcal{R}^*_{L,\mathrm{P}} > 6\bigl(\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_0)-\mathcal{R}^*_{L,\mathrm{P}}\bigr)+3r+3\epsilon\Bigr) \le 3e^{-\tau}.$$

7.5. Comparison of oracle inequalities ($\star\star\star$)
Apply Theorem 7.20 to the assumptions of Exercise 7.3, and show that Theorem 7.20 produces an oracle inequality that is sharper in n.

7.6. Different exponents in the regularization term of SVMs ($\star\star\star\star$)
Assume that L, H, P, and f_0 as well as M, B, V, ϑ, a, p, and B_0 are as in Theorem 7.23. Moreover, for $q \in [0,\infty)$ and $\lambda > 0$, consider the learning method that assigns to every $D \in (X \times Y)^n$ a function $f^{(q)}_{\mathrm{D},\lambda} \in H$ that solves

$$\min_{f\in H} \lambda\|f\|_H^q + \mathcal{R}_{L,\mathrm{D}}(f)\,.$$

Show that for all fixed $\tau > 0$, $n \ge 1$, and $\lambda > 0$, this learning method satisfies

$$\begin{aligned}\lambda\|f^{(q)}_{\mathrm{D},\lambda}\|_H^q+\mathcal{R}_{L,\mathrm{P}}(\widehat{f}^{(q)}_{\mathrm{D},\lambda})-\mathcal{R}^*_{L,\mathrm{P}} \le{}& 9\bigl(\lambda\|f_0\|_H^q + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}\bigr) + \frac{15B_0\tau}{n}\\ &+K\Bigl(\frac{a^{2pq}}{\lambda^{2p}n^q}\Bigr)^{\frac{1}{2q-2p-\vartheta q+\vartheta pq}}+3\Bigl(\frac{72V\tau}{n}\Bigr)^{\frac{1}{2-\vartheta}}\end{aligned}$$

with probability P^n not less than $1-3e^{-\tau}$, where $K \ge 1$ is a constant only depending on p, M, B, ϑ, and V.

Furthermore, assume that there exist constants $c > 0$ and $\beta \in (0,1]$ such that $A_2(\lambda) \le c\lambda^\beta$ for all $\lambda \ge 0$. In addition, suppose that L is Lipschitz continuous. Using Exercise 5.11 and Lemma A.1.7, show that (7.53) are the best learning rates this oracle inequality can provide. Can we make a conclusion regarding the "optimal" exponent q?

7.7. Eigenvalues vs. average entropy numbers ($\star\star\star$)
Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Assume that there exist constants $0 < p < 1$ and $a \ge 1$ such that the extended sequence of eigenvalues of the integral operator $T_{k,\mu}$ satisfies

$$\lambda_i(T_{k,\mu}) \le a\, i^{-\frac{1}{p}}\,, \qquad i \ge 1. \tag{7.68}$$

Show that there exists a constant $c_p > 0$ only depending on p such that

$$\mathbb{E}_{D\sim\mu^n} e_i(B_H, L_2(\mathrm{D})) \le c_p\, a\, \bigl(\min\{i,n\}\bigr)^{\frac{1}{2p}}\, i^{-\frac{1}{p}}\,, \qquad i,n \ge 1.$$

8

Support Vector Machines for Classification

Overview. *Classification is one of the main areas of application for support vector machines. This chapter presents key results on the generalization performance of SVMs when applied to this learning problem. In addition, we investigate how the choice of the loss influences both the number of support vectors we can typically expect and the ability of SVMs to estimate posterior probabilities.*

Prerequisites. *For the basic results on the generalization performance, we need Sections 2.1–2.3 on loss functions, Chapter 4 on kernels, Chapter 5 on infinite-sample SVMs, and the oracle inequalities from Chapter 6. The more advanced results in this direction also require Chapter 7. For estimating the number of support vectors and the posterior probabilities, we further need Sections 3.1–3.6 and 3.9.*

Usage. *Parts of this chapter are helpful in Section 10.3 on robustness of SVMs for classification and in Chapter 11, where practical aspects of SVMs are discussed.*

Binary classification is one of the central applications for machine learning methods in general and for SVMs in particular. In this chapter, we investigate features of SVMs when applied to binary classification. In particular, we consider the following questions:

- For what types of distributions do SVMs learn fast?
- How many support vectors can we expect?
- Can SVMs be used to estimate posterior probabilities?
- Which loss functions are a reasonable alternative to the hinge loss?

Obviously, the first question is *the* main question since the primary goal in almost every application dealing with classification is a good classification performance. However, in many applications, other features are also highly desired such as *a)* a low computational complexity in the training and/or the employment phase and *b)* the ability to estimate the probability $\eta(x)$ of a positive label y at point x. Now it is obvious that the number of support vectors has a direct influence on the time required to evaluate the SVM decision function, and therefore the second question addresses the computational complexity during the employment phase. Moreover, we will see in Chapter 11 that the number of support vectors also has a substantial influence on the time required to train the SVM, and therefore the second question actually

addresses the overall computational complexity. Finally, though the hinge loss is the loss function that is most often used in practice, it has some disadvantages. First, it is not differentiable, and hence certain optimization procedures cannot be applied to its SVM optimization problem. Second, we have seen in Chapters 2 and 3 that the minimizer of the hinge loss is $\operatorname{sign}(2\eta(x)-1)$, $x \in X$. However, this expression does not contain information about the size of $\eta(x)$, and therefore SVM decision functions obtained by using the hinge loss cannot be used to estimate $\eta(x)$. This discussion shows that a deliberative decision on whether or which SVM is used for a particular application requires at least answers to the questions above. Moreover, particular applications may require taking into account further aspects of learning methods such as robustness, considered in Section 10.3.

The rest of this chapter is organized as follows. In Section 8.1, we reformulate some basic oracle inequalities from Chapter 6. These oracle inequalities estimate the excess *classification* risk of SVMs using the hinge loss and can be used to develop data-dependent parameter selection strategies such as the one considered in Section 6.5. In Section 8.2, we will then focus on SVMs that use a Gaussian RBF kernel since these are the kernels that are most often used in practice. For such SVMs, we present a rather general assumption on the data-generating distribution P that describes the behavior of P in the vicinity of the decision boundary and that enables us to estimate the approximation error function. This is then used to analyze a training validation support vector machine (TV-SVM) that uses the validation set to determine both the regularization parameter and the kernel parameter. Section 8.3 then uses the more advanced oracle inequalities of Chapter 7 to improve the learning rates obtained in Section 8.2. To this end, another assumption on P is introduced that measures the amount of noise P has in the labeling process and that can be used to establish a variance bound for the hinge loss. In Section 8.4, we then investigate the sparseness of SVMs by presenting a lower bound on the number of support vectors. Here it will turn out that the Bayes classification risk is the key quantity that asymptotically controls the sparseness for SVMs using the hinge loss. In the last section, we will then consider SVMs for binary classification that use an alternative loss function. Here we first improve the general calibration inequalities obtained in Section 3.4 for distributions satisfying the low-noise assumption introduced in Section 8.3. Furthermore, we will establish a general lower bound on the sparseness of SVMs that use a margin-based loss. Finally, this lower bound is used to describe some properties of loss functions that prevent the decision functions from being sparse.

8.1 Basic Oracle Inequalities for Classifying with SVMs

The goal of this section is to illustrate how the basic oracle inequalities established in Section 6.4 can be used to investigate the classification performance of SVMs using the hinge loss.

Given a distribution P on $X \times Y$, we assume throughout this and the following sections that P^n denotes the canonical extension of n-fold product measure of P to the universal completion of the product σ-algebra on $(X \times Y)^n$. As in Section 6.4, this assumption makes it possible to ignore measurability questions.

Let us now begin with a reformulation of Theorem 6.24.

Theorem 8.1 (Oracle inequality for classification). *Let L be the hinge loss, $Y := \{-1, 1\}$, H be a separable RKHS with bounded measurable kernel k over X satisfying $\|k\|_\infty \leq 1$, and P be a distribution on $X \times Y$ such that H is dense in $L_1(\mathrm{P}_X)$. Then, for all $\lambda > 0$, $n \geq 1$, and $\tau > 0$, we have with probability P^n not less than $1 - e^{-\tau}$ that*

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} < A_2(\lambda) + \lambda^{-1}\left(\sqrt{\frac{8\tau}{n}} + \sqrt{\frac{4}{n}} + \frac{8\tau}{3n}\right),$$

where $A_2(\,\cdot\,)$ is the approximation error function with respect to L, H, and P.

Proof. Obviously, $L := L_{\text{hinge}}$ is a convex and Lipschitz continuous loss satisfying $L(y, 0) = 1$ for all $y \in Y$. Therefore, Theorem 6.24 shows that

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} < A_2(\lambda) + \lambda^{-1}\left(\sqrt{\frac{8\tau}{n}} + \sqrt{\frac{4}{n}} + \frac{8\tau}{3n}\right)$$

holds with probability P^n not less than $1 - e^{-\tau}$. Moreover, the hinge loss is also a P-integrable Nemitski loss of order 1 by Lemma 2.25, and hence Theorem 5.31 yields $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$. Consequently, we obtain by Theorem 2.31 that

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \;\leq\; \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H}\,. \qquad \square$$

Note that the left-hand side of the oracle inequality above considers the excess *classification* risk rather than the excess hinge risk, while the SVM is assumed to use the hinge loss. Consequently, if we can ensure that the right-hand side of the oracle inequality converges to 0, then the SVM is consistent with respect to the classification risk, and this finally justifies the use of the hinge loss as a surrogate for the classification loss.

We have seen in Section 6.4 that the oracle inequality above can sometimes be improved if the unit ball of the RKHS used has finite $\|\cdot\|_\infty$-covering numbers. The following theorem reformulates this result for polynomially growing covering numbers, i.e., polynomially decaying entropy numbers.

Theorem 8.2 (Classification with benign kernels). *Let L be the hinge loss, $Y := \{-1, 1\}$, X be a compact metric space, and H be the RKHS of a continuous kernel k over X with $\|k\|_\infty \leq 1$. Moreover, assume that there exist constants $a \geq 1$ and $p \in (0, 1]$ such that the dyadic entropy numbers satisfy*

$$e_i\big(\mathrm{id} : H \to C(X)\big) \;\leq\; a\, i^{-\frac{1}{2p}}\,, \qquad\qquad i \geq 1. \tag{8.1}$$

Then, for all distributions P *on* $X \times Y$*, for which* H *is dense in* $L_1(\mathrm{P}_X)$*, and all* $\lambda \in (0,1]$*,* $n \geq 1$*,* $\tau > 0$*, we have*

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} < A_2(\lambda) + 8\Big(\frac{a^{2p}}{\lambda^{1+p}\, n}\Big)^{1/(2+2p)} + \sqrt{\frac{8\tau+8}{\lambda n}}$$

with probability P^n *not less than* $1 - e^{-\tau}$*.*

Proof. Let us fix an $\varepsilon > 0$. Using Theorem 6.25, we see analogously to the proof of Theorem 8.1 that

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} < A_2(\lambda) + 4\varepsilon + \sqrt{\frac{8\tau+8\ln\big(2\mathcal{N}(B_H, \|\cdot\|_\infty, \lambda^{\frac{1}{2}}\varepsilon)\big)}{\lambda n}}$$

holds with probability P^n not less than $1 - e^{-\tau}$. Moreover, (8.1) implies

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \;\leq\; \ln(4) \cdot \Big(\frac{a}{\varepsilon}\Big)^{2p}, \qquad\qquad \varepsilon > 0\,,$$

by Lemma 6.21. Combining these estimates, optimizing the result with respect to ε by Lemma A.1.5, and using $(1+p)(4/p)^{p/(1+p)}(8\ln 4)^{1/(2+2p)} \leq 8$ then yields the assertion. □

Since the hinge loss is Lipschitz continuous, it is straightforward to check that the oracle inequalities above produce the learning rates we have discovered at the end of Section 6.4 and in Exercise 6.9. However, in this case, the learning rates are for the classification risk instead of the hinge risk. Moreover, note that we can easily use the oracle inequalities above to derive learning rates for data-dependent parameter selection strategies such as the TV-SVM. Since the oracle inequalities above consider the classification risk, we can, however, use either the clipped empirical hinge risk *or* the empirical classification risk in the validation step. We come back to this observation at the end of the following section, where we investigate a modified TV-SVM that also selects a kernel parameter in a data-dependent fashion.

8.2 Classifying with SVMs Using Gaussian Kernels

The Gaussian RBF kernel is one of the most often used kernels in practice. In this section, we derive learning rates for situations in which this kernel is used with different parameter values. To this end, we introduce some conditions on the data-generating distribution that describe their behavior near the "decision boundary". For such distributions, we then establish a bound on the approximation error function, which in turn will yield the learning rates. Finally, we discuss a method to select both the regularization parameter and the kernel parameter in a data-dependent, adaptive way.

Let us begin by presenting an oracle inequality for SVMs using the hinge loss and a Gaussian kernel.

Theorem 8.3 (Oracle inequality for Gaussian kernels). *Let L be the hinge loss, $Y := \{-1,1\}$, $X \subset \mathbb{R}^d$ be a compact subset, and $m \in \mathbb{N}$. Then there exists a constant $c_{m,d}(X) \geq 1$ such that, for all distributions P on $X \times Y$ and all fixed $\gamma, \lambda \in (0,1]$, $n \geq 1$, $\tau > 0$, we have with probability P^n not less than $1 - e^{-\tau}$ that*

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} < A_2^{(\gamma)}(\lambda) + c_{m,d}(X)\lambda^{-1/2}\bigl(\gamma^d\, n\bigr)^{-\frac{m}{2m+d}}\sqrt{\tau+1}\,.$$

Here $A_2^{(\gamma)}$ denotes the approximation error function with respect to L, P, and the Gaussian RBF kernel k_γ, and $f_{\mathrm{D},\lambda,\gamma}$ denotes the decision function of an SVM using L and k_γ.

Proof. Since X is compact, we may assume without loss of generality that X is a closed Euclidean ball. Then the assertion follows from combining Theorem 6.27 with Theorem 8.2. □

With some extra effort one can, in principle, determine a value for the constant $c_{m,d}(X)$. However, this requires us to find a constant in a bound on entropy numbers and hence we omit the details. Besides the constant $c_{m,d}(X)$, we also need to bound the approximation error function $A_2^{(\gamma)}$ in order to obtain learning rates from Theorem 8.3. Unfortunately, this requires us to impose assumptions on P since otherwise Theorem 8.3 would provide us with a uniform learning rate for classification, which by Corollary 6.7 is impossible.

In order to formulate such a condition, let us recall that "the" regular conditional probability $\mathrm{P}(\,\cdot\,|x)$ of a probability measure P on $X \times \{-1,1\}$ is only P_X-almost surely defined by (A.11). In other words, if we have two measurable functions $\eta_1, \eta_2 : X \to [0,1]$ for which the probability measures $\mathrm{P}_i(\,\cdot\,|x)$ defined by $\mathrm{P}_i(y=1|x) := \eta_i(x)$, $i = 1,2$, $x \in X$, satisfy (A.11), then η_1 and η_2 coincide up to a P_X-zero set. Conversely, any measurable modification of, say, η_1, on a P_X-zero set yields a regular conditional probability of P. Since in the following considerations we have to deal with this inherent ambiguity very carefully, we introduce the following definition.

Definition 8.4. *Let $Y := \{-1,1\}$, X be a measurable space, and P be a distribution on $X \times Y$. We say that a measurable function $\eta : X \to [0,1]$ is a **version of the posterior probability** of P if the probability measures $\mathrm{P}(\,\cdot\,|x)$ defined by $\mathrm{P}(y=1|x) := \eta(x)$, $x \in X$, form a regular conditional probability of P, i.e.,*

$$\mathrm{P}(A \times B) = \int_A \mathrm{P}(B|x)\, d\mathrm{P}_X(x)\,,$$

for all measurable sets $A \subset X$ and $B \subset Y$.

If a distribution P on $X \times Y$ has a smooth version of the posterior probability, then intuitively the set $\{x \in X : \eta = 1/2\}$ is the "decision boundary" that separates the class $\{x \in X : \eta(x) < 1/2\}$ of negatively labeled samples

from the class $\{x \in X : \eta(x) > 1/2\}$ of positively labeled samples. Intuitively, any classification method that uses a notion of distance on X to build its decision functions should learn well in regions that are not close to the decision boundary. This suggests that learning should be relatively easy for distributions that do not have a lot of mass in the vicinity of the decision boundary. Our next goal is to verify this intuition for SVMs using Gaussian kernels. To this end, we begin with the following definition that introduces a "distance to the decision boundary" without defining the decision boundary itself.

Definition 8.5. *Let (X, d) be a metric space, P be a distribution on $X \times \{-1, 1\}$, and $\eta : X \to [0, 1]$ be a version of its posterior probability. We write*

$$X_{-1} := \bigl\{x \in X : \eta(x) < 1/2\bigr\}\,,$$
$$X_1 := \bigl\{x \in X : \eta(x) > 1/2\bigr\}\,.$$

Then the associated ***version of the distance to the decision boundary*** *is the function $\Delta : X \to [0, \infty]$ defined by*

$$\Delta(x) := \begin{cases} d(x, X_1) & \text{if } x \in X_{-1}, \\ d(x, X_{-1}) & \text{if } x \in X_1, \\ 0 & \text{otherwise}\,, \end{cases} \tag{8.2}$$

where, as usual, $d(x, A) := \inf_{x' \in A} d(x, x')$.

In the following, we are mainly interested in the distance to the decision boundary for distributions defined on $X \times \{-1, 1\}$, where $X \subset \mathbb{R}^d$ is measurable. In this case, we will always assume that these subsets are equipped with the Euclidean metric.

One may be tempted to think that changing the posterior probability on a P_X-zero set has only marginal or even no influence on Δ. Unfortunately, quite the opposite is true. To illustrate this, let us consider the probability measure P on $\mathbb{R} \times \{-1, 1\}$ whose marginal distribution P_X is the uniform distribution on $[-1, 1]$ and for which $\eta(x) := \mathbf{1}_{[0,\infty)}(x)$, $x \in \mathbb{R}$, is a version of its posterior probability. Obviously this definition gives $X_{-1} = (-\infty, 0)$ and $X_1 = [0, \infty)$, and from these equations we immediately conclude that $\Delta(x) = |x|$, $x \in \mathbb{R}$, for the associated version of the distance to the decision boundary. Let us now change the posterior probability on the P_X-zero set $\mathbb{Q}$ by defining $\tilde{\eta}(x) := \eta(x)$ if $x \in \mathbb{R} \setminus \mathbb{Q}$ and $\tilde{\eta}(x) := 1 - \eta(x)$ if $x \in \mathbb{Q}$. Since $\mathrm{P}_X(\mathbb{Q}) = 0$, we easily see that $\tilde{\eta}$ is indeed another version of the posterior probability of P. However, for this version, we have $X_{-1} = ((-\infty, 0) \setminus \mathbb{Q}) \cup ([0, \infty) \cap \mathbb{Q})$ and $X_1 = ((-\infty, 0) \cap \mathbb{Q}) \cup ([0, \infty) \setminus \mathbb{Q})$, and therefore its associated distance to the decision boundary is given by $\Delta(x) = 0$ for all $x \in \mathbb{R}$. This example[1]

[1] Informally speaking, the example exploited the fact that sets that are "small" in a measure theoretic sense can be "big" in a topological sense. From this it becomes obvious that the described phenomenon can be observed for a variety of distributions on $\mathbb{R}^d$.

demonstrates that it is absolutely necessary to fix a version of the posterior probability when dealing with Δ.

Our next goal is to use the distance to the decision boundary Δ to describe the behavior of distributions P near the decision boundary. This behavior will be crucial when bounding the approximation error function for Gaussian kernels below. Let us begin by measuring the concentration of P_X near the decision boundary. To this end, we first note that $\{x \in X : \Delta(x) < t\}$ contains the set of points x that are either *a)* in $X_{-1} \cup X_1$ and geometrically close to the opposite class or *b)* satisfy $\eta(x) = 1/2$. Consequently, in the case $P_X(\{x \in X : \eta(x) = 1/2\}) = 0$, we see that $\{x \in X : \Delta(x) < t\}$ contains only the points (modulo a P_X-zero set) that are close to the opposite class. The following definition describes the size of this set.

Definition 8.6. *Let (X, d) be a metric space and P be a distribution on $X \times \{-1, 1\}$. We say that P has* ***margin exponent*** *$\alpha \in [0, \infty)$ for the version $\eta : X \to [0, 1]$ of its posterior probability if there exists a constant $c > 0$ such that the associated version Δ of the distance to the decision boundary satisfies*

$$P_X\big(\{x \in X : \Delta(x) < t\}\big) \;\le\; c\, t^{\alpha}\,, \qquad\qquad t \ge 0. \tag{8.3}$$

In the following, we sometimes say that P has margin exponent α for the version Δ of the distance to the decision boundary if (8.3) is satisfied. Moreover, we say that P has margin exponent α if there exists a version Δ of the distance to the decision boundary such that (8.3) is satisfied.

We will see later in this section that we are mainly interested in the behavior of P_X $(\{x \in X : \Delta(x) < t\})$ for $t \to 0$. In this case, an exponent $\alpha > 0$ in (8.3) describes how *fast* $P_X(\{x \in X : \Delta(x) < t\})$ converges to 0. Now note that for $\alpha > 0$ it is straightforward to check that $P_X(\{x \in X : \Delta(x) = 0\}) = 0$ holds. Since $\eta(x) = 1/2$ implies $\Delta(x) = 0$, we hence find

$$P_X\big(\{x \in X : \eta(x) = 1/2\}\big) = 0\,.$$

Informally speaking, (8.3) therefore measures the "size" of the set of points that are close to the opposite class. We refer to Figure 8.1 for an illustration of a distribution having a large margin exponent.

Obviously, every distribution has margin exponent $\alpha = 0$. Moreover, the margin exponent is monotone in the sense that every distribution P that has some margin exponent α also has margin exponent α' for all $\alpha' \in [0, \alpha]$. The following simple yet useful lemma shows that the margin exponent is invariant with respect to inclusions of the input space X.

Lemma 8.7. *Let $(\tilde{X}, d)$ be a metric space, $X \subset \tilde{X}$ be a measurable subset, and P be a distribution on $X \times Y$, where $Y := \{-1, 1\}$. Then there exists exactly one distribution $\tilde{P}$ on $\tilde{X} \times Y$, called the* ***canonical extension*** *of P, that satisfies the following two conditions:*

i) $\tilde{P}_X(A) = P_X(A \cap X)$ for all measurable $A \subset \tilde{X}$.

ii) $\tilde{\mathrm{P}}(y = 1|x) = \mathrm{P}(y = 1|x)$ *for* P_X*-almost all* $x \in X$*.*

Moreover, P *has margin exponent* $\alpha \in [0, \infty)$ *if and only if its canonical extension* $\tilde{\mathrm{P}}$ *has margin exponent* α*.*

Proof. Obviously, *i)* can be used to define $\tilde{\mathrm{P}}_X$, and it is furthermore clear that there is only one distribution on $\tilde{X}$ that satisfies *i)*. Moreover, *i)* implies that $\tilde{X} \setminus X$ is a $\tilde{\mathrm{P}}_X$-zero set and hence the behavior of $\tilde{\mathrm{P}}(y = 1|x)$ on $\tilde{X} \setminus X$ does not matter for the definition of $\tilde{\mathrm{P}}$. By using *ii)* as a definition for the posterior probability of $\tilde{\mathrm{P}}$, we then find the first assertion, where we use (A.11) to define $\tilde{\mathrm{P}}$. In order to prove the second assertion, we set $\tilde{\mathrm{P}}(y = 1|x) := 1/2$ for $x \in \tilde{X} \setminus X$. Then $\tilde{\mathrm{P}}(y = 1|x) \neq 1/2$ implies $x \in X$, and hence the classes $\tilde{X}_{-1}$ and $\tilde{X}_1$ of $\tilde{\mathrm{P}}$ are contained in X. The associated distances to the decision boundaries thus satisfy $\tilde{\Delta}(x) = \Delta(x)$ for all $x \in X$, and hence we finally find

$$\tilde{\mathrm{P}}_X\big(\{x \in \tilde{X} : \tilde{\Delta}(x) < t\}\big) = \mathrm{P}_X\big(\{x \in X : \Delta(x) < t\}\big)\,. \qquad \square$$

Let us now present some elementary examples of distributions having a strictly positive margin exponent.

Example 8.8 (Classes with strictly positive distance). Let X be a metric space, P be a distribution on $X \times Y$, and η be a version of the posterior probability for which the associated classes X_{-1} and X_1 have strictly positive distance, i.e., $d(X_{-1}, X_1) > 0$, and satisfy $\mathrm{P}_X(X_{-1} \cup X_1) = 1$. Then P has margin exponent α for all $\alpha > 0$.

To check this, let us write $t_0 := d(X_{-1}, X_1)$. Then we have $t_0 > 0$ and $\Delta(x) \geq t_0$ for all $x \in X_{-1} \cup X_1$. For $t \in (0, t_0]$, we thus find

$$\mathrm{P}_X\big(\{x \in X : \Delta(x) < t\}\big) = \mathrm{P}_X\big(\{x \in X_{-1} \cup X_1 : \Delta(x) < t\}\big) = 0\,.$$

Moreover, for $t > t_0$, we obviously have $\mathrm{P}_X(\{x \in X : \Delta(x) < t\}) \leq 1 < t_0^{-\alpha} t^{\alpha}$, and hence we obtain the assertion.

Finally, note that this example in particular includes distributions P on discrete metric spaces for which $\mathrm{P}(\{x \in X : \eta(x) = 1/2\}) = 0$. $\triangleleft$

Example 8.9 (Linear decision boundaries). Let $X \subset \mathbb{R}^d$ be a compact subset with strictly positive volume and P be a distribution on $X \times Y$ whose marginal distribution P_X is the uniform distribution. Moreover, assume that there exist a $w \in \mathbb{R}^d \setminus \{0\}$, a constant $b \in \mathbb{R}$, and a version η of the posterior probability such that the corresponding classes are given by $X_{-1} = \{x \in X : \langle w, x\rangle + b < 0\}$ and $X_1 = \{x \in X : \langle w, x\rangle + b > 0\}$. Then P has margin exponent $\alpha = 1$.

In order to check this, we first observe with the help of the rotation and translation invariance of the Lebesgue measure that we may assume without loss of generality that $w = e_1$ is the first vector of the standard ONB and $b = 0$. In addition, the compactness of X shows that there exists an $a > 0$ such that $X \subset [-a, a]^d$. Then we have

$$\begin{aligned}\{x \in X : \Delta(x) < t\} &= \{x \in X : -t < \langle e_1, x\rangle < t\} \\ &\subset \{x \in [-a,a]^d : -t < \langle e_1, x\rangle < t\},\end{aligned}$$

and since the volume of the last set is given by $a^{d-1}\min\{a,t\}$, the assertion becomes obvious.

Finally, note that P still has margin exponent $\alpha = 1$ if we assume that the classes X_{-1} and X_1 are "stripes", i.e., that they are described by finitely many parallel affine hyperplanes. ◁

Example 8.10 (Circular decision boundaries). Let $X \subset \mathbb{R}^d$ be a compact subset with a non-empty interior and P be a distribution on $X \times Y$ whose marginal distribution P_X is the uniform distribution. Moreover, assume that there exist an $x_0 \in \mathbb{R}^d$, an $r > 0$, and a version η of the posterior probability such that the corresponding classes are given by $X_{-1} = \{x \in X : \|x - x_0\| < r\}$ and $X_1 = \{x \in X : \|x - x_0\| > r\}$. Then P has margin exponent $\alpha = 1$.

To see this, we observe that we may assume without loss of generality that $x_0 = 0$. In addition, the compactness of X shows that there exists an $r_0 > 0$ such that $X \subset r_0 B_{\ell_2^d}$. Then we have

$$\begin{aligned}\{x \in X : \Delta(x) < t\} &= \{x \in X : r - t < \|x\|_2 < r + t\} \\ &\subset \{x \in r_0 B_{\ell_2^d} : r - t < \|x\|_2 < r + t\}.\end{aligned}$$

Since a simple calculation shows $(r+t)^d - (r-t)^d \le ct$ for a suitable constant $c > 0$ and all sufficiently small $t > 0$, we then obtain the assertion.

Finally note that P still has margin exponent $\alpha = 1$ if we assume that the classes X_{-1} and X_1 are described by finitely many circles. ◁

The previous two examples have not considered the *shape* of the input space X. However, this shape can have a substantial influence on the margin exponent. We refer to Exercise 8.3 for an example in this direction.

Let us finally consider an example of a distribution that does *not* have a strictly positive margin exponent.

Example 8.11 (Only trivial margin exponent). Assume that μ_{-1} is the uniform distribution on $[0,1]^2$ and that μ_1 is the uniform distribution on $\{0\} \times [0,1]$. Moreover, let $\eta := \mathbf{1}_{\{0\}\times[0,1]}$ and $P_X := (\mu_{-1} + \mu_1)/2$. Then the corresponding distribution P on $\mathbb{R}^2 \times \{-1,1\}$ has only margin exponent $\alpha = 0$. Indeed, for this version η of the posterior probability, we have $\Delta(x) = 0$ for all $x \in \{0\} \times [0,1]$, and it is easy to see that this cannot be changed by considering another version of the posterior probability. ◁

So far, we have only seen elementary examples of distributions having a non-trivial margin exponent. However, by combining these examples with the following lemma and the subsequent example, it is easy to see that the set of distributions having a non-trivial margin exponent is quite rich.

Lemma 8.12 (Images of inverse Hölder maps). *Let (X_1, d_1) and (X_2, d_2) be metric spaces and $\Psi : X_1 \to X_2$ be a measurable map whose image $\Psi(X_1)$ is measurable. Assume that Ψ is inverse Hölder continuous, i.e., there exist constants $c > 0$ and $\gamma \in (0, 1]$ such that*

$$d_1(x_1, x_1') \;\le\; c\, d_2^{\gamma}(\Psi(x_1), \Psi(x_1'))\,, \qquad x_1, x_1' \in X_1. \tag{8.4}$$

Let $Y := \{-1, 1\}$ and Q be a distribution on $X_1 \times Y$ that has some margin exponent $\alpha > 0$ for the version η_{Q} of its posterior probability. Furthermore, let P be a distribution on $X_2 \times Y$ for which there exists a constant $C \ge 1$ such that

$$\mathrm{P}_X(A) \;\le\; C\, \mathrm{Q}_X\bigl(\Psi^{-1}(A)\bigr) \tag{8.5}$$

for all measurable $A \subset X_2$. If P has a version η_{P} of its posterior probability such that

$$\eta_{\mathrm{P}}(\Psi(x_1)) = \eta_{\mathrm{Q}}(x_1)\,, \qquad x_1 \in X_1, \tag{8.6}$$

then P has margin exponent $\alpha\gamma$.

Proof. Without loss of generality, we may assume that $\eta_{\mathrm{P}}(x_2) = 1/2$ for all $x_2 \in X_2 \setminus \Psi(X_1)$. This assumption immediately implies $\Delta_{\mathrm{P}}(x_2) = 0$ for all $x_2 \in X_2 \setminus \Psi(X_1)$, but since (8.5) implies $\mathrm{P}_X(X_2 \setminus \Psi(X_1)) = 0$, we notice for later use that the behavior of Δ_{P} on $X_2 \setminus \Psi(X_1)$ has no influence on the margin exponent of P. Moreover, (8.4) implies that Ψ is injective and hence the inverse $\Psi^{-1} : \Psi(X_1) \to X_1$ of $\Psi : X_1 \to \Psi(X_1)$ exists. By (8.6), we conclude that $\eta_{\mathrm{P}}(x_2) = \eta_{\mathrm{Q}}(\Psi^{-1}(x_2)$ for all $x_2 \in \Psi(X_1)$. Let us now fix $x_2, x_2' \in X_2$ with $\eta_{\mathrm{P}}(x_2) < 1/2$ and $\eta_{\mathrm{P}}(x_2') > 1/2$. Then $\eta_{\mathrm{P}} \equiv 1/2$ on $X_2 \setminus \Psi(X_1)$ implies $x_2, x_2' \in \Psi(X_1)$, and hence we have $\eta_{\mathrm{Q}}(\Psi^{-1}(x_2)) < 1/2$ and $\eta_{\mathrm{Q}}(\Psi^{-1}(x_2')) > 1/2$. From this we conclude that

$$c\, d_2^{\gamma}(x_2, x_2') \ge d_1(\Psi^{-1}(x_2), \Psi^{-1}(x_2')) \ge \Delta_{\mathrm{Q}}(\Psi^{-1}(x_2))\,,$$

and hence $c\Delta_{\mathrm{P}}^{\gamma}(x_2) \ge \Delta_{\mathrm{Q}}(\Psi^{-1}(x_2))$. Repeating the argument above, we further see that this inequality hold not only for x_2 satisfying $\eta_{\mathrm{P}}(x_2) < 1/2$ but actually for all $x_2 \in \Psi(X_1)$. Combining this with our previous considerations and (8.5), we thus obtain

$$\begin{aligned}
\mathrm{P}_X\bigl(\{x_2 \in X_2 : \Delta_{\mathrm{P}}(x_2) < t\}\bigr) &= \mathrm{P}_X\bigl(\{x_2 \in \Psi(X_1) : \Delta_{\mathrm{P}}(x_2) < t\}\bigr) \\
&\le C\, \mathrm{Q}_X\bigl(\{x_1 \in X_1 : \Delta_{\mathrm{P}}(\Psi(x_1)) < t\}\bigr) \\
&\le C\, \mathrm{Q}_X\bigl(\{x_1 \in X_1 : \Delta_{\mathrm{Q}}(x_1) < ct^{\gamma}\}\bigr)\,.
\end{aligned}$$

Using the margin exponent of Q, we then find the assertion. □

Note that if the map Ψ in the lemma above is continuous and X_1 is a *complete* metric space, a simple Cauchy sequence argument combined with (8.4) shows that $\Psi(X_1)$ is a closed of subset X_2 and hence measurable.

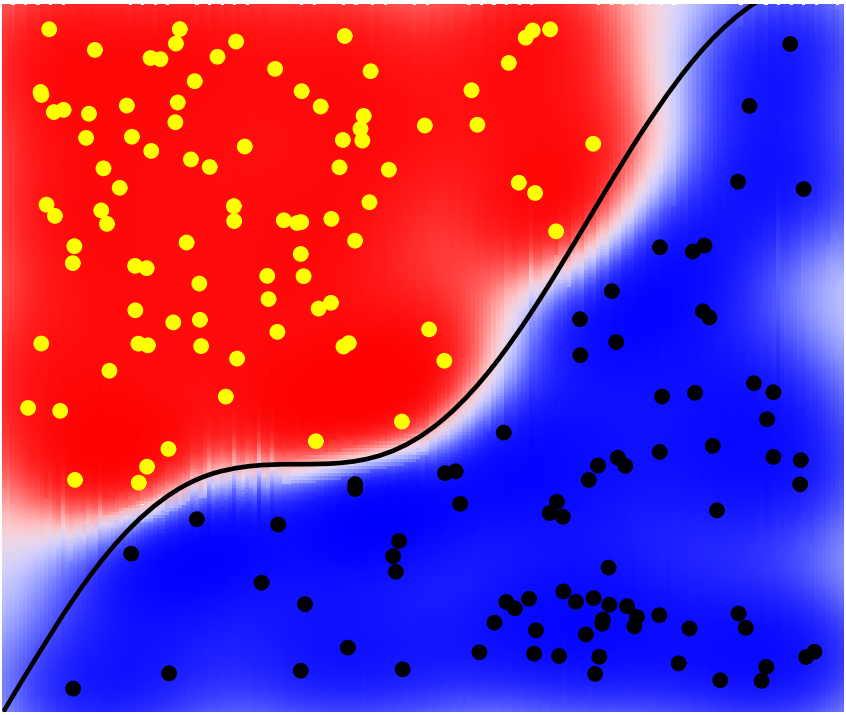


Fig. 8.1. Example of a distribution that has a large margin exponent. The red and blue areas are the negative and positive classes, respectively, and the decision boundary is located in the white area. The lighter colors indicate a low concentration of P_X, and hence only a few samples (indicated by yellow and black dots for positive and negative labels, respectively) can be found in this region. The black line shows an estimate of the decision boundary made by an SVM. Note that the larger the light area is, the larger the noise exponent is, and hence the better the SVM can estimate the true decision boundary. However, even in the absence of a light area, the corresponding distribution would have margin exponent $\alpha = 1$, if, for example, P_X had a bounded density.

If P_X is the image measure $\Psi(\mathrm{Q}_X)$ of Q_X under Ψ, then (8.5) becomes an equality with $C = 1$. However, this is by no means the only case where (8.5) can hold. For example, assume that P_X is only absolutely continuous with respect to $\Psi(\mathrm{Q}_X)$. Then (8.5) is obviously satisfied if the density of P_X with respect to $\Psi(\mathrm{Q}_X)$ is bounded. The next example illustrates such a situation.

Example 8.13 (Smooth transformations). Let $U, V \subset \mathbb{R}^d$ be two open non-empty sets and $\Psi : U \to V$ be a continuously differentiable map. Then recall that $\Psi(B)$ is measurable for all measurable $B \subset U$ and that a version of Sard's inequality (see, e.g., p. 313 in Floret, 1981) states that

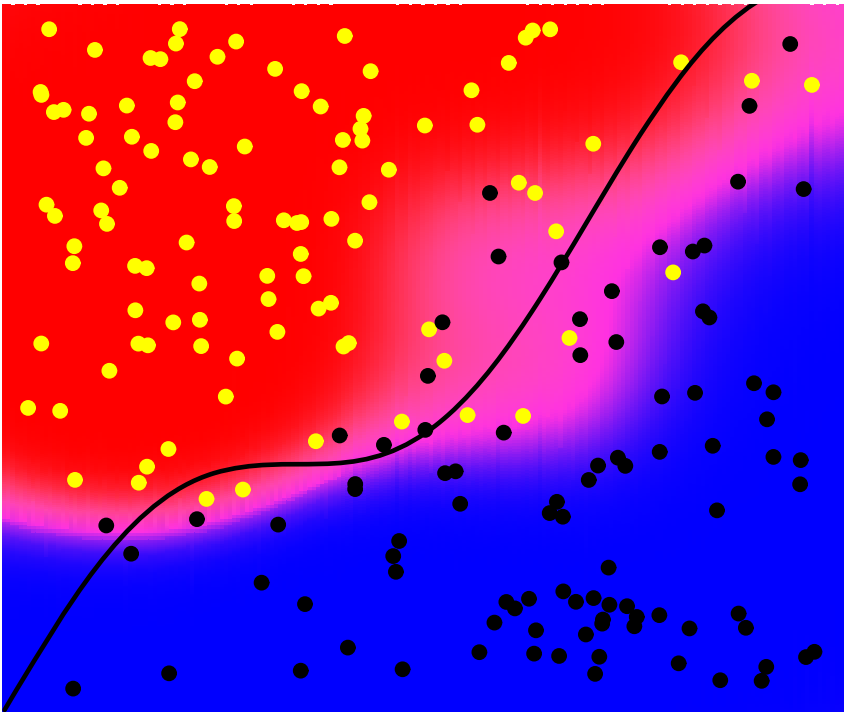


Fig. 8.2. Example of a distribution that has a large amount of noise around its decision boundary. The red and blue areas are the negative and positive classes, respectively, and the mixture of these colors indicates the amount of noise. Consequently, we see nearby samples with different labels (indicated by yellow and black dots for positive and negative labels, respectively) in this region. The black line shows an estimate of the decision boundary made by an SVM, whereas the true decision boundary is located in the middle of the purple region. Note that the larger the purple area around the decision boundary is, the larger the margin-noise exponent is. However, even there was no purple area the corresponding distribution would have margin-noise exponent $\alpha = 1$, if, for example, P_X had a bounded density.

$$\mathrm{vol}_d(\Psi(B)) \leq \int_B |\det \Psi'(x)| dx\,. \tag{8.7}$$

Let us further assume that Q is a distribution on $U \times Y$, where $Y := \{-1, 1\}$, such that $\operatorname{supp} \mathrm{Q}_X$ is bounded and $\mathrm{vol}_d(\Psi(\operatorname{supp} \mathrm{Q}_X)) > 0$. Then (8.7) yields $\mathrm{vol}_d(\operatorname{supp} \mathrm{Q}_X) > 0$. Let us assume for the sake of simplicity that Q_X is the uniform distribution on $\operatorname{supp} \mathrm{Q}_X$ and that P is a distribution on $V \times Y$ such that P_X is the uniform distribution on the compact set $\Psi(\operatorname{supp} \mathrm{Q}_X)$. Then (8.7) yields

$$\begin{aligned} \mathrm{P}_X(A) &\leq \frac{\mathrm{vol}_d(\operatorname{supp}\mathrm{Q}_X)}{\mathrm{vol}_d(\Psi(\operatorname{supp}\mathrm{Q}_X))} \int_U \mathbf{1}_A(\Psi(x))\,|\det \Psi'(x)|\mathrm{Q}_X(x) \\ &\leq \frac{\mathrm{vol}_d(\operatorname{supp}\mathrm{Q}_X)}{\mathrm{vol}_d(\Psi(\operatorname{supp}\mathrm{Q}_X))} \sup_{x\in\operatorname{supp}\mathrm{Q}_X} |\det \Psi'(x)| \cdot \mathrm{Q}_X(\Psi^{-1}(A)) \end{aligned}$$

for all measurable $A \subset V$. Since $\operatorname{supp}\mathrm{Q}_X$ is compact and $x \mapsto |\det \Psi'(x)|$ is continuous, we hence obtain (8.5).

Let us finally assume that Ψ is a diffeomorphism, i.e., Ψ is also bijective and $\Psi^{-1} : V \to U$ is continuously differentiable. If $V = \Psi(U)$ is convex, then the mean value theorem states that

$$\left\| \Psi^{-1}(x) - \Psi^{-1}(x') \right\|_2 \;\leq\; \sup_{t\in[0,1]} \left\| (\Psi^{-1})'(tx + (1-t)x') \right\|_2 \cdot \|x - x'\|_2$$

for all $x, x' \in V$. Since $\overline{\operatorname{co}\Psi(\operatorname{supp}\mathrm{Q}_X)} \subset V$ is compact, we thus obtain (8.4) for $\gamma := 1$, $X_1 := \operatorname{supp}\mathrm{Q}_X$, and $X_2 := \Psi(\operatorname{supp}\mathrm{Q}_X)$. Consequently, P has margin exponent α if Q has margin exponent α and (8.6) is satisfied. ◁

Our next goal is to show that small values of densities in the vicinity of the decision boundary improve the margin exponent. To this end, we say that the distributions P and Q on $X \times Y$ **generate the same classes** for the versions η_P and η_Q of their posterior probabilities of P and Q if $(2\eta_\mathrm{P} - 1)(2\eta_\mathrm{Q} - 1) > 0$. Obviously, in this case, the associated classes do coincide and hence the associated distances Δ_P and Δ_Q to the decision boundaries are equal. For such distributions, we can now prove the following lemma.

Lemma 8.14 (Low densities near the decision boundary). *Let (X, d) be a metric space and* Q *and* P *be two distributions on $X \times Y$ that generate the same classes for their posterior probabilities η_P and η_Q. We write Δ_Q for the associated distance to the decision boundary of* Q*. Furthermore, assume that P_X has a density $h : X \to [0,\infty)$ with respect to Q_X such that there exist constants $c > 0$ and $\gamma \in [0,\infty)$ satisfying*

$$h(x) \;\leq\; c\Delta_\mathrm{Q}^\gamma(x)\,, \qquad\qquad x \in X. \tag{8.8}$$

If Q *has margin exponent $\alpha \in [0,\infty)$ for Δ_Q, then* P *has margin exponent $\alpha + \gamma$.*

Proof. We have already seen before this lemma that there is a version Δ_P of the distance to the decision boundary of P that equals Δ_Q. For $t \geq 0$, we consequently obtain

$$\begin{aligned} \mathrm{P}_X\big(\{x \in X : \Delta_\mathrm{P}(x) < t\}\big) &= \int_{\Delta_\mathrm{Q}(x)<t} h(x)\, d\mathrm{Q}_X(x) \\ &\leq c\, t^\gamma\, \mathrm{Q}_X\big(\{x \in X : \Delta_\mathrm{Q}(x) < t\}\big)\,. \end{aligned}$$

From this we immediately obtain the assertion. □

For classifying with the hinge loss, we will see below that it is more suitable to measure the size of $\{x \in X : \Delta(x) < t\}$ by $|2\eta - 1|\mathrm{P}_X$ instead of P_X. This motivates the following definition.

Definition 8.15. *Let (X, d) be a metric space and P be a distribution on $X \times Y$. We say that P has* ***margin-noise exponent*** *$\beta \in [0, \infty)$ for the version $\eta : X \to [0, 1]$ of its posterior probability if there exists a constant $c \geq 1$ such that the distance to the decision boundary Δ associated to η satisfies*

$$\int_{\Delta(x)<t} |2\eta(x) - 1| \, d\mathrm{P}_X(x) \;\leq\; c\, t^{\beta}\,, \qquad\qquad t \geq 0. \tag{8.9}$$

Let us now investigate the relation between the margin exponent and the margin-noise exponent. Lemma 8.14 suggests that to this end we need to describe how the distance to the decision boundary influences the amount of noise. This is done in the following definition.

Definition 8.16. *Let (X, d) be a metric space, P be a distribution on $X \times Y$, and $\eta : X \to [0, 1]$ be a version of its posterior probability. We say that the associated* ***distance to the decision boundary Δ controls the noise by the exponent*** *$\gamma \in [0, \infty)$ if there exists a constant $c > 0$ such that*

$$|2\eta(x) - 1| \;\leq\; c\Delta^{\gamma}(x) \tag{8.10}$$

for P_X-almost all $x \in X$.

Note that since $|2\eta(x) - 1| \leq 1$ for all $x \in X$, condition (8.10) becomes trivial whenever $\Delta(x) \geq c^{-1/\gamma}$. Consequently, (8.10) is a condition that only considers points $x \in X$ with sufficiently small distance to the opposite class. In simple words, it states that $\eta(x)$ is close to the level $1/2$ of "complete noise" if x approaches the decision boundary. The following lemma, whose omitted proof is almost identical to that of Lemma 8.14, now relates the margin exponent to the margin-noise exponent.

Lemma 8.17. *Let X be a metric space and P be a distribution on $X \times Y$ that has margin exponent $\alpha \in [0, \infty)$ for the version η of its posterior probability. Assume that the associated distance to the decision boundary Δ controls the noise by the exponent $\gamma \in [0, \infty)$. Then P has margin-noise exponent $\alpha + \gamma$.*

Note that for $\alpha = 0$ the preceding lemma states that every distribution whose distance to the decision boundary controls the noise by some exponent $\gamma > 0$ has margin-noise exponent γ. In other words, distributions that have a high amount of noise around their decision boundary have a non-trivial margin-noise exponent. We refer to Figure 8.2 for an illustration of this situation. In addition, note that in the case $\gamma = 0$ the lemma states that distributions having some margin exponent $\alpha > 0$ also have margin-noise exponent α. Consequently, all considerations on the margin exponent made so

far have an immediate consequence for the margin-noise exponent. Finally, the general message of Lemma 8.17 is that for distributions satisfying both assumptions their exponents add up to become the margin-noise exponent.

With the help of the concepts above, we can now return to our initial goal of estimating the approximation error function for Gaussian kernels.

Theorem 8.18 (Approximation error of Gaussian kernels). *Let L be the hinge loss and P be a distribution on $\mathbb{R}^d \times \{-1,1\}$ that has margin-noise exponent $\beta \in (0,\infty)$ and whose marginal distribution P_X has tail exponent $\tau \in (0,\infty]$. Then there exist constants $c_{d,\tau} > 0$ and $\tilde{c}_{d,\beta} > 0$ such that for all $\gamma > 0$ and all $\lambda > 0$ there exists a function $f^* \in H_\gamma(\mathbb{R}^d)$ in the RKHS $H_\gamma(\mathbb{R}^d)$ of the Gaussian RBF kernel k_γ such that $\|f^*\|_\infty \le 1$ and*

$$\lambda \|f^*\|^2_{H_\gamma(\mathbb{R}^d)} + \mathcal{R}_{L,\mathrm{P}}(f^*) - \mathcal{R}^*_{L,\mathrm{P}} \le c_{d,\tau}\, \lambda^{\frac{\tau}{d+\tau}}\, \gamma^{-\frac{d\tau}{d+\tau}} \,+\, \tilde{c}_{d,\beta}\, c\, \gamma^\beta\,,$$

where c is the constant appearing in (8.9). In particular, we have

$$A_2^{(\gamma)}(\lambda) \;\le\; \max\bigl\{c_{d,\tau}, \tilde{c}_{d,\beta}\, c\bigr\} \cdot \bigl(\lambda^{\frac{\tau}{d+\tau}} \gamma^{-\frac{d\tau}{d+\tau}} + \gamma^\beta\bigr)\,.$$

Proof. Obviously, it suffices to prove the first assertion. Let $\eta : \mathbb{R}^d \to [0,1]$ be a version of the posterior probability of P such that its associated $\Delta(x)$, $x \in \mathbb{R}^d$, satisfies (8.9). We define X_{-1} and X_1 as in Definition 8.5 and further use the shorthand $B := \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ for the closed unit ball of $\mathbb{R}^d$. Let us fix a $\rho > 0$ such that $X_{-1} \cap \rho B \ne \emptyset$ and $X_1 \cap \rho B \ne \emptyset$. Then an easy consideration shows that $\Delta(x) \le 2\rho$ for all $x \in \rho B$. Finally, we define $f_\rho : \mathbb{R}^d \to [-1,1]$ by

$$f_\rho(x) := \mathbf{1}_{(X_{-1} \cup X_1) \cap 3\rho B}(x) \cdot \operatorname{sign}\bigl(2\eta(x) - 1\bigr)\,, \qquad\qquad x \in \mathbb{R}^d,$$

and $g_\rho : \mathbb{R}^d \to \mathbb{R}$ by $g_\rho := (\pi\gamma^2)^{-d/4} f_\rho$. Obviously, we have $g_\rho \in L_2(\mathbb{R}^d)$ with

$$\|g_\rho\|_{L_2(\mathbb{R}^d)} \le \left(\frac{9\rho^2}{\pi\gamma^2}\right)^{\frac{d}{4}} \sqrt{\mathrm{vol}_d(B)}\,, \tag{8.11}$$

where $\mathrm{vol}_d(B)$ denotes the volume of B. Let us now recall Lemma 4.45 and (4.43), which showed that $L_2(\mathbb{R}^d)$ is a feature space of $H_\gamma(\mathbb{R}^d)$ with canonical metric surjection $V_\gamma : L_2(\mathbb{R}^d) \to H_\gamma(\mathbb{R}^d)$ given by

$$V_\gamma g(x) = \left(\frac{4}{\pi\gamma^2}\right)^{\frac{d}{4}} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-x'\|_2^2}\, g(x')dx'\,, \qquad\qquad g \in L_2(\mathbb{R}^d),\, x \in \mathbb{R}^d.$$

By Theorem 4.21 and (8.11), we then obtain

$$\|V_\gamma g_\rho\|_{H_\gamma(\mathbb{R}^d)} \le \left(\frac{9\rho^2}{\pi\gamma^2}\right)^{\frac{d}{4}} \sqrt{\mathrm{vol}_d(B)}\,. \tag{8.12}$$

Moreover, $g_\rho = (\pi\gamma^2)^{-d/4} f_\rho$ together with $\|f_\rho\|_\infty \le 1$ and (A.3) implies

$$\left|V_\gamma g_\rho(x)\right| \le \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' = 1$$

for all $x \in \mathbb{R}^d$. Therefore, Theorem 2.31 yields

$$\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(V_\gamma g_\rho) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \;=\; \mathbb{E}_{\mathrm{P}_X}\big(\,|V_\gamma g_\rho - f^*_{L_{\text{class}},\mathrm{P}}| \cdot |2\eta - 1|\,\big). \quad (8.13)$$

In order to bound $|V_\gamma g_\rho - f^*_{L_{\text{class}},\mathrm{P}}|$ we fix an $x \in X_1 \cap \rho B$. Furthermore, we write $B(x,r) := \{x' \in \mathbb{R}^d : \|x-x'\|_2 < r\}$ for the *open* ball with radius r and center x. For $x' \in B(x, \Delta(x))$, we then have $\|x - x'\|_2 < \Delta(x)$, and thus we obtain $x' \in X_1$. Moreover, we also have $\|x'\|_2 \le \|x-x'\|_2 + \|x\|_2 < \Delta(x) + \rho \le 3\rho$, and therefore we conclude that

$$B(x, \Delta(x)) \subset X_1 \cap 3\rho B\,.$$

Similarly, for $x' \in X_{-1}$, we have $\Delta(x) \le \|x - x'\|_2$, and hence we obtain

$$X_{-1} \cap 3\rho B \;\subset\; X_{-1} \;\subset\; \mathbb{R}^d \backslash B(x, \Delta(x))\,.$$

Combining the definition of f with these two inclusions and (A.3) now yields

$$\begin{aligned}
V_\gamma g_\rho(x) &= \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-x'\|_2^2} f_\rho(x')dx' \\
&= \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \left(\int_{X_1 \cap 3\rho B} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' - \int_{X_{-1}\cap 3\rho B} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' \right) \\
&\ge \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \left(\int_{B(x,\Delta(x))} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' - \int_{\mathbb{R}^d\backslash B(x,\Delta(x))} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' \right) \\
&= 2\left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{B(x,\Delta(x))} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' - 1\,.
\end{aligned}$$

Since $V_\gamma g_\rho(x) \le 1$ and $f^*_{L_{\text{class}},\mathrm{P}}(x) = 1$, we thus obtain

$$\begin{aligned}
|V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| &= 1 - V_\gamma g_\rho(x) \\
&\le 2 - 2\left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{B(x,\Delta(x))} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' \\
&= 2 - 2\left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{B(0,\Delta(x))} e^{-2\gamma^{-2}\|x'\|_2^2} dx'\,.
\end{aligned}$$

Using the rotation invariance of $x' \mapsto e^{-2\gamma^{-2}\|x'\|_2^2}$ and $\Gamma(1+t) = t\Gamma(t)$, $t > 0$, we further find

$$\left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int\limits_{B(0,\Delta(x))} e^{-2\gamma^{-2}\|x'\|_2^2} dx' = \frac{d}{\Gamma(1+d/2)} \left(\frac{2}{\gamma^2}\right)^{\frac{d}{2}} \int_0^{\Delta(x)} e^{-2\gamma^{-2}r^2} r^{d-1} dr$$

$$= \frac{2}{\Gamma(d/2)} \int_0^{\sqrt{2}\Delta(x)\gamma^{-1}} e^{-r^2} r^{d-1} dr$$

$$= \frac{1}{\Gamma(d/2)} \int_0^{2\Delta^2(x)\gamma^{-2}} e^{-r} r^{d/2-1} dr\,.$$

Combining this equation with the previous estimate yields

$$|V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \le 2 - \frac{2}{\Gamma(d/2)} \int_0^{2\Delta^2(x)\gamma^{-2}} e^{-r} r^{d/2-1} dr$$

$$= \frac{2}{\Gamma(d/2)} \left(\int_0^\infty e^{-r} r^{d/2-1} dr - \int_0^{2\Delta^2(x)\gamma^{-2}} e^{-r} r^{d/2-1} dr \right)$$

$$= \frac{2}{\Gamma(d/2)} \int_0^\infty \mathbf{1}_{(2\Delta^2(x)\gamma^{-2},\infty)}(r)\, e^{-r} r^{d/2-1} dr$$

for all $x \in X_1 \cap \rho B$. Moreover, repeating the proof above for $x \in X_{-1} \cap \rho B$, we see that this estimate actually holds for all $x \in (X_{-1} \cup X_1) \cap \rho B$. Since $|2\eta(x) - 1| = 0$ for $x \notin X_{-1} \cup X_1$, we thus find

$$\begin{aligned}
&\int_{\rho B} |V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \cdot |2\eta(x) - 1|\, d\mathrm{P}_X(x) \\
&\le \frac{2}{\Gamma(d/2)} \int_{X_{-1}\cup X_1} \int_0^\infty \mathbf{1}_{(2\Delta^2(x)\gamma^{-2},\infty)}(r)\, e^{-r} r^{d/2-1} |2\eta(x) - 1|\, dr\, d\mathrm{P}_X(x) \\
&= \frac{2}{\Gamma(d/2)} \int_0^\infty e^{-r} r^{d/2-1} \int_{\mathbb{R}^d} \mathbf{1}_{[0,\gamma(r/2)^{1/2})}(\Delta(x)) \cdot |2\eta(x) - 1|\, d\mathrm{P}_X(x)\, dr \\
&\le \frac{2^{1-\beta/2}\, c\, \gamma^\beta}{\Gamma(d/2)} \int_0^\infty e^{-r} r^{(\beta+d)/2-1} dr \\
&= \frac{2^{1-\beta/2}\, c\, \Gamma((\beta+d)/2)}{\Gamma(d/2)}\, \gamma^\beta
\end{aligned}$$

by the definition of the margin-noise exponent. Using equation (8.13) together with $\|V_\gamma g_\rho\|_\infty \le 1$ and the tail exponent inequality (7.61) thus yields

$$\begin{aligned}
&\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(V_\gamma g_\rho) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \\
&\le \int_{\rho B} |V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \cdot |2\eta(x) - 1|\, d\mathrm{P}_X(x) + 2\mathrm{P}_X(\mathbb{R}^d \setminus \rho B) \\
&\le \frac{2^{1-\beta/2}\, c\, \Gamma((\beta+d)/2)}{\Gamma(d/2)}\, \gamma^\beta + 2\rho^{-\tau}\,. \qquad (8.14)
\end{aligned}$$

By combining this estimate with (8.12), we therefore obtain

$$\begin{aligned}
&\lambda \|V_\gamma g_\rho\|_{H_\gamma}^2 + \mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(V_\gamma g_\rho) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \\
&\leq \lambda \Big(\frac{9\rho^2}{\pi\gamma^2}\Big)^{\frac{d}{2}} \operatorname{vol}_d(B) + \frac{2^{1-\beta/2}\, c\, \Gamma((\beta+d)/2)}{\Gamma(d/2)}\, \gamma^\beta + 2\rho^{-\tau}\,. \qquad (8.15)
\end{aligned}$$

So far, we have found a g_ρ satisfying (8.15) only if ρ satisfies both $X_{-1} \cap \rho B \neq \emptyset$ and $X_1 \cap \rho B \neq \emptyset$. Our next goal is to find such a g_ρ for the remaining $\rho > 0$. To this end, let us first consider a $\rho > 0$ such that $X_{-1} \cap \rho B = \emptyset$ and $X_1 \cap \rho B = \emptyset$. For such ρ, we set $f_\rho := g_\rho := 0$, so that (8.12) and (8.13) are trivially satisfied. Moreover, for $x \in \rho B$, our assumption on ρ guarantees $x \notin X_{-1} \cup X_1$, which in turn yields $|2\eta(x) - 1| = 0$. Consequently, we find

$$\begin{aligned}
&\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(V_\gamma g_\rho) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \\
&\leq \int_{\rho B} |V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \cdot |2\eta(x) - 1|\, d\mathrm{P}_X(x) + 2\mathrm{P}_X(\mathbb{R}^d \backslash \rho B) \\
&\leq 2\rho^{-\tau}\,,
\end{aligned}$$

and hence (8.15) turns out to be true for the ρ and g_ρ considered. Let us now consider a $\rho \geq 1$ such that $X_{-1} \cap \rho B = \emptyset$ and $X_1 \cap \rho B \neq \emptyset$. In this case, we define $f_\rho := \mathbf{1}_{2\rho B}$ and $g_\rho := (\pi\gamma^2)^{-d/4} f_\rho$. Then (8.12) and (8.13) are obviously satisfied. Moreover, for $x \in \rho B$, we have $B(x,\rho) \subset 2\rho B$, and hence repeating our previous calculations yields

$$\begin{aligned}
V_\gamma g_\rho(x) = \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{2\rho B} e^{-2\gamma^{-2}\|x-y\|_2^2} dy &\geq \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{B(0,\rho)} e^{-2\gamma^{-2}\|y\|_2^2} dy \\
&= \frac{1}{\Gamma(d/2)} \int_0^{2\rho^2\gamma^{-2}} e^{-r} r^{d/2-1} dr\,.
\end{aligned}$$

For $x \in X_1 \cap \rho B$, it is then easy to conclude that

$$\begin{aligned}
|V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| &\leq 1 - \frac{1}{\Gamma(d/2)} \int_0^{2\rho^2\gamma^{-2}} e^{-r} r^{d/2-1} dr \\
&= \frac{1}{\Gamma(d/2)} \int_{2\rho^2\gamma^{-2}}^{\infty} e^{-r} r^{d/2-1} dr \\
&\leq \frac{2^{-\beta/2}\, \Gamma((\beta+d)/2)}{\Gamma(d/2)}\, \gamma^\beta\,,
\end{aligned}$$

where in the last step we used the last estimate of Lemma A.1.1. Since the constant c in (8.9) is assumed to be not smaller than 1, we then conclude that (8.15) holds. Finally, if $\rho \in (0,1)$, the latter inequality is satisfied for $g_\rho := 0$, and consequently we have found for all $\rho > 0$ a function g_ρ such that (8.15) holds. Minimizing (8.15) with respect to ρ now yields the assertion. □

Having a bound on the approximation error, we can now derive learning rates for λ_n and γ_n chosen *a priori* with the help of Theorem 8.1 and Theorem 8.3. However, it turns out that the optimal rates of such an approach

require knowledge about the distribution, namely the margin-noise exponent and the tail exponent. Since in practice such knowledge is not available, we skip the derivation of these rates and focus directly on data-dependent adaptive parameter selection strategies. The strategy we will focus on is a simple modification of the TV-SVM considered in Section 6.5. This modification determines not only the regularization parameter λ by a grid but also the kernel parameter γ. As in Section 6.5, we therefore need to know how much the approximation error function is affected by small changes of λ and γ. This is investigated in the following corollary.

Corollary 8.19. *Let* P *be a distribution on* $\mathbb{R}^d \times \{-1,1\}$ *that has margin-noise exponent* $\beta \in (0,\infty)$ *and whose* P_X *has tail exponent* $\tau \in (0,\infty]$*. Moreover, for* $\varepsilon > 0$ *and* $\delta > 0$*, we fix a finite* ε*-net* $\Lambda \subset (0,1]$ *and a finite* δ*-net* $\Gamma \subset (0,1]$*, respectively. Then, for all* $p > 0$*,* $q \geq 0$*, and all* $x \in (0,1]$*, we have*

$$\min_{\lambda\in\Lambda,\,\gamma\in\Gamma} \left(A_2^{(\gamma)}(\lambda) + x\lambda^{-p}\gamma^{-q} \right) \leq c \left(x^{\frac{\beta\tau}{\beta\tau+d\beta p+\beta p\tau+dp\tau+q\tau}} + \varepsilon^{\frac{\tau}{d+\tau}} + \delta^{\beta} \right),$$

where $c \geq 1$ *is a constant independent of* x*,* Λ*,* ε*,* Γ*, and* δ*.*

The proof of the preceding corollary actually yields a particular expression for the constant c, but since this expression is extremely complicated, we decided to omit the details.

Proof. Without loss of generality, we may assume that Λ and Γ are of the form $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$ and $\Gamma = \{\gamma_1, \ldots, \gamma_\ell\}$ with $\lambda_{i-1} < \lambda_i$ and $\gamma_{j-1} < \gamma_j$ for all $i = 2, \ldots, m$ and $\gamma = 2, \ldots, \ell$, respectively. Moreover, we fix a minimizer (λ^*, γ^*) of the function $(\lambda, \gamma) \mapsto \lambda^{\frac{\tau}{d+\tau}} \gamma^{-\frac{d\tau}{d+\tau}} + \gamma^\beta + x\lambda^{-p}\gamma^{-q}$ defined on $[0,1]^2$. Analogously to the proof of Lemma 6.30, we then see that there exist indexes $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, \ell\}$ such that $\lambda^* \leq \lambda_i \leq \lambda^* + 2\varepsilon$ and $\gamma^* \leq \gamma_j \leq \gamma^* + 2\delta$. By Theorem 8.18, we then obtain

$$\begin{aligned}
&\min_{\lambda\in\Lambda,\,\gamma\in\Gamma} \left(A_2^{(\gamma)}(\lambda) + x\lambda^{-p}\gamma^{-q} \right) \\
&\leq A_2^{(\gamma_j)}(\lambda_i) + x\lambda_i^{-p}\gamma_j^{-q} \\
&\leq c_1 \left(\lambda_i^{\frac{\tau}{d+\tau}} \gamma_j^{-\frac{d\tau}{d+\tau}} + \gamma_j^{\beta} \right) + x\,(\lambda^*)^{-p}(\gamma^*)^{-q} \\
&\leq c_1 \left((\lambda^* + 2\varepsilon)^{\frac{\tau}{d+\tau}} (\gamma^*)^{-\frac{d\tau}{d+\tau}} + (\gamma^* + 2\delta)^{\beta} \right) + x\,(\lambda^*)^{-p}(\gamma^*)^{-q} \\
&\leq c_2 \left((\lambda^*)^{\frac{\tau}{d+\tau}} (\gamma^*)^{-\frac{d\tau}{d+\tau}} + (\gamma^*)^{\beta} + x\,(\lambda^*)^{-p}(\gamma^*)^{-q} + \varepsilon^{\frac{\tau}{d+\tau}} + \delta^{\beta} \right) \\
&= c_2 \min_{\lambda,\gamma\in[0,1]} \left(\lambda^{\frac{\tau}{d+\tau}} \gamma^{-\frac{d\tau}{d+\tau}} + \gamma^{\beta} + x\lambda^{-p}\gamma^{-q} \right) + c_2\varepsilon^{\frac{\tau}{d+\tau}} + c_2\delta^{\beta},
\end{aligned}$$

where in the second to last step we used $\gamma^* \leq 1$, and where c_1 and c_2 are suitable constants independent of x, Λ, ε, Γ, and δ. Now the assertion follows from Lemma A.1.6. □

Let us now introduce the announced modification of the TV-SVM.

Definition 8.20. *Let $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be sequences of finite subsets $\Lambda_n, \Gamma_n \subset (0,1]$. For $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^d \times \{-1,1\})^n$, we define*

$$D_1 := ((x_1, y_1), \ldots, (x_m, y_m))$$
$$D_2 := ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n)),$$

where $m := \lfloor n/2 \rfloor + 1$ and $n \geq 3$. Then we use D_1 as a training set by computing the SVM decision functions

$$f_{\mathrm{D}_1,\lambda,\gamma} := \arg\min_{f \in H_\gamma} \lambda \|f\|_{H_\gamma}^2 + \mathcal{R}_{L_{\mathrm{hinge}},\mathrm{D}_1}(f), \qquad (\lambda,\gamma) \in \Lambda_n \times \Gamma_n,$$

and use D_2 to determine (λ,γ) by choosing a $(\lambda_{D_2}, \gamma_{D_2}) \in \Lambda_n \times \Gamma_n$ such that

$$\mathcal{R}_{L_{\mathrm{class}},\mathrm{D}_2}(f_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}) = \min_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L_{\mathrm{class}},\mathrm{D}_2}(f_{\mathrm{D}_1,\lambda,\gamma}).$$

A learning method that produces such decision functions $f_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}$ for all $D \in (X \times Y)^n$ is called a ***training validation support vector machine (TV-SVM)*** *with respect to Λ and Γ.*

Note that in the parameter selection step we consider the classification loss, although in principle we could have also used the hinge loss. Since the classification risk only considers the sign of a decision function and not its values, it is not necessary to clip $f_{\mathrm{D}_1,\lambda,\gamma}$ in this parameter selection step. Finally note that in general the pair $(\lambda_{D_2}, \gamma_{D_2})$, and thus also $f_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}$, is not uniquely determined.

The existence of a measurable TV-SVM with respect to Λ and Γ can be shown by elementary modifications of the proof of Lemma 6.29, which established the measurability of TV-SVMs with respect to Λ. We omit the details and leave the proof to the interested reader. Moreover, the oracle inequalities established in Theorems 6.31 and 6.32 can easily be adapted to the new TV-SVM, too. Again, we skip the details and only present the resulting consistency and learning rates.

Theorem 8.21. *Let $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be sequences of finite subsets $\Lambda_n, \Gamma_n \subset (0,1]$ such that Λ_n is an $n^{-1/2}$-net of $(0,1]$ and Γ_n is an $n^{-1/(2d)}$-net of $(0,1]$, respectively. Furthermore, assume that the cardinalities $|\Lambda_n|$ and $|\Gamma_n|$ grow polynomially in n. Then every measurable TV-SVM with respect to Λ and Γ is universally consistent. Moreover, for distributions on $\mathbb{R}^d \times \{-1,1\}$ that have margin-noise exponent $\beta \in (0,\infty)$ and whose P_X have tail exponent $\tau \in (0,\infty]$, such TV-SVMs learn with rate $n^{-\gamma}$ where*

$$\gamma := \begin{cases} \frac{\beta\tau}{4\beta\tau + 2d\beta + 2d\tau} & \text{if } \tau < \infty \\ \frac{\beta}{3\beta + 2d} + \rho & \text{if } \tau = \infty \end{cases}$$

and $\rho > 0$ is an arbitrarily small number.

Proof. Since the proof closely follows the lines of the proofs of Theorem 6.31 and 6.32, we only mention the main steps. To show the consistency, we need a simple modification of Theorem 6.32 to address the second set of parameter candidates Γ_n and the fact that $\lim_{\lambda \to 0} A_2^{(\gamma)}(\lambda) = 0$ for any fixed γ. Moreover, the first rate follows from combining such a modification of Theorem 6.32 with Corollary 8.19. For the last rate, we first establish an oracle inequality for the TV-SVM using Theorem 8.3 and a simple adaptation of the proof of Theorem 6.31. This oracle inequality is then combined with Corollary 8.19. □

8.3 Advanced Concentration Results for SVMs (*)

In this section, we improve the learning rates obtained in the previous section with the help of the advanced concentration results of Chapter 7. To this end, let us first recall that one of the key ingredients of that chapter was a *variance bound*. Consequently, our first goal is to establish such a bound for the hinge loss. We begin with the following definition.

Definition 8.22. *A distribution* P *on* $X \times \{-1, 1\}$ *is said to have* ***noise exponent*** $q \in [0, \infty]$ *if there exists a constant* $c > 0$ *such that*

$$\mathrm{P}_X\big(\{x \in X : |2\eta(x) - 1| < t\}\big) \;\leq\; (c\,t)^q\,, \qquad\qquad t \geq 0. \tag{8.16}$$

Note that we have a high amount of noise in the labeling process at $x \in X$ if $\eta(x)$ is close to $1/2$, i.e., if $|2\eta(x) - 1|$ is close to 0. Consequently, the noise exponent measures the size of the set of points that have a high noise in the labeling process. Obviously, every distribution has noise exponent $q = 0$, whereas noise exponent $q = \infty$ means that η is bounded away from the critical level $1/2$. In particular, noise-free distributions, i.e., distributions P with $\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}} = 0$, have noise exponent $q = \infty$. Moreover, if P has noise exponent q, then P also has noise exponent q' for all $q' < q$. In addition, note that the noise exponent does *not* locate the points x having high noise, i.e., it does not consider their closeness to the decision boundary. Finally, it is important to note that, unlike the concepts we considered in Section 8.2, the noise exponent does *not* depend on a specific version of the posterior probability. This makes it technically easier to combine the noise exponent with the previously introduced concepts such as the margin-noise exponent. The next lemma, which relates the noise exponent to the margin-noise exponent, illustrates this.

Lemma 8.23 (Relation between noise exponents). *Let* (X, d) *be a metric space and* P *be a distribution on* $X \times Y$ *that has noise exponent* $q \in [0, \infty)$*. Furthermore, assume that there exists a version* η *of its posterior probability such that the associated distance to the decision boundary* Δ *controls the noise by the exponent* $\gamma \in [0, \infty)$ *in the sense of Definition 8.16. Then* P *has margin exponent* $\alpha := \gamma q$ *and margin-noise exponent* $\beta := \gamma(q + 1)$*.*

Proof. We have $\{x \in X : \Delta(x) < t\} \subset \{x \in X : |2\eta(x) - 1| < ct^\gamma\}$, where the inclusion is only P_X-almost surely and $c > 0$ is the constant appearing in (8.10). From this it is easy to conclude that P has margin exponent γq. Combining this with Lemma 8.17 yields the second assertion. □

Theorem 8.24 (Variance bound for the hinge loss). *Let* P *be a distribution on* $X \times Y$ *that has noise exponent* $q \in [0, \infty]$. *Moreover, let* $f^*_{L,\mathrm{P}} : X \to [-1, 1]$ *be a fixed Bayes decision function for the hinge loss* L. *Then, for all measurable* $f : X \to \mathbb{R}$, *we have*

$$\mathbb{E}_\mathrm{P}(L \circ \widehat{f} - L \circ f^*_{L,\mathrm{P}})^2 \;\le\; 6\, c^{q/(q+1)} \left(\mathbb{E}_\mathrm{P}(L \circ \widehat{f} - L \circ f^*_{L,\mathrm{P}})\right)^{q/(q+1)},$$

where c *is the constant appearing in (8.16).*

Proof. Since the range of $\widehat{f}$ is contained in $[-1, 1]$, we may restrict our considerations to functions $f : X \to [-1, 1]$. Now observe that for such f we have $L(y, f(x)) = 1 - yf(x)$ for all $x \in X$, $y = \pm 1$, and hence we obtain

$$\big(L(y, f(x)) - L(y, f^*_{L,\mathrm{P}}(x))\big)^2 = \big(yf^*_{L,\mathrm{P}}(x) - yf(x)\big)^2 = \big(f(x) - f^*_{L,\mathrm{P}}(x)\big)^2$$

for all $x \in X$, $y = \pm 1$. From this we conclude that

$$\begin{aligned}
\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}})^2 &= \int_X |f - f^*_{L,\mathrm{P}}|^2 \, d\mathrm{P}_X \\
&\le 2\int_{|2\eta-1|\ge s} |f - f^*_{L,\mathrm{P}}| \, d\mathrm{P}_X + 2\int_{|2\eta-1|<s} |f - f^*_{L,\mathrm{P}}| \, d\mathrm{P}_X \\
&\le 2s^{-1}\int_X |f - f^*_{L,\mathrm{P}}| \cdot |2\eta - 1| d\mathrm{P}_X + 4\mathrm{P}_X(|2\eta - 1| < s) \\
&\le 2s^{-1}\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}}) + 4(c\, s)^q
\end{aligned}$$

for all $s > 0$, where in the last step we used Theorem 2.31 and the margin-exponent inequality. Optimizing over s by Lemma A.1.5 together with the estimate $(q + 1)2^{1/(q+1)}q^{-q/(q+1)} \le 3$ now yields the assertion. □

By combining the variance bound above with the analysis of Chapter 7, we can now formulate an oracle inequality for SVMs using Gaussian kernels.

Theorem 8.25 (Improved oracle inequality for Gaussian kernels). *Let* P *be a distribution on* $\mathbb{R}^d \times \{-1, 1\}$ *that has margin-noise exponent* $\beta \in (0, \infty)$ *and noise exponent* $q \in [0, \infty]$ *and whose* P_X *has tail exponent* $\tau \in (0, \infty]$. *Then, for all* $\varepsilon > 0$ *and all* $d/(d + \tau) < p < 1$, *there exists a constant* $K \ge 1$ *such that, for all fixed* $\varrho \ge 1$, $n \ge 1$, $\lambda \in (0, 1]$, *and* $\gamma \in (0, 1]$, *the SVM using the hinge loss* L *and a Gaussian RBF kernel with parameter* γ *satisfies*

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \le K\left(\lambda^{\frac{\tau}{d+\tau}}\gamma^{-\frac{d\tau}{d+\tau}} + \gamma^\beta + \varrho\left(n\lambda^p\gamma^{(1-p)(1+\varepsilon)d}\right)^{-\frac{q+1}{q+2-p}}\right)$$

with probability P^n *not less than* $1 - 3e^{-\varrho}$.

Proof. By Theorem 8.24, we have a variance bound of the form (7.36) for $\vartheta = q/(q+1)$, and the supremum bound (7.35) is obviously satisfied for $B = 2$. In addition, Theorem 7.34 together with Corollary 7.31 yields a bound (7.48) on the entropy numbers for the constant

$$a := c_{\varepsilon,p}\, \gamma^{-\frac{(1-p)(1+\varepsilon)d}{2p}},$$

where $d/(d+\tau) < p < 1$, $\varepsilon > 0$, and $c_{\varepsilon,p}$ is a constant only depending on ε and p. Finally, we set $f_0 := f^*$, where f^* is the function Theorem 8.18 provides. For $B_0 := 2$, Theorem 7.23 then yields the assertion. □

With the help of the oracle inequality above, we can now establish learning rates for SVMs using Gaussian kernels. For brevity's sake, we only mention the learning rates for the TV-SVM, but it should be clear from our previous considerations on this method that these learning rates match the fastest learning rates one could derive from Theorem 8.25.

Theorem 8.26 (Learning rates using Gaussian kernels). *Let $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be sequences of finite subsets $\Lambda_n, \Gamma_n \subset (0,1]$ such that Λ_n is an n^{-1}-net of $(0,1]$ and Γ_n is an $n^{-1/d}$-net of $(0,1]$, respectively. Assume that the cardinalities of Λ_n and Γ_n grow polynomially in n. Then the TV-SVM with respect to Λ and Γ is universally consistent. Moreover, for distributions P on $\mathbb{R}^d \times \{-1,1\}$ that have margin-noise exponent $\beta \in (0,\infty)$ and noise exponent $q \in [0,\infty]$, and whose P_X have tail exponent $\tau \in (0,\infty]$, the TV-SVM learns with rate*

$$n^{-\frac{\beta\tau(d+\tau)(q+1)}{\beta\tau^2(q+2)+d(d\tau+\beta d+2\beta\tau+\tau^2)(q+1)}+\rho},$$

where $\rho > 0$ is an arbitrarily small number.

Proof. The consistency was already established in Theorem 8.21. Moreover, analogously to the proof of Theorem 6.31 and Theorem 7.24, we see that a simple union bound together with Theorem 8.25 shows for $m := \lfloor n/2 \rfloor + 1$ that

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \le K\left(\lambda^{\frac{\tau}{d+\tau}}\gamma^{-\frac{d\tau}{d+\tau}} + \gamma^{\beta} + \varrho\left(m\lambda^p\gamma^{(1-p)(1+\varepsilon)d}\right)^{-\frac{q+1}{q+2-p}}\right)$$

holds with probability P^m not less than $1 - 3|\Lambda_n| \cdot |\Gamma_n| e^{-\varrho}$ for all $\lambda \in \Lambda_n$ and $\gamma \in \Gamma_n$ simultaneously. As in the proof of Theorem 7.24, we then use Theorem 7.2 to deal with the parameter selection step, and combining both shows that with probability P^n not less than $1 - e^{-2\varrho}$ we have

$$\begin{aligned}
&\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \\
&\le 12K \inf_{\lambda\in\Lambda_n,\gamma\in\Gamma_n}\left(\lambda^{\frac{\tau}{d+\tau}}\gamma^{-\frac{d\tau}{d+\tau}} + \gamma^{\beta} + \varrho_n\left(n\lambda^p\gamma^{(1-p)(1+\varepsilon)d}\right)^{-\frac{q+1}{q+2-p}}\right) \\
&\quad + \tilde{K}\left(\frac{\left(\varrho + \ln(1+|\Lambda_n|\cdot|\Gamma_n|)\right)}{n}\right)^{\frac{q+1}{q+2}}, \qquad (8.17)
\end{aligned}$$

where $\varrho_n := \bigl(\varrho + \ln(1+3|\Lambda_n| \cdot |\Gamma_n|)\bigr)$ and $\tilde{K}$ is a suitable constant only depending on q and the constant c appearing in (8.16). Moreover, for all $\lambda \in \Lambda_n$ and $\gamma \in \Gamma_n$, we have

$$\left(\frac{\bigl(\varrho+\ln(1+|\Lambda_n|\cdot|\Gamma_n|)\bigr)}{n}\right)^{\frac{q+1}{q+2}} \leq \varrho_n \bigl(n\lambda^p \gamma^{(1-p)(1+\varepsilon)d}\bigr)^{-\frac{q+1}{q+2-p}},$$

and hence we may omit the last term in (8.17) if we replace $12K$ by $12K + \tilde{K}$. Repeating the proof of Corollary 8.19 and choosing p and ε sufficiently close to $d/(d+\tau)$ and 0, respectively, then yields the assertion after some simple yet tedious calculations. □

In order to illustrate the learning rates above, we assume in the following that P_X has tail exponent $\tau = \infty$. In this case, the learning rate reduces to

$$n^{-\frac{\beta(q+1)}{\beta(q+2)+d(q+1)}+\rho}, \tag{8.18}$$

where $\rho > 0$ is an arbitrarily small number. Motivated by the examples in Section 8.2, we first assume that P has margin exponent $\alpha := 1$. Moreover, we assume a moderate noise exponent $q := 1$, and by setting $\gamma := 1$ we also assume a moderate control of the noise by the distance to the decision boundary. By Lemma 8.17, these assumptions yield a margin-noise exponent $\beta = 2$, and hence (8.18) reduces to

$$n^{-\frac{2}{3+d}+\rho}.$$

Obviously, this rate is never faster than $n^{-1/2}$, and for high input dimensions it is actually substantially worse.

Let us now consider a different scenario that is less dimension dependent. To this end, we assume that there exists version η of the posterior probability such that the associated distance to the decision boundary Δ controls the noise by the exponent $\gamma \in [0,\infty)$. Lemma 8.23 then shows that $\beta = \gamma(q+1)$ and hence (8.18) reduces to

$$n^{-\frac{\gamma(q+1)}{\gamma(q+2)+d}+\rho}.$$

For large q or γ, this rate is obviously rather insensitive to the input dimension, and in particular for $q \to \infty$ we obtain rates that are close to the rate n^{-1}. Moreover, note that large q reflects both a small amount of noise and, by Lemma 8.23, a low concentration of P_X near the decision boundary.

8.4 Sparseness of SVMs Using the Hinge Loss

We have seen in Theorem 5.5 that using a convex loss function there exists a unique SVM solution $f_{\mathrm{D},\lambda}$, which, in addition, is of the form

$$f_{\mathrm{D},\lambda} = \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i)\,, \tag{8.19}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ is a suitable vector of coefficients, and $D = ((x_1, y_1), \ldots, (x_n, y_n))$. Obviously, only **support vectors**, i.e., samples x_i whose coefficients α_i are non-zero, need to be considered when evaluating $f_{\mathrm{D},\lambda}(x)$ by (8.19), and hence the number of support vectors has a direct influence on the time required to evaluate $f_{\mathrm{D},\lambda}(x)$. Moreover, we will see in Section 11.2 that this number also has a substantial influence on the training time. Consequently, it is important to know whether we can expect **sparse** decision functions, i.e., decision functions for which not all samples are support vectors. The goal of this section is to present some results that estimate the typical number of support vectors when using the hinge loss. In the following section, we will then extend these considerations to general convex, classification calibrated, margin-based loss functions.

Let us begin by considering the (quadratic) optimization problem

$$\begin{array}{lll} \text{minimize} & \lambda\|f\|_H^2 + \frac{1}{n}\sum\limits_{i=1}^{n} \xi_i & \text{for } f \in H,\, \xi \in \mathbb{R}^n \\ \text{subject to} & y_i f(x_i) \geq 1 - \xi_i, & i = 1, \ldots, n \\ & \xi_i \geq 0, & i = 1, \ldots, n\,. \end{array} \tag{8.20}$$

It is obvious that a pair $(f^*, \xi^*) \in H \times \mathbb{R}^n$ with $\xi_i^* = \max\{0, 1 - y_i f^*(x_i)\}$ is a solution of (8.20) if and only if $f^* = f_{\mathrm{D},\lambda}$. Consequently, one can solve (8.20) in order to find the SVM decision function. We will see in Chapter 11 that in practice this quadratic optimization problem is usually solved by considering the dual problem

$$\begin{array}{lll} \text{maximize} & \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{4\lambda}\sum\limits_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) & \text{for } \alpha \in \mathbb{R}^n \\ \text{subject to} & 0 \leq \alpha_i \leq \frac{1}{n}, & i = 1, \ldots, n\,, \end{array} \tag{8.21}$$

where we note that in Example 11.3 the primal problem will first be rescaled. Note that this rescaling results in the differently scaled but otherwise identical dual problem (11.15), and hence we can consider (8.21) instead of the problem (11.15). In order to establish a lower bound on the number of support vectors, we now have to briefly discuss the relation between (8.20) and (8.21). To this end, we write

$$f^{(\alpha)} := \frac{1}{2\lambda} \sum_{i=1}^{n} y_i \alpha_i k(\,\cdot\,, x_i) \tag{8.22}$$

and $\xi_i^{(\alpha)} := \max\{0, 1 - y_i f^{(\alpha)}(x_i)\}$, where $\alpha \in \mathbb{R}^n$ is an arbitrary vector. Furthermore, we define

$$\begin{aligned}\operatorname{gap}(\alpha) &:= \lambda\|f^{(\alpha)}\|_H^2 + \frac{1}{n}\sum_{i=1}^n \xi_i^{(\alpha)} - \sum_{i=1}^n \alpha_i + \frac{1}{4\lambda}\sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j k(x_i, x_j) \\ &= 2\lambda\|f^{(\alpha)}\|_H^2 + \frac{1}{n}\sum_{i=1}^n \xi_i^{(\alpha)} - \sum_{i=1}^n \alpha_i \end{aligned} \tag{8.23}$$

for the difference between the value of the objective function of (8.20) and (8.21). Theorem A.6.28 together with Example 11.3 shows that if $\alpha^* \in \mathbb{R}^n$ is a solution of (8.21), then $(f^{(\alpha^*)}, \xi^{(\alpha^*)})$ is a solution of (8.20), i.e.,

$$f_{\mathrm{D},\lambda} = \frac{1}{2\lambda}\sum_{i=1}^n y_i \alpha_i^* k(\,\cdot\,, x_i)\,, \tag{8.24}$$

and we have $\operatorname{gap}(\alpha^*) = 0$. In other words, the SVM solution $f_{\mathrm{D},\lambda}$ can be obtained by solving the dual problem (8.21) and substituting the corresponding solution α^* into (8.22). Unfortunately, however, it is almost always impossible to find an *exact* solution α^* of (8.21). Consequently, let us assume that we have an $\alpha \in \mathbb{R}^n$ that is feasible for (8.21), i.e., $\alpha_i \in [0, 1/n]$ for all $i = 1, \ldots, n$. From the already mentioned fact that an exact solution α^* satisfies $\operatorname{gap}(\alpha^*) = 0$, we then conclude that

$$\begin{aligned}\operatorname{gap}(\alpha) &\geq \lambda\|f^{(\alpha)}\|_H^2 + \frac{1}{n}\sum_{i=1}^n \xi_i^{(\alpha)} - \sum_{i=1}^n \alpha_i^* + \frac{1}{4\lambda}\sum_{i,j=1}^n y_i y_j \alpha_i^* \alpha_j^* k(x_i, x_j) \\ &= \lambda\|f^{(\alpha)}\|_H^2 + \frac{1}{n}\sum_{i=1}^n \xi_i^{(\alpha)} - \lambda\|f^{(\alpha^*)}\|_H^2 - \frac{1}{n}\sum_{i=1}^n \xi_i^{(\alpha^*)}\,. \end{aligned}$$

Using the definitions of $f^{(\alpha)}$, $\xi^{(\alpha)}$, and $f^{(\alpha^*)} = f_{\mathrm{D},\lambda}$, we thus obtain

$$\lambda\|f^{(\alpha)}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f^{(\alpha)}) \;\leq\; \inf_{f \in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f) + \operatorname{gap}(\alpha)$$

for all $\alpha \in [0, 1/n]^n$. Consequently, if we have an $\alpha \in [0, 1/n]^n$ with $\operatorname{gap}(\alpha) \leq \epsilon$, where $\epsilon > 0$ is some predefined accuracy, then $f^{(\alpha)}$ is a decision function satisfying the ϵ-CR-ERM inequality (7.31), and hence the statistical analysis of Section 7.4 can be applied.

Our next goal is to estimate the number of non-zero coefficients of $f^{(\alpha)}$ when represented by (8.22). We begin with the following proposition.

Proposition 8.27 (Deterministic lower bound on the sparseness). *Let X be a non-empty set, k be a kernel on X, $D \in (X \times Y)^n$ be an arbitrary sample set, and $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. If α is feasible, i.e., $0 \leq \alpha_i \leq 1/n$ for all $i = 1, \ldots, n$, then we have*

$$\frac{|\{i : \alpha_i \neq 0\}|}{n} \;\geq\; 2\lambda\|f^{(\alpha)}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f^{(\alpha)}) - \operatorname{gap}(\alpha)\,.$$

Proof. By (8.23) and the definition of $\xi^{(\alpha)}$, we obtain

$$\frac{|\{i : \alpha_i \neq 0\}|}{n} \;=\; \sum_{\alpha_i \neq 0} \frac{1}{n} \;\geq\; \sum_{i=1}^{n} \alpha_i \;=\; 2\lambda \|f^{(\alpha)}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f^{(\alpha)}) - \operatorname{gap}(\alpha)\,.$$

□

Our next goal is to establish a probabilistic bound on the number of non-zero coefficients. To this end, we simplify the presentation by considering only *exact* solutions α^* and $f_{\mathrm{D},\lambda}$, though similar results also hold for approximate solutions. Note that in this case we have $\operatorname{gap}(\alpha^*) = 0$, and hence Proposition 8.27 yields

$$\frac{|\{i : \alpha_i^* \neq 0\}|}{n} \;\geq\; 2\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda})\,.$$

Consequently, we need a lower bound on the right-hand side of this estimate. A relatively simple lower bound is presented in the following proposition.

Proposition 8.28 (Probabilistic lower bound on the sparseness). *Let X be a compact metric space, H be the RKHS of a continuous kernel k on X with $\|k\|_\infty \leq 1$, and P be a probability measure on $X \times Y$. Then, for fixed $\lambda \in (0,1]$, $n \geq 1$, $\varepsilon > 0$, and $\tau > 0$, we have with probability P^n not less than $1 - e^{-\tau}$ that*

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) \;>\; \lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - 2\varepsilon - \sqrt{\frac{2\tau + 2\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \lambda^{1/2}\varepsilon)}{\lambda n}}\,.$$

Proof. We have $\lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) \leq \lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})$ by the definition of $f_{\mathrm{P},\lambda}$, and hence we obtain

$$\begin{aligned}
&\lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \lambda \|f_{\mathrm{D},\lambda}\|_H^2 - \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) \\
&\leq \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) \\
&\leq \sup_{f \in \lambda^{-1/2} B_H} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr) \\
&\leq \sup_{g \in \mathcal{F}_\varepsilon} \bigl(\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr) + 2\varepsilon\,,
\end{aligned}$$

where $\mathcal{F}_\varepsilon \subset \lambda^{-1/2} B_H$ is assumed to be an ε-net of $\lambda^{-1/2} B_H$ having cardinality $\mathcal{N}(\lambda^{-1/2} B_H, \|\cdot\|_\infty, \varepsilon) = \mathcal{N}(B_H, \|\cdot\|_\infty, \lambda^{1/2}\varepsilon)$. Moreover, for $g \in \lambda^{-1/2} B_H$, we have $\|g\|_\infty \leq \lambda^{-1/2}$, and hence we find

$$L(y, g(x)) = \max\{0, 1 - y g(x)\} \leq 1 + g(x) \leq 2\lambda^{-1/2}\,, \qquad y = \pm 1,\, x \in X.$$

By a union bound and Hoeffding's inequality, we thus obtain

$$\mathrm{P}^n\left(D \in (X \times Y)^n : \sup_{g \in \mathcal{F}_\varepsilon} \bigl(\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr) < \sqrt{\frac{2\tau}{\lambda n}}\right) \geq 1 - |\mathcal{F}_\varepsilon|\, e^{-\tau}\,.$$

Combining this estimate with the first one then yields the assertion. □

In order to illustrate how the results above can be combined, let us assume for a moment that we use a *fixed* RKHS H whose covering numbers satisfy the usual assumption

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \;\leq\; a\varepsilon^{-2p}\,, \qquad\qquad \varepsilon > 0,$$

where $a, p > 0$ are some constants independent of ε. For $\lambda \in (0,1]$, $\tau > 0$, $n \geq 1$, and $\varepsilon := (\frac{p}{2})^{1/(1+p)}(\frac{2a}{n})^{1/(2+2p)}\lambda^{-1/2}$, Proposition 8.28 then shows that

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) > \mathcal{R}^*_{L,\mathrm{P},H} - 4\Big(\frac{a}{\lambda^{1+p}n}\Big)^{\frac{1}{2+2p}} - \sqrt{\frac{2\tau}{\lambda n}}$$

holds with probability P^n not less than $1-e^{-\tau}$. Combining this estimate with Proposition 8.27, we find that

$$\frac{|\{i : \alpha_i^*(D) \neq 0\}|}{n} \;\geq\; \mathcal{R}^*_{L,\mathrm{P},H} - 4\Big(\frac{a}{\lambda^{1+p}n}\Big)^{\frac{1}{2+2p}} - \sqrt{\frac{2\tau}{\lambda n}}$$

holds with probability P^n not less than $1 - e^{-\tau}$, where $(\alpha_1^*(D), \ldots, \alpha_n^*(D))$ is an exact solution of the dual problem (8.21) defined by the sample set $D = ((x_1, y_1), \ldots, (x_n, y_n))$. In particular, if we use an *a priori* chosen regularization sequence $(\lambda_n) \subset (0,1]$ satisfying both $\lambda_n \to 0$ and $\lambda_n^{1+p}n \to \infty$ and an RKHS H such that $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$, then the SVM is (classification) consistent as we have seen around (6.22), *and* for all $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathrm{P}^n\left(D \in (X \times Y)^n : \frac{|\{i : \alpha_i^*(D) \neq 0\}|}{n} \;\geq\; \mathcal{R}^*_{L,\mathrm{P}} - \varepsilon\right) = 1\,. \tag{8.25}$$

In other words, the number of support vectors tends to be not smaller than $\mathcal{R}^*_{L,\mathrm{P}}$. Moreover, recalling the calculations in Example 2.4 and the proof of Theorem 2.31, we find $\mathcal{R}^*_{L,\mathrm{P}} = 2\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}}$, i.e., asymptotically the number of support vectors is not smaller than twice the Bayes classification risk. Finally, one can generalize this result to *data-dependent* methods of choosing λ such as the one considered in Section 6.5. The details can be found in Exercise 8.7.

Note that the results above describe the number of support vectors when we use the dual problem (8.21) for finding $f_{\mathrm{D},\lambda}$. However, we will see in the next section that if the RKHS H is universal and P_X is an atom-free distribution, then we almost surely have a unique representation (8.19), and consequently the above lower bounds on the number of support vectors hold for *every* algorithm producing $f_{\mathrm{D},\lambda}$. On the other hand, for *finite-dimensional* RKHSs H, the number of support vectors does depend on the algorithm. We refer to Exercise 8.6 for details.

8.5 Classifying with other Margin-Based Losses (*)

Although the hinge loss is the most commonly used loss for binary classification with SVMs, it is by no means the only choice. Indeed we have seen in

Section 3.4 that there are various other (convex) margin-based loss functions that are calibrated to the classification loss. Moreover, by Theorem 3.34, these losses are automatically uniformly classification calibrated and hence they are excellent alternatives if the sole objective is classification. However, some of these losses possess nice additional features that are important for some applications. For example, the (truncated) least squares loss and the logistic loss enable us to estimate posterior probabilities (see Example 3.66 for the latter), and hence these losses can be interesting whenever such an estimate is relevant for the application at hand. In addition, different loss functions lead to different optimization problems, and it is possible that some of these optimization problems promise algorithmic advantages over the one associated with the hinge loss. Consequently, there are situations in which the hinge loss may *not* be the first choice and alternatives are desired. In this section, we investigate some features of such alternatives so that an informed decision can be made. In particular, we will revisit calibration inequalities, establish a lower bound on the number of support vectors, and describe a relation between sparseness and the possibility of estimating the posterior probability.

Let us begin by recalling Theorem 3.36, which showed that a convex margin-based loss L is classification calibrated if and only if its representing function $\varphi : \mathbb{R} \to [0,\infty)$ is differentiable at 0 with $\varphi'(0) < 0$. Moreover, Theorem 3.34 showed that in this case L is actually uniformly classification calibrated and that the uniform calibration function is given by (3.41). Now assume for simplicity that this calibration function satisfies $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) \geq c_p \varepsilon^p$ for some constants $c_p > 0$, $p \geq 1$, and all $\varepsilon \in [0,1]$. Then Theorem 3.22 yields

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \;\leq\; c_p^{-1/p} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\bigr)^{1/p} \tag{8.26}$$

for all distributions P on $X \times Y$ and all $f : X \to \mathbb{R}$. The next result shows that (8.26) can be improved if $p > 1$ and P has a non-trivial noise exponent.

Theorem 8.29 (Inequality for classification calibrated losses). *Let L be a convex, margin-based, and classification calibrated loss whose uniform calibration function satisfies $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) \geq c_p \varepsilon^p$ for some constants $c_p > 0$, $p > 1$, and all $\varepsilon \in [0,1]$. Moreover, let* P *be a distribution on $X \times Y$ that has some noise exponent $q \in [0,\infty]$. Then, for all measurable $f : X \to \mathbb{R}$, we have*

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \;\leq\; 2 c_p^{-\frac{q+1}{q+p}} c^{\frac{q(p-1)}{q+p}} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\bigr)^{\frac{q+1}{q+p}},$$

where c is the constant appearing in the noise exponent inequality (8.16).

Proof. We have seen in Remark 3.35 that $L_{\mathrm{P}} : X \times \mathbb{R} \to [0,\infty)$ defined by

$$L_{\mathrm{P}}(x,t) := \bigl|2\eta(x) - 1\bigr| \cdot \mathbf{1}_{(-\infty,0]}\bigl((2\eta(x) - 1)\operatorname{sign} t\bigr), \qquad x \in X, t \in \mathbb{R},$$

is a detection loss with respect to $A := \{(x,t) \in X \times \mathbb{R} : (2\eta(x) - 1)\operatorname{sign} t \leq 0\}$ and $h(x) = |2\eta(x) - 1|$, $x \in X$, where $\eta(x) := \mathrm{P}(y = 1|x)$. Moreover, (3.9) shows for the inner excess risks that

$$\mathcal{C}_{L_{\mathrm{P}},x}(t) - \mathcal{C}^*_{L_{\mathrm{P}},x} \;=\; \mathcal{C}_{L_{\mathrm{class}},\eta(x)}(t) - \mathcal{C}^*_{L_{\mathrm{class}},\eta(x)}\,, \qquad x \in X,\, t \in \mathbb{R},$$

i.e., L_{P} describes the classification goal for the distribution P. In particular, we have $\delta_{\max,L_{\mathrm{P}},L}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = \delta_{\max,L_{\mathrm{class}},L}(\varepsilon, \mathrm{P}(\,\cdot\,|x))$ for all $x \in X$, $\varepsilon \in [0,\infty]$, and hence we obtain

$$\delta_{\max,L_{\mathrm{P}},L}\bigl(\varepsilon, \mathrm{P}(\,\cdot\,|x), x\bigr) \;\geq\; \delta^{**}_{\max,L_{\mathrm{class}},L}(\varepsilon, \mathcal{Q}_Y) \;\geq\; c_p \varepsilon^p$$

for all $x \in X$ and $\varepsilon \in [0,1]$. Since $\|h\|_\infty \leq 1$, we conclude that

$$\begin{aligned} B(s) &:= \bigl\{x \in X : \delta_{\max,L_{\mathrm{P}},L}\bigl(h(x), \mathrm{P}(\,\cdot\,|x), x\bigr) < s\,h(x)\bigr\} \\ &\subset \bigl\{x \in X : |2\eta(x) - 1| < (s/c_p)^{1/(p-1)}\bigr\} \end{aligned}$$

for all $s > 0$. Using the noise exponent, we thus obtain

$$\begin{aligned} \int_X \mathbf{1}_{B(s)} h \, d\mathrm{P}_X &\leq (s/c_p)^{1/(p-1)} \mathrm{P}_X\bigl(\bigl\{x \in X : |2\eta(x) - 1| < (s/c_p)^{1/(p-1)}\bigr\}\bigr) \\ &\leq \Bigl(c^{\frac{q(p-1)}{q+1}} \cdot \frac{s}{c_p}\Bigr)^{\frac{q+1}{p-1}} \end{aligned}$$

for all $s > 0$. Now the assertion follows from Theorem 3.28. □

In Table 3.1, we see that the (truncated) least squares L loss satisfies $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) \geq \varepsilon^2$, and hence the inequality above reduces to

$$\mathcal{R}_{L_{\mathrm{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}} \;\leq\; 2c^{\frac{q}{q+2}} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\bigr)^{\frac{q+1}{q+2}}\,. \tag{8.27}$$

Moreover, for the logistic loss for classification, we have $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) \geq \varepsilon^2/2$ and hence it satisfies (8.27) if we replace the factor 2 by 4. Finally, note that both factors can be improved by a more careful analysis in the proof of Theorem 3.28.

Let us now consider the number of support vectors we may expect using different convex margin-based loss functions. To this end, recall again that Theorem 5.5 showed that the unique SVM solution $f_{\mathrm{D},\lambda}$ is of the form

$$f_{\mathrm{D},\lambda} = \sum_{i=1}^n \alpha_i k(\,\cdot\,, x_i)\,, \tag{8.28}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ is a suitable vector of coefficients and $D = ((x_1, y_1), \ldots, (x_n, y_n))$. Let us assume for a moment that we have $x_i \neq x_j$ for all $i \neq j$, where we note that this is almost surely satisfied if D is sampled from a distribution P with $\mathrm{P}(\{x\}) = 0$ for all $x \in X$. In addition, we assume throughout this section that the kernel k is strictly positive definite, where we recall that this assumption is necessary (but in general not sufficient) for universal consistency by Corollary 5.34. From linear algebra, we then know that the kernel matrix $K := (k(x_j, x_i))_{i,j=1}^n$ has P^n-almost surely full rank, and by considering the system of linear equations

$$f_{\mathrm{D},\lambda}(x_j) = \sum_{i=1}^{n} \alpha_i k(x_j, x_i)\,, \qquad j = 1, \ldots, n\,,$$

we see that the representation (8.28) is P^n-almost surely unique, i.e., there exists exactly one $\alpha \in \mathbb{R}^n$ satisfying (8.28). In the following, we thus write

$$\#SV(f_{\mathrm{D},\lambda}) := |\{i : \alpha_i \neq 0\}|\,,$$

where we note that this quantity is only P^n-almost surely defined.

Let us now assume that L is a convex margin-based loss with representing function φ, i.e., $L(y,t) = \varphi(yt)$, $y = \pm 1$, $t \in \mathbb{R}$. Since the subdifferential of L is given by $\partial L(y,\cdot) = y\partial\varphi(y\,\cdot\,)$ for $y = \pm 1$, Corollary 5.12 then shows that the almost surely uniquely defined coefficients α_i in (8.28) satisfy

$$\alpha_i \in -\frac{y_i}{2\lambda n}\partial\varphi\big(y_i f_{\mathrm{D},\lambda}(x_i)\big)\,, \qquad i = 1, \ldots, n\,. \tag{8.29}$$

In particular, if φ is differentiable, we have

$$\alpha_i = -\frac{\varphi'(y_i f_{\mathrm{D},\lambda}(x_i))}{2\lambda n y_i}\,, \qquad i = 1, \ldots, n\,.$$

Now (8.29) shows that x_i must be a support vector if $0 \notin \partial\varphi(y_i f_{\mathrm{D},\lambda}(x_i))$. This immediately leads to the following preliminary result.

Proposition 8.30 (No sparseness without global minimum). *Let L be a convex, margin-based loss with representing function φ. Moreover, let k be a strictly positive definite kernel on X and P be a distribution on $X \times Y$ such that $\mathrm{P}(\{x\}) = 0$ for all $x \in X$. If φ does not have a global minimum, then for P^n-almost all $D \in (X \times Y)^n$ we have $\#SV(f_{\mathrm{D},\lambda}) = n$ for all $\lambda > 0$.*

Proof. If φ does not have a global minimum, then $0 \notin \partial\varphi(t)$ for all $t \in \mathbb{R}$ and hence (8.29) yields the assertion. □

Note that the logistic loss for classification satisfies the assumptions of the preceding proposition and hence we can in general not expect to obtain sparse decision functions when using this loss.

Because of Proposition 8.30, it remains to investigate convex margin-based losses L whose representing functions φ do have at least one global minimum. Note that by Lemma 2.23 the latter assumption is satisfied if and only if L can be clipped, and hence these loss functions enjoy the advanced oracle inequalities of Section 7.4. Now recall that Lemma 3.64 together with Lemma 3.60 showed that such losses are also self-calibrated for *all* distributions Q on $\{-1, 1\}$. Consequently, if D is a training set such that $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})$ is close to $\mathcal{R}^*_{L,\mathrm{P}}$, then $f_{\mathrm{D},\lambda}$ is close to the set-valued function $x \mapsto \mathcal{M}_{L,\eta(x)}(0^+)$ of exact minimizers, where, as usual, $\eta(x) := \mathrm{P}(y = 1|x)$. This suggests that x_i must be a support vector of the representation above whenever $0 \notin \partial\varphi(y_i\mathcal{M}_{L,\eta(x_i)}(0^+))$. Our next goal is to verify this intuition. We begin with a simple lemma that collects some useful properties of $\mathcal{M}_{L,\eta}(0^+)$.

Lemma 8.31. *Let L be a convex, classification calibrated, and margin-based loss that can be clipped. Then, for all $t \in \mathbb{R}$ and $\eta \in [0,1]$, the following statements are true:*

i) The set $\mathcal{M}_{L,\eta}(0^+)$ is a bounded, closed interval whenever $\eta \in (0,1)$.
ii) If $\partial L(1,t) \cap [0,\infty) \neq \emptyset$, then we have $t > 0$.
iii) We have $0 \notin \partial L(1,t) \cap \partial L(-1,t)$.
iv) If $t \in \mathcal{M}_{L,\eta}(0^+)$, then there exist $s^+ \in \partial L(1,t) \cap (-\infty, 0]$ and $s^- \in \partial L(-1,t) \cap [0,\infty)$ such that $\eta s^+ + (1-\eta)s^- = 0$.
v) If $t \in \mathcal{M}_{L,\eta}(0^+)$, we have $\min \partial \mathcal{C}_{L,\eta'}(t) < 0$ for all $\eta' > \eta$.
vi) The set-valued map $\eta \mapsto \mathcal{M}_{L,\eta}(0^+)$ is a monotone operator, i.e., we have $\sup \mathcal{M}_{L,\eta}(0^+) \leq \inf \mathcal{M}_{L,\eta'}(0^+)$ for all $\eta' \in [0,1]$ with $\eta' > \eta$.

Proof. i). Let $\varphi : \mathbb{R} \to [0,\infty)$ be the function representing L. By Theorem 3.36, we know $L'(1,0) = \varphi'(0) < 0$ and, since φ is convex, we conclude that $\lim_{t\to-\infty} \varphi(t) = \infty$. For $\eta \in (0,1)$, the latter implies $\lim_{t\to\pm\infty} \mathcal{C}_{L,\eta}(t) = \infty$, and hence $\mathcal{M}_{L,\eta}(0^+)$ is bounded. The remaining assertions follow from the continuity and convexity of $t \mapsto \mathcal{C}_{L,\eta}(t)$.

ii). Assume we had $t \leq 0$. Then the monotonicity of subdifferentials together with Theorem 3.36 would yield $s \leq \varphi'(0) < 0$ for all $s \in \partial L(1,t)$.
iii). Assume that there exists a $t \in \mathbb{R}$ such that $0 \in \partial L(1,t) \cap \partial L(-1,t)$. By *ii)*, we then see that $0 \in \partial L(1,t)$ implies $t > 0$, while $0 \in \partial L(-1,t) = -\partial L(1,-t)$ implies $t < 0$.

iv). For a given $t \in \mathcal{M}_{L,\eta}(0^+)$, there exist an $s^+ \in \partial L(1,t)$ and an $s^- \in \partial L(-1,t)$ with $0 = \eta s^+ + (1-\eta)s^-$. If $\eta = 1$, we find $s^+ = 0$. By *ii)* this implies $t > 0$, and, since $s^- \in \partial L(-1,t) = -\partial L(1,-t)$, we thus have $s^- > 0$ by *ii)*. The case $\eta = 0$ can be treated analogously. Finally, if $0 < \eta < 1$, we have $(1-\eta)s^- = -\eta s^+$, which leads to either $s^- \leq 0 \leq s^+$ or $s^+ \leq 0 \leq s^-$. Moreover, *ii)* shows that $s^+ \geq 0$ implies $t > 0$ and that $s^- \leq 0$ implies $t < 0$, and hence we conclude that $s^+ \leq 0 \leq s^-$.

v). Without loss of generality, we may assume $\eta < 1$. Let us fix $s^+ \in \partial L(1,t)$ and $s^- \in \partial L(-1,t)$ according to *iv)*. Then we find $s^+ - s^- < 0$ by *iii)* and hence

$$s \;:=\; \eta' s^+ + (1-\eta')s^- \;<\; \eta s^+ + (1-\eta)s^- \;=\; 0\,.$$

Finally, the linearity of subdifferentials yields $s \in \partial \mathcal{C}_{L,\eta'}(t)$.

vi). Let $0 \leq \eta < \eta' \leq 1$ as well as $t \in \mathcal{M}_{L,\eta}(0^+)$ and $t' \in \mathcal{M}_{L,\eta'}(0^+)$. By *v)*, we find an $s \in \partial \mathcal{C}_{L,\eta'}(t)$ with $s < 0$. Then we obtain $t' \geq t$ since otherwise the monotonicity of subdifferentials implies $s' \leq s < 0$ for all $s' \in \partial \mathcal{C}_{L,\eta'}(t')$, which in turn contradicts $t' \in \mathcal{M}_{L,\eta'}(0^+)$. □

The next lemma shows that $0 \notin \partial\varphi(t)$ implies $0 \notin \partial\varphi(s)$ for all s that are sufficiently close to t. This fact will be crucial for verifying our intuition since in general we cannot guarantee $f_{\mathrm{D},\lambda} \in \mathcal{M}_{L,\eta(x)}(0^+)$.

Lemma 8.32. *Let L be a convex, classification calibrated, and margin-based loss that can be clipped and* P *be a distribution on $X \times Y$. For $\varepsilon \geq 0$, we define*

$$S_\varepsilon := \left\{(x,y) \in X \times Y : 0 \notin \partial L\big(y, \mathcal{M}_{L,\eta(x)}(0^+) + \varepsilon B_{\mathbb{R}}\big)\right\}.$$

Then we have $S_\varepsilon \subset S_{\varepsilon'}$ for all $\varepsilon > \varepsilon' \geq 0$. Moreover, we have

$$\bigcup_{\varepsilon>0} S_\varepsilon = \left\{(x,y) \in X \times Y : 0 \notin \partial L\big(y, \mathcal{M}_{L,\eta(x)}(0^+)\big)\right\}.$$

Proof. Since the first assertion is trivial, it suffices to prove $S_0 \subset \bigcup_{\varepsilon>0} S_\varepsilon$. Obviously, this follows once we have established

$$\bigcap_{\varepsilon>0} \bigcup_{\delta\in[-\varepsilon,\varepsilon]} \bigcup_{t\in\mathcal{M}_{L,\eta}(0^+)} \partial L(y, t+\delta) \subset \bigcup_{t\in\mathcal{M}_{L,\eta}(0^+)} \partial L(y,t) \tag{8.30}$$

for all $\eta \in [0,1]$, $y = \pm 1$. Let us fix an element s from the set on the left-hand side of (8.30). Then, for all $n \in \mathbb{N}$, there exist $\delta_n \in [-1/n, 1/n]$ and $t_n \in \mathcal{M}_{L,\eta}(0^+)$ with $s \in \partial L(y, t_n + \delta_n)$. If (t_n) is unbounded, we obtain $\eta \in \{0,1\}$ by *i)* of Lemma 8.31. Furthermore, in this case we find $t_n + \delta_n \in \mathcal{M}_{L,\eta}(0^+)$ for all sufficiently large n since $\mathcal{M}_{L,\eta}(0^+)$ is an interval by the convexity of L. Hence we have shown (8.30) in this case. On the other hand, if (t_n) is bounded, there exists a subsequence (t_{n_k}) of (t_n) converging to a $t_0 \in \mathbb{R}$, and since $\mathcal{M}_{L,\eta}(0^+)$ is closed, we find $t_0 \in \mathcal{M}_{L,\eta}(0^+)$. Now let us fix an $\varepsilon > 0$. By Proposition A.6.14, we find that

$$s \in \partial L(y, t_{n_k} + \delta_{n_k}) \subset \partial L(y, t_0) + \varepsilon B_{\mathbb{R}}$$

for a sufficiently large k. This yields

$$s \in \bigcap_{\varepsilon>0} \big(\partial L(y,t_0) + \varepsilon B_{\mathbb{R}}\big),$$

and thus we finally obtain $s \in \partial L(y,t_0)$ by the compactness of $\partial L(y,t_0)$. □

Before we establish a lower bound on the number of support vectors, we need another elementary lemma.

Lemma 8.33. *Let L be a convex, classification calibrated, and margin-based loss that can be clipped and* P *be a distribution on $X \times Y$. For a function $f : X \to \mathbb{R}$ and $\varepsilon > 0$, we write*

$$A(f,\varepsilon) := \left\{(x,y) \in X \times Y : \operatorname{dist}\big(f(x), \mathcal{M}_{L,\eta(x)}(0^+)\big) \geq \varepsilon\right\}.$$

Then, for all $\varepsilon > 0$, all $f, g \in \mathcal{L}_\infty(X)$ satisfying $\|f - g\|_\infty \leq \varepsilon$, and all $(x,y) \in S_{2\varepsilon} \setminus A(g,\varepsilon)$, we have

$$0 \notin \partial\varphi(yf(x)).$$

Proof. We fix an element $(x,y) \in S_{2\varepsilon}$ such that $(x,y) \notin A(g,\varepsilon)$. Then we have $\operatorname{dist}(g(x), \mathcal{M}_{L,\eta(x)}(0^+)) < \varepsilon$, and thus there exists a $t \in \mathcal{M}_{L,\eta(x)}(0^+)$ such that $|g(x) - t| < \varepsilon$. This implies $\operatorname{dist}(f(x), \mathcal{M}_{L,\eta(x)}(0^+)) \leq |f(x) - t| < 2\varepsilon$ by the triangle inequality, i.e., we have $f(x) \in \mathcal{M}_{L,\eta(x)}(0^+) + 2\varepsilon B_{\mathbb{R}}$. The definition of $S_{2\varepsilon}$ together with $\partial L(y,t) = y\partial\varphi(yt)$ then yields the assertion. □

With these preparations, we can now establish the announced lower bound on the number of support vectors.

Theorem 8.34 (Probabilistic lower bound on the sparseness). *Let L be a convex, classification calibrated, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. Moreover, let P be a distribution on $X \times Y$ such that $\mathrm{P}(\{x\}) = 0$ for all $x \in X$. We define*

$$\mathcal{S}_{L,\mathrm{P}} := \mathrm{P}\Big(\big\{(x,y) \in X \times Y : 0 \notin \partial\varphi(y\mathcal{M}_{L,\eta(x)}(0^+))\big\}\Big)\,.$$

Furthermore, let k be a bounded measurable and strictly positive definite kernel on X whose RKHS H is separable and satisfies $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$. We define $h(\lambda) := \lambda^{-1}|L|_{\lambda^{-1/2},1}$ for $\lambda > 0$. Then, for all $\varepsilon > 0$, there exists a $\delta > 0$ and a $\lambda_0 \in (0,1]$ such that for all $\lambda \in (0,\lambda_0]$ and all $n \geq 1$ the corresponding SVM satisfies

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \#SV(f_{\mathrm{D},\lambda}) \geq (\mathcal{S}_{L,\mathrm{P}} - \varepsilon)n\Big) \;\geq\; 1 - 2e^{-\frac{\delta^2 n}{18h^2(\lambda)}}\,. \quad (8.31)$$

Note that, roughly speaking, $\mathcal{S}_{L,\mathrm{P}}$ describes the probability of samples (x,y) for which for no Bayes decision function $f^*_{L,\mathrm{P}}$ is the value $yf^*_{L,\mathrm{P}}(x)$ a minimizer of the representing function φ. For example, for the squared hinge loss L, we saw in Exercise 3.1 that

$$\mathcal{M}_{L,\eta}(0^+) = \begin{cases} (-\infty,-1] & \text{if } \eta = 0 \\ \{2\eta - 1\} & \text{if } 0 < \eta < 1 \\ [1,\infty) & \text{if } \eta = 1\,. \end{cases}$$

Moreover, we clearly have $\{t \in \mathbb{R} : 0 \notin \partial\varphi(t)\} = (-\infty,1)$, and hence we obtain $0 \notin \partial\varphi(\pm\mathcal{M}_{L,\eta}(0^+))$ for all $\eta \in (0,1)$. Moreover, for $\eta \in \{0,1\}$, we have $0 \in \partial\varphi(\pm\mathcal{M}_{L,\eta}(0^+))$, and thus we find

$$\mathcal{S}_{L,\mathrm{P}} = \mathrm{P}_X\big(\{x \in X : 0 < \eta(x) < 1\}\big)\,. \quad (8.32)$$

Interestingly, we will see in Theorem 8.36 that for *differentiable* φ the right-hand side of (8.32) is always a lower bound on $\mathcal{S}_{L,\mathrm{P}}$.

Proof. Without loss of generality, we may assume $\varphi(0) \leq 1$ and $\|k\|_\infty \leq 1$. Obviously, $\lambda \mapsto h(\lambda)$ is a decreasing function on $(0,\infty)$, and hence we have $h(\lambda) \geq h(1)$ for all $\lambda \in (0,1]$. Furthermore, note that $\partial L(y,t) = y\partial\varphi(yt)$ yields $\mathcal{S}_{L,\mathrm{P}} = \mathrm{P}(S_0)$, where S_0 is defined in Lemma 8.32. Let us fix an $\varepsilon \in (0,1]$. By

Lemma 8.32, there then exists a $\delta > 0$ such that $\delta \leq 3h(1)\varepsilon$ and $\mathrm{P}(S_{2\delta}) \geq \mathrm{P}(S_0) - \varepsilon$. Moreover, $\lim_{\lambda\to 0} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) = \mathcal{R}^*_{L,\mathrm{P}}$ together with Theorem 3.61 shows that there exists a $\lambda_0 \in (0,1]$ such that $\mathrm{P}(A(f_{\mathrm{P},\lambda},\delta)) \leq \varepsilon$ for all $\lambda \in (0,\lambda_0]$. Let us fix a $\lambda \in (0,\lambda_0]$ and an $n \geq 1$. Without loss of generality, we may additionally assume

$$\frac{n}{h^2(\lambda)} \geq \frac{18}{\delta^2} \tag{8.33}$$

since otherwise the left-hand side of (8.31) is negative. For $D \in (X \times Y)^n$, we further write

$$\mathcal{E}_{D,\delta} := \bigl|\bigl\{i : (x_i,y_i) \in S_{2\delta}\backslash A(f_{\mathrm{P},\lambda},\delta)\bigr\}\bigr| .$$

Hoeffding's inequality and $\mathrm{P}(S_{2\delta}\backslash A(f_{\mathrm{P},\lambda},\delta)) \geq \mathcal{S}_{L,\mathrm{P}} - 2\varepsilon$ then yield

$$\mathrm{P}^n\Bigl(D \in (X\times Y)^n : \mathcal{E}_{D,\delta} > (\mathcal{S}_{L,\mathrm{P}} - 3\varepsilon)n\Bigr) \;\geq\; 1 - e^{-2\varepsilon^2 n} \;\geq\; 1 - e^{-\frac{\delta^2 n}{18h^2(\lambda)}} ,$$

where in the last step we used that $\delta \leq 3h(1)\varepsilon$ implies $\delta \leq 6h(\lambda)\varepsilon$. Moreover, as in the proof of Theorem 6.24, we have

$$\mathrm{P}^n\left(D \in (X\times Y)^n : \|f_{\mathrm{D},\lambda} - f_{\mathrm{P},\lambda}\|_H < h(\lambda)\left(\sqrt{\frac{2\tau}{n}} + \sqrt{\frac{1}{n}} + \frac{4\tau}{3n}\right)\right) \geq 1 - e^{-\tau}$$

for all $\tau > 0$. We define τ by $\delta = 3h(\lambda)\sqrt{2\tau/n}$. Then $\delta \leq 3h(1)$ together with $h(1) \leq h(\lambda)$ implies $2\tau/n \leq 1$, which in turn yields $\frac{4\tau}{3n} \leq \sqrt{2\tau/n}$. In addition, (8.33) implies $\tau \geq 1$, and hence we have $\sqrt{1/n} \leq \sqrt{2\tau/n}$. Combining these estimates, we conclude that

$$\mathrm{P}^n\bigl(D \in (X\times Y)^n : \|f_{\mathrm{D},\lambda} - f_{\mathrm{P},\lambda}\|_H < \delta\bigr) \geq 1 - e^{-\frac{\delta^2 n}{18h^2(\lambda)}} ,$$

and hence we have

$$\mathrm{P}^n\bigl(D : \|f_{\mathrm{D},\lambda} - f_{\mathrm{P},\lambda}\|_H < \delta \text{ and } \mathcal{E}_{D,\delta} > (\mathcal{S}_{L,\mathrm{P}} - 3\varepsilon)n\bigr) \geq 1 - 2e^{-\frac{\delta^2 n}{18h^2(\lambda)}} .$$

Let us now fix such a $D \in (X \times Y)^n$. For an index i such that $(x_i,y_i) \in S_{2\delta}\backslash A(f_{\mathrm{P},\lambda},\delta)$, Lemma 8.33 then shows that $0 \notin \partial\varphi(y_i f_{\mathrm{D},\lambda}(x_i))$, and hence x_i must be a support vector. Moreover, the estimate above shows that there are at least $(\mathcal{S}_{L,\mathrm{P}} - 3\varepsilon)n$ such indexes. □

In order to illustrate the preceding result assume for simplicity that we use an *a priori* fixed sequence (λ_n) of regularization parameters. If this sequence satisfies both $\lambda_n \to 0$ and $h^2(\lambda_n)/n \to 0$, then the right-hand side of (8.31) converges to 1 and hence the fraction of support vectors tends to be not smaller than $\mathcal{S}_{L,\mathrm{P}}$. Here we note that for the hinge loss, for example, the latter condition on (λ_n) reduces to $\lambda_n^2 n \to \infty$, whereas for the (truncated) least squares loss the condition becomes $\lambda_n^3 n \to \infty$. Finally, it is obvious that

the same result holds for data-dependent choices of λ that asymptotically respect the constraints imposed on (λ_n).

The following proposition presents a general lower bound on $\mathcal{S}_{L,\mathrm{P}}$. Together with Theorem 8.34, it shows that the sparseness of general SVMs is related to the Bayes classification risk.

Proposition 8.35. *Let L be a convex, classification calibrated, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. Then, for all distributions P with $\mathrm{P}_X(\{x \in X : \eta(x) = 1/2\}) = 0$, we have*

$$\mathcal{S}_{L,\mathrm{P}} \;\geq\; \mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}}\,.$$

Proof. We first show that $0 \notin \partial\varphi(\mathcal{M}_{L,\eta}(0^+)) \cap \partial\varphi(-\mathcal{M}_{L,\eta}(0^+))$ for all $\eta \neq 1/2$. To this end, let us fix an $\eta \in [0,1]$ such that $0 \in \partial\varphi(\mathcal{M}_{L,\eta}(0^+)) \cap \partial\varphi(-\mathcal{M}_{L,\eta}(0^+))$. Then there exist $t, t' \in \mathcal{M}_{L,\eta}(0^+)$ such that $0 \in \partial\varphi(t)$ and $0 \in \partial\varphi(-t')$. From part *ii)* of Lemma 8.31, we conclude that $t' < 0 < t$ and hence we find $H(\eta) = 0$, where $H(\eta)$ is defined in (3.37). From this we conclude that $\eta = 1/2$ by part *i)* of Lemma 3.33. Now the assertion follows from

$$\begin{aligned}\mathcal{S}_{L,\mathrm{P}} &= \mathrm{P}\Big(\big\{(x,y) \in X \times Y : 0 \notin \partial\varphi(y\mathcal{M}_{L,\eta(x)}(0^+))\big\}\Big)\\ &\geq \int_{\eta\neq 1/2} \min\{\eta, 1-\eta\}\, d\mathrm{P}_X\end{aligned}$$

and the formula for $\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}}$ given in Example 2.4. □

It is relatively straightforward to show that Proposition 8.35 is sharp for the hinge loss. Combining this with (8.25), we conclude that in general Theorem 8.34 is *not* sharp.

Our next goal is to illustrate the connection between certain desirable properties of L and the asymptotic sparseness of the resulting SVM decision functions. Our first result in this direction shows that differentiable losses L do *not*, in general, lead to sparse decision functions.

Theorem 8.36 (No sparseness for differentiable losses). *Let L be a convex, classification calibrated, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. If φ is differentiable, then for all distributions P we have*

$$\mathcal{S}_{L,\mathrm{P}} \;\geq\; \mathrm{P}_X\big(\{x \in X : 0 < \eta(x) < 1\}\big)\,.$$

Proof. In order to prove the assertion, it obviously suffices to show

$$0 \;\notin\; \partial\varphi\big(\mathcal{M}_{L,\eta}(0^+)\big) \cup \partial\varphi\big(-\mathcal{M}_{L,\eta}(0^+)\big)\,, \qquad \eta \in (0,1).$$

Let us assume the converse, i.e., that there exists an $\eta \in (0,1)$, a $y \in Y$, and a $t \in \mathcal{M}_{L,\eta}(0^+)$ such that $0 \in \partial\varphi(yt)$. Without loss of generality, we may assume $y = 1$. Since L is differentiable, we thus have $\partial L(1,t) = \{0\}$. Therefore, $0 \in \partial\mathcal{C}_{L,\eta}(t)$ implies $0 \in \partial L(-1,t)$, which contradicts part *iii)* of Lemma 8.31. □

Note that certain fast training algorithms (see, e.g., Chapelle, 2007, and the references therein) require the differentiability of the loss. Unfortunately, however, the preceding theorem in conjunction with Theorem 8.34 shows that this possible advantage is paid for by non-sparse decision functions, i.e., by a possible disadvantage during the application phase of the decision function obtained.

At the beginning of this section, we mentioned that another reason for using a loss other than the hinge is the desire to estimate the posterior probability. Our next goal is to show that producing such an estimate is in conflict with the sparseness of the decision functions. We begin with two technical lemmas.

Lemma 8.37. *Let L be a convex, classification calibrated, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. We define $t_0 := \inf\{t \in \mathbb{R} : 0 \in \partial\varphi(t)\}$. Then we have $0 < t_0 < \infty$ and $0 \in \partial\varphi(t_0)$.*

Proof. By Theorem 3.36, we have $t_0 > 0$, and since φ has a global minimum, we further find $t_0 < \infty$. In order to show the third assertion, we observe that by the definition of t_0 there exists a sequence $(s_n) \subset [0,\infty)$ such that $s_n \to 0$ and $0 \in \partial\varphi(t_0 + s_n)$ for all $n \geq 1$. Let us fix an $\varepsilon > 0$. By Proposition A.6.14, there then exists a $\delta > 0$ such that $\partial\varphi(t_0 + \delta B_{\mathbb{R}^d}) \subset \partial\varphi(t_0) + \varepsilon B_{\mathbb{R}^d}$ and for this δ there obviously exists an $n_0 \in \mathbb{N}$ such that $|s_n| \leq \delta$ for all $n \geq n_0$. We conclude that $0 \in \partial\varphi(t_0 + s_n) \subset \partial\varphi(t_0) + \varepsilon B_{\mathbb{R}^d}$, and since $\partial\varphi(t_0)$ is compact, we find $0 \in \partial\varphi(t_0)$ by letting $\varepsilon \to 0$. □

Lemma 8.38. *Let L be a convex, classification calibrated, and margin-based loss function whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. We define t_0 as in Lemma 8.37 and the function $\gamma : \mathbb{R} \to \mathbb{R}$ by*

$$\gamma(t) := \frac{d^+\varphi(-t)}{d^+\varphi(-t) + d^-\varphi(t)}\,, \qquad t \in \mathbb{R}\,,$$

where d^+ and d^- denote the left and the right derivatives defined in front of Lemma A.6.15. Then we have $1/2 \leq \gamma(t) \leq 1$ for all $t \in [0,t_0]$. Moreover, for all $\eta \in [0,1]$ and all $t \in [0,t_0]$ we have

$$\begin{aligned} \eta \geq \gamma(t) \qquad &\Longleftrightarrow \qquad d^-\mathcal{C}_{L,\eta}(t) \leq 0\,, \\ \eta > \gamma(t) \qquad &\Longleftrightarrow \qquad d^-\mathcal{C}_{L,\eta}(t) < 0\,. \end{aligned}$$

Finally, $d^+\mathcal{C}_{L,\eta}(t_0) \geq 0$ for all $\eta \in [0,1]$, and for $\eta \in [0,1)$ we actually have $d^+\mathcal{C}_{L,\eta}(t_0) > 0$.

Proof. Let us fix a $t \in [0,t_0]$. Then we have $-t \leq 0 < t_0$, and hence both $d^+\varphi(-t)$ and $d^-\varphi(t)$ are negative by the definition of t_0 and the monotonicity of subdifferentials. A simple algebraic transformation thus shows that $\eta \geq \gamma(t)$ if and only if $0 \geq \eta d^-\varphi(t) - (1-\eta)d^+\varphi(-t)$. In addition, we have

$$d^-\mathcal{C}_{L,\eta}(t) = \eta d^-\varphi(t) + (1-\eta)d^-\varphi(-\cdot)(t) = \eta d^-\varphi(t) - (1-\eta)d^+\varphi(-t)\,,$$

and hence we obtain the first equivalence. In addition, a straightforward modification of the preceding proof shows that the equivalence also holds with strict inequalities. For the proof of $\gamma(t) \leq 1$, we first observe that both $d^+\varphi(-t)$ and $d^-\varphi(t)$ are strictly negative for $t \in [0, t_0]$, and hence we find $\gamma(t) \leq 1$. Furthermore, we have $d^-\varphi(t) \geq d^+\varphi(-t)$ by the convexity of φ, and since both derivatives are strictly negative we conclude that $\gamma(t) \geq 1/2$ for all $t \in [0, t_0]$. In order to show the last assertion, we observe by Lemma 8.37 that $0 \in \partial\varphi(t_0) = [d^-\varphi(t_0), d^+(t_0)]$, and hence we have $d^-\varphi(-t_0) \leq d^-(t_0) \leq 0 \leq d^+(t_0)$. Since

$$d^+\mathcal{C}_{L,\eta}(t_0) = \eta d^+\varphi(t_0) - (1-\eta)d^-\varphi(-t_0)\,,$$

we then obtain $d^+\mathcal{C}_{L,\eta}(t_0) \geq 0$. Finally, recall that Theorem 3.36 together with $t_0 \geq 0$ shows that $d^-\varphi(-t_0) < 0$, and hence the argument above yields $d^+\mathcal{C}_{L,\eta}(t_0) > 0$ whenever $\eta \neq 1$. □

With the help of these lemmas, we can now establish a theorem that describes the conflict between the goals of having sparse decision functions and estimating the posterior probability.

Theorem 8.39 (Sparseness vs. estimating posterior probabilities). *Let L be a convex, classification calibrated, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0, \infty)$ has a global minimum. Let t_0 and γ be defined as in Lemma 8.37 and Lemma 8.38, respectively. Then, for all $\eta \in [0, 1]$ and $y := \operatorname{sign}(2\eta - 1)$, we have*

$$\begin{aligned} 1 - \gamma(t_0) < \eta < \gamma(t_0) \quad &\Longleftrightarrow \quad 0 \notin \partial\varphi(y\mathcal{M}_{L,\eta}(0^+))\,, \\ \eta \notin \{0,1\} \cup [1-\gamma(t_0), \gamma(t_0)] \quad &\Longrightarrow \quad y\mathcal{M}_{L,\eta}(0^+) = \{t_0\}\,, \\ \eta \in \{0,1\} \quad &\Longrightarrow \quad t_0 \in y\mathcal{M}_{L,\eta}(0^+)\,. \end{aligned}$$

Moreover, if the restriction $\varphi_{|(-t_0,t_0)}$ is differentiable and strictly convex, then for all $1-\gamma(t_0) < \eta < \gamma(t_0)$ there exists a $t^*_\eta \in (-t_0, t_0)$ with $\{t^*_\eta\} = \mathcal{M}_{L,\eta}(0^+)$ and $\gamma(t^*_\eta) = \eta$. In addition, the restriction

$$\gamma_{|(-t_0,t_0)} : (-t_0, t_0) \to \big(1-\gamma(t_0), \gamma(t_0)\big)$$

is bijective, continuous, and increasing.

Proof. Since $\mathcal{M}_{L,\eta}(0^+) = -\mathcal{M}_{L,1-\eta}(0^+)$, we observe that we may restrict our considerations to $\eta \in [1/2, 1]$ for the proofs of the first three assertions.

Let us begin with the equivalence. For $\eta \geq \gamma(t_0)$, we find $d^-\mathcal{C}_{L,\eta}(t_0) \leq 0 \leq d^+\mathcal{C}_{L,\eta}(t_0)$ by Lemma 8.38. In other words, we have $0 \in \partial\mathcal{C}_{L,\eta}(t_0)$ and hence $t_0 \in \mathcal{M}_{L,\eta}(0^+)$. Combining this with $0 \in \partial(t_0)$ from Lemma 8.37, we conclude that $0 \in \partial\varphi(\mathcal{M}_{L,\eta}(0^+))$. Conversely, if $\eta \in [1/2, \gamma(t_0))$, we find $d^-\mathcal{C}_{L,\eta}(t_0) > 0$ by Lemma 8.38 and hence $0 \notin \partial\mathcal{C}_{L,\eta}(t_0)$, i.e., $t_0 \notin \mathcal{M}_{L,\eta}(0^+)$. On the other

hand, we have just seen that $t_0 \in \mathcal{M}_{L,\gamma(t_0)}(0^+)$, and since $\eta' \mapsto \mathcal{M}_{L,\eta'}(0^+)$ is a monotone operator by Lemma 8.31, we conclude that $\mathcal{M}_{L,\eta}(0^+) \subset (-\infty, t_0)$. By the definition of t_0, we further have $0 \notin \partial\varphi(t)$ for all $t < t_0$, and hence we find $0 \notin \partial\varphi\bigl(\mathcal{M}_{L,\eta}(0^+)\bigr)$, i.e., we have shown the equivalence.

In order to show the next implication, recall that for $\eta \geq \gamma(t_0)$ we have already seen that $t_0 \in \mathcal{M}_{L,\eta}(0^+)$, and hence it remains to show that $\mathcal{M}_{L,\eta}(0^+) \subset \{t_0\}$ whenever $\gamma(t_0) < \eta < 1$. To show the latter, we first observe that $d^-\mathcal{C}_{L,\eta}(t_0) < 0 < d^+\mathcal{C}_{L,\eta}(t_0)$ by Lemma 8.38. Now assume that $\mathcal{M}_{L,\eta}(0^+)$ contains a $t^* \neq t_0$. If $t^* < t_0$, the monotonicity of $t \mapsto \partial\mathcal{C}_{L,\eta}(t)$ implies $s < 0$ for all $s \in \partial\mathcal{C}_{L,\eta}(t^*)$, which in turn contradicts $t^* \in \mathcal{M}_{L,\eta}(0^+)$. Analogously, we see that $t^* > t_0$ is impossible. The third implication is trivial.

In order to show the remaining assertions, we fix an η with $1-\gamma(t_0) < \eta < \gamma(t_0)$. Then we saw in the first part of the proof that $\mathcal{M}_{L,\eta}(0^+) \subset (-t_0, t_0)$. Let us begin by showing that $\mathcal{M}_{L,\eta}(0^+)$ is a singleton. To this end, we fix $t, t' \in \mathcal{M}_{L,\eta}(0^+)$. Since $\varphi_{|(-t_0,t_0)}$ is differentiable, we then have

$$\begin{aligned} \eta\varphi'(t) - (1-\eta)\varphi'(-t) &= 0\,, \\ \eta\varphi'(t') - (1-\eta)\varphi'(-t') &= 0\,, \end{aligned}$$

and by simple algebra we thus find $\eta(\varphi'(t)-\varphi'(t')) = (1-\eta)(\varphi'(-t)-\varphi'(-t'))$. Let us now assume that $t > t'$. Then the strict convexity of $\varphi_{|(-t_0,t_0)}$ shows both $\varphi'(t) > \varphi'(t')$ and $\varphi'(-t) < \varphi'(-t')$, which clearly contradicts the equality above. Consequently, $\mathcal{M}_{L,\eta}(0^+)$ is indeed a singleton. Now $0 \in \partial\mathcal{C}_{L,\eta}(t^*_\eta)$ together with the definition of γ, the differentiability of $\varphi_{|(-t_0,t_0)}$, and some simple transformations yields $\gamma(t^*_\eta) = \eta$, i.e., $\gamma_{|(-t_0,t_0)}$ is surjective. Conversely, assume that we have a $t \in (-t_0, t_0)$ with $\gamma(t) = \eta$. Then the definition of γ together with the differentiability of $\varphi_{|(-t_0,t_0)}$ and some simple transformations yields $0 \in \partial\mathcal{C}_{L,\eta}(t)$, i.e., $t \in \mathcal{M}_{L,\eta}(0^+) = \{t^*_\eta\}$. Consequently, $\gamma_{|(-t_0,t_0)}$ is also injective. Finally, $\gamma_{|(-t_0,t_0)}$ is continuous since convex, differentiable functions are continuously differentiable by Proposition A.6.14, and $\gamma_{|(-t_0,t_0)}$ is increasing since $\eta \mapsto \mathcal{M}_{L,\eta}(0^+) = \{t^*_\eta\}$ is a monotone operator by Lemma 8.31. □

The preceding theorem is quite technical and merely illuminates the situation. Therefore let us finally illustrate its consequences. To this end, we define

$$G(\eta) := \eta\mathbf{1}_{0\notin\partial\varphi(\mathcal{M}_{L,\eta}(0^+))} + (1-\eta)\mathbf{1}_{0\notin\partial\varphi(\mathcal{M}_{L,1-\eta}(0^+))}\,, \qquad \eta \in [0,1].$$

The definition of $\mathcal{S}_{L,\mathrm{P}}$ together with $-\mathcal{M}_{L,\eta}(0^+) = \mathcal{M}_{L,1-\eta}(0^+)$ then yields

$$\begin{aligned} \mathcal{S}_{L,\mathrm{P}} &= \int_X \eta(x)\mathbf{1}_{0\notin\partial\varphi(\mathcal{M}_{L,\eta(x)}(0^+))} + (1-\eta(x))\mathbf{1}_{0\notin\partial\varphi(-\mathcal{M}_{L,\eta(x)}(0^+))}\,d\mathrm{P}_X(x) \\ &= \int_X G(\eta(x))\,d\mathrm{P}_X(x)\,. \end{aligned}$$

Now the equivalence presented in Theorem 8.39 shows that $G(\eta) = 1$ for all $\eta \in (1-\gamma(t_0), \gamma(t_0))$, and consequently we cannot expect the set

$$\bigl\{x \in X : |2\eta(x) - 1| < 2\gamma(t_0) - 1\bigr\} \tag{8.34}$$

to contribute to the sparseness of the decision function. Let us now assume for a moment that $\varphi_{|(-t_0,t_0)}$ is also differentiable and strictly convex. Then the last part of Theorem 8.39 shows that $\gamma(f^*_{L,\mathrm{P}}(x)) = \eta(x)$ for all $x \in X$ with $-t_0 < f^*_{L,\mathrm{P}}(x) < t_0$. By Corollary 3.62 and the continuity of $\gamma_{|(-t_0,t_0)}$, we thus see that we can estimate $\eta(x)$ whenever x is contained in the set in (8.34). In other words, *if we wish to estimate the posterior probability in regions with high noise, we cannot expect the decision function to have a sparse representation in those regions.* Conversely, the second and third implications of Theorem 8.39 show that on the complement of the set in (8.34), i.e., on the set

$$\bigl\{x \in X : |2\eta(x) - 1| \geq 2\gamma(t_0) - 1\bigr\}, \tag{8.35}$$

it is impossible to estimate the posterior probability using the self-calibration of L. However, Lemma 8.37 shows $0 \in \partial\varphi(t_0)$, and hence we may at least hope to get sparseness on the set in (8.35). In other words, *in regions with low noise, we cannot estimate the posterior probability if we also wish to have sparse representations in these regions.* Finally, note that we can only estimate the posterior probability in regions where the noise is below a certain level, i.e., *if we wish to estimate posterior probability in regions with low noise, we also have to estimate posterior probability in regions with high noise.* We refer to Exercise 8.8 for a loss that only estimates η for noise below a certain level.

8.6 Further Reading and Advanced Topics

Binary classification is one of the oldest problems in machine learning, and consequently there exist a variety of different approaches. A rigorous mathematical treatment of many important methods developed up to the mid 1990s can be found in the book by Devroye *et al.* (1996). In particular, classical methods such as nearest neighbor, histogram rules, and kernel rules are considered in great detail. Moreover, aspects of both classical algorithms and new developments on classification are contained in almost every book on machine learning, so we only mention the recent books by Hastie *et al.* (2001) and Bishop (2006), which give a broad overview of different techniques. Finally, a thorough survey on recent developments on the statistical analysis of classification methods was compiled by Boucheron *et al.* (2005).

The bound on the approximation error function for Gaussian kernels was shown by Steinwart and Scovel (2007) for a slightly more complicated version of the margin-noise exponent. Since at first glance, their concept appears to be closely tailored to the Gaussian kernel, we decided to revise their work. Moreover, it appears that the margin or margin-noise exponent is a somewhat natural concept when dealing with the approximation error function for the hinge loss and *continuous* kernels. Indeed, if we have some noise around the decision boundary, then the minimizer of the hinge loss that we have to

approximate is a step function. Since such functions cannot be uniformly approximated by continuous functions, we have to make some error. Intuitively, this error is larger the closer one gets to the decision boundary. Clearly a low concentration of P_X (and a high noise if (8.13) is used to estimate the excess hinge risk) in this region helps to control the overall error. Finally, Vert and Vert (2006) presented another condition on P that makes it possible to bound the approximation error function of Gaussian kernels.

The notion of the noise exponent goes back to Mammen and Tsybakov (1999) and Tsybakov (2004). Its relation to (a slightly more complicated version of) the margin-noise exponent was described by Steinwart and Scovel (2007). The latter authors also proved the variance bound for the hinge loss; however, the first results in this direction had been shown by Blanchard *et al.* (2008), Massart and Nédélec (2006), and Tarigan and van de Geer (2006). Finally, the resulting improved learning rates for the TV-SVM were found by Steinwart *et al.* (2007) though their results required a substantially finer grid.

It is presently unknown whether the rates obtained are optimal in a minmax sense, i.e., whether there cannot exist a classifier that learns faster than, for example, (8.18) for *all* distributions on $[0,1]^d \times Y$ having a fixed margin-noise exponent β and a fixed noise exponent q. However, learning rates that are faster than those presented are possible under more restrictive assumptions on the data-generating distribution. For example, Steinwart (2001) established *exponentially* fast learning rates for SVMs using either the hinge or the truncated least squares loss if the classes have strictly positive distance from each other and the classification risk is zero. Koltchinskii and Beznosova (2005) generalized this result to distributions having noise exponent $q = \infty$ and a Lipschitz continuous version of the posterior probability. Finally, Audibert and Tsybakov (2007) showed that under certain circumstances such fast rates are also possible for so-called *plug-in* rules.

As in the previous chapters, the parameter selection method discussed is only meant to be an illustration of how the tools work together. We refer to Sections 6.6 and 11.3, where other methods are discussed and references to the literature are given.

One can show that the lower bound (8.25) on the number of support vectors is sometimes sharp for SVMs that use the hinge loss. Namely, Steinwart (2004) proved that, using a Gaussian kernel with *fixed* width γ, there exists a sequence (λ_n) of regularization parameters such that

$$\lim_{n\to\infty} \mathrm{P}^n\left(D \in (X \times Y)^n : \left| \frac{|\#SV(f_{\mathrm{D},\lambda_n})}{n} - 2\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}} \right| < \varepsilon \right) = 1$$

for all $\varepsilon > 0$ and all distributions P on $B_{\ell_2^d} \times Y$ whose marginal distributions P_X are absolutely continuous with respect to the Lebesgue measure. However, no particular properties of this sequence were specified, and it is unclear whether this result remains true for certain data-dependent choices of λ (and γ). Steinwart (2004) further showed that the lower bound (8.31) is sharp for

the truncated least squares loss and a fixed Gaussian kernel in the sense that there exists a sequence (λ_n) of regularization parameters such that

$$\lim_{n\to\infty} \mathrm{P}^n\left(D \in (X \times Y)^n : \left|\frac{\#SV(f_{\mathrm{D},\lambda_n})}{n} - \mathcal{S}_{L,P}\right| < \varepsilon\right) = 1\,, \qquad \varepsilon > 0.$$

Here again, P_X is assumed to be absolutely continuous with respect to the Lebesgue measure. Finally, Steinwart (2004) showed in the same sense that $\frac{\#SV(f_{\mathrm{D},\lambda_n})}{n}$ may converge to 1 for the least squares loss. These considerations were extended by Bartlett and Tewari (2004, 2007) to convex, margin-based, classification calibrated losses of the form $\varphi(t) := h((t_0 - t)_+)$, where h is assumed to be continuously differentiable and convex. Namely, they showed under the above-mentioned assumptions of Steinwart (2004) that

$$\lim_{n\to\infty} \mathrm{P}^n\left(D \in (X \times Y)^n : \left|\frac{\#SV(f_{\mathrm{D},\lambda_n})}{n} - \mathbb{E}_{x\sim \mathrm{P}_X} G(\eta(x))\right| < \varepsilon\right) = 1\,,$$

where $G(\eta) := \min\{\eta, 1-\eta\}(1-\eta_0)^{-1}\mathbf{1}_{[0,1-\eta_0]\cup[\eta_0,1]} + \mathbf{1}_{(1-\eta_0,\eta_0)}$ and $\eta_0 := h'(2t_0)/(h'(0)+h'(2t_0))$. Finally, in the presentation of Section 8.4, we followed Steinwart (2004), whereas Section 8.5 is a simplified compilation from both Steinwart (2003) and Bartlett and Tewari (2004, 2007).

There have been some attempts to make SVM decision functions sparser. For example, Wu *et al.* (2005) added a constraint to the primal problem that enforces sparseness. Since their resulting optimization problem is no longer convex, the algorithmic approach is, however, more complicated. A different approach is taken by Bakır *et al.* (2005), who edit the training set in order to remove the samples that are identified as mislabeled. This idea is based on the observation that mislabeled samples will (hopefully) be considered as such by the SVM, and in this case they will be support vectors. Finally, Keerthi *et al.* (2006) consider an SVM that uses the squared hinge loss. To force the decision function to be sparse, they propose to optimize the primal optimization problem over a subset of samples while greedily adding samples to this subset. Finally, references to further approaches can be found in these articles.

The density level detection (DLD) scenario introduced in Example 2.9 can be used for anomaly detection purposes when no labeled data are given. We saw in Section 3.8 that despite these missing labels the DLD learning scenario can be viewed as a binary classification problem and that the classification loss can be used as a surrogate loss. Since, for example, the hinge loss is in turn a surrogate for the classification loss, we can then directly use the hinge loss as a surrogate for the DLD loss function. The details of this approach have been worked out by Steinwart *et al.* (2005) and Scovel *et al.* (2005). A different approach to this problem is taken by the so-called *one-class SVM* proposed by Schölkopf *et al.* (2001a). Remarkably, Vert and Vert (2006) established both consistency and learning rates if this method uses Gaussian kernels with widths depending on the sample size. A learning scenario closely

related to the DLD problem is that of learning minimal volume sets, which was recently investigated by Scott and Nowak (2006). For further work on these learning problems, we refer to the references in the above-mentioned articles.

8.7 Summary

In this chapter, we applied the theory we developed in the previous chapters to analyze the learning performance of SVMs for binary classification. We mainly focused on SVMs using the hinge loss and a Gaussian kernel whose width can change with either the data set size or the data set itself. The key element in this analysis was the margin or margin-noise exponent that described the behavior of the data-generating distribution near the decision boundary. In particular, we saw that, for distributions that have a low concentration and a high noise level in the vicinity of the decision boundary, the approximation error function for Gaussian kernels was relatively small, which in turn resulted in favorable learning rates. We then analyzed a simple strategy for selecting both the regularization parameter and the kernel parameter in a data-dependent way. As in the previous two chapters, it turned out that this strategy is adaptive in the sense that without knowledge about the distribution P it achieves the fastest learning rates that the underlying oracle inequalities can provide. We then improved our analysis by introducing the noise exponent that measures the amount of high noise in the labeling process. This noise exponent implied a variance bound for the hinge loss that in turn could be plugged into our advanced statistical analysis of Chapter 7. The resulting learning rates for the data-dependent parameter selection strategy were sometimes as fast as n^{-1}; however, the exact rates depended heavily on the properties of P.

We then analyzed the typical number of support vectors SVM decision functions have. Here it turned out that, using the hinge loss, the fraction of support vectors tends to be lower bounded by twice the Bayes classification risk. We concluded that we cannot expect extremely sparse decision functions when the underlying distribution has a large Bayes risk.

Since in some situations the hinge loss does not fit the needs of the application, we finally considered alternative surrogate loss functions. Here it turned out that convex margin-based surrogates that do not have a global minimum never lead to sparse decision functions. A typical example of this class of losses is the logistic loss for classification. Moreover, these loss functions are not clippable either, and hence the advanced statistical analysis of Chapter 7 does not apply. For margin-based losses that do have a global minimum, the message was more complicated. First, they do enjoy the advanced statistical analysis of Chapter 7; and second, the established lower bound for the fraction of support vectors is in general smaller than 1. However, we saw examples where this lower bound is not sharp, and hence these results only

suggest that there is a chance of obtaining sparse decision functions. Moreover, for differentiable losses, which are used in some algorithmic approaches, the lower bound actually equals 1 under relatively realistic assumptions on the distribution. Consequently, possible algorithmic advances in the training phase are paid for by disadvantages in the employment phase of the decision function. Finally, we showed that noise levels that may contribute to the sparseness of the decision function cannot be estimated by the decision function. In other words, the ability of estimating the posterior probability by the decision function is paid for by higher costs for the evaluation of the decision function.

8.8 Exercises

8.1. An intuitive illustration for the decision boundary ($\star$)
Let (X, d) be a metric space and P be a distribution on $X \times Y$ that has a *continuous* version of its posterior probability. Show that the corresponding classes X_{-1} and X_1 are open. Furthermore, give some examples where Δ coincides with the intuitive notion of the distance to the decision boundary.

8.2. Checkerboard distributions ($\star$)
Let $m \geq 2$ be a fixed integer and P be the distribution on $[0,1]^d \times Y$ whose marginal distribution P_X is the uniform distribution. Assume that there exists a version of the posterior probability such that every $x = (x_1, \ldots, x_d) \in [0,1]^d$ belongs to the class X_j, where

$$j := \prod_{i=1}^{d} (-1)^{\lceil m x_i \rceil}$$

and $\lceil t \rceil := \min\{n \in \mathbb{N} : n \geq t\}$, $t \in [0, \infty)$. Show that P has margin exponent $\alpha := 1$. Modify P_X so that P has margin exponent α for all $\alpha > 0$.

8.3. Margin exponents and different shapes ($\star$)
For fixed $p > 0$, define $X := \{(x_1, x_2) \in [-1,1]^2 : |x_2| \leq |x_1|^p\}$. Moreover, let P be a distribution on $X \times Y$ whose marginal distribution P_X is the uniform distribution on X. Assume that there exists a version η of the posterior probability such that $\eta(x) > 1/2$ if $x_1 > 0$ and $\eta(x) < 1/2$ if $x_1 < 0$. Draw a picture of the classes X_{-1} and X_1 and show that P has margin exponent $\alpha = 1 + p$.

8.4. Effective classes ($\star\star$)
Show that the margin exponent does not change if we consider the *effective classes* $X_{-1} \cap \operatorname{supp} \mathrm{P}_X$ and $X_1 \cap \operatorname{supp} \mathrm{P}_X$ in the definition of the distance to the decision boundary.

8.5. Another relation between margin and noise exponent ($\star$)
Let (X, d) be a metric space and P be a distribution on $X \times Y$ that has margin exponent $\alpha \in [0, \infty)$. Furthermore, assume that there exist constants $c > 0$ and $\gamma \in [0, \infty)$ and a version η of the posterior probability such that the associated distance to the decision boundary satisfies

$$\Delta(x) \;\le\; c\,|2\eta(x) - 1|^{\gamma}$$

for P_X-almost all $x \in X$. Show that P has noise exponent $q := \alpha\gamma$.

8.6. Sparse decision functions for finite-dimensional RKHSs ($\star$)
Let H be a finite-dimensional RKHS. Show that there always exists a representation (8.19) such that $|\{i : \alpha_i \neq 0\}| \le \dim H$.

8.7. Sparsity for TV-SVMs ($\star\star\star$)
Let X be a compact metric space, L be the hinge loss, H be the RKHS of a universal kernel k over X with $\|k\|_\infty \le 1$, and $a \ge 1$ and $p \in (0, 1]$ be constants with

$$e_i\bigl(\mathrm{id} : H \to C(X)\bigr) \;\le\; a\, i^{-\frac{1}{2p}}\,, \qquad\qquad i \ge 1.$$

Moreover, let $(\varepsilon_n) \subset (0, 1)$ be a sequence with $\varepsilon_n \to 0$ and $\varepsilon_n^{1+p} n \to \infty$. In addition, let $\Lambda = (\Lambda_n)$ be a sequence of ε_n-nets having cardinality growing polynomially in n.

i) Show that the corresponding TV-SVM satisfies the lower bound (8.25) on the number of support vectors.
ii) Generalize this result to TV-SVM using Gaussian kernels with data-dependent kernel parameters.

8.8. Partially estimating the posterior probability ($\star\star\star$)
For a fixed $a \in [0, 1)$, define $\varphi_a : \mathbb{R} \to [0, \infty)$ by $\varphi_a(t) := (1 - t)^2 - a^2$ if $t \le 1 - a$ and $\varphi_a(t) := 0$ otherwise. Show that the margin-based loss represented by φ_a can be used to estimate the posterior probability for noise levels $|2\eta - 1| < 1 - a$. Compare your findings with Theorem 8.39 and discuss the possibility of sparseness.

9

Support Vector Machines for Regression

Overview. *Regression is, besides classification, one of the main areas where support vector machines are applied. This chapter presents results on the learning properties of SVMs when applied to regression problems such as estimating conditional means, medians, or quantiles.*

Prerequisites. *Knowledge of loss functions, kernels, and stability of infinite-sample SVMs is needed from Chapters 2 and 4 and Section 5.3, respectively. Some results from measure theory, integration, and functional analysis from the appendix are used.*

In this chapter, we investigate under which conditions support vector machines are able to learn in the sense of L-risk consistency in regression problems with emphasis on the case of an unbounded output space. An introduction into SVMs for regression problems is given in Section 9.1. Section 9.2 considers the case of general loss functions. Section 9.3 covers the special case of SVMs designed to estimate conditional quantile functions, and Section 9.4 contains some numerical results for such SVMs.

9.1 Introduction

The goal in non-parametric regression is to estimate a functional relationship between an input random variable X and an output random variable Y under the assumption that the joint distribution P of (X, Y) is (almost) completely *unknown*. In order to solve this problem, we assume the existence of a set of observations (x_i, y_i) from independent and identically distributed random variables (X_i, Y_i), $i = 1, \ldots, n$, all of which have the distribution P on $X \times Y$ with corresponding Borel σ-algebra, where $Y \subset \mathbb{R}$ is Polish and e.g. $X = \mathbb{R}^d$. Denote by D the corresponding empirical distribution. The aim is to build a predictor $f : X \to \mathbb{R}$ on the basis of these observations such that $f(X)$ is a "good" approximation of Y. In this chapter, we investigate SVMs for regression problems, defined as minimizers of the regularized L-risk

$$f_{\mathrm{P},\lambda} = \arg\inf_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|_H^2 \,. \tag{9.1}$$

Replacing P by D in (9.1) gives the corresponding empirical problem. Following the interpretation that a small risk is desired, one tries to find a predictor

whose risk is close to the optimal risk given by the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$ (see Definition 2.3). This is a much stronger requirement than convergence in probability of $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n})$ to $\mathcal{R}_{L,\mathrm{P},H} := \inf_{f\in H} \mathcal{R}_{L,\mathrm{P}}(f)$, $n \to \infty$, because it is not obvious whether $\mathcal{R}_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$ even for large Hilbert spaces H.

In this chapter, we will assume that the loss function L is chosen in advance. If a suitable choice of L is not known for a specific application, Chapter 3, on surrogate loss functions, may be helpful in choosing one. We restrict attention in this chapter to convex loss functions because in this case we are not faced with computationally NP-hard problems and we know that $f_{\mathrm{P},\lambda}$ exists and is unique. We have to fix a reproducing kernel Hilbert space H with corresponding kernel k and a regularization constant $\lambda > 0$. However, we are additionally faced with the question of how to specify the output space Y. The obvious choice in Chapter 8, on binary classification problems, was $Y = \{-1, +1\}$. In regression problems, the choice of Y is less obvious and depends on the application. Two cases are important: the *unbounded case*, where $Y = \mathbb{R}$ or Y is equal to an unbounded interval, and the *bounded case*, where Y is a bounded interval, say $[a, b]$ with $-\infty < a < b < \infty$. Although in some regression problems the output variable can only take values in some bounded interval, suitable numbers for a and b are often unknown. In this situation, many practitioners prefer to choose the unbounded case.[1]

Of course, a natural question is whether the risk $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n})$ actually tends to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$ if n increases. If $\mathcal{R}_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$ and the output space Y is bounded, this can be assured by concentration inequalities, and the techniques from Chapters 6 and 7 are applicable. However, these techniques do not work for the unbounded case. Therefore, in this chapter attention is restricted to the unbounded case and we will present some results on the L-risk consistency of SVMs in regression problems for this more challenging situation.

Traditionally, most research on non-parametric regression considered the least squares loss $L(y, t) := (y - t)^2$ mainly because of historical and computational reasons. However, both from a theoretical and from a practical point of view, there are situations in which a different loss function is more appropriate. We mention three cases:

i) Regression problems not described by least squares loss. It is well-known that the least squares risk is minimized by the conditional mean of Y given x; see Example 2.6. However, in many situations one is actually not interested in this mean but for example in the conditional median instead. Now recall that the conditional median is the minimizer of $\mathcal{R}^*_{L,\mathrm{P}}$, where L is the absolute value loss (i.e., $L(y, t) := |y - t|$), and the same statement holds for conditional quantiles if one replaces the absolute value loss by

[1] In many parametric regression models, the output variable is assumed to have a Gaussian, Gamma, or log-Gaussian distribution; hence $Y = \mathbb{R}$ in the first situation and $Y = (0, \infty)$ in the last two cases.

an *asymmetric* variant known as the pinball loss treated in Example 2.43; see also Proposition 3.9.

ii) Surrogate losses. If the conditional distributions of Y given x are known to be symmetric, basically all distance-based loss functions of the form $L(y,t) = \psi(y-t)$, where $\psi : \mathbb{R} \to [0,\infty)$ is convex, symmetric, and has its only minimum at 0, can be used to estimate the conditional mean of Y given x; see Section 3.7. In this case, a less steep surrogate such as the absolute value loss, Huber's loss, or the logistic loss may be more suitable if one expects outliers in the y-direction, as we will discuss in Chapter 10.

iii) Algorithmic aspects. If the goal is to estimate the conditional median of Y given x, then the ϵ-insensitive loss given by $L_\epsilon(y,t) = \max\{|y-t| - \epsilon, 0\}$, $y,t \in \mathbb{R}$, $\epsilon \in (0,\infty)$, promises algorithmic advantages in terms of sparseness compared with the absolute loss; see Chapter 11 for details.

In Section 9.2, a general result on L-risk consistency of SVMs based on a distance-based loss function $L(y,t) = \psi(y-t)$ is given. Section 9.3 covers SVMs based on the pinball loss for quantile regression and Section 9.4 gives some numerical results for this particular case. In Section 9.5 median regression based on the ϵ-insensitive loss is considered.

9.2 Consistency

In this section, we assume that $L : Y \times \mathbb{R} \to [0,\infty)$ is a convex distance-based loss in the sense of Definition 2.32. In other words, we have a representing function $\psi : \mathbb{R} \to [0,\infty)$ with $\psi(0) = 0$ and $L(t,y) = \psi(y-t)$ for all $y,t \in \mathbb{R}$. Recall that L-risk consistency is defined as the convergence in probability of

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}\,, \quad n \to \infty,$$

where (λ_n) is a suitably chosen data-independent null sequence of regularization parameters with $\lambda_n > 0$. For technical reasons, we will additionally assume that the reproducing kernel Hilbert space H is separable and that its kernel is measurable, so the canonical feature map Φ becomes measurable by Lemma 4.25. We can now formulate our main result on the learnability of SVMs under a natural tail assumption on P. Note that no symmetry assumption on L is made.

Theorem 9.1. *Let X be a complete measurable space, $Y \subset \mathbb{R}$ be closed, L be a continuous, convex, distance-based loss function of growth type $p \in [1,\infty)$, and $H \subset L_p(\mathrm{P}_X)$ be a dense, separable RKHS with a bounded and measurable kernel k and canonical feature map $\Phi : X \to H$. We write $p^* := \max\{2p, p^2\}$ and fix a sequence (λ_n) of positive numbers with $\lambda_n \to 0$ and $\lambda_n^{p^*} n \to \infty$. Then*

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}\,, \quad n \to \infty, \tag{9.2}$$

in probability for all $|D| = n$ and for all distributions $\mathrm{P} \in \mathcal{M}_1(X \times Y)$ with $|\mathrm{P}|_p < \infty$.

Recall that the RKHS of a Gaussian RBF kernel is dense in $L_p(\mathrm{P}_X)$ by Theorem 4.63. Hence this RKHS fulfills the assumption of Theorem 9.1 if $X = \mathbb{R}^d$.

In order to prove Theorem 9.1, we need the following lemma to bound the probability of $|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})| \leq \varepsilon$ for $|D| \to \infty$.

Lemma 9.2. *Let Z be a measurable space, P be a distribution on Z, H be a separable Hilbert space, and $g : Z \to H$ be a measurable function with $\|g\|_q := (\mathbb{E}_\mathrm{P}\|g\|_H^q)^{1/q} < \infty$ for some $q \in (1,\infty)$. We write $q^* := \min\{1/2, 1/q'\}$, where q' fulfills $\frac{1}{q} + \frac{1}{q'} = 1$. Then there exists a universal constant $c_q > 0$ such that, for all $\varepsilon > 0$ and all $n \geq 1$, we have*

$$\mathrm{P}^n \left((z_1,\ldots,z_n) \in Z^n \,:\, \left\| \frac{1}{n}\sum_{i=1}^n g(z_i) - \mathbb{E}_\mathrm{P} g \right\|_H \geq \varepsilon \right) \leq c_q \left(\frac{\|g\|_q}{\varepsilon\, n^{q^*}} \right)^q .$$

For the proof of Lemma 9.2, we have to recall some basics from local Banach space theory and Rademacher sequences. We refer to Section 7.3 and Section A.8 for details. A sequence of independent, symmetric $\{-1,+1\}$-valued random variables (ε_i) is called a *Rademacher sequence*. Now let E be a separable Banach space, (X_i) be an i.i.d. sequence of E-valued random variables with expectation 0, and (ε_i) be a Rademacher sequence that is independent from the sequence (X_i). The distribution of ε_i is denoted by ν. Using the symmetrization argument given in Theorem A.8.1, we have for all $1 \leq p < \infty$ and all $n \geq 1$ that

$$\mathbb{E}_{\mathrm{P}^n} \left\| \sum_{i=1}^n X_i \right\|^p \leq 2^p\, \mathbb{E}_{\mathrm{P}^n} \mathbb{E}_{\nu^n} \left\| \sum_{i=1}^n \varepsilon_i X_i \right\|^p , \tag{9.3}$$

where the left expectation is with respect to the product distribution P^n of $(X_1,\ldots,X_n)$, whereas the right expectation is also with respect to the product distribution ν^n of $(\varepsilon_1,\ldots,\varepsilon_n)$. Furthermore, for $n = 2$, we obtain

$$\begin{aligned}
\mathbb{E}_{\nu^2} \|\varepsilon_1 x_1 + \varepsilon_2 x_2\|^2 &= \frac{1}{2}\left(\|x_1 + x_2\|^2 + \|x_1 - x_2\|^2 \right) \\
&= \frac{1}{2}\left(\|x_1\|^2 + 2\langle x_1, x_2\rangle + \|x_2\|^2 + \|x_1\|^2 - 2\langle x_1, x_2\rangle + \|x_2\|^2 \right) \\
&= \|x_1\|^2 + \|x_2\|^2 .
\end{aligned}$$

An induction over n therefore shows that

$$\mathbb{E}_{\nu^n} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|^2 = \sum_{i=1}^n \|x_i\|^2 \tag{9.4}$$

for all $n \geq 1$ and all finite sequences $x_1,\ldots,x_n$. Furthermore, Kahane's inequality (see Theorem A.8.3) ensures that

$$\left(\mathbb{E}_{\nu^n}\Big\|\sum_{i=1}^n \varepsilon_i x_i\Big\|^p\right)^{1/p} \le c_{p,q}\left(\mathbb{E}_{\nu^n}\Big\|\sum_{i=1}^n \varepsilon_i x_i\Big\|^q\right)^{1/q}$$

for all $p, q \in (0,\infty)$, $n \ge 1$, all Banach spaces E, all $x_1, \ldots, x_n \in E$, and constants $c_{p,q}$ only depending on p and q. Now we can proceed with the following proof.

Proof of Lemma 9.2. Define $h : Z^n \to H$,

$$h(z_1, \ldots, z_n) := \frac{1}{n}\sum_{i=1}^n g(z_i) - \mathbb{E}_{\mathrm{P}} g\,, \qquad (z_1, \ldots, z_n) \in Z^n.$$

Markov's inequality yields

$$\mathrm{P}^n\big(\|h\|_H \ge \varepsilon\big) \le \varepsilon^{-q}\,\mathbb{E}_{\mathrm{P}^n}\|h\|_H^q\,.$$

Hence it remains to estimate $\mathbb{E}_{\mathrm{P}^n}\|h\|_H^q$. By (9.3), we have

$$\mathbb{E}_{\mathrm{P}^n}\left\|\sum_{i=1}^n g(Z_i) - \mathbb{E}_{\mathrm{P}} g\right\|_H^q \le 2^q\,\mathbb{E}_{\mathrm{P}^n}\mathbb{E}_{\nu^n}\left\|\sum_{i=1}^n \varepsilon_i\big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big)\right\|_H^q . \tag{9.5}$$

If $q \in (1,2]$, we obtain with Kahane's inequality, see Theorem A.8.3, that

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}^n}\|h\|_H^q &\le 2^q n^{-q}\,\mathbb{E}_{\mathrm{P}^n}\mathbb{E}_{\nu^n}\left\|\sum_{i=1}^n \varepsilon_i\big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big)\right\|_H^q \\
&\le 2^q n^{-q}\,\mathbb{E}_{\mathrm{P}^n}\left(\mathbb{E}_{\nu^n}\left\|\sum_{i=1}^n \varepsilon_i\big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big)\right\|_H^2\right)^{q/2} \\
&= 2^q n^{-q}\,\mathbb{E}_{\mathrm{P}^n}\left(\sum_{i=1}^n \big\|\big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big\|_H^2\right)^{q/2} \\
&\le 2^q n^{-q}\,\mathbb{E}_{\mathrm{P}^n}\sum_{i=1}^n \big\|\big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big\|_H^q \\
&\le 4^q n^{1-q}\,\mathbb{E}_{\mathrm{P}}\,\|g\|_H^q\,.
\end{aligned}$$

From this we obtain the assertion for $q \in (1,2]$. Now assume that $q \in (2,\infty)$. By (9.5) and Kahane's inequality, there is a universal constant $c_q > 0$ with

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}^n} \|h\|_H^q &\leq 2^q n^{-q} \mathbb{E}_{\mathrm{P}^n} \mathbb{E}_{\nu^n} \left\| \sum_{i=1}^n \varepsilon_i \big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g \big) \right\|_H^q \\
&\leq c_q n^{-q} \mathbb{E}_{\mathrm{P}^n} \left(\mathbb{E}_{\nu^n} \left\| \sum_{i=1}^n \varepsilon_i \big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g \big) \right\|_H^2 \right)^{q/2} \\
&\leq c_q n^{-q} \mathbb{E}_{\mathrm{P}^n} \left(\sum_{i=1}^n \left\| g(Z_i) - \mathbb{E}_{\mathrm{P}} g \right\|_H^2 \right)^{q/2} \\
&\leq c_q n^{-q} \left(\sum_{i=1}^n \big(\mathbb{E}_{\mathrm{P}} \|g(Z_i) - \mathbb{E}_{\mathrm{P}} g\|_H^q \big)^{2/q} \right)^{q/2} \\
&\leq 2^q c_q n^{-q/2} \, \mathbb{E}_{\mathrm{P}} \|g\|_H^q \,,
\end{aligned}$$

where (9.4) is used in the third step. The assertion follows for $q \in (2, \infty)$. □

Proof of Theorem 9.1. Because $H \subset L_p(\mathrm{P}_X)$ is dense, we obtain from Lemma 2.38 (i) and from the assumption $|\mathrm{P}|_p < \infty$ that L is a P-integrable Nemitski loss of order $p \in [1, \infty)$. Hence Theorem 5.31 gives

$$\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}} \,. \tag{9.6}$$

To avoid handling too many constants, let us assume $\|k\|_\infty = 1$, $|\mathrm{P}|_p = 1$, and $c := c_{L,p} := 2^{-(p+2)}$ for the upper order constant of L. This yields $\mathcal{R}_{L,\mathrm{P}}(0) \leq 1$. Furthermore, we assume without loss of generality that $\lambda_n \leq 1$ for all $n \geq 1$. Using (5.4), we obtain $\|f_{\mathrm{P},\lambda_n}\|_\infty \leq \|f_{\mathrm{P},\lambda_n}\|_H \leq \lambda_n^{-1/2}$. It follows from Lemma 5.15 that

$$\lim_{\lambda \to 0} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{\mathrm{P},\lambda}) = \mathcal{R}^*_{L,\mathrm{P},H}$$

because $A_2(\lambda)$ is continuous and $A_2(0) = 0$. Combining this with (9.6) yields

$$\lim_{\lambda_n \to 0} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) = \mathcal{R}^*_{L,\mathrm{P}} \,.$$

Therefore, we obtain L-risk consistency if we can show that

$$|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n})| \to 0$$

holds in probability for $n \to \infty$.

For $n \in \mathbb{N}$ and $\lambda_n > 0$, let $h_n : X \times Y \to \mathbb{R}$ be the function obtained by Corollary 5.11. Our assumptions give for $p' := p/(p-1)$ that there is a constant $c(p, L)$ such that

$$\|h_n\|_{L_{p'}(\mathrm{P})} \leq c(p, L) \lambda_n^{-(p-1)/2} \,. \tag{9.7}$$

Moreover, for $g \in H$ with $\|f_{\mathrm{P},\lambda_n} - g\|_H \leq 1$, we have

$$\|g\|_\infty \;\leq\; \|f_{\mathrm{P},\lambda_n}\|_\infty + \|f_{\mathrm{P},\lambda_n} - g\|_\infty \;\leq\; 2\lambda_n^{-1/2} \,.$$

First, we consider the case $p > 1$. By Lemma 2.38 (ii) with $q := p - 1$, there exists a constant $c_{p,L} > 0$ only depending on L and p such that

$$\begin{aligned}
&\bigl|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| \\
&\leq c_{p,L}\left(|\mathrm{P}|_{p-1}^{p-1} + \|f_{\mathrm{P},\lambda_n}\|_\infty^{p-1} + \|g\|_\infty^{p-1} + 1\right) \|f_{\mathrm{P},\lambda_n} - g\|_\infty \\
&\leq \tilde{c}_{p,L}\,\lambda_n^{-(p-1)/2}\,\|f_{\mathrm{P},\lambda_n} - g\|_H
\end{aligned} \tag{9.8}$$

for all measurable $g \in H$ with $\|f_{\mathrm{P},\lambda_n} - g\|_H \leq 1$. Let $\varepsilon \in (0,1]$ and $D \in (X \times Y)^n$ be a training set of length n with empirical distribution D such that

$$\|\mathbb{E}_\mathrm{P} h_n \Phi - \mathbb{E}_\mathrm{D} h_n \Phi\|_H \leq \lambda_n^{(p+1)/2}\,\varepsilon / \tilde{c}_{p,L}\,. \tag{9.9}$$

Then Corollary 5.11 gives $\|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D},\lambda_n}\|_H \leq \lambda_n^{(p-1)/2}\,\varepsilon/\tilde{c}_{p,L} \leq 1$ for n large enough, and hence (9.8) yields

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n})\bigr| \leq \varepsilon\,. \tag{9.10}$$

Second, we consider the special case $p = 1$. The loss function L is by assumption convex and of upper growth type 1 and therefore Lipschitz continuous by Lemma 2.36 (iv). Hence we obtain

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| \;\leq\; |\psi|_1\,\|f_{\mathrm{P},\lambda_n} - g\|_\infty \;\leq\; |\psi|_1\,\|f_{\mathrm{P},\lambda_n} - g\|_H \tag{9.11}$$

for all $g \in H$ with $\|f_{\mathrm{P},\lambda_n} - g\|_H \leq 1$. As L is of growth type $p \in [1,\infty)$ we have $|\psi|_1 > 0$. Now let $\varepsilon \in (0, |\psi|_1^{-1}]$ and $D \in (X \times Y)^n$ be a training set of length n with corresponding empirical distribution D such that

$$\|\mathbb{E}_\mathrm{P} h_n \Phi - \mathbb{E}_\mathrm{D} h_n \Phi\|_H \leq \lambda_n\,\varepsilon\,/\,|\psi|_1\,. \tag{9.12}$$

Corollary 5.11 and (9.12) give $\|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D},\lambda_n}\|_H \leq \varepsilon/|\psi|_1 \leq 1$ such that (9.11) yields the validity of (9.10) also for the case $p = 1$.

Let us now estimate the probability of D satisfying (9.9). Define $q := p/(p-1)$ if $p > 1$ and $q := 2$ if $p = 1$. Then we have $q^* := \min\{1/2, 1-1/q\} = \min\{1/2, 1/p\} = p/p^*$. Further define $c := \max\{c(p,L), \tilde{c}_{p,L}, |\psi|_1\}$. Combining Lemma 9.2 and (9.7) yields

$$\begin{aligned}
&\mathrm{P}^n\bigl(D \in (X \times Y)^n \,:\, \|\mathbb{E}_\mathrm{P} h_n \Phi - \mathbb{E}_\mathrm{D} h_n \Phi\|_H \leq c^{-1}\lambda_n^{(p+1)/2}\varepsilon\bigr) \\
&\geq 1 - \hat{c}_{p,L}\left(\frac{\|h_n\|_q}{\varepsilon\,\lambda_n^{(p+1)/2}\,n^{q^*}}\right)^q \geq 1 - \hat{c}_{p,L}\left(\frac{c(p,L)}{\varepsilon\,\lambda_n^p\,n^{p/p^*}}\right)^q,
\end{aligned} \tag{9.13}$$

where $\hat{c}_{p,L}$ is a constant only depending on p and L. Now using $\lambda_n^p n^{p/p^*} = (\lambda_n^{p^*} n)^{p/p^*} \to \infty$, if $n \to \infty$, we find that the probability of sample sets D satisfying (9.9) converges to 1 if $|D| = n \to \infty$. As we have seen above, this implies that (9.10) holds true with probability tending to 1. Now, since $\lambda_n \to 0$, we additionally have $|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}}| \leq \varepsilon$ for all sufficiently large values of n, and hence we obtain the assertion of L-risk consistency. □

Remark 9.3. Note that Theorem 9.1 in particular shows that the SVM for regression using the least squares loss function is *weakly universally consistent* in the sense of Györfi *et al.* (2002, p.13). Furthermore, it is worthwhile to note that under the assumptions above on L, H, and (λ_n) we can even characterize the distributions P for which SVMs based on (9.1) are L-risk consistent. Indeed, if $|\mathrm{P}|_p = \infty$, then SVMs are trivially L-risk consistent for P whenever $\mathcal{R}_{L,\mathrm{P}} = \infty$. Conversely, if $|\mathrm{P}|_p = \infty$ and $\mathcal{R}_{L,\mathrm{P}} < \infty$, then SVMs cannot be L-risk consistent for P since $\mathcal{R}_{L,\mathrm{P}}(f) = \infty$ for all $f \in H$. ◁

Remark 9.4. In some sense, it seems natural to consider only consistency for distributions satisfying the tail condition $|\mathrm{P}|_p < \infty$ as was done for example in Györfi *et al.* (2002) for least squares methods. In this sense, Theorem 9.1 gives consistency for all reasonable distributions. Note that the characterization above shows that SVMs are in general *not* robust against small violations of this tail assumption in regression problems. Indeed, let P be a distribution with $|\mathrm{P}|_p < \infty$ and $\tilde{\mathrm{P}}$ be a distribution with $|\tilde{\mathrm{P}}|_p = \infty$ and $\mathcal{R}_{L,\tilde{\mathrm{P}}}(f^*) < \infty$ for some $f^* \in L_p(\mathrm{P})$ (see Problem 2.6). Then every mixture distribution $\mathrm{Q}_\varepsilon := (1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}}$, $\varepsilon \in (0,1)$, satisfies both $|\mathrm{Q}_\varepsilon|_p = \infty$ and $\mathcal{R}_{L,\mathrm{Q}_\varepsilon} < \infty$. Thus an SVM for regression defined by (9.1) is not consistent for any of the small perturbations Q_ε of P, while it is consistent for the distribution P. Hence some integrability conditions for the robustness results in Section 10.3 and Section 10.4 seem to be necessary if Y is unbounded, see also Remark 10.19(iii). ◁

9.3 SVMs for Quantile Regression

This section gives a mathematical justification for using support vector machines based on the pinball loss function $L := L_{\tau\text{-pin}}$ for quantile regression. The results can informally be described in the following manner.

i) SVMs based on the pinball loss allow the non-parametric estimation of conditional quantiles.(i)
ii) Such SVMs are L-risk consistent under weak assumptions on P and k.
iii) If these SVMs are L-risk consistent, then the empirical decision function $f_{\mathrm{D},\lambda}$ approximates the conditional quantile function.

Consider a random sample (x_i, y_i), $1 \le i \le n$, from independent and identically distributed random variables (X_i, Y_i) each with unknown probability distribution P on some measurable space $X \times Y$ with corresponding Borel σ-algebra. For technical reasons, we assume throughout this section that $Y \subset \mathbb{R}$ is closed and is thus Polish. Recall that in this case P can be split up into the marginal distribution P_X and the regular conditional probability $\mathrm{P}(\cdot\,|\,x)$, $x \in X$, on Y, which exists by Ulam's theorem (Theorem A.3.15); see also Lemma A.3.16.

The goal of quantile regression is to estimate the (set-valued) *τ-quantile function*

$$F^*_{\tau,\mathrm{P}}(x) := \bigl\{q \in \mathbb{R} : \mathrm{P}(Y \le q \,|\, x) \ge \tau \text{ and } \mathrm{P}(Y \ge q \,|\, x) \ge 1-\tau\bigr\},\ x \in X, \tag{9.14}$$

where $\tau \in (0,1)$ is a fixed constant. For conceptual simplicity, we assume throughout this section that $F^*_{\tau,\mathrm{P}}(x)$ consists of singletons,[2] so that there exists a unique conditional quantile function $f^*_{\tau,\mathrm{P}} : X \to \mathbb{R}$ defined by $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$, $x \in X$; see Proposition 3.9. Now recall that the distance-based pinball loss function $L : Y \times \mathbb{R} \to [0,\infty)$ is given by

$$L(y,t) = \psi_\tau(r) = \begin{cases} (\tau-1)r & \text{if } r < 0, \\ \tau r & \text{if } r \ge 0, \end{cases}$$

where $r := y - t$ (see Example 2.43) has the property that $f^*_{\tau,\mathrm{P}}(x)$ is a τ-quantile, $\tau \in (0,1)$, of $\mathrm{P}(Y \,|\, x)$ for all $x \in X$ if and only if $f^*_{\tau,\mathrm{P}}(x)$ minimizes the inner L-risk of $\mathrm{P}(y|x)$,

$$\int_Y L(y, f^*_{\tau,\mathrm{P}}(x))\, d\mathrm{P}(y|x) = \inf_{q(x)\in\mathbb{R}} \int_Y L(y,q(x))\, d\mathrm{P}(y|x)\,. \tag{9.15}$$

It is easy to check that the pinball loss function has for each $\tau \in (0,1)$ the following properties (see Problem 9.1): ψ_τ is strictly convex, $\psi_\tau(0) = 0$, $\lim_{|r|\to\infty} \psi_\tau(r) = \infty$, and ψ_τ is Lipschitz continuous with Lipschitz constant $|\psi_\tau|_1 = \max\{\tau, 1-\tau\} \le 1$. Furthermore, we have

$$\min\{\tau, 1-\tau\}\,|r| \le \psi_\tau(r) \le |\psi_\tau|_1\,|r|\,, \quad r \in \mathbb{R}. \tag{9.16}$$

Koenker and Bassett (1978) proposed the estimator

$$\hat{f}_\tau = \arg\inf_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \psi_\tau(y_i - x_i^{\mathsf{T}}\theta)$$

if $f^*_{\tau,\mathrm{P}}$ is assumed to be a linear function and $X = \mathbb{R}^d$. Schölkopf *et al.* (2000, p. 1216) and Takeuchi *et al.* (2006) proposed to relax the assumption of a linear quantile function by using the kernel trick.

Definition 9.5. *Let X be a measurable space, $Y \subset \mathbb{R}$, and $\mathrm{P} \in \mathcal{M}_1(X \times Y)$. Furthermore, let L be the pinball loss function with $\tau \in (0,1)$, H be a separable RKHS of a measurable kernel k with canonical feature map $\Phi : X \to H$, and $\lambda > 0$. A support vector machine for quantile regression is defined by*

$$f_{\mathrm{P},\lambda} := \arg\inf_{f\in H} \mathbb{E}_\mathrm{P}\, L(Y, f(X)) + \lambda\|f\|_H^2.$$

For any fixed data set $D = \{(x_i, y_i), 1 \le i \le n\} \subset X \times Y$, we thus obtain the estimator $f_{D,\lambda}$. We obtain $f_{D,\lambda} = \hat{f}_\tau$ if we choose the linear kernel $k(x,x') := \langle x, x'\rangle$ and $\lambda := 0$. We can now formulate our result on L-risk consistency for SVMs for quantile regression.

[2] This assumption can be relaxed; see Theorem 3.61 and Corollary 3.65.

Theorem 9.6. *Let X be a complete measurable space, $Y \subset \mathbb{R}$ be closed, L be the pinball loss with $\tau \in (0,1)$, and H be a separable RKHS of a bounded measurable kernel k on X such that H is dense in $L_1(\mu)$ for all distributions μ on X. Let (λ_n) be a sequence of strictly positive numbers with $\lambda_n \to 0$.*

i) If $\lambda_n^2 n \to \infty$, then

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}\,, \quad n \to \infty, \tag{9.17}$$

in probability for all $|D| = n$ and for all $\mathrm{P} \in \mathcal{M}_1(X \times Y)$ with $|\mathrm{P}|_1 < \infty$.
ii) If $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$, then (9.17) holds even almost surely.

We mention that Theorem 9.6(i) is a direct consequence of Theorem 9.1 for $p = 1$, but the following proof uses a better concentration inequality because the pinball loss function is Lipschitz continuous.

Proof. (i). To avoid handling too many constants, let us assume $\|k\|_\infty = 1$. This implies $\|f\|_\infty \le \|k\|_\infty \|f\|_H \le \|f\|_H$ for all $f \in H$. Now we use the Lipschitz continuity of $L_{\tau\text{-pin}}$, $|\psi_\tau|_1 \le 1$, and Lemma 2.19 to obtain

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| \le |\psi_\tau|_1 \, \|f_{\mathrm{P},\lambda_n} - g\|_H\,, \ g \in H. \tag{9.18}$$

For $n \in \mathbb{N}$ and $\lambda_n > 0$, we now write $h_{\tau,n} : X \times Y \to \mathbb{R}$ for the function h obtained by Corollary 5.11. Let $\Phi : X \to H$ be the canonical feature map. We have $f_{\mathrm{P},\lambda_n} = -(2\lambda_n)^{-1}\mathbb{E}_\mathrm{P} h_{\tau,n}\Phi$, and for all distributions Q on $X \times Y$ with $|\mathrm{Q}|_1 < \infty$, we have

$$\|f_{\mathrm{P},\lambda_n} - f_{\mathrm{Q},\lambda_n}\|_H \le \lambda_n^{-1} \, \|\mathbb{E}_\mathrm{P} h_{\tau,n}\Phi - \mathbb{E}_\mathrm{Q} h_{\tau,n}\Phi\|_H .$$

Note that $\|h_{\tau,n}\|_\infty \le |\psi_\tau|_1$. Moreover, let $\varepsilon \in (0,1)$ and D be a training set of n data points and empirical distribution D such that

$$\|\mathbb{E}_\mathrm{P} h_{\tau,n}\Phi - \mathbb{E}_\mathrm{D} h_{\tau,n}\Phi\|_H \le \lambda_n \varepsilon. \tag{9.19}$$

Then Corollary 5.11 gives $\|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D}_n,\lambda_n}\|_H \le \varepsilon$ and hence (9.18) yields

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n})\bigr| \le \|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D},\lambda_n}\|_H \le \varepsilon. \tag{9.20}$$

Let us now estimate the probability of D satisfying (9.19). To this end, we first observe that $\lambda_n n^{1/2} \to \infty$ implies that for all sufficiently large n we have $\lambda_n \varepsilon \ge n^{-1/2}$. Moreover, Corollary 5.11 shows $\|h_{\tau,n}\|_\infty \le 1$, and our assumption $\|k\|_\infty = 1$ thus yields $\|h_{\tau,n}\Phi\|_\infty \le 1$. Consequently, Hoeffding's inequality in Hilbert spaces (see Theorem 6.15) yields for $B = 1$ and $\xi = \frac{3}{8}\varepsilon^2\lambda_n^2 n/(\varepsilon\lambda_n + 3)$ the bound

$$\begin{aligned}
&\mathrm{P}^n\bigl(D \in (X \times Y)^n \,:\, \|\mathbb{E}_\mathrm{P} h_{\tau,n}\Phi - \mathbb{E}_\mathrm{D} h_{\tau,n}\Phi\|_H \le \lambda_n\varepsilon\bigr) \\
&\ge \mathrm{P}^n\Bigl(D \in (X \times Y)^n \,:\, \|\mathbb{E}_\mathrm{P} h_{\tau,n}\Phi - \mathbb{E}_\mathrm{D} h_{\tau,n}\Phi\|_H \le \bigl(\sqrt{2\xi}+1\bigr)n^{-1/2} + \frac{4\xi}{3n}\Bigr) \\
&\ge 1 - \exp\Bigl(-\frac{3}{8}\cdot\frac{\varepsilon^2\lambda_n^2 n}{\varepsilon\lambda_n + 3}\Bigr)
\end{aligned} \tag{9.21}$$

for all sufficiently large values of n. Note that (9.21) is stronger than (9.13). Using $\lambda_n n^{1/2} \to \infty$, $\lambda_n \to 0$, and the same argumentation as in the proof of Theorem 9.1, we obtain that (9.20) holds true with probability tending to 1 and that $|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}}| \le \varepsilon$ for all sufficiently large n, which gives the assertion.

(ii). In order to show the second assertion, we define $\varepsilon_n := (\ln(n+1))^{-1/2}$, and $\delta_n := \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon_n$, $n \ge 1$. Moreover, for an infinite sample $D_\infty := ((x_1,y_1),(x_2,y_2),\ldots) \in (X \times Y)^\infty$, we write $D_n := ((x_1,y_1),\ldots,(x_n,y_n))$. With these notations, we define

$$A_n := \left\{ D_\infty \in (X \times Y)^\infty : \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}} > \delta_n \right\}, \qquad n \in \mathbb{N}.$$

Now, our estimates above together with $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$ yield

$$\sum_{n\in\mathbb{N}} \mathrm{P}^\infty(A_n) \;\le\; \sum_{n\in\mathbb{N}} \exp\left(-\frac{3}{8} \cdot \frac{\varepsilon_n^2 \lambda_n^2 n}{\varepsilon_n \lambda_n + 3}\right) < \infty,$$

and hence we obtain by the Borel-Cantelli lemma (Lemma A.4.7) that

$$\mathrm{P}^\infty\left(\left\{ D_\infty \in (X \times Y)^\infty \,\middle|\, \exists\, n_0\, \forall\, n \ge n_0 : \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}} \le \delta_n \right\}\right) = 1.$$

The assertion follows because $\lambda_n \to 0$ implies $\delta_n \to 0$. □

In order to formulate our result on consistency of the estimated quantile function itself, we need some additional notations. Let $f, g : X \to \mathbb{R}$ be measurable functions. We write

$$\|f\|_{L_0(\mathrm{P}_X)} := \mathbb{E}_{\mathrm{P}_X} \min\{1, |f(X)|\} \tag{9.22}$$

and define $d(f,g) := \|f - g\|_{L_0(\mathrm{P}_X)}$. Note that d is a *translation-invariant* metric on the space of all measurable functions defined on X and that d describes the convergence in probability P_X (see Problem 9.2).

The next result shows that f_{D,λ_n} approximates the conditional quantile function in terms of $\|\cdot\|_{L_0(\mathrm{P}_X)}$.

Theorem 9.7. *Let X be a complete measurable space, $Y \subset \mathbb{R}$ be closed, L be the pinball loss with $\tau \in (0,1)$, H be a separable RKHS of a bounded measurable kernel k on X such that H is dense in $L_1(\mu)$ for all distributions μ on X, and (λ_n) be a sequence of strictly positive numbers with $\lambda_n \to 0$.*

i) If $\lambda_n^2 n \to \infty$, then

$$\|f_{\mathrm{D},\lambda_n} - f^*_{\tau,\mathrm{P}}\|_{L_0(\mathrm{P}_X)} \to 0 \tag{9.23}$$

in probability for $n \to \infty$ for all $\mathrm{P} \in \mathcal{M}_1(X \times Y)$ with $|\mathrm{P}|_1 < \infty$.

ii) If $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$, then (9.23) holds even almost surely.

Proof. (i). We have already seen in Theorem 9.6(i) that f_{D,λ_n} satisfies $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}$ in probability for $n \to \infty$. The existence of a unique minimizer $f^*_{\tau,\mathrm{P}}$ is guaranteed by the general assumption of this section, and Corollary 3.62 yields the assertion.

(ii). Combine Theorem 9.6(ii) with Corollary 3.62. □

It is interesting to note that the assumption $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$ is only needed to formulate Theorem 9.7 in terms of $\|\cdot\|_{L_0(\mathrm{P}_X)}$. However, Theorem 3.63 provides a framework to replace $\|\cdot\|_{L_0(\mathrm{P}_X)}$ by a more general notion of closeness if the assumption $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$ is violated. Note that Theorem 9.6 established for $\tau = 1/2$ the convergence in probability of

$$\mathbb{E}_\mathrm{P}|Y - f_{\mathrm{D}_n,\lambda_n}(X)| - \mathbb{E}_\mathrm{P}|Y - f^*_{\tau,\mathrm{P}}(X)| \to 0, \quad n \to \infty, \tag{9.24}$$

which naturally raises the question of whether we have the convergence of

$$\mathbb{E}_\mathrm{P}|f_{\mathrm{D}_n,\lambda_n}(X) - f^*_{\tau,\mathrm{P}}(X)| \to 0, \quad n \to \infty \tag{9.25}$$

in probability. Of course, the inverse triangle inequality $\big||a| - |b|\big| \le |a - b|$ immediately shows that (9.25) implies (9.24), but since for general $a, b, c \in \mathbb{R}$ the inequality $|a - c| - |b - c| \ge |a - c|$ is false, we conjecture that without additional assumptions on P the convergence in (9.25) does not follow from (9.24). However, Example 3.67 shows that we can actually replace $\|\cdot\|_{L_0(\mathrm{P}_X)}$ by some (quasi)-norm $\|\cdot\|_{L_p(\mathrm{P}_X)}$ for certain distributions P. By (3.74), we have the following inequality for distributions P of $\mathcal{Q}^\alpha_\tau$-type. Assume the function $b : X \to [0, \infty)$ defined by $b(x) := c_{\mathrm{P}(\cdot\,|x)}$, $x \in X$, where $c_{\mathrm{P}(\cdot\,|x)}$ is determined by (3.73). If b satisfies $b^{-1} \in L_p(\mathrm{P}_X)$ for some $p \in (0, \infty]$, then

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_q(\mathrm{P}_X)} \le \sqrt{2}\,\|b^{-1}\|^{1/2}_{L_p(\mathrm{P}_X)} \left(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\right)^{1/2}$$

for all functions $f : X \to \mathbb{R}$ such that $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \le 2^{-\frac{p+2}{p+1}} \alpha^{\frac{2p}{p+1}}$, where $f^*_{L,\mathrm{P}}(x) = f^*_{\tau,\mathrm{P}}$ and $q := \frac{p}{p+1}$.

Another interesting question is whether we can establish convergence rates in Theorem 9.6 or Theorem 9.7. It is well-known in statistical learning theory that such convergence rates require additional assumptions on the distribution P due to the no-free-lunch theorem (see Corollary 6.8), for example in terms of the approximation properties of H with respect to $f^*_{\tau,\mathrm{P}}$. Moreover, the techniques used in the proofs of Theorem 9.6 and Theorem 9.7 are tuned to provide consistency under rather minimal assumptions on X, Y, P, and H, but in general these techniques are too weak to obtain good convergence results. Some results on learning rates of SVMs for quantile regression are given by Steinwart and Christmann (2008).

9.4 Numerical Results for Quantile Regression

In this section, we will illustrate that SVMs for quantile regression with different values of the quantile level τ can offer valuable information that is not obtainable by just considering one regression function for the center (conditional mean or conditional median).[3] It will also be shown by a small simulation that the asymptotic results obtained in the previous section can offer reasonable approximations for small to moderate sample sizes.

[3] Portions of this section are based on material originally published in "A. Christmann and I. Steinwart (2008), 'Consistency of kernel based quantile

Example: LIDAR Data Set

Let us start with a simple example for the application of SVMs for quantile regression. We analyze data concerning the so-called LIDAR technique. LIDAR is the abbreviation of LIght Detection And Ranging. This technique uses the reflection of laser-emitted light to detect chemical compounds in the atmosphere. We consider the logarithm of the ratio of light received from two laser sources as the response variable $Y =$ `logratio`, whereas the single explanatory variable $X =$ `range` is the distance traveled before the light is reflected back to its source. We refer to Ruppert *et al.* (2003) for more details on this data set.

A scatterplot of the data set consisting of $n = 221$ observations is shown in Figure 9.1 together with the fitted curves based on SVMs using the pinball loss function using the Gaussian RBF kernel for the median and the lower and upper 5 percent quantiles. The KBQR clearly shows that the relationship between both variables is non-linear, almost constant for values of `range` below 550 and decreasing for higher values of `range`. However, KBQR also shows that the variability of `logratio` is non-constant and much greater for values of `range`, say, above 600 than for values below this.

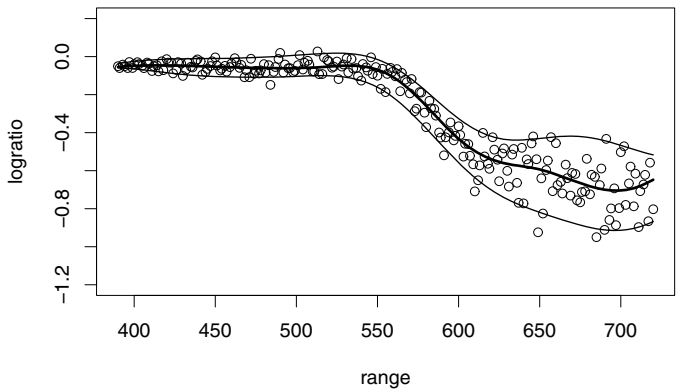


Fig. 9.1. LIDAR data set ($n = 221$). SVMs for quantile regression based on the Gaussian RBF kernel with $\gamma^2 = 0.5$ and $\lambda = \frac{1}{700} n^{-1/3}$ (resulting from a grid search). Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

Simulation Results

Now we describe a small simulation and its results to investigate how well the asymptotic results derived in Section 9.3 on the consistency of SVMs for

quantile regression work for small to moderate sample sizes. We consider $n \in \{221, 1000, 4000\}$ and use the same parameter settings for the Gaussian RBF kernel as in the previous subsection; i.e., $\gamma^2 = 0.5$ and $\lambda_n = \frac{1}{700} n^{-1/3}$. The number of replications in the simulation was set to 1000. For each replication $\ell \in \{1, \ldots, 1000\}$, we independently generated n data points $x_i^{(\ell)}$ for `range` from a continuous uniform distribution with support $(390, 720)$. Furthermore, for each replication, we generated n data points $y_i^{(\ell)}$ for `logratio` according to independent normal distributions with conditional expectations

$$\mu(x_i^{(\ell)}) := -0.05 - 0.7\big(1 + \exp(-(x_i^{(\ell)} - 600)/10)\big)^{-1}$$

and conditional variances

$$\sigma^2(x_i^{(\ell)}) := \big(0.01 + 0.1\big(1 + \exp(-(x_i^{(\ell)} - 600)/50)\big)\big)^2,$$

respectively. The true conditional τ-quantile curves are thus given by $f^*_{\tau,\mathrm{P}}(x) = \mu(x) + u_\tau \sigma(x)$, $\tau \in (0, 1)$, where u_τ defines the τ-quantile of a normal distribution with mean 0 and variance 1. The curves for the conditional medians and the conditional lower and upper 5 percent quantiles are shown in Figure 9.2 to illustrate that this model generates data sets similar to the LIDAR data set; see Figure 9.1.

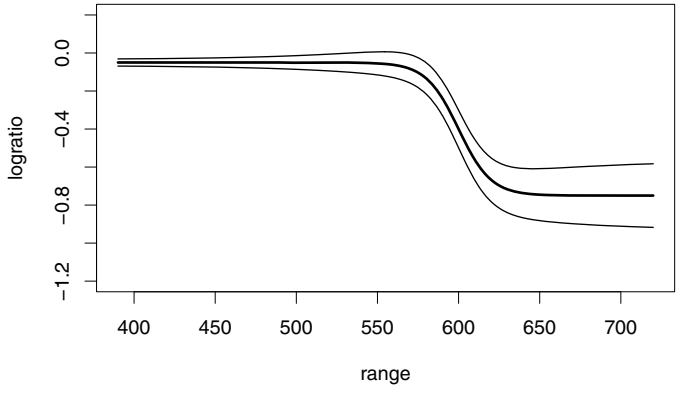


Fig. 9.2. True quantile regression curves for the simulation. Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

We use two criteria to measure how well the KBQR estimates approximate the true conditional quantiles. Our first criterion is

$$\mathrm{IMSE}_\tau := \frac{1}{1000} \sum_{\ell=1}^{1000} \frac{1}{n} \sum_{i=1}^{n} \Big(f_{\mathrm{D}_n, \lambda_n}(x_i^{(\ell)}) - f^*_{\tau,\mathrm{P}}(x_i^{(\ell)}) \Big)^2 ,$$

which is an empirical version of the integrated mean squared error. To measure the worst-case behavior of the KBQR estimates, we use the criterion

$$\text{mBias}_\tau := \frac{1}{1000} \sum_{\ell=1}^{1000} \max_{1 \le i \le n} \left| f_{\text{D}_n, \lambda_n}(x_i^{(\ell)}) - f^*_{\tau,\text{P}}(x_i^{(\ell)}) \right| ,$$

which is an empirical version of the maximum bias. The random number generation and the plots were made with the statistical software `R` (R Development Core Team, 2006). The program `mySVM` (Rüping, 2000) was used for the computation of the KBQR estimates.

The boxplots[4] given in Figures 9.3 and 9.4 show that SVMs based on the pinball loss function perform quite well with respect to both criteria under the circumstances considered, because both criteria have relatively small values and their values decrease with increasing sample sizes. A considerable improvement is obtained by increasing the sample size by a factor of around 4. The boxplots also show that the variability of the estimated conditional median is considerably smaller than the variability of the estimated conditional quantiles for $\tau \in \{0.05, 0.95\}$. The simulations indicate that the consistency results derived in Section 9.3 can be useful even for moderate sample sizes.

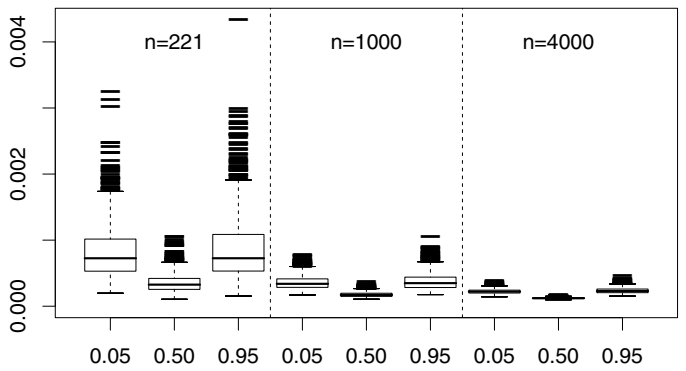


Fig. 9.3. Simulation results for the criterion IMSE_τ for SVMs for quantile regression based on the Gaussian RBF kernel with $\gamma^2 = 0.5$ and $\lambda_n = \frac{1}{700} n^{-1/3}$. Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

[4] The box is defined by the 0.25 and 0.75 quantiles, and the median is the line inside the box. The whiskers give additional information about the variability. Values outside the whiskers are shown as small lines.

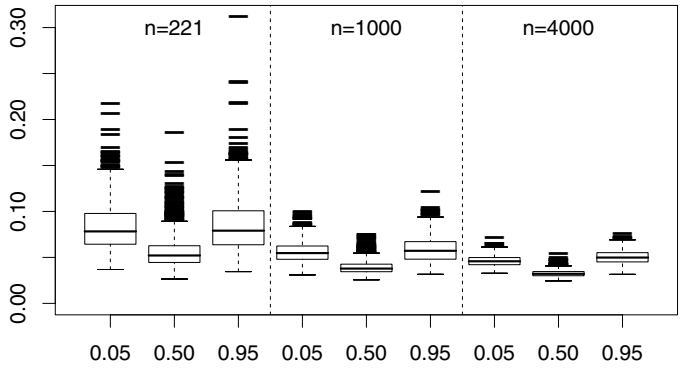


Fig. 9.4. Simulation results for the criterion mBias_τ for SVMs for quantile regression based on the Gaussian RBF kernel with $\gamma^2 = 0.5$ and $\lambda_n = \frac{1}{700} n^{-1/3}$. Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

9.5 Median Regression with the eps-Insensitive Loss (*)

In this section, we give simple conditions for the distribution P that guarantee that the set of exact minimizers of support vector machines based on the ϵ-insensitive loss function contains only one function. This fact is at first glance surprising because this loss function equals zero in the entire interval $[-\epsilon, +\epsilon]$.

In this section, we will assume that P is a distribution on $X \times Y$, where X is an arbitrary set and $Y \subset \mathbb{R}$ is closed. Further, we assume that the σ-algebra on X is complete with respect to the marginal distribution P_X of P; i.e., every subset of a P_X-zero set is contained in the σ-algebra. Since the latter can always be ensured by increasing the original σ-algebra in a suitable manner, we note that the assumption of a complete σ-algebra on X is no restriction at all. Recall from Section 3.7 that a distribution $\mathrm{Q} \in \mathcal{M}_1(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is symmetric, if there exists some $c \in \mathbb{R}$ such that $\mathrm{Q}(c + A) = \mathrm{Q}(c - A)$ for all $A \in \mathcal{B}(\mathbb{R})$ with $A \subset [0, \infty)$.

Theorem 9.8. *Let* P *be a distribution on* $X \times \mathbb{R}$ *that has a unique median* $f^*_{1/2,\mathrm{P}}$. *Further, assume that all conditional distributions* $\mathrm{P}(\,\cdot\,|\,x)$, $x \in X$, *are atom-free and symmetric. If for an* $\epsilon > 0$ *the conditional distributions have a positive mass for intervals around* $f^*_{1/2,\mathrm{P}} \pm \epsilon$, *then* $f^*_{1/2,\mathrm{P}}$ *is the only minimizer of* $\mathcal{R}_{L,\mathrm{P}}(\cdot)$ *where* L *is the* ϵ*-insensitive loss.*

The proof of Theorem 9.8 follows immediately from the following lemma.

Lemma 9.9. *Let* Q *be a symmetric, atom-free distribution on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *with median* $q^* = 0$. *Then, for* $\epsilon > 0$ *and* L *the* ϵ*-insensitive loss, we have*

$$\mathcal{C}_{L,\mathrm{Q}}(0) = \mathcal{C}^*_{L,\mathrm{Q}} = 2 \int_\epsilon^\infty \mathrm{Q}([s, \infty)),$$

and if $\mathcal{C}_{L,\mathrm{Q}}(0) < \infty$*, we further have*

$$\mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(0) = \int_{\epsilon-t}^{\epsilon} \mathrm{Q}([s,\epsilon])\,ds + \int_{\epsilon}^{\epsilon+t} \mathrm{Q}([\epsilon,s])\,ds, \qquad \text{if } t \in [0,\epsilon],$$

$$\begin{aligned}\mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(\epsilon) &= \int_0^{t-\epsilon} \mathrm{Q}([s,\infty))\,ds - \int_{2\epsilon}^{\epsilon+t} \mathrm{Q}([s,\infty))\,ds \\ &\quad +2\int_0^{t-\epsilon} \mathrm{Q}([0,s])\,ds \geq 0, \qquad \text{if } t > \epsilon.\end{aligned}$$

In particular, if $Q([\epsilon-\delta, \epsilon+\delta]) = 0$ *for some* $\delta > 0$*, then* $\mathcal{C}_{L,\mathrm{Q}}(\delta) = \mathcal{C}^*_{L,\mathrm{Q}}$.

Proof. Because $L(y,t) = L(-y,-t)$ for all $y, t \in \mathbb{R}$, we only have to consider $t \geq 0$. Recall that, given a distribution Q on $\mathbb{R}$ and a *non-negative* measurable function $g : X \to [0,\infty)$, we have

$$\int_{\mathbb{R}} g\,d\mathrm{Q} = \int_0^{\infty} \mathrm{Q}(g \geq s)\,ds\,; \tag{9.26}$$

see Lemma A.3.11. For later use, we note that for $0 \leq a \leq b \leq \infty$ equation (9.26) yields

$$\int_a^b y\,d\mathrm{Q}(y) = a\mathrm{Q}([a,b]) + \int_a^b \mathrm{Q}([s,b])\,ds\,. \tag{9.27}$$

Moreover, the definition of L implies

$$\mathcal{C}_{L,\mathrm{Q}}(t) = \int_{-\infty}^{t-\epsilon} t - y - \epsilon\,d\mathrm{Q}(y) + \int_{t+\epsilon}^{\infty} y - \epsilon - t\,d\mathrm{Q}(y)\,.$$

Using the symmetry of Q yields

$$-\int_{-\infty}^{t-\epsilon} y\,d\mathrm{Q}(y) = \int_{\epsilon-t}^{\infty} y\,d\mathrm{Q}(y),$$

and hence we obtain

$$\begin{aligned}\mathcal{C}_{L,\mathrm{Q}}(t) &= \int_0^{t-\epsilon} \mathrm{Q}((-\infty, t-\epsilon])\,ds - \int_0^{t+\epsilon} \mathrm{Q}([t+\epsilon,\infty))\,ds \\ &\quad + \int_{\epsilon-t}^{t+\epsilon} y\,d\mathrm{Q}(y) + 2\int_{t+\epsilon}^{\infty} y\,d\mathrm{Q}(y)\,.\end{aligned} \tag{9.28}$$

Let us first consider the case $t \geq \epsilon$. Then the symmetry of Q yields

$$\int_{\epsilon-t}^{t+\epsilon} y\,d\mathrm{Q}(y) = \int_{t-\epsilon}^{t+\epsilon} y\,d\mathrm{Q}(y),$$

and hence (9.27) implies

$$\mathcal{C}_{L,\mathrm{Q}}(t) = \int_0^{t-\epsilon} \mathrm{Q}([\epsilon - t, \infty))ds + \int_0^{t-\epsilon} \mathrm{Q}([t-\epsilon, t+\epsilon])\, ds + \int_{t-\epsilon}^{t+\epsilon} \mathrm{Q}([s, t+\epsilon])\, ds$$
$$+2\int_{t+\epsilon}^{\infty} \mathrm{Q}([s,\infty))\, ds + \int_0^{t+\epsilon} \mathrm{Q}([t+\epsilon,\infty))\, ds.$$

Using

$$\int_{t-\epsilon}^{t+\epsilon} \mathrm{Q}([s, t+\epsilon))\, ds = \int_0^{t+\epsilon} \mathrm{Q}([s, t+\epsilon))\, ds - \int_0^{t-\epsilon} \mathrm{Q}([s, t+\epsilon))\, ds,$$

we further obtain

$$\int_{t-\epsilon}^{t+\epsilon} \mathrm{Q}([s, t+\epsilon))\, ds + \int_0^{t+\epsilon} \mathrm{Q}([t+\epsilon,\infty))\, ds + \int_{t+\epsilon}^{\infty} \mathrm{Q}([s,\infty))\, ds$$
$$= \int_0^{\infty} \mathrm{Q}([s,\infty))\, ds - \int_0^{t-\epsilon} \mathrm{Q}([s, t+\epsilon))\, ds.$$

From this and

$$\int_0^{t-\epsilon} \mathrm{Q}([t-\epsilon, t+\epsilon])\, ds - \int_0^{t-\epsilon} \mathrm{Q}([s, t+\epsilon])\, ds = -\int_0^{t-\epsilon} \mathrm{Q}([s, t-\epsilon])\, ds,$$

it follows that $\mathcal{C}_{L,\mathrm{Q}}(t)$ equals

$$-\int_0^{t-\epsilon} \mathrm{Q}([s, t-\epsilon])\, ds + \int_0^{t-\epsilon} \mathrm{Q}([\epsilon - t, \infty))\, ds + \int_{t+\epsilon}^{\infty} \mathrm{Q}([s,\infty))\, ds + \int_0^{\infty} \mathrm{Q}([s,\infty))\, ds.$$

The symmetry of Q implies

$$\int_0^{t-\epsilon} \mathrm{Q}([\epsilon - t, t-\epsilon])\, ds = 2\int_0^{t-\epsilon} \mathrm{Q}([0, t-\epsilon])\, ds,$$

such that

$$-\int_0^{t-\epsilon} \mathrm{Q}([s, t-\epsilon])\, ds + \int_0^{t-\epsilon} \mathrm{Q}([\epsilon - t, \infty))\, ds$$
$$= 2\int_0^{t-\epsilon} \mathrm{Q}([0, s))\, ds + \int_0^{t-\epsilon} \mathrm{Q}([s,\infty))\, ds.$$

This and

$$\int_{t+\epsilon}^{\infty} \mathrm{Q}([s,\infty))\, ds + \int_0^{\infty} \mathrm{Q}([s,\infty))\, ds = 2\int_{t+\epsilon}^{\infty} \mathrm{Q}([s,\infty))\, ds + \int_0^{t+\epsilon} \mathrm{Q}([s,\infty))\, ds$$

shows that $\mathcal{C}_{L,\mathrm{Q}}(t)$ equals

$$2\int_0^{t-\epsilon} \mathrm{Q}([0, s))\, ds + \int_0^{t-\epsilon} \mathrm{Q}([s,\infty))\, ds + 2\int_{t+\epsilon}^{\infty} \mathrm{Q}([s,\infty))\, ds + \int_0^{t+\epsilon} \mathrm{Q}([s,\infty))\, ds.$$

By

$$\int_0^{t-\epsilon} \mathrm{Q}([s,\infty))\,ds + \int_0^{t+\epsilon} \mathrm{Q}([s,\infty))\,ds = 2\int_0^{t-\epsilon} \mathrm{Q}([s,\infty))\,ds + \int_{t-\epsilon}^{t+\epsilon} \mathrm{Q}([s,\infty))\,ds,$$

we obtain

$$\mathcal{C}_{L,\mathrm{Q}}(t) = 2\int_0^{t-\epsilon} \mathrm{Q}([0,\infty))\,ds + 2\int_{t+\epsilon}^{\infty} \mathrm{Q}([s,\infty))\,ds + \int_{t-\epsilon}^{t+\epsilon} \mathrm{Q}([s,\infty))\,ds$$

if $t \geq \epsilon$. Let us now consider the case $t \in [0,\epsilon]$. Analogously, we get from (9.28) that $\mathcal{C}_{L,\mathrm{Q}}(t)$ equals

$$\begin{aligned}&\int_0^{\epsilon-t} \mathrm{Q}([\epsilon-t,t+\epsilon])\,ds + \int_{\epsilon-t}^{\epsilon+t} \mathrm{Q}([s,t+\epsilon])\,ds + 2\int_{\epsilon+t}^{\infty} \mathrm{Q}([s,\infty))\,ds\\ &+2\int_0^{\epsilon+t} \mathrm{Q}([\epsilon+t,\infty))\,ds - \int_0^{\epsilon-t} \mathrm{Q}([\epsilon-t,\infty))\,ds - \int_0^{\epsilon+t} \mathrm{Q}([\epsilon+t,\infty))\,ds.\end{aligned}$$

Combining this with

$$\int_0^{\epsilon-t} \mathrm{Q}([\epsilon-t,t+\epsilon])\,ds - \int_0^{\epsilon-t} \mathrm{Q}([\epsilon-t,\infty))\,ds = -\int_0^{\epsilon-t} \mathrm{Q}([\epsilon+t,\infty))\,ds$$

and

$$\int_0^{\epsilon+t} \mathrm{Q}([\epsilon+t,\infty))\,ds - \int_0^{\epsilon-t} \mathrm{Q}([\epsilon+t,\infty))\,ds = \int_{\epsilon-t}^{\epsilon+t} \mathrm{Q}([\epsilon+t,\infty))\,ds,$$

we get

$$\begin{aligned}\mathcal{C}_{L,\mathrm{Q}}(t) &= \int_{\epsilon-t}^{\epsilon+t} \mathrm{Q}([\epsilon+t,\infty))\,ds + \int_{\epsilon-t}^{\epsilon+t} \mathrm{Q}([s,t+\epsilon])\,ds + 2\int_{\epsilon+t}^{\infty} \mathrm{Q}([s,\infty))\,ds\\ &= \int_{\epsilon-t}^{\epsilon+t} \mathrm{Q}([s,\infty))\,ds + 2\int_{\epsilon+t}^{\infty} \mathrm{Q}([s,\infty))\,ds\\ &= \int_{\epsilon-t}^{\infty} \mathrm{Q}([s,\infty))\,ds + \int_{\epsilon+t}^{\infty} \mathrm{Q}([s,\infty))\,ds.\end{aligned}$$

Hence

$$\mathcal{C}_{L,\mathrm{Q}}(0) = 2\int_{\epsilon}^{\infty} \mathrm{Q}([s,\infty))\,ds.$$

The expressions for $\mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(0)$, $t \in (0,\epsilon]$, and $\mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(\epsilon)$, $t > \epsilon$, given in Lemma 9.9 follow by using the same arguments. Hence one exact minimizer of $\mathcal{C}_{L,\mathrm{Q}}(\cdot)$ is the median $t^* = 0$. The last assertion is a direct consequence of the formula for $\mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(0)$ in the case $t \in (0,\epsilon]$. □

9.6 Further Reading and Advanced Topics

For additional information, we refer to Poggio and Girosi (1990), Wahba (1990), Vapnik (1995, 1998), Schölkopf and Smola (2002), and the references cited by these authors. For ν-support vector regression, which is strongly related to support vector regression based on the ϵ-insensitive loss function but allows the tube width to adapt automatically to the data, see Schölkopf *et al.* (2000), Smola and Schölkopf (2004), and Chen *et al.* (2005). For a brief description of kernel ridge regression and Gaussian processes, see Cristianini and Shawe-Taylor (2000, Section 6.2). We refer to Wahba (1999) for the relationship between SVMs and Gaussian processes.

Section 9.2 on the L-risk consistency of SVMs for regression was based on Christmann and Steinwart (2007). An unbounded output space Y instead of a bounded one makes proofs of L-risk consistency of SVMs harder. This is also true for investigating robustness properties; see Chapter 10. This was one reason why we treated Nemitski loss functions in Chapters 2 and 5. Equation (9.4) shows that the reproducing kernel Hilbert space H is a Banach space with type and cotype two. For details, we refer to Diestel *et al.* (1995).

Quantile regression for linear models was proposed by Koenker and Bassett (1978). A recent textbook on this topic is Koenker (2005). We refer to Koenker (1986) for strong consistency of regression quantiles and related empirical processes, He and Liang (2000) for quantile regression in errors-in-variables models, Portnoy (2003) for censored regression quantiles, and Koenker and Xiao (2006) for quantile autoregression. SVMs for quantile regression were proposed by Schölkopf *et al.* (2000, p. 1216) and Takeuchi *et al.* (2006). The latter article also describes algorithmic aspects for the efficient computation of SVMs for quantile regression. It is possible that the estimated conditional quantile functions intersect for different values of $\tau \in (0,1)$. For a discussion of this crossing problem and how to overcome this unfavorable property, we refer to Example 11.9, He (1997), and Takeuchi *et al.* (2006). For non-parametric generalizations of quantile regression based on splines, we refer to Koenker *et al.* (1994) and He and Ng (1999).

Sections 9.3 and 9.4 on the L-risk consistency of SVMs for quantile regression are based on Christmann and Steinwart (2008). We conjecture[5] that it is possible to get rid of the assumption $|\mathrm{P}|_1 < \infty$ if one changes the regularized minimization problem to

$$f_{\mathrm{P},\lambda} := \arg \inf_{f \in H} \mathbb{E}_{\mathrm{P}} L^*(Y, f(X)) + \lambda \|f\|_H^2 \,,$$

where $L^*(y,t) := L(y,t) - L(y,0)$. However, loss functions that can take on negative values are not treated in this textbook because many practitioners do not accept negative losses.

Section 9.5 on the uniqueness of the SVM solution based on the ϵ-insensitive loss function is based on Steinwart and Christmann (2008).

[5] As a result of discussions with Ursula Gather and Xuming He

For non-parametric regression with constraints such as monotonicity or convexity, we refer to Smola and Schölkopf (1998), Takeuchi *et al.* (2006), Hall and Huang (2001), and Dette *et al.* (2006).

9.7 Summary

This chapter gave a mathematical justification for the informal notion that support vector machines are "able to learn" in non-parametric regression models. It was shown that SVMs are L-risk consistent for a broad class of convex loss functions under weak assumptions even for the case of an unbounded output space. Support vector machines based on the pinball loss function lead to kernel-based quantile regression. SVMs based on this loss function are L-risk consistent under weak assumptions on P and k. Furthermore, if this SVM is L-risk consistent, we also obtained a consistent estimator for the conditional quantile function. Finally, conditions were derived under which SVMs based on the ϵ-insensitive loss function, which is equal to zero in the interval $[-\epsilon, +\epsilon]$, allow a *unique* estimation of the conditional median function.

9.8 Exercises

9.1. Pinball loss ($\star$)
Prove that the pinball loss function $L_{\tau\text{-pin}}(y-t) = \psi_\tau(y-t)$, $y, t \in \mathbb{R}$, has for each $\tau \in (0,1)$ the following properties.

i) ψ_τ is strictly convex and satisfies both $\psi_\tau(0) = 0$ and $\lim_{|r|\to\infty} \psi_\tau(r) = \infty$.
ii) ψ_τ is Lipschitz continuous with Lipschitz constant $|\psi_\tau|_1 = \max\{\tau, 1-\tau\}$.
iii) For all $r \in \mathbb{R}$, we have $\min\{\tau, 1-\tau\}\,|r| \le \psi_\tau(r) \le |\psi_\tau|_1\,|r|$.

9.2. Translation-invariant metric ($\star\star$)
Let P be a distribution on $X \times Y$ with $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ both closed sets. Let $f, g : X \to \mathbb{R}$ be measurable functions. Define

$$\|f\|_{L_0(\mathrm{P}_X)} := \|f\|_0 := \mathbb{E}_{\mathrm{P}_X} \min\bigl\{1, |f(X)|\bigr\}$$

and $d(f,g) := \|f - g\|_0$; see (9.22). Show that d is a *translation-invariant* metric on the space of all measurable functions defined on X and d describes the convergence in probability P_X.

Hint: Use Chebyshev's inequality.

9.3. Least squares loss ($\star\star$)
Specialize Theorem 9.1 for the least squares loss function. Investigate in which sense the conditional mean function is approximated.

9.4. Comparison least squares loss and logistic loss ($\star\star$)
Compare the results of the previous exercise concerning L_{LS} with the results for the logistic loss function for the special case of a symmetric conditional distribution of Y given x. Consider the bounded and the unbounded cases.

9.5. Consistency for Lipschitz-continuous loss functions ($\star\star$)
Generalize Theorem 9.6 based on the pinball loss function $L_{\tau\text{-pin}}$ to general Lipschitz-continuous loss functions.

9.6. Quantile regression ($\star\star\star\star$)
Generalize the results of Section 9.3 to the case of non-unique quantiles.

9.7. A variance bound for the pinball loss ($\star\star\star$)
For fixed $\tau \in (0,1)$, let L be the τ-pinball loss. Moreover, let $Y := [-1,1]$, X be a measurable space, and P be a distribution on $X \times Y$ such that $\mathrm{P}(\cdot\,|x)$ is of type $\mathcal{Q}_\tau^\alpha$ for some $\alpha > 0$ and all $x \in X$. In other words, (3.73) is satisfied for $\mathrm{Q} := \mathrm{P}(\cdot\,|x)$, $x \in X$. We write $b(x) := c_{\mathrm{P}(\cdot\,|x)}$, $x \in X$, for the corresponding constants in (3.73) and assume that $b \in L_q(\mathrm{P}_X)$ for some $q \in (0,\infty]$. Show that there exists a constant $V \geq 1$ such that

$$\mathbb{E}_{\mathrm{P}}(L \circ \widehat{f} - L \circ f^*_{\tau,\mathrm{P}})^2 \leq V \cdot \bigl(\mathbb{E}_{\mathrm{P}}(L \circ \widehat{f} - L \circ f^*_{\tau,\mathrm{P}})\bigr)^{\vartheta} \tag{9.29}$$

for all measurable $f : X \to \mathbb{R}$, where $\vartheta := q/(2q+2)$, $\widehat{t} := \max\{-1, \min\{1,t\}\}$ denotes the clipping operation (2.14) at $M := 1$, and $L \circ f(x,y) := L(y, f(x))$.

Hint: For clipped functions with "small" excess risk, use the Lipschitz-continuity and (3.74), whereas for clipped functions with "large" excess risk, use the fact that the left-hand side of (9.29) is never larger than 4.

9.8. Oracle inequality for quantile regression ($\star\star\star$)
For fixed $\tau \in (0,1)$, let L be the τ-pinball loss. Moreover, let $Y := [-1,1]$, X be a measurable space, P be a distribution on $X \times Y$, and H be the separable RKHS of a measurable kernel. Assume that (9.29) holds and that there are constants $p \in (0,1)$ and $a \geq 1$ such that the entropy numbers satisfy

$$e_i(\mathrm{id} : H \to L_2(\mathrm{P}_X)) \leq a\, i^{-1/(2p)}\,, \qquad i \geq 1.$$

Show that there is a constant K only depending on p, ϑ, and V such that for all $\tau, \lambda > 0$, and $n \geq 1$ we have with probability not less than $1 - 3e^{-\tau}$ that

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} \leq 9A_2(\lambda) + \frac{15\tau}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}} + K\Bigl(\frac{a^{2p}}{\lambda^p n}\Bigr)^{c(p,\vartheta)} + K\tau n^{-\frac{1}{2-\vartheta}}\,,$$

where $c(p,\vartheta) := 1/(2-p-\vartheta+\vartheta p)$ and $A_2(\lambda)$ denotes the approximation error function from Definition 5.14. Use this oracle inequality to derive L-risk learning rates for SVMs based on the pinball loss in the sense of the discussion following Theorem 7.23. Moreover, interpret these rates with the help of (3.74) if P satisfies the assumptions of Exercise 9.7, and discuss the adaptivity of a TV-SVM. Finally, apply the discussion in Section 5.6 if H is a Sobolev space and $f^*_{\tau,\mathrm{P}}$ is contained in another Sobolev space.

Hint: Use Theorem 7.23 together with Corollary 7.31.

10

Robustness

Overview. *So far, we have assumed that the pairs of input and output variables are independently generated by the same probability distribution* P*. In this chapter, we investigate robustness properties of support vector machines for classification and regression. Here we will determine the stability of SVMs if the distribution is not* P *but some other distribution* Q *close to* P*. For example,* Q *can be the empirical distribution generated by a data set or a mixture distribution* $\mathrm{Q} = (1-\varepsilon)\mathrm{P}+\varepsilon\tilde{\mathrm{P}}$*, where* $\tilde{\mathrm{P}}$ *can be any distribution. Using methods from robust statistics, we obtain conditions on the loss function and the kernel that guarantee that SVMs are stable in a neighborhood around* P*.*

Prerequisites. *Knowledge of loss functions, kernels, and the stability of infinite-sample SVMs is needed from Chapter 2, Section 3.9, Chapter 4, and Section 5.3. Some results from measure and integration theory, statistics, and functional analysis from the appendix are used.*

In this chapter, we argue that robustness is an important aspect for statistical machine learning. This chapter can be seen as a continuation of Section 5.3, which investigated the stability of infinite-sample SVMs. It will be shown that SVMs also have—besides other good properties—the advantage of being robust if the loss function L and the kernel k are carefully chosen. Weak conditions on L and k are derived that guarantee good robustness of SVM methods not only for some fixed parametric class of distributions—say Gaussian distributions—but for large classes of probability distributions. In the sense specified by Hadamard (1902), support vector machines are hence well-posed problems. Hadamard believed that well-posed mathematical problems should have the property that there exists a unique solution that additionally depends on the data continuously.

In previous chapters of the book, statistical properties of SVMs were derived such as existence, uniqueness, and L-risk consistency. These properties were derived under the model assumption that the pairs (X_i, Y_i), $i = 1, \ldots, n$, are *independent* and *identically distributed* random variables on some space $X \times Y$ each with the (totally unknown) probability distribution $\mathrm{P} \in \mathcal{M}_1$, where $\mathcal{M}_1$ denotes the set of all probability distributions on $X \times Y$. This is surely a weak assumption if it is compared with parametric models that assume that P is an element of a specified *finite-dimensional* subset of $\mathcal{M}_1$.

From a theoretical and an applied point of view, it is nevertheless worthwhile to investigate the possible impact of model violations on $f_{\mathrm{P},\lambda}$, $f_{\mathrm{D},\lambda}$, and the corresponding L-risks. The results of this chapter can be used to choose the loss function L and the kernel k such that the corresponding SVM is robust.

If not otherwise stated, we assume in this chapter that $X = \mathbb{R}^d$, $Y = \mathbb{R}$, and $d \in \mathbb{N}$.

The rest of the chapter is organized as follows. Section 10.1 gives a motivation for investigating robustness properties of SVM methods. Section 10.2 describes general concepts of robust statistics: qualitative robustness, influence functions, the related notions of gross error sensitivity, the sensitivity curve, and maxbias, and breakdown points. In the following two sections, we investigate in detail robustness properties of SVMs for classification and regression and clarify the role of the loss function and the kernel. Section 10.5 treats a simple but powerful strategy based on independent subsampling for calculating SVMs from huge data sets for which current numerical algorithms might be too slow. It will be shown that this strategy offers robust and consistent estimations if the SVM used is itself robust and consistent.

10.1 Motivation

Why Is Robustness Important?

In almost all cases, statistical models are only approximations to the true random process that generated a given data set.[1] Hence it is necessary to investigate how deviations may influence the results. J. W. Tukey, one of the pioneers of robust statistics, mentioned already in 1960 (Hampel *et al.*, 1986, p. 21):

> *A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians.*

The main aims of robust statistics are the description of the structure best fitting the *bulk* of the data and the identification for further treatment of points deviating from this structure or deviating substructures, see Hampel *et al.* (1986). An informal description of a good robust method is as follows.

i) A good robust method offers results with a reasonably high quality when the data set was in fact generated by the assumed model.

[1] This is especially true in data mining; see Chapter 12.

ii) If the strict model assumptions are violated, then the results of a robust method are only influenced in a bounded way by a few data points that deviate grossly from the structure of the bulk of the data set or by many data points that deviate only mildly from the structure of the bulk of the data set.

Although these two considerations are for data sets (i.e., for the finite-sample case), it will become clear that asymptotic considerations are also helpful for investigating robustness properties.

Support vector machines are non-parametric methods and make no specific assumptions on the distribution P. Nevertheless, the robustness issue is important also for SVMs because the two classical assumptions that all data points are generated *independently* by the *same distribution* can be violated in practice. One reason is that outliers often occur in real data sets. *Outliers* can be described as data points that "*are far away ... from the pattern set by the majority of the data*"; see Hampel *et al.* (1986, p. 25). Let us consider a simple two-dimensional example. Figure 10.1 shows a simulated data set from a bivariate Gaussian distribution $N(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1.0 & 1.2 \\ 1.2 & 2.0 \end{pmatrix},$$

and with four artificial outliers. Of course, μ and Σ will usually be unknown in applications. The ellipses cover 95 percent (dashed) and 99.99 percent (dotted) of the mass of this bivariate Gaussian distribution. The points $a = (4, 5)$ and $d = (5, -5)$ are extreme in the x- and y-directions, $b = (4, 0)$ is only extreme in the x-direction, and $c = (1.5, -2)$ is non-extreme in both directions but deviates strongly from the pattern set by the majority of the data. It is often relatively easy to detect outliers that are extreme in at least one component, in this case the points a, b, and d. However, outliers are often very hard to identify in high-dimensional data sets due to the curse of high dimensionality and because it is not feasible to consider all one-dimensional projections. One reason is that such outliers can be non-extreme in every component and in many lower-dimensional subspaces but are extreme if all components and the (unknown) pattern set by the majority of the data are taken into account at the same time; see the point c in Figure 10.1. With respect to this bivariate Gaussian distribution, c is even more extreme than a because c is not even an element of the ellipse containing 99.99 percent of the distribution mass of the specified bivariate Gaussian distribution. This can also be seen by computing the so-called *Mahalanobis distance*, which is defined by

$$\sqrt{(z-\mu)^\mathsf{T} \Sigma^{-1} (z-\mu)}, \quad z \in \mathbb{R}^2.$$

The Mahalanobis distance is 4.0 for a but equals 5.3 for c and 15.0 for d. In contrast to that, the Euclidean distance is not helpful here. The Euclidean distance of c to μ is only 2.5, which is smaller than the Euclidean distance

of a to μ, which equals 6.4. Outliers like c and d usually have a large impact on non-robust methods and can make the whole statistical analysis useless if a non-robust method is used; see, e.g., Rousseeuw (1984), Rousseeuw and van Zomeren (1990), and Davies and Gather (1993).

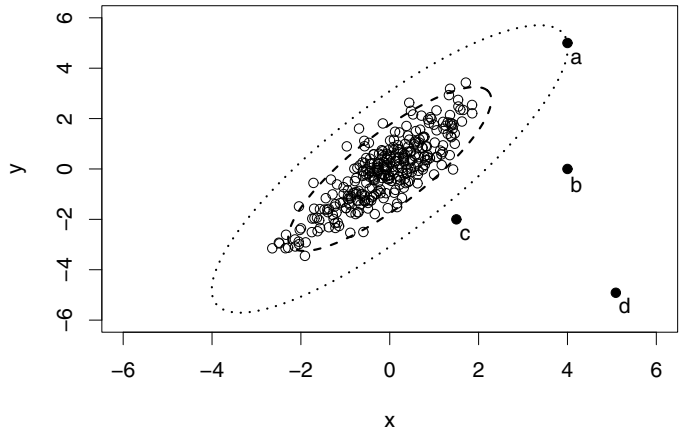


Fig. 10.1. Types of outliers.

There are many reasons for the occurrence of outliers or extreme values. *Typing errors* are often present if the data points are reported manually. *Gross errors* are errors due to a source of deviation that acts only occasionally but is quite powerful. Another reason might be that the whole data set or a part of it was actually generated by another distribution, say a Student's t-distribution or a generalized Pareto distribution, under which extreme values are much more likely than under the assumed model, say the class of Gaussian distributions. It can happen that outliers are even correlated, which contradicts the classical assumption that the observations in the data set are generated in an independent manner.

One might ask whether it is necessary to pay attention to outliers or gross errors. This is definitely true, as the number of patents on methods claiming reliable outlier detection shows. We would like to give three reasons for the importance of outliers:

i) Outliers do occur in practice. There are often no or virtually no gross errors in high-quality data, but 1 percent to 10 percent of gross errors in routine data seem to be more the rule than the exception; see Hampel *et al.* (1986, pp. 27ff.). The data quality is sometimes far from being optimal, especially in data mining problems, as will be explained in Chapter 12.

ii) Outliers may unfortunately have a high impact on the results if methods are used that do *not* bound the impact of outliers.

iii) Outliers may be interesting in their own right because they show a different behavior than the bulk of the data. This might even indicate some novelty worth considering in a detailed subsequent analysis. It is well-known that outlier identification based on robust methods is much safer than outlier identification based on non-robust methods.

It is worth mentioning that, from a robustness point of view, the occurrence of outliers is only one of several possible deviations from the assumed model. Obviously, it is in general *not* the goal to *model* the occurrence of typing errors or gross errors because it is unlikely that they will occur in the same manner for other data sets that will be collected in the future.

Let us summarize. The classical assumptions made by SVMs of independent random vectors (X_i, Y_i), $i = 1, \ldots, n$, each having the same distribution P, are weak but nevertheless not always fulfilled. Hence the question arises which impact small distortions of P may have on SVMs. Of course, the same question arises for empirical SVMs, where the unknown distribution P is replaced by the empirical distribution D. We will later use neighborhoods of P or D in the metric space of probability distributions to specify precisely what is meant by small distortions.

What Are the Goals of Robust Statistics?

In a nutshell, robust statistics investigates the impact that violations of the statistical model, outliers, and gross errors can have on the results of estimation, testing, or prediction methods and develops methods such that the impact is bounded.

Figure 10.2 sketches the idea of robustness from a somewhat more mathematical point of view. Assume that (X_i, Y_i), $i = 1, \ldots, n$, are independent and identically distributed random variables on some space $X \times Y \subset \mathbb{R}^d \times \mathbb{R}$ with unknown probability distribution P. Denote the empirical distribution corresponding to a data set $D_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ by $\mathrm{D} = \mathrm{D}_n$. The sequence $(\mathrm{D}_n)_{n \in \mathbb{N}}$ converges almost surely by Theorem A.4.11 to the true data-generating distribution P if the sample size n converges to infinity. Assume that the statistical method of interest can be written as $S(\mathrm{D}_n)$ for any possible data set D_n and more general, as $S(\mathrm{P})$ for any distribution P, where

$$S : \mathrm{P} \mapsto S(\mathrm{P}) \tag{10.1}$$

is a measurable function specifying the statistical method. In this chapter, we will assume that the functions considered are measurable. In our case, we will have

$$S(\mathrm{P}) := f_{\mathrm{P},\lambda} = \arg\min_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|_H^2 \,, \tag{10.2}$$

where $\lambda \geq 0$. A robust method S should be bounded and smooth in a neighborhood of P. Let us assume for a moment that continuity or a bounded

directional derivative is meant by smoothness. Let $\mathcal{M}_1$ be the set of all probability measures on $X \times Y$ and a corresponding σ-algebra, d_1 be a metric on $\mathcal{M}_1$, and d_2 be a metric on the space of induced probability measures of $S(\mathrm{P})$, $\mathrm{P} \in \mathcal{M}_1$. Then we expect from a robust method that for all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$d_1(\mathrm{Q}, \mathrm{P}) < \delta \quad \Rightarrow \quad d_2(S(\mathrm{Q}), S(\mathrm{P})) < \varepsilon,$$

or that the derivative of $S(\mathrm{P})$ in the direction of $S(\mathrm{Q})$ is bounded. The smoothness property should in particular be valid for any sequence $(\mathrm{P}_n)_{n\in\mathbb{N}}$ of probability distributions converging to P. A special case are sequences of empirical distributions $(\mathrm{D}_n)_{n\in\mathbb{N}}$ converging to P. Recall that P is unknown. Hence it is essential that the function S has this smoothness property not only for one particular distribution P but for a large subclass of $\mathcal{M}_1$.

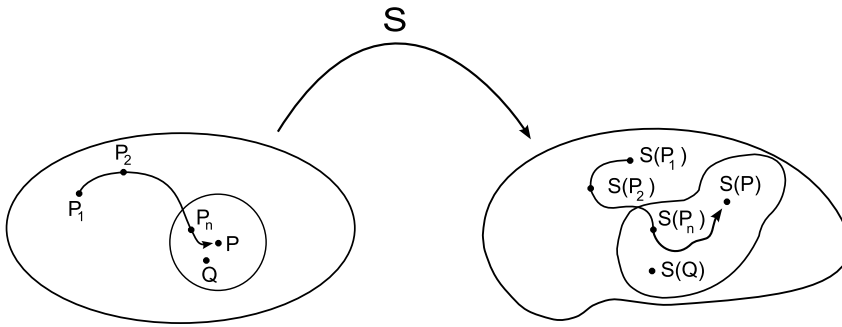


Fig. 10.2. Sketch: reasoning of robustness of $S(\mathrm{P})$. Left: P, a δ-neighborhood of P, and $\mathcal{M}_1$. Right: $S(\mathrm{P})$, an ε-neighborhood of $S(\mathrm{P})$, and the space of all probability measures of $S(\mathrm{P})$ for $\mathrm{P} \in \mathcal{M}_1$.

Let us consider two rather simple examples: the mean and the median. Let $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, $n \in \mathbb{N}$. The (empirical) mean is defined by

$$\bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i \,, \tag{10.3}$$

and the (empirical) median is given by

$$\mathrm{median}(y) := \begin{cases} y_{(\ell:n)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(y_{(\ell:n)} + y_{((\ell+1):n)}\right) & \text{if } n \text{ is even}\,, \end{cases} \tag{10.4}$$

where $\ell = \lfloor (n+1)/2 \rfloor$, and $y_{(1:n)} \le \ldots \le y_{(n:n)}$ denote the ordered values of $\{y_i, i = 1, \ldots, n\}$. Here we use the usual way of making the median unique for

n even. Now assume that the data points y_i are the observations of independent and identically distributed random variables Y_i on $(\mathbb{R}, \mathcal{B})$, $i = 1, \ldots, n$. Clearly, $\bar{Y} := \frac{1}{n} \sum_{i=1}^{n} Y_i$ is the solution of

$$\bar{Y} = \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta)^2 \,, \tag{10.5}$$

and the median is a solution of

$$\text{median}(Y) = \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - \theta| \,. \tag{10.6}$$

The mean and median are also solutions of a special class of empirical risk minimization problems. This can easily be seen if we use a reproducing kernel Hilbert space H defined via a *linear* kernel $k(x, x') = \langle x, x' \rangle$, the least squares loss or the L_1 loss function, which is identical to the pinball loss function for $\tau = \frac{1}{2}$, and a data set containing n data points $z_i = (x_i, y_i) \in \mathbb{R}^2$ with $x_i \equiv 1$, $i = 1, \ldots, n$. Assume that $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$, are independent and identically distributed random variables on $(\mathbb{R}^2, \mathcal{B}^2)$ with distribution $\mathrm{P} \in \mathcal{M}_1$ such that the marginal distribution P_X of X_i is equal to the Dirac distribution $\delta_{\{1\}}$. Under these assumptions, we obtain

$$S(\mathrm{D}) := f_{\mathrm{D},0} = \arg\min_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \tag{10.7}$$

with the corresponding infinite-sample problem

$$S(\mathrm{P}) := f_{\mathrm{P},0} = \arg\min_{f \in H} \mathbb{E}_{\mathrm{P}}\, L\big(Y, f(X)\big) \,. \tag{10.8}$$

Therefore, we can consider the mean and median either as function values $S_n((z_1, \ldots, z_n))$ of a function

$$S_n : \big(\mathbb{R}^{2n}, \mathcal{B}(\mathbb{R}^{2n})\big) \to (\mathbb{R}, \mathcal{B}) \tag{10.9}$$

(see (10.3) and (10.4)) or as function values $S(\mathrm{D}_n)$ of a function S, where $\mathrm{D}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}$ denotes the empirical distribution and

$$S : \mathcal{M}_1(\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B}) \,, \tag{10.10}$$

$$S(\mathrm{P}) = \arg\min_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|_H^2 \,, \tag{10.11}$$

where $\lambda = 0$ (see (10.7) and (10.8)). As far as we know, this functional approach was first used by von Mises (1937), and it is now a standard and powerful tool for investigating robustness properties of statistical methods. Robustness of an estimator S_n can then be defined via properties of the function S by requiring continuity, differentiability, or boundedness of S in P or in neighborhoods of P. This will be described in the next section.

10.2 Approaches to Robust Statistics

This section contains the necessary preliminaries to investigate robustness properties of SVMs in Sections 10.3 to 10.5. A reader familiar with robust statistics may skip this section and go directly to Section 10.3. A reader not familiar with robust statistics can find additional information also in Appendix A.4.

In the statistical literature, many different criteria have been proposed to define robustness in a mathematical way. The definitions of qualitative robustness and breakdown points are given, but we will mainly restrict attention to influence functions together with the related notions of gross error sensitivity, sensitivity curve, and maxbias in the subsequent sections.

Qualitative Robustness

As we explained in the previous section, neighborhoods and distances of probability measures are important in robust statistics. Let us therefore start with the definition of the Prohorov metric, which is needed for the definition of qualitative robustness. Let (Z, τ_Z) be a Polish space with complete metric d_Z. For any set $A \subset Z$, we define the *closed δ-neighborhood* of a set A by

$$A^\delta := \bigl\{x \in Z : \inf_{y \in A} d_Z(x,y) \le \delta\bigr\}. \tag{10.12}$$

Definition 10.1. *Let (Z, τ_Z) be a Polish space, $\mathrm{P} \in \mathcal{M}_1(Z)$ be a probability measure on Z, and $\varepsilon > 0$, $\delta > 0$. Then the set*

$$N^{Pro}_{\varepsilon,\delta}(\mathrm{P}) := \bigl\{\mathrm{Q} \in \mathcal{M}_1(Z) : \mathrm{Q}(A) \le \mathrm{P}(A^\delta) + \varepsilon \ \text{ for all } A \in \mathcal{B}(Z)\bigr\} \tag{10.13}$$

is called a ***Prohorov neighborhood*** *of* P*. We write $N^{Pro}_{\varepsilon}(\mathrm{P})$ instead of $N^{Pro}_{\varepsilon,\varepsilon}(\mathrm{P})$. The* ***Prohorov metric*** *between two probability distributions P_1, $\mathrm{P}_2 \in \mathcal{M}_1(Z)$ is defined by*

$$d_{\mathrm{Pro}}(\mathrm{P}_1, \mathrm{P}_2) := \inf\bigl\{\varepsilon > 0 : \mathrm{P}_1(A) \le \mathrm{P}_2(A^\varepsilon) + \varepsilon \ \text{ for all } A \in \mathcal{B}(Z)\bigr\}. \tag{10.14}$$

The Prohorov neighborhood has the property that it has a component reflecting statistical uncertainty (i.e.; the dimensionless term ε) and a component reflecting the occurrence of rounding errors (i.e.; the term δ that, however, is not dimensionless). The Prohorov metric defines a metric on $\mathcal{M}_1(Z)$ and metricizes the weak* topology in $\mathcal{M}_1(Z)$; see Theorem A.4.19 and Theorem A.4.20. The weak* topology of $\mathcal{M}_1$ can also be metricized by other metrics; see Theorem A.4.22 for the bounded Lipschitz metric. For the special case $Z = \mathbb{R}$, the Lévy metric (see Problem 10.4) is often easier to use than the Prohorov metric because the Lévy metric is based on cumulative distribution functions.

Let (Z, τ_Z) and (W, τ_W) be Polish spaces with complete metrics d_Z and d_W that metricize the weak* topologies. Define a metric on Z^n by $d_{Z^n}(z, z') =$

$\max_{1\le i\le n} d_Z(z_i, z_i')$, where $z = (z_1, \ldots, z_n) \in Z^n$ and $z' = (z_1', \ldots, z_n') \in Z^n$. A data set consisting of n data points $z_i \in Z$, $i = 1, \ldots, n$, will be denoted by D_n, and the corresponding empirical measure will be denoted by D_n. Let $Z_1, \ldots, Z_n$ be independent and identically distributed random variables, each with probability distribution $\mathrm{P} \in \mathcal{M}_1(Z)$. We will denote the empirical distribution of a random sample from P of size n by P_n because it will often be important in this section to be precise which distribution generated the data set. The set of all empirical distributions based on n points is denoted by $\mathcal{M}_{1n}(Z) = \left\{ \frac{1}{n} \sum_{i=1}^n \delta_{z_i} : z_i \in Z, i = 1, \ldots, n \right\}$. Furthermore, let $S_n : Z^n \to W$, $n \in \mathbb{N}$, be a sequence of random functions. The distribution of S_n will be denoted by P_{S_n} if Z_i has distribution P. Often there exists a measurable function $S : \mathcal{M}_1(Z) \to W$ such that $S_n = S(\mathrm{P}_n)$ for all $\mathrm{P}_n \in \mathcal{M}_{1n}(Z)$ and all $n \in \mathbb{N}$. The following definition goes back to Hampel (1968) and was generalized by Cuevas (1988) for Polish spaces.

Definition 10.2. *Let (Z, τ_Z) and (W, τ_W) be Polish spaces and $(S_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions, where $S_n : Z^n \to W$, $S_n(Z_1, \ldots, Z_n) \in W$, and $Z_1, \ldots, Z_n$ are independent and identically distributed according to $\mathrm{P} \in \mathcal{M}_1(Z)$. The sequence $(S_n)_{n\in\mathbb{N}}$ is called* ***qualitatively robust*** *at P if*

$$\forall \varepsilon > 0 \; \exists \delta > 0 : \left\{ d_{\mathrm{Pro}}(\mathrm{P}, \mathrm{Q}) < \delta \;\Rightarrow\; d_{\mathrm{Pro}}\big(\mathrm{P}_{S_n}, \mathrm{Q}_{S_n}\big) < \varepsilon, \forall\, n \in \mathbb{N} \right\}. \quad (10.15)$$

Qualitative robustness, which is defined as equicontinuity of the distributions of the statistic as the sample size changes, is hence closely related to continuity of the statistic viewed as a function in the weak* topology; see Theorems A.4.26 and A.4.27.

Remark 10.3. The arithmetic mean $S_n(Y_1, \ldots, Y_n) := \frac{1}{n} \sum_{i=1}^n Y_i$ of real-valued random variables can be written as $S_n(Y_1, \ldots, Y_n) = \mathbb{E}_{\mathrm{P}_n}(Y)$, where $\mathrm{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$. By the law of large numbers (see Theorems A.4.8 and A.4.9), $\mathbb{E}_{\mathrm{P}_n}(Y)$ converges in probability or almost surely to $\mathbb{E}_{\mathrm{P}}(Y)$ if mild assumptions are satisfied. Furthermore, the mean is the uniformly minimum variance unbiased estimator of the expectation for Gaussian distributions. However, *the mean is neither continuous nor qualitatively robust at any distribution!* This follows from the fact that in every Prohorov neighborhood of every $\mathrm{P} \in \mathcal{M}_1(\mathbb{R})$ there are probability distributions that have an infinite expectation or for which the expectation does not exist. Consider for example the mixture distribution $\mathrm{Q} = (1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}}$, where $\tilde{\mathrm{P}}$ is a Cauchy distribution and $\varepsilon > 0$ is positive but sufficiently small. Hence the mean is extremely sensitive with respect to such violations of the distributional assumption. This non-robustness property of the expectation operator has two consequences.

i) SVMs are in general not qualitatively robust if $X \times Y$ is unbounded.
ii) If $X \times Y$ is unbounded, certain P-integrability conditions for the loss function will be necessary for our results on robustness properties of SVMs given in Sections 10.3 and 10.4. ◁

Qualitative robustness has the disadvantage that it does not offer arguments on how to choose among different qualitative robust procedures. However, this can be done by the following approaches.

Influence Function and Related Measures

Let (X_i, Y_i), $i = 1, \ldots, n$, be independent and identically distributed random variables on some measurable space $(Z, \mathcal{B}(Z))$, for example $Z = X \times Y \subset \mathbb{R}^d \times \mathbb{R}$. We will consider a function S that assigns to every probability distribution P an element $S(\mathrm{P})$ of a given Banach space E. In the case of SVMs, we have $E = H$ and $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$ or $E = \mathbb{R}$ and $S(\mathrm{P}) = \inf_{f \in H} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)$.

Recall that qualitative robustness is related to *equicontinuity* of the sequence of estimators $(S_n)_{n \in \mathbb{N}}$ with respect to the Prohorov distance of the corresponding probability distributions. In contrast, the influence function proposed by Hampel (1968, 1974) is related to *differentiation* of S. Denote the Dirac distribution at z by δ_z.

Definition 10.4. *The* ***influence function*** $\mathrm{IF} : Z \to E$ *of* $S : \mathcal{M}_1(Z) \to E$ *at a point* z *for a distribution* $\mathrm{P} \in \mathcal{M}_1(Z)$ *is given by*

$$\mathrm{IF}(z; S, \mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{S\big((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z\big) - S(\mathrm{P})}{\varepsilon} \tag{10.16}$$

in those $z \in Z$ *where the limit exists.*

If the influence function $\mathrm{IF}(z; S, \mathrm{P})$ exists and is continuous and linear, then the function $S : \mathrm{P} \mapsto S(\mathrm{P})$ is Gâteaux differentiable in the direction of the mixture distribution $\mathrm{Q} := (1-\varepsilon)\mathrm{P} + \varepsilon\delta_z$; see also Figure 10.2. Note that Gâteaux differentiation is weaker than Hadamard differentiation (or compact differentiation) and Fréchet differentiation; see Averbukh and Smolyanov (1967, 1968), Fernholz (1983), and Rieder (1994) for details.

The influence function has the interpretation that it measures the impact of an (infinitesimal) small amount of contamination of the original distribution P in the direction of a Dirac distribution located in the point z on the theoretical quantity of interest $S(\mathrm{P})$. Therefore, it is desirable that a statistical method has a *bounded* influence function. If different methods have a bounded influence function, the one with a lower bound is considered to be more robust.

If S fulfills some regularity conditions such as Fréchet differentiability (see Clarke, 1983, 1986; Bednarski *et al.*, 1991), it can be linearized near P in terms of the influence function via

$$S(\mathrm{P}^*) = S(\mathrm{P}) + \int \mathrm{IF}(z; S, \mathrm{P})\, d(\mathrm{P}^* - \mathrm{P})(z) + \ \ldots \ ,$$

where P^* is a probability measure in a neighborhood of P. According to Huber (1981, p. 72), *"it is not enough to look at the influence function at the model*

distribution only; we must also take into account its behavior in a neighborhood of the model." Therefore, we will investigate the influence function of $f_{\mathrm{P},\lambda}$ of SVMs not only at a single fixed distribution P. We will show that the influence function of SVMs can be bounded for large sets of distributions (see Sections 10.3 and 10.4).

Definition 10.5. *The* ***sensitivity curve*** $SC_n : Z \to E$ *of an estimator* $S_n : Z^n \to E$ *at a point* $z \in Z$ *given a data set* $z_1, \ldots, z_{n-1}$ *is defined by*

$$\mathrm{SC}_n(z; S_n) = n\big(S_n(z_1, \ldots, z_{n-1}, z) - S_{n-1}(z_1, \ldots, z_{n-1})\big)\,. \tag{10.17}$$

The sensitivity curve was proposed by J. W. Tukey and its properties are discussed by Hampel *et al.* (1986, p. 93). The sensitivity curve measures the impact of just one additional data point z on the empirical quantity of interest (i.e., on the estimate S_n).

Consider an estimator S_n defined via $S(\mathrm{D}_n)$, where $\mathrm{D}_n \in \mathcal{M}_1(Z)$ denotes the empirical distribution of the data points $z_1, \ldots, z_n$. Denote the empirical distribution of $z_1, \ldots, z_{n-1}$ by D_{n-1}. Then we have for $\varepsilon_n = \frac{1}{n}$ that

$$\mathrm{SC}_n(z; S_n) = \frac{S\big((1-\varepsilon_n)\mathrm{D}_{n-1} + \varepsilon_n \delta_z\big) - S(\mathrm{D}_{n-1})}{\varepsilon_n}\,. \tag{10.18}$$

Therefore, the sensitivity curve can be interpreted as a finite-sample version of the influence function. Let us consider for illustration purposes a univariate parametric location problem and a data set $z_i = (x_i, y_i)$, where $x_i = 1$ for $1 \le i \le n$. Figure 10.3 shows the sensitivity curve of the mean and that of the median for a univariate parametric location problem where P is set to the standard Gaussian distribution with Lebesgue density $f(y) = (2\pi)^{-1/2} e^{-y^2/2}$, $y \in \mathbb{R}$. It is obvious that the impact of a single extreme value y_i on the estimated location parameter increases linearly with $|y_i| \to \infty$ if the mean is used but that the median has a *bounded* sensitivity curve. Hence the median is more robust than the mean with respect to the sensitivity curve.

The following notion of (unstandardized) gross error sensitivity allows to compare the robustness of different statistical methods.

Definition 10.6. *Let* E *be a Banach space with norm* $\|\cdot\|_E$*. The* ***gross error sensitivity*** *of a function* $S : \mathcal{M}_1(Z) \to E$ *at a probability distribution* P *is defined by*

$$\gamma_u^*(S, \mathrm{P}) := \sup_{z \in Z} \|\mathrm{IF}\big(z; S, \mathrm{P}\big)\|_E \tag{10.19}$$

if the influence function exists.

Maxbias

Of theoretical as well as practical importance is the notion of maxbias, which measures the maximum bias $S(\mathrm{Q}) - S(\mathrm{P})$ within a neighborhood of probability

distributions Q near P. There are several ways to define a neighborhood of P by using appropriate metrics on $\mathcal{M}_1(Z)$. Besides the Prohorov metric, the so-called contamination neighborhood, defined below, is quite common in robust statistics, although such neighborhoods do not metricize the weak* topology on $\mathcal{M}_1(Z)$.

Definition 10.7. *Let* $\mathrm{P} \in \mathcal{M}_1(Z)$ *and* $\varepsilon \in [0, 1/2)$. *A* ***contamination neighborhood*** *or* ***gross error neighborhood*** *of* P *is given by*

$$N_\varepsilon(\mathrm{P}) = \left\{(1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}} : \tilde{\mathrm{P}} \in \mathcal{M}_1(Z)\right\}.$$

Let S *be a function mapping from* $\mathcal{M}_1(Z)$ *into a Banach space* E *with norm* $||\cdot||_E$. *The* ***maxbias*** *(or supremum bias) of* S *at the distribution* P *with respect to the contamination neighborhood* $N_\varepsilon(\mathrm{P})$ *is defined by*

$$\mathrm{maxbias}(\varepsilon; S, \mathrm{P}) = \sup_{\mathrm{Q} \in N_\varepsilon(\mathrm{P})} \|S(\mathrm{Q}) - S(\mathrm{P})\|_E .$$

A contamination neighborhood has several nice properties. *(i)* It allows a good interpretation because it contains mixture distributions Q with respect to P and some other distribution $\tilde{\mathrm{P}}$ specifying the type of contamination. This can be seen as follows. Fix $\varepsilon \in [0, 1]$. Define n i.i.d. Bernoulli distributed random variables $\xi_1, \ldots, \xi_n$ such that $\varepsilon = \mathrm{P}(\xi_i = 1) = 1 - \mathrm{P}(\xi_i = 0)$. Then define n i.i.d. random functions $\zeta_1, \ldots, \zeta_n$, each with probability distribution P. Define also n i.i.d. random functions $\zeta_1^*, \ldots, \zeta_n^*$ each with probability distribution $\tilde{\mathrm{P}}$. It follows that the random functions

$$Z_i := \begin{cases} \zeta_i, & \text{if } \xi_i = 0, \\ \zeta_i^*, & \text{if } \xi_i = 1, \end{cases} \qquad 1 \le i \le n,$$

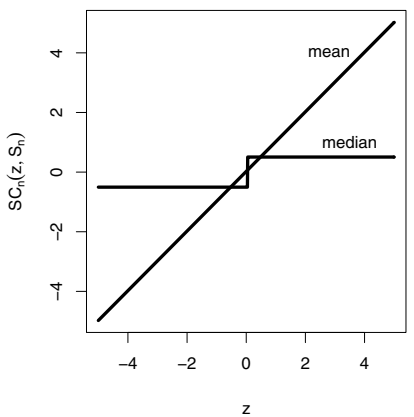


Fig. 10.3. Sensitivity curve of mean and median for a simulated data set with $n = 100$ data points generated from the standard Gaussian distribution $N(0, 1)$.

are i.i.d. each with distribution $\mathrm{Q} = (1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}}$. If $\varepsilon \in [0, 0.5)$, we therefore expect that the pattern described by the majority of n data points generated by Q will follow P and the expected percentage of outliers with respect to P is at most ε. *(ii)* The contamination neighborhood is related to the influence function; see Definition 10.4. *(iii)* It is often easier to deal with this set of distributions than with other neighborhoods.

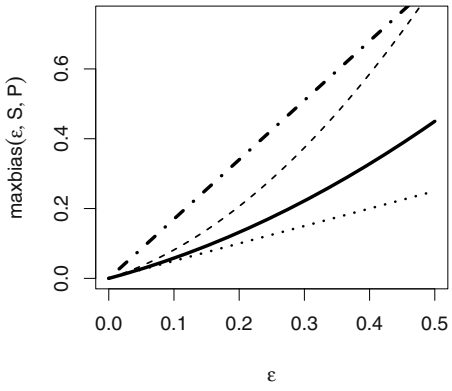


Fig. 10.4. Sketch of relationships between influence function, bias, and maxbias. Define $\mathrm{Q}_\varepsilon = (1-\varepsilon)\mathrm{P} + \varepsilon\delta_z$, $\varepsilon \in [0, 0.5]$. Dotted line: $\varepsilon \|\mathrm{IF}(z; S, \mathrm{P})\|_E$; solid line: $\|S(\mathrm{Q}_\varepsilon) - S(\mathrm{P})\|_E$; dashed line: maxbias$(\varepsilon; S, \mathrm{P})$; dotdashed line: linear upper bound for maxbias$(\varepsilon; S, \mathrm{P})$.

Figure 10.4 illustrates the relationships between influence function, bias, and maxbias. Consider a mixture distribution $\mathrm{Q}_\varepsilon = (1-\varepsilon)\mathrm{P} + \varepsilon\delta_z$, $\varepsilon \in [0, 0.5]$. The dotted line has the slope $\|\mathrm{IF}(z; S, \mathrm{P})\|_E$ and offers an approximation of the bias $\|S(\mathrm{Q}_\varepsilon) - S(\mathrm{P})\|_E$ for small values of ε. The bias is of course below the maxbias$(\varepsilon; S, \mathrm{P})$. The gross error sensitivity $\gamma_u^*(S, \mathrm{P})$ offers—in contrast to $\|\mathrm{IF}(z; S, \mathrm{P})\|_E$—a *uniform* slope (i.e., the slope is valid for all points $z \in X \times Y$). The dotdashed line gives an upper bound for the maxbias where the bound is linear in ε. In the following two sections such quantities will be derived for SVMs for classification and regression.

Breakdown Points

Breakdown points measure the worst-case behavior of a statistical method. The following definition was proposed by Donoho and Huber (1983).

Definition 10.8. *Let E be a Banach space and $D_n = \{z_1, \ldots, z_n\}$ be a data set with values in $Z \subset \mathbb{R}^d$. The* ***finite-sample breakdown point*** *of an E-valued statistic $S(D_n)$ is defined by*

$$\varepsilon_n^*(S, D_n) = \max\left\{ \frac{m}{n} \,:\, \text{Bias}(m; S, D_n) \text{ is finite} \right\},$$

where

$$\text{Bias}(m; S, D_n) = \sup_{D_n'} \|S(D_n') - S(D_n)\|_E$$

and the supremum is over all possible samples D_n' that can be obtained by replacing any *m of the original data points by* arbitrary *values in Z.*

Remark 10.9. It is possible to define a variant of the breakdown point via the maxbias by

$$\varepsilon^*(S, \mathrm{P}) := \sup\{\varepsilon > 0 :\ \text{maxbias}(\varepsilon; S, \mathrm{P}) < \infty\}\,. \tag{10.20}$$

There are variants of the asymptotic breakdown point where the Prohorov metric is replaced by other metrics (e.g., Lévy, bounded Lipschitz, Kolmogorov, total variation). One can also define an asymptotic breakdown point based on gross error neighborhoods. ◁

10.3 Robustness of SVMs for Classification

In this section, we consider robustness properties of classifiers

$$S(\mathrm{P}) := f_{\mathrm{P},\lambda} = \arg\inf_{f \in H} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f),$$

where L is a margin-based loss function; see Definition 2.24. Throughout this section, we define $Y := \{-1, +1\}$. First, we give sufficient conditions for the existence of the influence function. Then we show that the influence function is bounded under weak conditions. Most of our results in this section are valid for *any* distribution $\mathrm{P} \in \mathcal{M}_1(X \times Y)$, where $X = \mathbb{R}^d$, $d \in \mathbb{N}$. Therefore, they are also valid for the special case of empirical distributions $\mathrm{D} = \mathrm{D}_n = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$; i.e.; for any given data set consisting of n data points and for the empirical regularized risks defined by

$$S(\mathrm{D}) := f_{\mathrm{D},\lambda} = \arg\inf_{f \in H} \mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f)\,.$$

The proofs given in this and the following section make use of some general results from functional analysis and probability theory. Before we formulate the robustness results, let us therefore recall some basic notions from calculus in (infinite-dimensional) Banach spaces (see Lemma A.5.15). Let $G : E \to F$ be a function between two Banach spaces E and F. Then G is *(Fréchet) differentiable in $x_0 \in E$* if there exists a bounded linear operator $A : E \to F$ and a function $\varphi : E \to F$ with $\varphi(x)/\|x\| \to 0$ for $x \to 0$ such that

$$G(x_0 + x) - G(x_0) \;=\; Ax + \varphi(x) \tag{10.21}$$

for all $x \in E$. It turns out that A is uniquely determined by (10.21). Hence we write $G'(x) := \frac{\partial G}{\partial E}(x) := A$. The function G is called *continuously differentiable* if the function $x \mapsto G'(x)$ exists on E and is continuous. Analogously we define continuous differentiability on open subsets of E.

We also have to recall the notion of *Bochner integrals*; see Section A.5.4. We restrict attention to the reproducing kernel Hilbert space H since this is the only space we need in this section. Let H be a separable RKHS of a bounded, measurable kernel k on X with canonical feature map $\Phi : X \to H$; i.e.; $\Phi(x) = k(x, \cdot)$. Note that Φ is measurable due to Lemma 4.25. Furthermore, let $\mathrm{P} \in \mathcal{M}_1$ and $h : Y \times X \to \mathbb{R}$ be a measurable and P-integrable function. Then the Bochner integral $\mathbb{E}_\mathrm{P} h(Y, X)\Phi(X)$ is an element of H. In our special situation, we can also interpret this integral as an element of the dual space H' by the Fréchet-Riesz Theorem A.5.12; i.e.; $\mathbb{E}_\mathrm{P} h(Y, X)\Phi(X)$ acts as a bounded linear functional on H via $w \mapsto \langle \mathbb{E}_\mathrm{P} h(Y, X)\Phi(X), w\rangle$. Finally, we will consider Bochner integrals of the form $\mathbb{E}_\mathrm{P} h(Y, X)\langle \Phi(X), \cdot\rangle\Phi(X)$ that define bounded linear operators on H by the function $w \mapsto \mathbb{E}_\mathrm{P} h(Y, X)\langle \Phi(X), w\rangle\Phi(X)$. We sometimes write $L \circ f$ instead of $L(Y, f(X))$ and Φ instead of $\Phi(X)$ to shorten the notation if misunderstandings are unlikely. We use this kind of notation also for derivatives of L. The Dirac distribution in z is denoted by δ_z.

Existence of the Influence Function

We can now establish our first two results for smooth margin-based loss functions and a bounded continuous kernel. The first theorem covers, for example, the Gaussian RBF kernel.

Theorem 10.10. *Let $Y = \{-1, +1\}$ and $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex and twice continuously differentiable margin-based loss function with representing function φ. Furthermore, let $X \subset \mathbb{R}^d$ be a closed subset, H be an RKHS of a bounded continuous kernel on X with canonical feature map Φ, and $\mathrm{P} \in \mathcal{M}_1$. Then the influence function of $S(\mathrm{P}) := f_{\mathrm{P},\lambda}$ exists for all $z = (x, y) \in X \times Y$ and is given by*

$$\mathrm{IF}(z; S, \mathrm{P}) = \mathbb{E}_\mathrm{P}[\varphi'\big(Y f_{\mathrm{P},\lambda}(X)\big)K^{-1}\Phi(X)] - \varphi'\big(y f_{\mathrm{P},\lambda}(x)\big)K^{-1}\Phi(x)\,, \quad (10.22)$$

where $K : H \to H$ defined by $K = 2\lambda\,\mathrm{id}_H + \mathbb{E}_\mathrm{P}\varphi''\big(Y f_{\mathrm{P},\lambda}(X)\big)\langle \Phi(X), \cdot\rangle\Phi(X)$ denotes the Hessian of the regularized risk.

Proof. Our analysis relies heavily on the function $G : \mathbb{R} \times H \to H$ defined by

$$G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z}\varphi'\big(Y f(X)\big)\Phi(X) \quad (10.23)$$

and $K = \frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$. Note that for $\varepsilon \notin [0, 1]$ the H-valued expectation is with respect to a signed measure; see Definition A.3.2.

Let us first recall that the solution $f_{\mathrm{P},\lambda}$ exists by Theorem 5.2 and Corollary 5.3. Additionally, we have $\|f_{\mathrm{P},\lambda}\|_H \leq \bigl(\varphi(0)/\lambda\bigr)^{1/2}$, where $\varphi : \mathbb{R} \to [0,\infty)$ is the function representing L by $L(y,t) = \varphi(yt)$, $y \in Y$, $t \in \mathbb{R}$. By using Lemma 2.21 and (5.7), we obtain

$$G(\varepsilon, f) = \frac{\partial \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}(f)}{\partial H}\,, \qquad \varepsilon \in [0,1]. \tag{10.24}$$

Furthermore, the function $f \mapsto \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}(f)$ is convex for all $\varepsilon \in [0,1]$, and (10.24) shows that we have $G(\varepsilon, f) = 0$ if and only if $f = f_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}$. Our aim is to show the existence of a differentiable function $\varepsilon \mapsto f_\varepsilon$ defined on a small interval $[-\delta,\delta]$ for some $\delta > 0$ that satisfies $G(\varepsilon, f_\varepsilon) = 0$ for all $\varepsilon \in [-\delta,\delta]$. Once we have shown the existence of this function, we immediately obtain

$$\mathrm{IF}(z; S, \mathrm{P}) = \frac{\partial f_\varepsilon}{\partial \varepsilon}(0)\,.$$

For the existence of $\varepsilon \mapsto f_\varepsilon$, we have to check by the implicit function theorem in Banach spaces (see Theorem A.5.17) that G is continuously differentiable and that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is invertible. Let us start with the first: an easy computation shows

$$\frac{\partial G}{\partial \varepsilon}(\varepsilon, f) = -\mathbb{E}_{\mathrm{P}}\varphi'\bigl(Yf(X)\bigr)\Phi(X) + \mathbb{E}_{\delta_z}\varphi'\bigl(Yf(X)\bigr)\Phi(X)\,. \tag{10.25}$$

Note that $\varphi'' \circ f$ is bounded because φ'' is continuous and $f \in H$ is bounded. Similar to (10.24), we thus find

$$\frac{\partial G}{\partial H}(\varepsilon, f) = 2\lambda\,\mathrm{id}_H + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z}\varphi''\bigl(Yf(X)\bigr)\langle\Phi(X), \cdot\rangle\Phi(X)\,. \tag{10.26}$$

Since H has a bounded kernel, it is a simple routine to check that both partial derivatives are continuous. This together with the continuity of G ensures that G is continuously differentiable; see Theorem A.5.16.

In order to show that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is invertible, it suffices to show by the Fredholm Alternative (see Theorem A.5.5) that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is *injective* and that $A : H \mapsto H$ defined by

$$Ag := \mathbb{E}_{\mathrm{P}}\varphi''\bigl(Yf_{\mathrm{P},\lambda}(X)\bigr)g(X)\Phi(X)\,, \qquad g \in H,$$

is a *compact operator*. To show the compactness, we need some facts from measure theory. Since $X \subset \mathbb{R}^d$ is assumed to be closed, it is a Polish space; see the examples listed after Definition A.2.11. Furthermore, Borel probability measures on Polish spaces are regular by Ulam's theorem; see Theorem A.3.15. Therefore, such probability measures can be approximated from inside by compact sets; see Definition A.3.14. In our situation, this means that for all $n \geq 1$ there exists a compact measurable subset $X_n \subset X$ with $\mathrm{P}_X(X_n) \geq 1 - \frac{1}{n}$, where P_X denotes the marginal distribution of P with respect to X. Now define a sequence of operators $A_n : H \to H$ by

$$A_n g := \int_{X_n \times Y} \varphi''\big(y f_{\mathrm{P},\lambda}(x)\big) g(x) \Phi(x)\, d\mathrm{P}(x,y), \qquad g \in H.$$

Note that, if X is compact, we can of course choose $X_n = X$, which implies $A_n = A$. Let us now show that all operators A_n are compact.

By the definition of A_n and Theorem A.5.22, there exists a constant $c > 0$ depending on λ, φ'' and k such that for all $g \in B_H$ we have

$$A_n g \in c \cdot \overline{\operatorname{aco} \Phi(X_n)}\,, \tag{10.27}$$

where $\operatorname{aco} \Phi(X_n)$ denotes the absolute convex hull of $\Phi(X_n)$ and the closure is with respect to $\|\cdot\|_H$. This shows that A_n is compact, $n \in \mathbb{N}$. In order to see that the operator A is compact, it therefore suffices to show $\|A_n - A\|_H \to 0$ w.r.t. the operator norm for $n \to \infty$. However, the latter convergence can be easily checked using $\mathrm{P}_X(X_n) \geq 1 - \frac{1}{n}$.

It remains to prove that A is injective. For $g \neq 0$, we find

$$\begin{aligned}
\big\langle (2\lambda\, \mathrm{id}_H + A)g, (2\lambda\, \mathrm{id}_H + A)g \big\rangle &= 4\lambda^2 \langle g, g\rangle + 4\lambda \langle g, Ag\rangle + \langle Ag, Ag\rangle \\
&> 4\lambda \langle g, Ag\rangle \\
&= 4\lambda \big\langle g, \mathbb{E}_{\mathrm{P}} \varphi''\big(Y f_{\mathrm{P},\lambda}(X)\big) g(X) \Phi(X) \big\rangle \\
&= 4\lambda \mathbb{E}_{\mathrm{P}} \varphi''\big(Y f_{\mathrm{P},\lambda}(X)\big) g^2(X) \\
&\geq 0\,.
\end{aligned}$$

Here the last equality is due to the fact that $B\mathbb{E}_{\mathrm{P}} h = \mathbb{E}_{\mathrm{P}} Bh$ for all E-valued functions h and bounded linear operators B due to (A.32). The last inequality is true since the second derivative of a convex and twice-differentiable function is nonnegative. Obviously, the estimate above shows that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda}) = 2\lambda\, \mathrm{id}_H + A$ is injective.

Finally, we use the equality $\mathrm{IF}(z; S, \mathrm{P}) = \frac{\partial f_\varepsilon}{\partial \varepsilon}(0)$ to derive a formula for the influence function, where $\varepsilon \mapsto f_\varepsilon$ is the function implicitly defined by $G(\varepsilon, f) = 0$ such that the implicit function theorem A.5.17 gives

$$\mathrm{IF}(z; S, \mathrm{P}) = -K^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{\mathrm{P},\lambda})\,, \tag{10.28}$$

where $K := \frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$. Now combine (10.28) with (10.25) and (10.26). □

The next remark follows immediately from Theorem 10.10.

Remark 10.11. If the assumptions of Theorem 10.10 are fulfilled, then a convex margin-based loss function with φ' and φ'' bounded in combination with a bounded continuous kernel yields a bounded influence function of $f_{\mathrm{P},\lambda}$. Hence, the class of Lipschitz-continuous loss functions is of primary interest from the viewpoint of robust statistics. SVMs based on the logistic loss yield a bounded influence function, if k is bounded and continuous. SVMs based on the least squares loss or the AdaBoost loss yield unbounded influence functions.

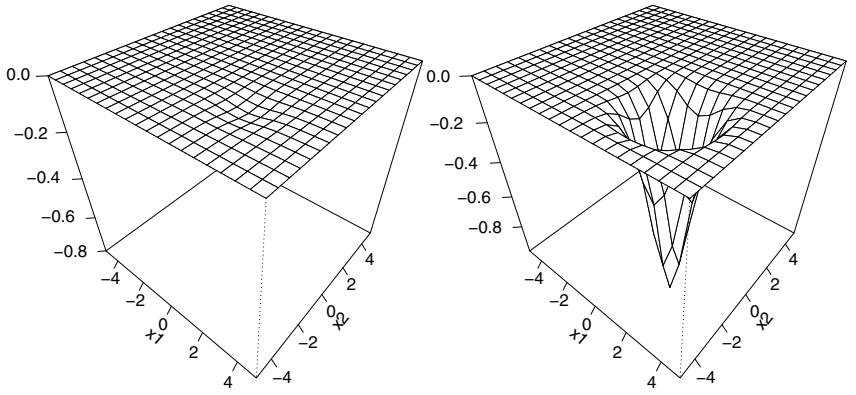


Fig. 10.5. Plot of the function $\varphi'\big(yf_{\mathrm{P},\lambda}(x)\big)\Phi(x)$ for $L = L_{\text{c-logist}}$, where the point mass contamination is δ_z, $z := (x, y) = (2, -2, y)$ and $\mathrm{P}(Y = +1 \mid X = x) = 0.982$. Left subplot: $y = +1$. Right subplot: $y = -1$.

Remark 10.12. The influence function derived in Theorem 10.10 depends on the point $z = (x, y)$, where the point mass contamination takes place only via the term

$$\varphi'\big(yf_{\mathrm{P},\lambda}(x)\big)\Phi(x)\,. \tag{10.29}$$

This function is illustrated in Figure 10.5 for the special case of kernel logistic regression (i.e.; $L = L_{\text{c-logist}}$) in combination with a Gaussian RBF kernel and $\mathrm{P}(Y = 1|X = x) = 1/\big(1 + e^{-f(x)}\big)$, $f(x) := -x_1 + x_2$, $(x_1, x_2) \in \mathbb{R}^2$. The left subplot clearly shows that the quantity $\varphi'\big(yf_{\mathrm{P},\lambda}(x)\big)\Phi(x)$ is approximately zero for this combination of L and k if the highly probable value $z = (x, y) = (2, -2, 1)$ is considered. Of special interest is the right subplot, which shows that the improbable value $z = (x, y) = (2, -2, -1)$ affects the influence function via the quantity $\varphi'\big(yf_{\mathrm{P},\lambda}(x)\big)\Phi(x)$ only in a *smooth, bounded, and local* manner. Smoothness is achieved by choosing L and k continuous and differentiable. Boundedness is achieved by using a bounded kernel in combination with a loss function having a bounded first derivative φ'. A local impact instead of a more global impact due to a Dirac distribution is achieved by an appropriate bounded and non-linear kernel such as the Gaussian RBF kernel. Note that polynomial kernels with $m \geq 1$ are unbounded if $X = \mathbb{R}^d$. ◁

In practice, the set X is often a bounded and closed subset of $\mathbb{R}^d$ and hence compact. In this case, the existence of the influence function can be shown without the assumption that the kernel is bounded, and hence the following result also covers polynomial kernels.

Corollary 10.13. *Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex and twice continuously differentiable margin-based loss function. Furthermore, let $X \subset \mathbb{R}^d$ be*

compact, H be an RKHS of a continuous kernel on X, and $\mathrm{P} \in \mathcal{M}_1$. Then the influence function of $f_{\mathrm{P},\lambda}$ exists for all $z \in X \times Y$.

Proof. Every compact subset of $\mathbb{R}^d$ is closed, and continuous kernels on compact subsets are bounded. Hence the assertion follows directly from Theorem 10.10. □

Remark 10.14. By a straightforward modification of the proof of Corollary 10.13, we actually find that the special Gâteaux derivative of $S : \mathrm{P} \mapsto f_{\mathrm{P},\lambda}$ exists for *every* direction; i.e.,

$$\lim_{\varepsilon \downarrow 0} \frac{f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}}{\varepsilon}$$

exists for all $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1$ provided that the assumptions of Theorem 10.10 hold. This is interesting from the viewpoint of applied statistics because a point mass contamination is just one kind of contamination that can occur in practice. ◁

Bounds for the Bias and the Influence Function

As explained in Section 10.2, a desirable property of a robust statistical method $S(\mathrm{P})$ is that S has a bounded influence function. In this section, we investigate whether it is possible to obtain an SVM $S : \mathrm{P} \mapsto f_{\mathrm{P},\lambda}$ with a bounded influence function where the bound is *independent* of $z \in X \times Y$ and $\mathrm{P} \in \mathcal{M}_1$. We will also consider the question of whether the sensitivity curve or the maxbias can be bounded in a similar way.

For the formulation of our results, we need to recall that the norm of total variation of a signed measure μ on a Banach space E is defined by

$$\|\mu\|_{\mathcal{M}} := |\mu|(E) := \sup \left\{ \sum_{i=1}^{n} |\mu(E_i)| : E_1, \ldots, E_n \text{ is a partition of } E \right\} .$$

The following theorem bounds the difference quotient in the definition of the influence function for classifiers based on $f_{\mathrm{P},\lambda}$. For practical applications, this result is especially important because it also gives an upper bound for the bias in gross error neighborhoods. In particular, it states that the influence function of such classifiers is uniformly bounded whenever it exists and that the sensitivity curve is uniformly bounded, too.

Theorem 10.15. *Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex margin-based loss function. Furthermore, let $X \subset \mathbb{R}^d$ be closed and H be an RKHS of a bounded, continuous kernel k with canonical feature map $\Phi : X \to H$.*

i) For all $\lambda > 0$, there exists a constant $c(L, k, \lambda) > 0$ explicitly given by (10.33) such that for all distributions $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1$ we have

$$\left\| \frac{f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}}{\varepsilon} \right\|_H \leq c(L, k, \lambda) \, \|\mathrm{Q} - \mathrm{P}\|_{\mathcal{M}} , \quad \varepsilon > 0 . \tag{10.30}$$

ii) An upper bound for the maxbias of $f_{\mathrm{P},\lambda}$ is given by

$$\operatorname{maxbias}(\varepsilon; f, \mathrm{P}) \;\le\; \varepsilon\, c(L,k,\lambda) \sup_{\mathrm{Q}\in\mathcal{M}_1} \|\mathrm{Q}-\mathrm{P}\|_{\mathcal{M}} \;\le\; 2c(L,k,\lambda)\cdot\varepsilon\,,$$

where $\varepsilon \in (0, 1/2)$.

iii) If the influence function of $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$ exists, then it is bounded by

$$\|\mathrm{IF}(z; S, \mathrm{P})\|_H \le 2c(L,k,\lambda)\,, \qquad z \in X \times Y\,, \tag{10.31}$$

and the gross error sensitivity is bounded by

$$\gamma_u^*(S,\mathrm{P}) \le 2c(L,k,\lambda)\,. \tag{10.32}$$

Proof. i). Recall that every convex function on $\mathbb{R}$ is locally Lipschitz-continuous (see Lemma A.6.5). Combining Lemma 2.25 with $\mathcal{R}_{L,\mathrm{P}}(0) = \varphi(0) < \infty$, where φ satisfies $L(y,t) = \varphi(yt)$ for all $y \in Y$ and $t \in \mathbb{R}$, shows that the assumptions of Corollary 5.10 are fulfilled. Let us define $B_\lambda := \|k\|_\infty \sqrt{\mathcal{R}_{L,\mathrm{P}}(0)/\lambda}$, where $\|k\|_\infty := \sup_{x\in X}\sqrt{k(x,x)} < \infty$. Let us fix $\mathrm{P} \in \mathcal{M}_1$. Then Corollary 5.10 guarantees the existence of a bounded measurable function $h : X \times Y \to \mathbb{R}$ such that for all $\tilde{\mathrm{P}} \in \mathcal{M}_1$ we have

$$\|f_{\mathrm{P},\lambda} - f_{\tilde{\mathrm{P}},\lambda}\|_H \le \lambda^{-1}\|\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\tilde{\mathrm{P}}} h\Phi\|_H\,.$$

Now define $\tilde{\mathrm{P}} = (1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}$, where $\mathrm{Q} \in \mathcal{M}_1$. Let $|L_{|Y\times[-c,c]}|_1$ denote the Lipschitz constant of L restricted to $Y \times [-c,c]$, $c > 0$. Hence, we have

$$\begin{aligned}
\varepsilon^{-1}\|f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} - f_{\mathrm{P},\lambda}\|_H &\le (\varepsilon\lambda)^{-1}\|\mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} h\Phi - \mathbb{E}_{\mathrm{P}} h\Phi\|_H \\
&= \lambda^{-1}\|\mathbb{E}_{\mathrm{Q}} h\Phi - \mathbb{E}_{\mathrm{P}} h\Phi\|_H \\
&\le c(L,k,\lambda)\|\mathrm{Q}-\mathrm{P}\|_{\mathcal{M}}\,,
\end{aligned}$$

where

$$c(L,k,\lambda) = \lambda^{-1}\,\|k\|_\infty\,|L_{|Y\times[-B_\lambda,B_\lambda]}|_1\,. \tag{10.33}$$

We obtain (10.30). The parts *ii)* and *iii)* follow from $\|\mathrm{Q}-\mathrm{P}\|_{\mathcal{M}} \le 2$. □

Note that Theorem 10.15 applies to almost all margin-based loss functions of practical interest because differentiability of φ is not assumed. Special cases are the loss functions hinge, kernel logistic, modified Huber, least squares, truncated least squares, and AdaBoost.

Remark 10.16. The preceding theorem also gives uniform bounds for Tukey's sensitivity curve. Consider the special case where P is equal to the empirical distribution of $(n-1)$ data points (i.e.; $\mathrm{D}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1}\delta_{(x_i,y_i)}$) and that Q is equal to the Dirac measure $\delta_{(x,y)}$ in some point $(x,y) \in X \times Y$. Let $\varepsilon = \frac{1}{n}$. Combining (10.17) and (10.18), we obtain for $\varepsilon > 0$ under the assumptions of Theorem 10.15 the inequality

$$n\|f_{(1-\varepsilon)\mathrm{D}_{n-1}+\varepsilon\delta_{(x,y)},\lambda} - f_{\mathrm{D}_{n-1},\lambda}\|_H \le c(L,k,\lambda)\ \|\delta_{(x,y)} - \mathrm{D}_{n-1}\|_{\mathcal{M}}\,. \quad (10.34)$$

Note that the bound of the bias presented in Corollary 5.10 relies on the Hilbert norm $\left\| \mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\bar{\mathrm{P}}} h\Phi \right\|_H$, whereas the bounds of the bias given by Theorem 10.15 and by (10.34) use the norm of total variation $\|\mathrm{Q} - \mathrm{P}\|_{\mathcal{M}}$.

Furthermore, Theorem 10.15 shows for empirical distributions D_n that the maxbias of $f_{\mathrm{D}_n,\lambda}$ in an ε-contamination neighborhood $N_\varepsilon(\mathrm{D}_n)$ is at most $2c(L,k,\lambda)\varepsilon$, where $\varepsilon \in (0, 1/2)$. In other words, the upper bound for the maxbias increases *at most linearly* with slope $2c(L,k,\lambda)$ with respect to the mixing proportion ε; see also Figure 10.4. ◁

Remark 10.17. Consider $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1$ having densities p, q with respect to some σ-finite measure ν. Then, Theorem 10.15 also gives bounds of the influence function and the sensitivity curve in terms of the Hellinger metric $H(\mathrm{P},\mathrm{Q}) = (\int (p^{1/2} - q^{1/2})^2\, d\nu)^{1/2}$ because we have $\|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}} \le 2\, H(\mathrm{P},\mathrm{Q}) \le 2\, \|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}}^{1/2}$; see Witting (1985). ◁

Note that the *bounds* for the difference quotient and the influence function in Theorem 10.15 converge to infinity if λ converges to 0 and $\|\mathrm{Q} - \mathrm{P}\|_{\mathcal{M}} > 0$. However, λ converging to 0 has the interpretation that misclassifications are penalized by constants proportional to $\frac{1}{\lambda}$ tending to ∞. Decreasing values of λ therefore correspond to a decreasing amount of robustness, which is to be expected. The regularizing quantity λ hence has two roles.

i) It controls the *penalization of misclassification errors.*
ii) It controls the *robustness properties* of $f_{\mathrm{P},\lambda}$.

We would like to mention that for the robustness properties of $f_{\mathrm{P},\lambda}$, the ratio $\frac{1}{\lambda}$ plays a role similar to the tuning constant for Huber-type M-estimators. Let us consider the Huber-type M-estimator (Huber, 1964) in a univariate location model where all data points are realizations from n independent and identically distributed random variables (X_i, Y_i) with some cumulative distribution function $\mathrm{P}\big(Y_i \le y \,|\, X_i = x\big) = F(y - \theta)$, $y \in \mathbb{R}$, where $\theta \in \mathbb{R}$ is unknown. Huber's robust M-estimator with tuning constant $M \in (0, \infty)$ has an influence function proportional to $\varphi_M(z) = \max\big\{-M, \min\{M, z\}\big\}$; see Hampel *et al.* (1986, pp. 104ff.). For all $M \in (0, \infty)$, the influence function is bounded by $\pm M$. However, the bound tends to $\pm\infty$ if $M \to \infty$, and Huber's M-estimator with $M = \infty$ is equal to the non-robust mean, which has an unbounded influence function for Gaussian distributions. Therefore, the quantity $\frac{1}{\lambda}$ for SVMs plays a role similar to the tuning constant M for Huber-type M-estimators. The same argumentation is true for Huber-type M-estimators in classification and in regression.

Christmann and Steinwart (2004) derived some results for the influence function for the more general case where not only $f_{\mathrm{P},\lambda}$ but also a real-valued offset term $b_{\mathrm{P},\lambda}$ must be estimated.

Empirical Results for SVMs

The question arises as to how the theoretical results for $f_{\mathrm{P},\lambda}$ given above are related to empirical results for finite sample sizes.

Let us therefore complement the previous theoretical results by a few numerical results for a training data set with a relatively small sample size. We consider the slightly more general case where an offset term also must be estimated (i.e.; $S(\mathrm{P}) = (f_{\mathrm{P},\lambda}, b_{\mathrm{P},\lambda}) \in H \times \mathbb{R}$) because many software tools for SVMs use an offset term and because the theoretical results given above were for the case without an intercept term; see Christmann and Steinwart (2004) for additional theoretical results. Let $\mathrm{D} = \mathrm{D}_n$ be the empirical distribution of a training data set with n data points $(x_i, y_i) \in \mathbb{R} \times \{-1, +1\}$, $i = 1, \ldots, n$. We will investigate the impact that an additional data point can have on the support vector machine with an offset term $b \in \mathbb{R}$ for pattern recognition. The replacement of one of the n data points can be treated in the same manner. An analogous investigation for the case without offset gave results similar to those described in this section. We generated a training data set with $n = 500$ data points x_i from a bivariate normal distribution with expectation $\mu = (0, 0)$ and covariance matrix Σ. The variances were set to 1, whereas the covariances were set to 0.5. The responses y_i were generated from a classical logistic regression model with $\theta = (-1, 1)$ and $b = 0.5$ such that

$$\mathrm{P}(Y = 1|x) = \frac{1}{1 + e^{-(\langle x,\theta\rangle + b)}}, \quad x \in \mathbb{R}^2.$$

We consider two popular kernels: a Gaussian RBF kernel with parameter $\gamma > 0$ and a linear kernel. Appropriate values for γ and for the constant C (or λ) are important for the SVM and are often determined by cross-validation. The computations were done using the software SVM^{light} developed by Joachims (1999). A cross-validation based on the leave-one-out error for the training data set was carried out by a two-dimensional grid search consisting of 13×10 grid points. As a result of the cross-validation, the tuning parameters for the SVM with RBF kernel were set to $\gamma = 2$ and $\lambda = \frac{1}{4n}$. The leave-one-out error for the SVM with a linear kernel turned out to be stable over a broad range of values for C. We used $\lambda = \frac{1}{2n}$ in the computations for the linear kernel. For $n = 500$, this results in $\lambda = 5 \times 10^{-4}$ for the RBF kernel and $\lambda = 0.001$ for the linear kernel. Please note that such small values of λ will result in relatively large bounds.

Figure 10.6 shows the sensitivity curves of $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ if we add a single point $z = (x, y)$ to the original data set, where $x := (x_1, x_2) = (6, 6)$ and $y = +1$. The additional data point has a local and smooth impact on $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ with a peak in a neighborhood of x if one uses the RBF kernel. For a linear kernel, the impact is approximately linear. The reason for this different behavior of the SVM with different kernels becomes clear from Figure 10.7, where plots of $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ are given for the original data set and for the modified data set, which contains the additional data point z. Please note

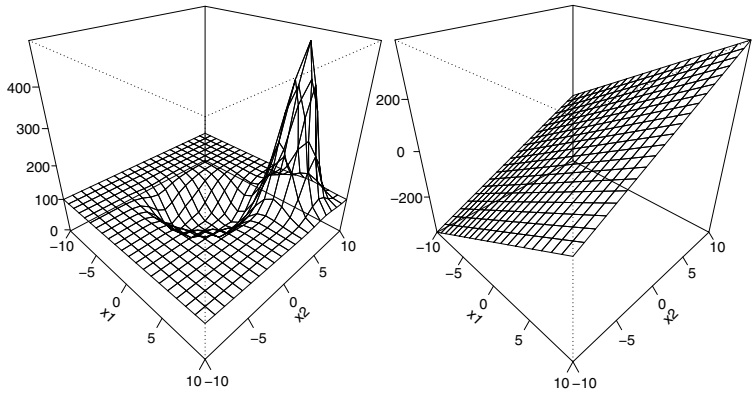


Fig. 10.6. Sensitivity function of $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ if the additional data point z is located at $z = (x, y)$, where $x = (6, 6)$ and $y = +1$. Left: RBF kernel. Right: linear kernel.

that the RBF kernel yields $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ approximately equal to zero outside a central region, as almost all data points are lying inside the central region. Comparing the plots of $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ based on the RBF kernel for the modified data set with the corresponding plot for the original data set, it is obvious that the additional smooth peak is due to the new data point located at $x = (6, 6)$ with $y = +1$. It is interesting to note that although the estimated functions $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ for the original data set and for the modified data set based on the SVM with the linear kernel look quite similar, the sensitivity curve is similar to an affine hyperplane that is affected by the value of z. This allows the interpretation that just a single data point can have an impact on $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$ estimated by an SVM with a linear kernel over a broader region than for an SVM with an RBF kernel.

Now we study the impact of an additional data point $z = (x, y)$, where $y = +1$, on the percentage of classification errors and on the fitted y-value for z. We vary z over a grid in the x-coordinates. Figure 10.8 shows that the percentage of classification errors is approximately constant outside the central region, which contains almost all data points if a Gaussian RBF kernel was used. For the SVM with a linear kernel, the percentage of classification errors tends to be approximately constant in one affine half-space but changes in the other half-space. The response of the additional data point was correctly estimated by $\hat{y} = +1$ outside the central region if a Gaussian RBF kernel is used; see Figure 10.9. In contrast, using a linear kernel results in estimated responses $\hat{y} = +1$ or $\hat{y} = -1$ of the additional data point depending on the affine half-space in which the x-value of z is lying.

Finally, let us study the impact of an additional data point located at $z = (x, y)$, where $y = +1$, on the estimated parameters $\hat{b}$ and $\hat{\theta}$ of the parametric logistic regression model; see Figure 10.10. We vary z over a grid in the

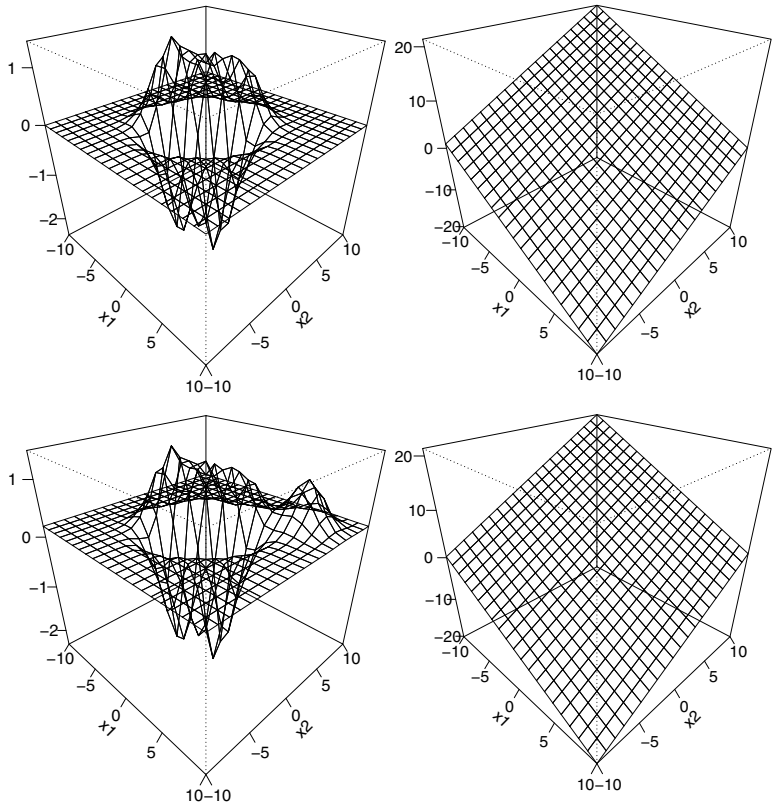


Fig. 10.7. Plot of $f_{\mathrm{D},\lambda} + b_{\mathrm{D},\lambda}$. Upper left: RBF kernel, original data set. Upper right: linear kernel, original data set. Lower left: RBF kernel, modified data set. Lower right: linear kernel, modified data set. The modified data set contains the additional data point $z = (x, y)$, where $x = (6, 6)$ and $y = +1$.

x-coordinates in the same manner as before. As the plots for $\hat{\theta}_1$ and $\hat{\theta}_2$ look very similar, we only show the latter. Note that the axes are not identical in Figure 10.10 due to the kernels. The sensitivity curves for the slopes estimated by the SVM with an RBF kernel are similar to a hyperplane outside the central region, which contains almost all data points. In the central region, there is a smooth transition between regions with higher sensitivity values and regions with lower sensitivity values. The sensitivity curves for the slopes of the SVM with a linear kernel are flat in one affine half-space but change approximately linearly in the other affine half-space. This behavior also occurs for the sensitivity curve of the offset by using a linear kernel. In contrast, the sensitivity curve of the offset based on an SVM with an RBF kernel shows a

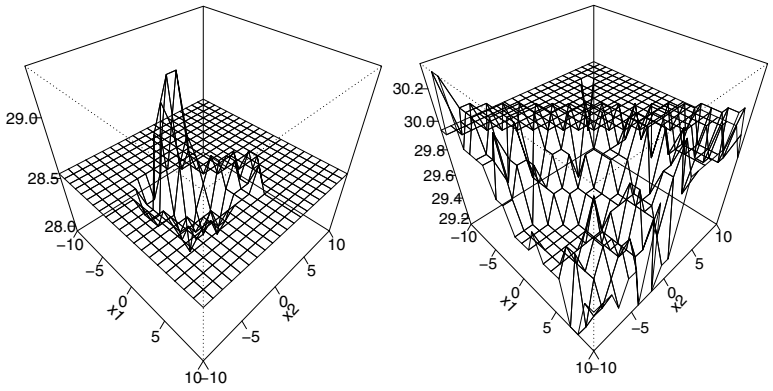


Fig. 10.8. Percentage of classification errors if one data point $z = (x, 1)$ is added to the original data set, where x varies over the grid. Left: RBF kernel. Right: linear kernel.

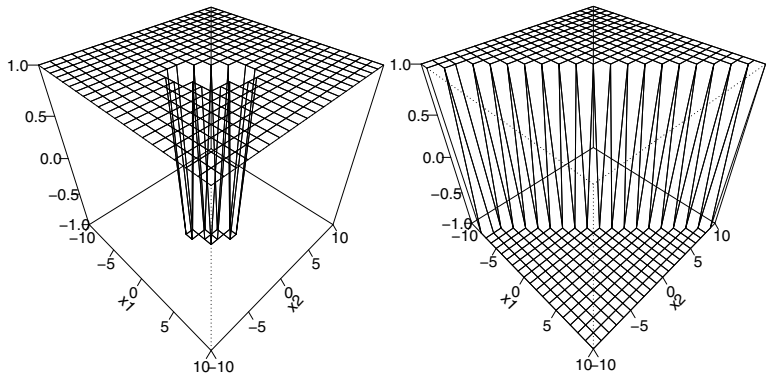


Fig. 10.9. Fitted y-value for a new observation if one data point $z = (x, 1)$ is added to the original data set, where x varies over the grid. Left: RBF kernel. Right: linear kernel.

smooth but curved shape outside the region containing the majority of the data points.

10.4 Robustness of SVMs for Regression (*)

In this section, we will investigate the existence and boundedness of the influence function of the mapping $S : \mathrm{P} \mapsto f_{\mathrm{P},\lambda}$ for the regression case. Some results for the related robustness measures sensitivity curve, gross error sensitivity, and maxbias will also be given. This section has relationships with

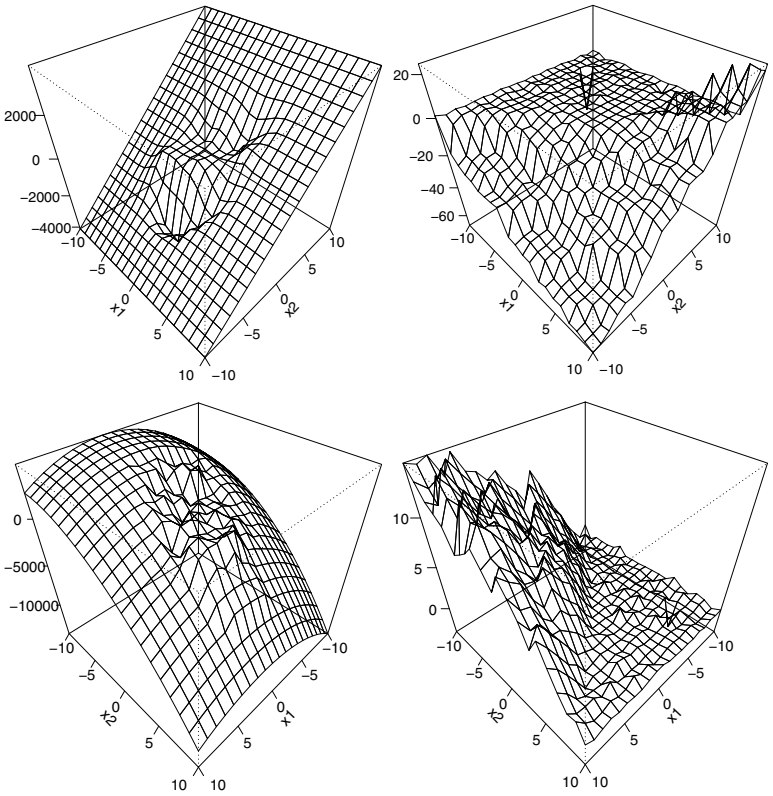


Fig. 10.10. Sensitivity functions for $\hat{\theta}$ and $\hat{\beta}$. Upper left: sensitivity function for $\hat{\theta}_2$, RBF kernel. Upper right: sensitivity function for $\hat{\theta}_2$, linear kernel. Lower left: sensitivity function for $\hat{\beta}$, RBF kernel. Lower right: sensitivity function for $\hat{\beta}$, linear kernel. Note that the axes differ in the four subplots due to improved visibility.

Section 5.3 on the stability of infinite sample versions of SVMs. We assume in this section that $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$.

Existence of the Influence Function

The first result shows that the influence function of $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$ exists if the loss function is convex and twice continuously differentiable and if the kernel is bounded and continuous. The proof of this result is based on the application of the implicit function theorem A.5.17. In contrast to the proof of Theorem 10.10 for the classification case, we are now usually faced with *unbounded* label sets Y such that we need some additional arguments. In particular, the relationship between the growth type of the loss function and the tail behavior

of the distribution P will turn out to be important. This was one reason why we investigated Nemitski losses in Chapter 2.

Theorem 10.18. *Let $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be closed, $\mathrm{P} \in \mathcal{M}_1$, H be an RKHS of a bounded continuous kernel k on X with canonical feature map $\Phi : X \to H$, and $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex P-integrable Nemitski loss function that has partial derivatives $L' = \partial_2 L$ and $L'' = \partial_{22} L$ such that $|L'|$ and L'' are P-integrable Nemitski loss functions. Then the influence function of $f_{\mathrm{P},\lambda}$ exists for all $z := (x,y) \in X \times Y$ and we have*

$$\mathrm{IF}(z; S, \mathrm{P}) = \mathbb{E}_{\mathrm{P}} L'\big(Y, f_{\mathrm{P},\lambda}(X)\big) K^{-1} \Phi(X) - L'\big(y, f_{\mathrm{P},\lambda}(x)\big) K^{-1} \Phi(x), \quad (10.35)$$

where $K : H \to H$, $K = 2\lambda \operatorname{id}_H + \mathbb{E}_{\mathrm{P}} L''\big(Y, f_{\mathrm{P},\lambda}(X)\big) \langle \Phi(X), \cdot \rangle \Phi(X)$ denotes the Hessian of the regularized risk.

Proof. Define the function $G : \mathbb{R} \times H \to H$ by

$$G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P} + \varepsilon\delta_z} L'\big(Y, f(X)\big) \Phi(X)$$

for all $\varepsilon \in \mathbb{R}$, $f \in H$. Let us first check that G is well-defined. Since $\|k\|_\infty < \infty$, we have $\|f\|_H < \infty$ for all $f \in H$. As in the proof of Lemma 2.17, we get $\mathbb{E}_{\mathrm{P}} |L'(Y, f(X))| < \infty$ for all $f \in H$. Note that $\sup_{x \in X} \|\Phi(x)\|_H = \|k\|_\infty < \infty$. Therefore, the H-valued integral used in the definition of G is defined for all $\varepsilon \in \mathbb{R}$ and all $f \in H$. Note that for $\varepsilon \notin [0,1]$ the H-valued integral is with respect to a signed measure. Now we obtain

$$G(\varepsilon, f) = \frac{\partial \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}}{\partial H}(f)\,, \quad \varepsilon \in [0,1]\,; \qquad (10.36)$$

see Lemma 2.21 and the discussion at the beginning of Section 5.2. Since the mapping $f \mapsto \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}(f)$ is convex and continuous for all $\varepsilon \in [0,1]$, equation (10.36) shows that we have

$$G(\varepsilon, f) = 0 \quad \Longleftrightarrow \quad f = f_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}$$

for such values of ε. Our aim is to show the existence of a differentiable function $\varepsilon \mapsto f_\varepsilon$ defined on a small interval $(-\delta, \delta)$ for some $\delta > 0$ that satisfies $G(\varepsilon, f_\varepsilon) = 0$ for all $\varepsilon \in (-\delta, \delta)$ because the existence of such a function guarantees

$$\mathrm{IF}(z; S, \mathrm{P}) = \frac{\partial f_\varepsilon}{\partial \varepsilon}(0)\,.$$

To this end, recall that the implicit function theorem A.5.17 for Banach spaces guarantees the existence of this function provided that *(i)* G is continuously differentiable and that *(ii)* $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is invertible.

Let us start with part *(i)*. A straightforward calculation shows that

$$\frac{\partial G}{\partial \varepsilon}(\varepsilon, f) = -\mathbb{E}_{\mathrm{P}} L'\big(Y, f(X)\big) \Phi(X) + L'\big(y, f(x)\big) \Phi(x), \qquad (10.37)$$

and a computation slightly more involved than in the proof of Lemma 2.21 gives

$$\frac{\partial G}{\partial H}(\varepsilon, f) = 2\lambda \, \mathrm{id}_H + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z} L''\big(Y, f(X)\big)\langle \Phi(X), \cdot\rangle \Phi(X) =: K. \quad (10.38)$$

In order to prove that $\frac{\partial G}{\partial \varepsilon}$ is continuous, we fix $\varepsilon \in \mathbb{R}$ and a sequence $(f_n)_{n\in\mathbb{N}}$ such that $f_n \in H$, $n \in \mathbb{N}$, and $\lim_{n\to\infty} f_n = f \in H$. Since $\|k\|_\infty < \infty$, the sequence $(f_n)_{n\in\mathbb{N}}$ is uniformly bounded. By the continuity of L' and because $|L'|$ is a P-integrable Nemitski loss function, there exists a bounded measurable function $g : Y \to \mathbb{R}$ with $L'(y, f_n(x)) \le L'(y, g(y))$ for all $n \ge 1$ and all $(x, y) \in X \times Y$. For the mapping $v(y) := L(y, g(y))$, $y \in Y$, we get $v \in L_1(\mathrm{P})$, and therefore an application of the dominated convergence theorem A.5.21 for Bochner integrals gives the continuity of $\frac{\partial G}{\partial \varepsilon}$. The continuity of G and $\frac{\partial G}{\partial H}$ can be shown analogously. Using Theorem A.5.16, we conclude that G is continuously differentiable, and *(i)* is shown.

Now let us prove *(ii)*. In order to show that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is invertible it suffices to show by the Fredholm Alternative A.5.5 that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda})$ is *injective* and that

$$Ag := \mathbb{E}_{\mathrm{P}} L''\big(Y, f_{\mathrm{P},\lambda}(X)\big) g(X)\Phi(X), \qquad g \in H,$$

defines a *compact operator* on H.

To show the compactness of the operator A, recall that X and Y are Polish spaces since we assumed that X and Y are closed subsets of $\mathbb{R}^d$ and $\mathbb{R}$, respectively; see the examples listed after Definition A.2.11. Furthermore, Borel probability measures on Polish spaces are regular by Ulam's Theorem A.3.15. Therefore, they can be approximated from inside by compact sets; see Definition A.3.14. Hence there exists a sequence of compact subsets $X_n \times Y_n \subset X \times Y$ with $\mathrm{P}(X_n \times Y_n) \ge 1 - \frac{1}{n}$, $n \in \mathbb{N}$. Let us also define a sequence of operators $A_n : H \to H$ by

$$A_n g := \int_{X_n} \int_{Y_n} L''\big(y, f_{\mathrm{P},\lambda}(x)\big)\, \mathrm{P}(dy|x)\, g(x)\Phi(x)\, d\mathrm{P}_X(x), \quad g \in H. \quad (10.39)$$

Note that if $X \times Y$ is compact, we can choose $X_n \times Y_n := X \times Y$, which implies $A = A_n$. Let us now show that all A_n are compact operators. To this end, we first observe for g in the unit ball B_H and $x \in X$ that

$$\begin{aligned} h_g(x) &:= \int_{Y_n} L''\big(y, f_{\mathrm{P},\lambda}(x)\big)|g(x)|\, \mathrm{P}(dy|x) \\ &\le \|k\|_\infty \int_{Y_n} |L''\big(y, f_{\mathrm{P},\lambda}(x)\big)|\, \mathrm{P}(dy|x) =: h(x), \quad n \in \mathbb{N}, \end{aligned}$$

because L'' is a P-integrable Nemitski loss function. Therefore, we have $h \in L_1(\mathrm{P}_X)$, which implies $h_g \in L_1(\mathrm{P}_X)$ with $\|h_g\|_1 \le \|h\|_1$ for all $g \in B_H$. Consequently, $d\mu_g := h_g d\mathrm{P}_X$ and $d\mu := h d\mathrm{P}_X$ are finite measures. By Theorem A.5.22, we hence obtain

$$A_n g = \int_{X_n} \operatorname{sign} g(x)\Phi(x)\, h_g(x)\, d\mathrm{P}_X(x) = \int_{X_n} \operatorname{sign} g(x)\Phi(x)\, d\mu_g(x)$$
$$\in \mu_g(X_n)\,\overline{\operatorname{aco}\Phi(X_n)} \subset \mu_g(X_n)\,\overline{\operatorname{aco}\Phi(X_n)}\,, \qquad g \in H,$$

where $\operatorname{aco}\Phi(X_n)$ denotes the absolute convex hull of $\Phi(X_n)$, and the closure is with respect to $\|\cdot\|_H$. Now, using the continuity of Φ, we see that $\Phi(X_n)$ is compact and hence so is the closure of $\operatorname{aco}\Phi(X_n)$. This shows that A_n is a compact operator. In order to see that A is compact, it therefore suffices to show that $\|A_n - A\| \to 0$ with respect to the operator norm for $n \to \infty$. Recalling that the convexity of L implies $L'' \geq 0$, the desired convergence follows from $\mathrm{P}(X_n \times Y_n) \geq 1 - \frac{1}{n}$, $L'' \circ f_{\mathrm{P},\lambda} \in L_1(\mathrm{P})$, and

$$\begin{aligned}
\|A_n g - Ag\|_H &= \Big\| \int_{(X\times Y)\setminus(X_n\times Y_n)} L''\big(y, f_{\mathrm{P},\lambda}(x)\big) g(x)\Phi(x)\, d\mathrm{P}(x,y) \Big\|_H \\
&\leq \int_{(X\times Y)\setminus(X_n\times Y_n)} L''\big(y, f_{\mathrm{P},\lambda}(x)\big)\, |g(x)|\, \|\Phi(x)\|_H\, d\mathrm{P}(x,y) \\
&\leq \|k\|_\infty^2\, \|g\|_H \int_{(X\times Y)\setminus(X_n\times Y_n)} L''\big(y, f_{\mathrm{P},\lambda}(x)\big)\, d\mathrm{P}(x,y)\,.
\end{aligned}$$

Let us now show that $\frac{\partial G}{\partial H}(0, f_{\mathrm{P},\lambda}) = 2\lambda \operatorname{id}_H + A$ is injective. For $g \in H\setminus\{0\}$, we obtain

$$\begin{aligned}
\big\langle (2\lambda \operatorname{id}_H + A)g, (2\lambda \operatorname{id}_H + A)g \big\rangle &= 4\lambda^2\langle g, g\rangle + 4\lambda\langle g, Ag\rangle + \langle Ag, Ag\rangle \\
&> 4\lambda\langle g, Ag\rangle \\
&= 4\lambda\big\langle g, \mathbb{E}_\mathrm{P} L''\big(Y, f_{\mathrm{P},\lambda}(X)\big) g(X)\Phi(X)\big\rangle \\
&= 4\lambda \mathbb{E}_\mathrm{P} L''\big(Y, f_{\mathrm{P},\lambda}(X)\big) g^2(X) \\
&\geq 0,
\end{aligned}$$

which shows the injectivity and hence *(ii)*.

The implicit function theorem A.5.17 guarantees that the function $\varepsilon \mapsto f_\varepsilon$ is differentiable on $(-\delta, \delta)$ if $\delta > 0$ is small enough. Furthermore, (10.37) and (10.38) yield for $z = (x, y) \in X \times Y$ that

$$\begin{aligned}
\mathrm{IF}(z; S, \mathrm{P}) &= \frac{\partial f_\varepsilon}{\partial \varepsilon}(0) = -K^{-1} \circ \frac{\partial G}{\partial \varepsilon}(0, f_{\mathrm{P},\lambda}) \\
&= K^{-1}\big(\mathbb{E}_\mathrm{P}\big(L'(Y, f_{\mathrm{P},\lambda}(X))\Phi(X)\big)\big) - L'\big(y, f_{\mathrm{P},\lambda}(x)\big) K^{-1}\Phi(x). \quad \square
\end{aligned}$$

Remark 10.19. Theorem 10.18 contains a tail assumption on P to ensure that $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) < \infty$. Recall that $\mathcal{R}_{L,\mathrm{P}} : L_p(\mathrm{P}_X) \to [0, \infty)$ is well-defined and continuous if L is a P-integrable Nemitski loss function of order $p \in [1, \infty)$, see Lemma 2.17. Taking Remark 10.3 into account, a tail assumption on P seems to be quite natural for *unbounded* sets Y. We see three ways to avoid this tail assumption on P, but all of them seem to be unsatisfactory. *(i)* One can restrict attention to bounded sets Y, but this is often unrealistic for regression

problems and does not make sense from the viewpoint of robust statistics. *(ii)* One can replace the convex loss function by a bounded non-convex loss function, but in general this yields non-convex risk functions and hence problems with respect to the existence and uniqueness of $f_{\mathrm{P},\lambda}$. Further, the advantage of computational efficiency often vanishes because the numerical problems can turn from convex problems into NP-hard ones. *(iii)* The redefinition of the minimization problem $\inf_{f\in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda\|f\|_H^2$ into

$$\inf_{f\in H} \mathbb{E}_{\mathrm{P}} L^*(X, Y, f(X)) + \lambda\|f\|_H^2,$$

where $L^* : X \times Y \times \mathbb{R}$, $L^*(x, y, t) := L(x, y, t) - L(x, y, 0)$, seems to be helpful for Hölder-continuous loss functions, which includes of course the important class of Lipschitz-continuous loss functions, but $L^*(x, y, f(x))$ can be *negative* for some values of $(x, y, f(x))$. ◁

Remark 10.20. The proof of Theorem 10.18 can be modified in order to replace point mass contaminations δ_z by arbitrary contaminations $\mathrm{Q} \in \mathcal{M}_1$ if L, $|L'|$, and L'' are Q-integrable Nemitski loss functions. ◁

Remark 10.21. From a robustness point of view, one is mainly interested in statistical methods with *bounded influence functions*. It is worth mentioning that Theorem 10.18 not only ensures the existence of the influence function $\mathrm{IF}(z; S, \mathrm{P})$ but also indicates how to guarantee its boundedness. Indeed, (10.35) shows that the only term of the influence function that depends on the point mass contamination δ_z is

$$-L'\big(y, f_{\mathrm{P},\lambda}(x)\big)K^{-1}\Phi(x)\,. \tag{10.40}$$

Hence a combination of a bounded continuous kernel with a convex loss function with L' being bounded assures a bounded influence function. Hence, we have, analogous to the results obtained in Section 10.3, that the class of Lipschitz-continuous loss functions are of primary interest from the viewpoint of robust statistics. ◁

The next result gives influence functions for distance-based loss functions.

Corollary 10.22. *Let $X = \mathbb{R}^d$, $Y = \mathbb{R}$, and H be an RKHS of a bounded continuous kernel k on X with canonical feature map $\Phi : X \to H$. Further, let $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex distance-based loss function with representing function $\psi : \mathbb{R} \to [0, \infty)$ having partial derivatives $L' = \partial_2 L$ and $L'' = \partial_{22} L$ such that L, $|L'|$, and L'' are P-integrable Nemitski loss functions and $\|b\|_{L_1(\mathrm{P})} < \infty$. Then the following statements hold.*

i) The influence function of $f_{\mathrm{P},\lambda}$ exists for all $z := (x, y) \in X \times Y$ and

$$\mathrm{IF}(z; S, \mathrm{P}) = -\mathbb{E}_{\mathrm{P}}\psi'\big(Y - f_{\mathrm{P},\lambda}(X)\big)K^{-1}\Phi(X) + \psi'\big(y - f_{\mathrm{P},\lambda}(x)\big)K^{-1}\Phi(x), \tag{10.41}$$

where $K : H \to H$, $K = 2\lambda\,\mathrm{id}_H + \mathbb{E}_{\mathrm{P}}\psi''\big(Y - f_{\mathrm{P},\lambda}(X)\big)\langle\Phi(X), \cdot\rangle\Phi(X)$.

ii) The influence function $\mathrm{IF}(z; S, \mathrm{P})$ *is bounded in* z *if* L *is Lipschitz-continuous.*

Proof. Theorem 10.18 yields that $\mathrm{IF}(z; S, \mathrm{P})$ exists and is bounded provided $L'(\cdot, f_{\mathrm{P},\lambda}(x)) : \mathbb{R} \to \mathbb{R}$ is bounded for all $x \in X$. For distance-based loss functions, we hence immediately obtain the assertions. □

Remark 10.23.

i) Let us emphasize that it is *not* sufficient to choose a bounded continuous kernel alone to obtain a bounded influence function: the loss function also must be chosen appropriately.

ii) Corollary 10.22 shows that the SVM based on the least squares loss, which has some nice computational properties, as will be shown in Chapter 11, is a method with an *unbounded* influence function. Therefore, an ad hoc one-step reweighted version of the SVM based on the least squares loss can be interesting from a robustness point of view, see Suykens *et al.* (2002) and Debruyne *et al.* (2007).

iii) In contrast, the logistic loss function provides a robust method with a *bounded* influence function if we use it in combination with a Gaussian RBF kernel. ◁

Example 10.24. For illustration purposes, let us consider a Gaussian RBF kernel and the loss functions $L_{\text{r-logist}}$ and L_{LS}. Note that the partial derivative with respect to the last argument of $L_{\text{r-logist}}$ is bounded, whereas the corresponding partial derivative of L_{LS} is unbounded. We expect by (10.40) and (10.41) that SVMs based on $L_{\text{r-logist}}$ will give more robust results than SVMs using L_{LS}. Figure 10.11 shows that this is indeed true. The small data set listed in Table 10.1 contains 100 data points $(x_i, y_i) \in \mathbb{R}^2$ of a baby's daily milk consumption.[2] The upper subplot shows that both SVMs offer similar fits to the original data set. Now let us assume that the tired father made a typing error when he reported the daily measurement during the night for day number 40: he reported $y_{40} = 7.4$ instead of $y_{40} = .74$, which is an obvious mistake. This single typing error has a strong impact on $f_{\mathrm{D},\lambda}$ based on L_{LS} in a broad interval. On the other hand, the impact of this extreme y-value on $f_{\mathrm{D},\lambda}$ based on $L_{\text{r-logist}}$ is much smaller due to the boundedness of L'. ◁

Bounds for the Influence Function

Unfortunately, Theorem 10.18 and Corollary 10.22 require a twice continuously differentiable loss function and therefore they cannot be used to investigate SVMs based on, for example, the ϵ-insensitive loss, Huber's loss, or the pinball loss function, which are not Fréchet differentiable in one or two

[2] From one of the authors

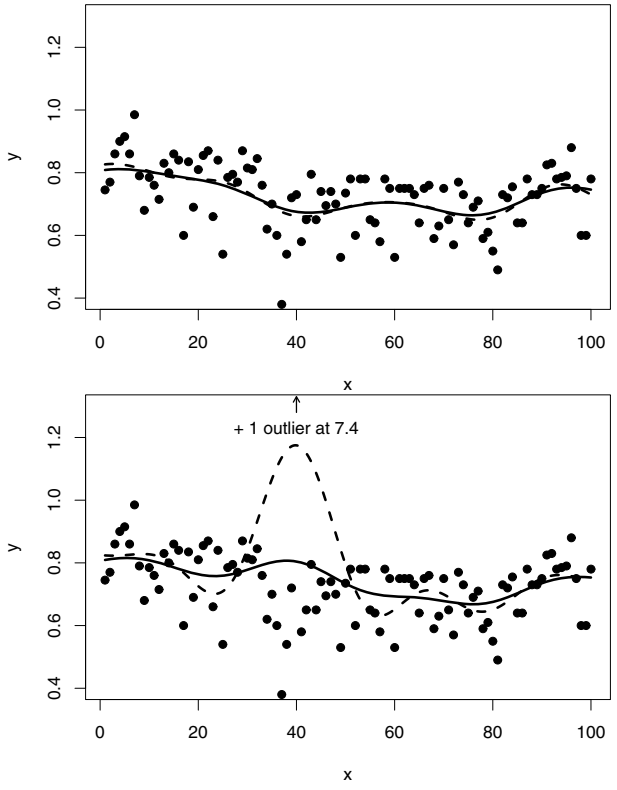


Fig. 10.11. Daily milk consumption. The upper subplot displays the fitted regression curves $f_{\mathrm{D},\lambda}(x)$ for the original data set. The lower subplot displays the fitted regression curves $f_{\mathrm{D},\lambda}(x)$ for the data set containing one extreme value. The curves were fitted using the loss functions $L_{\text{r-logist}}$ (solid) and L_{LS} (dashed), respectively.

points.[3] The next three theorems give bounds for the difference quotient used in the definition of the influence function and apply to convex Nemitski loss functions of some order p. Hence these results partially resolve the problem above for non-Fréchet differentiable loss functions. For practical purposes, the following results may even be more interesting than the results for the influence function because they give bounds for the bias and do not consider an infinitesimally small amount of contamination. Note that the following three results show that *upper bounds for the bias* under gross error contamination models increase *at most linearly* with respect to the mixing proportion ε because the constants c used in the bounds do not depend on ε.

[3] Christmann and Van Messem (2008) derived an analogon to Theorem 10.18 using Bouligand derivatives for SVMs based on non-smooth Lipschitz-continuous losses.

Table 10.1. Data set: daily milk consumption of a baby. The measurements are listed in liters.

Day	Milk	Day	Milk	Day	Milk	Day	Milk	Day	Milk
1	0.745	2	0.770	3	0.860	4	0.900	5	0.915
6	0.860	7	0.985	8	0.790	9	0.680	10	0.785
11	0.760	12	0.715	13	0.830	14	0.800	15	0.860
16	0.840	17	0.600	18	0.835	19	0.690	20	0.810
21	0.855	22	0.870	23	0.660	24	0.840	25	0.540
26	0.785	27	0.795	28	0.770	29	0.870	30	0.815
31	0.810	32	0.845	33	0.760	34	0.620	35	0.700
36	0.600	37	0.380	38	0.540	39	0.720	40	0.740
41	0.580	42	0.650	43	0.795	44	0.650	45	0.740
46	0.695	47	0.740	48	0.700	49	0.530	50	0.735
51	0.780	52	0.600	53	0.780	54	0.780	55	0.650
56	0.640	57	0.580	58	0.780	59	0.750	60	0.530
61	0.750	62	0.750	63	0.750	64	0.730	65	0.640
66	0.750	67	0.760	68	0.590	69	0.630	70	0.750
71	0.650	72	0.570	73	0.770	74	0.730	75	0.640
76	0.690	77	0.710	78	0.590	79	0.610	80	0.550
81	0.490	82	0.730	83	0.720	84	0.755	85	0.640
86	0.640	87	0.780	88	0.730	89	0.730	90	0.750
91	0.825	92	0.830	93	0.780	94	0.785	95	0.790
96	0.880	97	0.750	98	0.600	99	0.600	100	0.780

Theorem 10.25. *Let* $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1$ *and* $L : Y \times \mathbb{R} \to [0, \infty)$ *be a convex* P*-integrable and* Q*-integrable Nemitski loss function of order* $p \in [1, \infty)$*. Furthermore, let* k *be a bounded, measurable kernel on* X *with separable RKHS* H *and canonical feature map* $\Phi : X \to H$*. Then, for all* $\lambda > 0$ *and* $\varepsilon \in [0, 1]$*, we have*

$$\left\| f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda} \right\|_H \leq c_{\mathrm{P},\mathrm{Q}}\, \varepsilon\,, \tag{10.42}$$

where

$$c_{\mathrm{P},\mathrm{Q}} := \frac{\|b\|_{L_1(\mathrm{P})} + \|b\|_{L_1(\mathrm{Q})} + 2^{2p+1} c B_\lambda^p}{(\lambda\, \mathcal{R}_{L,\mathrm{P}}(0))^{1/2}}\,,$$

$B_\lambda := \|k\|_\infty (\mathcal{R}_{L,\mathrm{P}}(0)/\lambda)^{1/2}$*, and* $b : X \times Y \to [0, \infty)$ *is a function satisfying the Nemitski condition* (2.9)*. If* $\mathrm{Q} = \delta_{(x,y)}$ *and if* $\mathrm{IF}((x, y); S, \mathrm{P})$ *exists, then*

$$\|\mathrm{IF}((x, y); S, \mathrm{P})\|_H \leq c_{\mathrm{P},\delta_{(x,y)}}\,, \quad (x, y) \in X \times Y \tag{10.43}$$

and

$$\gamma_u^*(S, \mathrm{P}) \leq \frac{\|b\|_{L_1(\mathrm{P})} + \sup_{(x,y)\in X\times Y} b(x, y) + 2^{2p+1} c B_\lambda^p}{(\lambda\, \mathcal{R}_{L,\mathrm{P}}(0))^{1/2}}\,.$$

Proof. Fix $\varepsilon \in [0, 1]$ and define $\tilde{\mathrm{P}} := (1 - \varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}$. By Theorem 5.9, there exists a bounded, measurable function $h : X \times Y \to \mathbb{R}$ independent of ε and Q such that

$$\begin{aligned}\|f_{\mathrm{P},\lambda} - f_{\tilde{\mathrm{P}},\lambda}\|_H &\le \lambda^{-1} \|\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} h\Phi\|_H \\ &= \varepsilon\, \lambda^{-1} \|\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\mathrm{Q}} h\Phi\|_H \\ &\le \varepsilon\, \lambda^{-1} \|k\|_\infty \big(\|h\|_{L_1(\mathrm{P})} + \|h\|_{L_1(\mathrm{Q})}\big) \\ &\le \varepsilon\, \lambda^{-1} \|k\|_\infty \big(\|b\|_{L_1(\mathrm{P})} + \|b\|_{L_1(\mathrm{Q})} + 2c|4B_\lambda|^p\big)/B_\lambda\,,\end{aligned}$$

where (5.16) is used in the last inequality. This gives the assertion (10.42). The inequality (10.43) follows from the definition of the influence function and the fact that the constant $c_{\mathrm{P},\mathrm{Q}}$ does not depend on the mixture proportion ε. For the special case $\mathrm{Q} = \delta_{(x,y)}$ we have $\|b\|_{L_1(\mathrm{Q})} = b(x,y)$ and hence we obtain bounds for the difference quotient used in the definition of the influence function if we divide the bound by ε. □

If we restrict our attention to distance-based loss functions of growth type $p \ge 1$, and many loss functions used in practice are distance-based, we are able to obtain stronger results.

Theorem 10.26. *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss function of upper growth type $p > 1$ with representing function $\psi : \mathbb{R} \to [0,\infty)$ and P, $\mathrm{Q} \in \mathcal{M}_1$ such that L is P-integrable and Q-integrable. Furthermore, let k be a bounded, measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. Then, for all $\lambda > 0$ and for $\varepsilon \in [0,1]$, we have*

$$\|f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}\|_H \le c_{\mathrm{P},\mathrm{Q}}\, \varepsilon\,, \tag{10.44}$$

where

$$c_{\mathrm{P},\mathrm{Q}} = \tilde{c}_3\, \lambda^{-1} \|k\|_\infty \big(|\mathrm{P}-\mathrm{Q}|_{p-1}^{p-1} + \|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}} \left(1 + \lambda^{(1-p)/2} \|k\|_\infty^{p-1} |\mathrm{P}|_p^{p(p-1)/2}\right)\big)$$

and the constant $\tilde{c}_3 = \tilde{c}_3(L,p,k,\lambda) \ge 0$. If $\mathrm{Q} = \delta_{(x,y)}$ and if the influence function exists, then

$$\|\mathrm{IF}((x,y); S, \mathrm{P})\|_H \le c_{\mathrm{P},\delta_{(x,y)}}\,, \quad (x,y) \in X \times Y, \tag{10.45}$$

and the gross error sensitivity fulfills

$$\gamma_u^*(S,\mathrm{P}) \le \sup_{(x,y)\in X\times Y} c_{\mathrm{P},\delta_{(x,y)}}\,. \tag{10.46}$$

Proof. Corollary 5.11 gives the existence of a measurable function $h : X \times Y \to \mathbb{R}$ with

$$\|f_{\mathrm{P},\lambda} - f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}\|_H \le \frac{\varepsilon}{\lambda} \|\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\mathrm{Q}} h\Phi\|_H\,.$$

As L is a loss function of upper growth type $p > 1$, there exists $c > 0$ such that

$$\lambda \|f_{\mathrm{P},\lambda}\|_H^2 \le \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{\mathrm{P},\lambda}) \le \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(0) \le c\big(|\mathrm{P}|_p^p + 1\big)\,.$$

This yields for some $\tilde{c}_1 > 0$ the inequalities

$$\|f_{\mathrm{P},\lambda}\|_\infty \le \|k\|_\infty \|f_{\mathrm{P},\lambda}\|_H \le (\tilde{c}_1/\lambda)^{1/2}\|k\|_\infty |\mathrm{P}|_p^{p/2}\,.$$

Using Corollary 5.11 and (5.28) from its proof, we obtain for some $\tilde{c}_2 > 0$ that

$$\begin{aligned}|h(x,y)| &\le 4^p c_L \max\bigl\{1, |y - f_{\mathrm{P},\lambda}(x)|^{p-1}\bigr\} \\ &\le 4^p c_L \bigl(1 + |y|^{p-1} + |f_{\mathrm{P},\lambda}(x)|^{p-1}\bigr) \\ &\le \tilde{c}_2 \left(1 + |y|^{p-1} + \lambda^{(1-p)/2}\|k\|_\infty^{p-1}\,|\mathrm{P}|_p^{p(p-1)/2}\right), \quad (x,y) \in X \times Y.\end{aligned}$$

It follows that

$$\begin{aligned}&\|f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}\|_H \;\le\; \varepsilon\,\lambda^{-1}\,\|k\|_\infty\,\mathbb{E}_{|\mathrm{P}-\mathrm{Q}|}|h| \\ &\le \varepsilon\,\tilde{c}_3\,\lambda^{-1}\|k\|_\infty \bigl(|\mathrm{P}-\mathrm{Q}|_{p-1}^{p-1} + \|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}}\left(1 + \lambda^{(1-p)/2}\|k\|_\infty^{p-1}|\mathrm{P}|_p^{p(p-1)/2}\right)\bigr)\,,\end{aligned}$$

where $\tilde{c}_3$ depends on L but not on ε. This gives the assertion in (10.44). The inequalities (10.45) and (10.46) follow immediately from the fact that $c_{\mathrm{P},\mathrm{Q}}$ does not depend on ε. □

Recall that Lipschitz-continuous, distance-based, convex loss functions are of upper growth type $p = 1$ due to Lemma 2.36*ii)*. Such loss functions are therefore *not* covered by the previous theorem, but by the next one. Important special cases of the next theorem are the ϵ-insensitive loss and the pinball loss.

Theorem 10.27. *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a Lipschitz-continuous, convex, distance-based loss function with representing function $\psi : \mathbb{R} \to [0,\infty)$, and $\mathrm{P},\ \mathrm{Q} \in \mathcal{M}_1$ with $|\mathrm{P}|_1 < \infty$ and $|\mathrm{Q}|_1 < \infty$. Furthermore, let k be a bounded and measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. Then, for all $\lambda > 0$ and $\varepsilon \in [0,1]$, we have*

$$\bigl\|f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}\bigr\|_H \le c_{\mathrm{P},\mathrm{Q}}\,\varepsilon\,,$$

where

$$c_{\mathrm{P},\mathrm{Q}} = \lambda^{-1}\,\|k\|_\infty\,|\psi|_1\,\|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}}.$$

If $\mathrm{Q} = \delta_{(x,y)}$ and if the influence function of $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$ exists, then

$$\|\mathrm{IF}((x,y);S,\mathrm{P})\|_H \le c_{\mathrm{P},\delta_{(x,y)}}\,, \quad (x,y) \in X \times Y,$$

and the gross error sensitivity fulfills

$$\gamma_u^*(S,\mathrm{P}) \le \lambda^{-1}\,\|k\|_\infty\,|\psi|_1 \sup_{(x,y)\in X\times Y} \|\mathrm{P} - \delta_{(x,y)}\|_{\mathcal{M}} \le 2\lambda^{-1}\,\|k\|_\infty\,|\psi|_1\,.$$

In particular, the H-norm of the sensitivity curve is bounded by a uniform constant independent of the sample size n, independent of (x,y), and valid for any empirical distribution $\mathrm{D}_n \in \mathcal{M}_1$:

$$\|\mathrm{SC}_n((x,y);f_{\mathrm{D}_n,\lambda})\|_H \le 2\lambda^{-1}\,\|k\|_\infty\,|\psi|_1\,, \quad (x,y) \in X \times Y. \tag{10.47}$$

Proof. Corollary 5.10 guarantees that there exists a bounded measurable function $h : X \times Y \to \mathbb{R}$ such that $\|h\|_\infty \le |L|_{B_\lambda,1} \le |\psi|_1$ and

$$\|f_{\mathrm{P},\lambda} - f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}\|_H \le \varepsilon\lambda^{-1}\|\mathbb{E}_\mathrm{P} h\Phi - \mathbb{E}_\mathrm{Q} h\Phi\|_H \,.$$

It follows that

$$\|f_{\mathrm{P},\lambda} - f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}\|_H \le \varepsilon\lambda^{-1}\|k\|_\infty \mathbb{E}_{|\mathrm{P}-\mathrm{Q}|}|h| \le \varepsilon\lambda^{-1}\|k\|_\infty |\psi|_1 \, \|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}} \,.$$

This gives the assertion. □

Remark 10.28.

i) The combination of a Lipschitz-continuous loss function with $|\psi|_1 = 1$ and a Gaussian RBF kernel with $\|k\|_\infty = 1$ is of particular interest for applications. The H-norm of the sensitivity curve is then uniformly bounded by $\frac{2}{\lambda}$ for all data sets and for all points (x, y) by Theorem 10.27.

ii) The previous three theorems also offer bounds for $\mathrm{maxbias}(\varepsilon; S, \mathrm{P})$ over contamination neighborhoods provided attention is restricted to distributions $\mathrm{Q} \in N_\varepsilon(\mathrm{P})$ satisfying the tail conditions given in those theorems; see Problem 10.7. ◁

Comparison of SVMs with M-Estimation (*)

Let us now compare the influence function of SVMs with the influence function of M-estimators in *linear* regression models. A linear regression model assumes that (X_i, Y_i) are independent and identically distributed random variables according to $\mathrm{P} \in \mathcal{M}_1(X \times Y)$, $X = \mathbb{R}^d$, $Y = \mathbb{R}$, with regular conditional distribution such that

$$\mathbb{E}_\mathrm{P}(Y|x) := \int_Y y \, d\mathrm{P}(y|x) = x^\mathsf{T}\theta \,,$$

where $\theta \in \Theta := \mathbb{R}^d$ is unknown. Such linear models are quite popular in many areas of applied statistics and in data mining. Obviously, linear regression is a special case of SVMs: we use $\lambda = 0$ in combination with a linear kernel $k(x, x') := \langle x, x' \rangle$, $x, x' \in \mathbb{R}^d$. Let us assume for reasons of simplicity that the scale parameter $\sigma \in (0, \infty)$ of the linear regression model is known, say $\sigma = 1$. Note that $\sigma^2 = \mathrm{Var}_\mathrm{P}(Y|x)$ for Gaussian distributions. The function $S : \mathcal{M}_1 \to \mathbb{R}^d$ corresponding to an M-estimator is the solution of

$$\mathbb{E}_\mathrm{P} \, \eta(X, Y - X^\mathsf{T} S(\mathrm{P}))X = 0 \,, \tag{10.48}$$

where the odd function $\eta(x, \cdot)$ is continuous for $x \in \mathbb{R}^d$ and $\eta(x, u) \ge 0$ for all $x \in \mathbb{R}^d$, $u \in [0, \infty)$. Almost all M-estimators for linear regression proposed in the literature may be written in the form $\eta(x, u) = \psi\big(v(x)u\big)w(x)$, where $\psi : \mathbb{R} \to \mathbb{R}$ is usually continuous, bounded, and increasing and $w : \mathbb{R}^d \to [0, \infty)$, $v : \mathbb{R}^d \to [0, \infty)$ are weight functions. The functions ψ, w,

and v are chosen in advance by the user or may result as solutions to certain robust optimization problems. An important subclass of M-estimators are those of Mallows type, where η is of the form $\eta(x,u) = \psi(u)w(x)$. Note that defining M-estimators as solutions of (10.48) is more general than defining M-estimators via optimization problems. The influence function of $S(\mathrm{P}) = \theta$ in the point $z = (x,y)$ at a distribution $\mathrm{P} \in \mathcal{M}_1$ is given by

$$\mathrm{IF}(z; S, \mathrm{P}) = K^{-1}(\eta, \mathrm{P})\, \eta(x, y - x^{\mathsf{T}} S(\mathrm{P}))\, x \in \mathbb{R}^d\,, \tag{10.49}$$

where $K(\eta, \mathrm{P}) := \mathbb{E}_{\mathrm{P}}\, \eta'(X, Y - X^{\mathsf{T}} S(\mathrm{P}))XX^{\mathsf{T}}$. An important difference between SVMs and M-estimators in linear regression is that $\mathrm{IF}(z; S, \mathrm{P}) \in \mathbb{R}^d$ in (10.49), but $\mathrm{IF}(z; S, \mathrm{P}) \in H$ in (10.35). In other words, the influence function of an M-estimator is only a function if z varies, whereas the influence function of SVMs is an H-valued function already for any fixed value of z. For linear kernels, there exists an isomorphism between the RKHS H and $\mathbb{R}^d$, but this is not true for many other kernels (e.g., for Gaussian RBF kernels).

A comparison of the influence functions for SVMs (see (10.35) and (10.41)) with the influence function of M-estimators given by (10.49) yields that both influence functions nevertheless have a similar structure. The function $K = K(L'', k, \mathrm{P})$ for SVMs and the matrix $K(\eta, \mathrm{P})$ for M-estimation do not depend on z. The terms in the influence functions depending on $z = (x,y)$, where the point mass contamination δ_z occurs, are a product of two factors. The first factors, measuring the outlyingness in the y-direction, are

$-L'(y, f_{\mathrm{P},\lambda}(x))$	for SVMs,
$\psi(v(x)(y - x^{\mathsf{T}}\theta))$	for general M-estimation,
$\psi'(y - f_{\mathrm{P},\lambda}(x))$	for SVMs with distance-based loss, and
$\psi(y - x^{\mathsf{T}}\theta)$	for Mallows type M-estimation.

SVMs based on a distance-based loss function and Mallows type M-estimators use first factors that only depend on the residuals. The second factors are

$K^{-1}\Phi(x)$	for SVMs, and
$w(x)x$	for M-estimation.

Therefore, the second factors do not depend on y and measure the outlyingness in the x-direction. Note that $K^{-1}\Phi(x)$ takes values in the RKHS H in the case of SVMs, whereas $w(x)x \in \mathbb{R}^d$ for M-estimation.

Concluding, one can say that there is a natural connection between SVMs and M-estimators for linear regression, although M-estimators have no penalty term (i.e., $\lambda = 0$) and are estimating a vector in $\mathbb{R}^d$, whereas SVMs with nonlinear kernels estimate a function in a reproducing kernel Hilbert space that can have an infinite dimension.

10.5 Robust Learning from Bites (*)

In this section, we investigate a simple method called robust learning from bites (RLB) based on independent subsampling. The main goal of the method

is to broaden the applicability of robust SVMs for huge data sets where classical algorithms to compute $f_{\mathrm{D},\lambda}$ for the whole data set might be too slow.

The idea of RLB is quite simple. Consider a huge data set $D = D_n$ containing data points $(x_i, y_i) \in X \times Y$, $i = 1, \ldots, n$. First split the huge data set D randomly into disjoint subsets S_b with $|S_b| = n_b$, where $1 \leq b \leq B$ and $B \in \mathbb{N}$ much smaller than n. Then use robust SVMs for each subset and aggregate the SVM results in a robust manner.

In this section, we will assume that $X \subset \mathbb{R}^d$, $d \in \mathbb{N}$, $Y \subset \mathbb{R}$, and n is large. We will further assume that $\min_{1 \leq b \leq B} n_b \to \infty$ as $n \to \infty$.

Definition 10.29. *Let $D = \big((x_1, y_1), \ldots, (x_n, y_n)\big) \in X \times Y^n$, $n \in \mathbb{N}$. Consider a random partition of D into B non-empty and disjoint subsets, i.e.,*

$$D = S_1 \uplus \ldots \uplus S_B \,,$$

where $S_b \subset D$, $n_b := |S_b|$, $n = \sum_{b=1}^{B} n_b$, $b \in \{1, \ldots, B\}$, $B \in \{1, \ldots, n\}$, $B \ll n$. Let $f_{\mathrm{S}_b,\lambda}$ be the SVM decision functions based on the subsamples S_b, $b = 1, \ldots, B$, and let $g : H^B \to H$ and $g^* : \mathbb{R}^B \to \mathbb{R}$ be measurable functions.

*i) An **RLB estimator of type I** is defined by*

$$f_{\mathrm{D},\lambda,B}^{RLB,I} = g(f_{\mathrm{S}_1,\lambda}, \ldots, f_{\mathrm{S}_B,\lambda}) \,.$$

*ii) An **RLB estimator of type II** is given by*

$$f_{\mathrm{D},\lambda,B}^{RLB,II}(x) = g^*(f_{\mathrm{S}_1,\lambda}(x), \ldots, f_{\mathrm{S}_B,\lambda}(x)) \,, \qquad \forall\, x \in X.$$

Remark 10.30.

i) Obviously, RLB estimates can be defined not only based on SVMs.
ii) An RLB estimator of type I can obviously be used to define an RLB estimator of type II.
iii) An RLB estimator of type II does not necessarily define an RLB estimator of type I because the related function g^* does not necessarily correspond to a measurable and H-valued function g.
iv) The class of RLB estimators of type I and, due to *ii)*, the class of RLB estimators of type II are non-empty because for g equal to the mean we have $g(f_{\mathrm{S}_1,\lambda}, \ldots, f_{\mathrm{S}_B,\lambda}) = \frac{1}{B} \sum_{b=1}^{B} f_{\mathrm{S}_b,\lambda} \in H$. ◁

From our point of view, RLB is mainly a numerical algorithm to allow the computation of $f_{\mathrm{D},\lambda}$ or the predictions $f_{\mathrm{D},\lambda}(x)$ for huge data sets. Nevertheless, it will turn out that RLB estimators have some nice properties. Clearly, an RLB estimator of type II is of interest if only predictions are necessary.

Table 10.2 summarizes the three main steps of RLB. Let us now restrict attention to location estimators in the aggregation step belonging to the flexible class of L-estimators that use non-negative weights $a_{B,b} \in [0, 1]$ with $\sum_{b=1}^{B} a_{B,b} = 1$; see Huber (1981, Section 3.3) for the general case. L-estimators are *linear* combinations of order statistics (hence their name has nothing to

Table 10.2. Principle of RLB.

Step 1: Construct bites.
Split data set D randomly into B disjoint subsets S_b of sample sizes $n_b \approx \lfloor n/B \rfloor$, $n_b > 1$, $b = 1, \ldots, B$, $B \ll n$.
Step 2: Fit bites.
for $(b = 1, \ldots, B)$
{ Compute the (robust) estimator $f_{\mathrm{S}_b,\lambda}$ based on bite S_b. }
Step 3: Aggregate predictions.
Fix weights $a_{B,b} \in [0,1]$ with $\sum_{b=1}^{B} a_{B,b} = 1$.
RLB type I: compute $f_{\mathrm{D},\lambda,B}^{RLB,I} = \sum_{b=1}^{B} a_{B,b} f_{\mathrm{S}_b,\lambda}$.
RLB type II: compute $f_{\mathrm{D},\lambda,B}^{RLB,II}(x_i) = \sum_{b=1}^{B} a_{B,b} f_{\mathrm{S}_b,\lambda}(x_i)$, $x_i \in X$.
If RLB type II and g^* is the median:
(i) Compute $f_{\mathrm{D},\lambda,B}^{RLB,II}(x_i) = \mathrm{median}_{1 \le b \le B} f_{\mathrm{S}_b,\lambda}(x_i)$, $x_i \in X$.
(ii) Compute distribution-free $(1-\alpha)$ confidence intervals for $f_{\mathrm{P},\lambda,B}^{RLB,II}(x_i)$ based on the pair (r,s) of order statistics; i.e. $[f_{\mathrm{S}_{(r:B)},\lambda}(x), f_{\mathrm{S}_{(s:B)},\lambda}(x)]$.

do with the loss function L). L-estimators in the aggregation step are defined by

$$f_{\mathrm{D},\lambda,B}^{RLB,II}(x) = \sum_{b=1}^{B} a_{B,b} f_{\mathrm{S}_{(\mathrm{b:B})},\lambda}(x)\,, \tag{10.50}$$

where

$$\{f_{\mathrm{S}_{(\mathrm{b:B})},\lambda}(x) : b = 1, \ldots, B\} = \{f_{\mathrm{S}_b,\lambda}(x) : b = 1, \ldots, B\}$$

and

$$f_{\mathrm{S}_{(1:B)},\lambda}(x) \le f_{\mathrm{S}_{(2:B)},\lambda}(x) \le \ldots \le f_{\mathrm{S}_{(B:B)},\lambda}(x)\,, \quad x \in X, \tag{10.51}$$

denote the order statistics of $f_{\mathrm{S}_b,\lambda}(x)$, $b = 1, \ldots, B$. Such L-estimators are obviously convex combinations of the B estimators computed for the bites. Important L-estimators for location are given in the following example.

Example 10.31.

i) The *α-trimmed means*, $\alpha \in [0, 1/2)$, use

$$a_{B,b} = \begin{cases} \frac{1}{B - 2\lfloor \alpha B \rfloor} & \text{if } b \in \{\lfloor \alpha B \rfloor + 1, \ldots, B - \lfloor \alpha B \rfloor\}, \\ 0 & \text{otherwise.} \end{cases}$$

ii) The *mean* is obtained for $\alpha = 0$ (i.e., $a_{B,b} = \frac{1}{B}$) and we obtain an RLB estimator of type I.

iii) The median uses

$$a_{B,b} = \begin{cases} \frac{1}{2} & \text{if } B \text{ is even and } b \in \left\{\frac{B}{2}, \frac{B}{2}+1\right\} \\ 0 & \text{if } B \text{ is even and } b \notin \left\{\frac{B}{2}, \frac{B}{2}+1\right\} \\ 1 & \text{if } B \text{ is odd and } b = \frac{B+1}{2} \\ 0 & \text{if } B \text{ is odd and } b \neq \frac{B+1}{2}\,, \end{cases}$$

which is the limiting case as $\alpha \to 1/2$. ◁

Our leading examples will be convex combinations of $f_{\mathrm{S}_b,\lambda}$ and the median. Of course, other robust estimators can be used instead (e.g., M-estimators, S-estimators, or Hodges-Lehmann-type R-estimators); see Huber (1981, p. 63).

If B is large, precision estimates can additionally be obtained by computing standard deviations of the predictions $f^{RLB}_{\mathrm{D},\lambda,B}(x)$ using the central limit theorem (see Theorem A.4.10) or by applying versions of the law of the iterated logarithm (see, e.g., Einmahl and Li, 2008). However, in general we favor an alternative distribution-free method based on the median. If B is small or if it is unknown whether $f^{RLB}_{\mathrm{D},\lambda,B}(x)$ has a finite variance, one can construct distribution-free confidence intervals for the median of $f^{RLB}_{\mathrm{D},\lambda,B}(x)$ based on special order statistics, as the following well-known result shows.

Theorem 10.32. *Let $B \in \mathbb{N}$, $0 \le r < s \le B$, and $\tau \in (0,1)$. Let $Z_1, \ldots, Z_B$ be independent and identically distributed real-valued random variables and denote the τ-quantile by $q_\tau := \inf\{z \in \mathbb{R} : \mathrm{P}(Z_1 \le z) \ge \tau\}$. Then the corresponding order statistics $Z_{(1:B)} \le \ldots \le Z_{(B:B)}$ satisfy*

$$\mathrm{P}(Z_{(r:B)} \le q_\tau \le Z_{(s:B)}) \ge \sum_{i=r}^{s-1} \binom{B}{i} \tau^i (1-\tau)^{B-i} \ge \mathrm{P}(Z_{(r:B)} < q_\tau < Z_{(s:B)})\,.$$

Proof. Let $b \in \{0, \ldots, B\}$, $z \in \mathbb{R}$, and define $p_z := \mathrm{P}(Z_1 \le z)$. We have

$$\mathrm{P}(Z_{(b:B)} \le z) = \mathrm{P}\Big(\sum_{i=1}^{B} \mathbf{1}_{(-\infty,z]}(Z_i) \ge b\Big) = \sum_{i=b}^{B} \binom{B}{i} p_z^i (1-p_z)^{B-i}\,.$$

Recall that the incomplete beta function is given by

$$I_\tau(a,c) := \int_0^\tau t^{a-1}(1-t)^{c-1}\,dt \,\Big/ \int_0^1 t^{a-1}(1-t)^{c-1}\,dt, \quad a,c \in [0,\infty).$$

Using partial integration we obtain for $b \in \{0, \ldots, B\}$ that $I_\tau(b, B-b+1) = \sum_{i=b}^{B} \binom{B}{i} \tau^i (1-\tau)^{B-i}$. Using $\mathrm{P}(Z_{(r:B)} \le q_\tau \le Z_{(s:B)}) = \mathrm{P}(Z_{(r:B)} \le q_\tau) - \mathrm{P}(Z_{(s:B)} < q_\tau)$ and the definition of q_τ, we obtain

$$\mathrm{P}(Z_{(r:B)} \le q_\tau \le Z_{(s:B)}) \ge I_\tau(r, B-r+1) - I_\tau(s, n-s+1)\,.$$

The second inequality follows by similar arguments. □

Table 10.3 lists some values of B and the corresponding pair (r, s) of order statistics determining the confidence interval $[f_{S_{(r:B)},\lambda}(x), f_{S_{(s:B)},\lambda}(x)]$. The lower bound of the actual confidence level which is $0.5^B \sum_{j=r}^{s} \binom{B}{j}$ is also listed. The actual level of the confidence intervals can differ from $1 - \alpha$ for small values of B, see Table 10.3. The last column in Table 10.3 lists the value of $\min\{r - 1, B - s\}/B$, which is the finite-sample breakdown point for the distribution-free confidence interval for the median. For example, if $B = 17$, the fifth and the thirteenth order statistics yield a distribution-free confidence interval at the 95 percent level for the median without any further distributional assumption. Because the results of the four lowest and the four highest predictions are not considered, the breakdown point of this confidence interval is $4/17 = 0.235$.

Table 10.3. Selected pairs (r, s) of order statistics for non-parametric confidence intervals for the median.

Confidence Level $1 - \alpha$	B	r	s	Lower Bound of Actual Confidence Level	$\min\{r - 1, B - s\}/B$
0.90	8	2	7	0.930	0.125
	10	2	9	0.979	0.100
	18	6	13	0.904	0.278
	30	11	20	0.901	0.333
	53	21	33	0.902	0.377
	71	29	43	0.904	0.394
	104	44	61	0.905	0.413
0.95	9	2	8	0.961	0.111
	10	2	9	0.979	0.100
	17	5	13	0.951	0.235
	37	13	25	0.953	0.324
	51	19	33	0.951	0.353
	74	29	46	0.953	0.378
	101	41	61	0.954	0.396
0.99	10	1	10	0.998	0.000
	12	2	11	0.994	0.083
	26	7	20	0.991	0.231
	39	12	28	0.991	0.282
	49	16	34	0.991	0.306
	73	26	48	0.990	0.342
	101	38	64	0.991	0.366

If the robust estimator is based on hyperparameters (e.g., kernel parameters or the constant ϵ for SVMs based on the ϵ-insensitive loss) and if their values must be determined from the data set itself, a common approach is to split huge data sets into three parts for training, validation, and testing.

The training data set is used to estimate the quantity of interest for a given set of hyperparameters. The validation data set is used to determine good values for the hyperparameters by optimizing an appropriate goodness-of-fit criterion or by minimizing the generalization error. Finally, the test data set is used to estimate the goodness-of-fit criterion or the generalization error for new data points.

General Properties of RLB

The estimators $f_{\mathrm{S}_b,\lambda}$ from the B bites are stochastically independent because they are computed from disjoint parts of the data set. Denote the number of available CPUs by c and let k_B be the smallest integer that is not smaller than B/c. The computation time and the memory space for RLB can be obviously approximated in the following way.

i) Computation time, c CPUs. Assume that the computation time of the estimator $f_{\mathrm{D},\lambda}$ for a data set with n observations and d explanatory variables is of order $O(g(n,d))$, where g is some positive function. Then the computation time of RLB with B bites of subsample size $n_b \approx n/B$ is approximately of order $O(k_B g(n/B,d))$.

ii) Memory space, c CPUs. Assume that the estimator $f_{\mathrm{D},\lambda}$ for a data set with n observations and d explanatory variables needs memory space and hard disk space of order $O(g_1(n,d))$ and $O(g_2(n,d))$, respectively, where g_1 and g_2 are positive functions. Then the computation of RLB with B bites of subsample size $n_b \approx n/B$ needs memory space and hard disk space approximately of order $O(cg_1(n/B,d))$ and $O(cg_2(n/B,d))$, respectively.

The next result shows that RLB estimators inherit the usual consistency properties from the original estimators.

Lemma 10.33 (Convergence). *Consider an RLB estimator $f^{RLB,I}_{\mathrm{D},\lambda,B}$ of type I based on a convex combination with $a_{B,b} \in [0,1]$ and $\sum_{b=1}^{B} a_{B,b} = 1$.*

i) If $\mathbb{E}_{\mathrm{P}}(f_{\mathrm{S}_b,\lambda}) = \mathbb{E}_{\mathrm{P}}(f_{\mathrm{D}_n,\lambda})$, $b \in \{1,\ldots,B\}$, then $\mathbb{E}_{\mathrm{P}}(f^{RLB,I}_{\mathrm{D},\lambda,B}) = \mathbb{E}_{\mathrm{P}}(f_{\mathrm{D}_n,\lambda})$.

ii) Assume that $f_{\mathrm{D}_n,\lambda}$ converges in probability (or almost surely) to $f_{\mathrm{P},\lambda}$ if $n \to \infty$, B is fixed, and $\min_{1\le b\le B} n_b \to \infty$. If g is continuous, then $f^{RLB,I}_{\mathrm{D}_n,\lambda,B}$ converges in probability (or almost surely) to $f_{\mathrm{P},\lambda}$ if $n \to \infty$.

iii) Assume that $n_b^{1/2}(f_{\mathrm{S}_b,\lambda} - f_{\mathrm{P},\lambda})$ converges in distribution to a multivariate Gaussian distribution $N(0,\Sigma)$ if $\min_{1\le b\le B} n_b \to \infty$, where $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite, and that B is fixed. If g is continuous, then $n^{1/2}(f^{RLB,I}_{\mathrm{D}_n,\lambda,B} - f_{\mathrm{P},\lambda})$ converges in distribution to a multivariate Gaussian distribution $N(0,\Sigma)$, $n \to \infty$.

Proof. i) follows from the linearity of the expectation operator. The parts *ii)* and *iii)* follow immediately from the continuity of g. □

The next result shows that an RLB estimator based on SVMs is a kernel-based estimator, too.

Theorem 10.34. *Consider the estimators* $f_{\mathrm{D}_n,\lambda}(x) = \sum_{i=1}^n \alpha_i k(x,x_i)$ *and* $f_{\mathrm{S}_b,\lambda}(x) = \sum_{(x_i,y_i)\in S_b} \alpha_{i,b} k(x,x_i)$, $x \in X$, $b = 1,\ldots,B$. *Then the RLB estimator* $f^{RLB,II}_{\mathrm{D}_n,\lambda,B}(x)$ *with weights* $a_{B,b} \in [0,1]$, $\sum_{b=1}^B a_{B,b} = 1$, *and* B *fixed, is a kernel-based estimator and satisfies*

$$f^{RLB,II}_{\mathrm{D}_n,\lambda,B}(x) = \sum_{i=1}^n \alpha_{i,RLB}\, k(x,x_i) \tag{10.52}$$

$$= \sum_{i \in SV(S_1)\cup\ldots\cup SV(S_B)} \alpha_{i,RLB}\, k(x,x_i), \qquad x \in X, \tag{10.53}$$

where $SV(S_b)$ *denotes the indexes of support vectors in* $f_{\mathrm{S}_b,\lambda}$ *and* $\alpha_{i,RLB} = \sum_{b=1}^B a_{B,b}\,\alpha_{i,b}$, $(x_i,y_i) \in D_n$.

Proof. By assumption, each bite S_b is fitted with an SVM having the decision function

$$f_{\mathrm{S}_b,\lambda}(x) = \sum_{i\in S_b} \alpha_{i,b} k(x,x_i), \qquad x \in X.$$

Because the bites S_b, $b = 1,\ldots,B$, are disjoint and B is fixed, the RLB estimator with weights $a_{B,b}$ has the representation

$$f^{RLB,II}_{\mathrm{D}_n,\lambda,B}(x) = \sum_{b=1}^B a_{B,b} \sum_{i\in S_b} \alpha_{i,b}\, k(x,x_i) \tag{10.54}$$

$$= \sum_{i=1}^n \sum_{b=1}^B a_{B,b}\,\alpha_{i,b}\, k(x,x_i)\,, \quad x \in X, \tag{10.55}$$

which gives the assertion. □

If all support vectors in $S_1,\ldots,S_B$ are different, we have $\alpha_{i,RLB} = a_{B,b}\,\alpha_{i,b}$ in (10.53). Now, we investigate the number of support vectors of RLB estimators based on SVMs.

Theorem 10.35. *Under the assumptions of Theorem 10.34, an RLB estimator* $f^{RLB,I}_{\mathrm{D}_n,\lambda,B}$ *has the following properties.*

i) The number of support vectors (i.e., $\alpha_{i,RLB} \neq 0$*) of* $f^{RLB,I}_{\mathrm{D}_n,\lambda,B}$ *is given by*

$$\#SV(f^{RLB,I}_{\mathrm{D}_n,\lambda,B}) = \Big| \bigcup_{b\in\{1,\ldots,B:\, a_{B,b}>0\}} SV(S_b) \Big|. \tag{10.56}$$

ii) Let $Y = \{-1,+1\}$, $B \in \mathbb{N}$ *be fixed,* $\min\{n_1,\ldots,n_B\} \to \infty$, *and consider weights* $a_{B,b} \in (0,1)$ *with* $\sum_{b=1}^{B} a_{B,b} = 1$. *Under the assumptions of Theorem 8.34, we have the probabilistic lower bound on the sparseness:*

$$\mathrm{P}^n\left(D_n \in (X\times Y)^n : \#SV(f^{RLB,I}_{\mathrm{D}_n,\lambda,B}) \geq \frac{\mathcal{S}_{L,\mathrm{P}}-\varepsilon}{n}\right) \geq \prod_{b=1}^{B}\left(1-e^{-\frac{\delta^2 n_b}{18h^2(\lambda)B}}\right). \tag{10.57}$$

Proof. i) follows immediately from (10.53). Now let us consider *ii)*. Using the probabilistic lower bound on the sparseness given in Theorem 8.34, we see that $f_{\mathrm{D}_n,\lambda}$ satisfies

$$\mathrm{P}^n\Big(D_n \in (X\times Y)^n : \#SV(f_{\mathrm{D}_n,\lambda}) \geq (\mathcal{S}_{L,\mathrm{P}}-\varepsilon)n\Big) \;\geq\; 1-e^{-\frac{\delta^2 n}{18h^2(\lambda)}}\,. \tag{10.58}$$

The bites S_b, $b=1,\ldots,B$, are independent and identically distributed by the construction of RLB and we have $n = Bn_b$. Hence

$$\begin{aligned}
&\mathrm{P}^n\left(D_n \in (X\times Y)^n : \#\mathrm{SV}\big(f^{RLB,I}_{\mathrm{D}_n,\lambda,B}\big) \geq (\mathcal{S}_{L,\mathrm{P}}-\varepsilon)n\right)\\
&= \mathrm{P}^n\Big((S_1,\ldots,S_B) \in (X\times Y)^n : \#\mathrm{SV}\big(f^{RLB,I}_{\mathrm{D}_n,\lambda,B}\big) \geq \sum_{b=1}^{B}(\mathcal{S}_{L,\mathrm{P}}-\varepsilon)n_b\Big)\\
&\geq \mathrm{P}^n\left(\forall\, S_b \in (X\times Y)^{n_b}, b=1,\ldots,B \,:\, \#\mathrm{SV}\big(f_{\mathrm{S}_b,\lambda_{n_b}}\big) \geq (\mathcal{S}_{L,\mathrm{P}}-\varepsilon)n_b\right)\\
&= \prod_{b=1}^{B}\mathrm{P}^{n_b}\left(S_b \in (X\times Y)^{n_b} : \#\mathrm{SV}\big(f_{\mathrm{S}_b,\lambda_{n_b}}\big) \geq (\mathcal{S}_{L,\mathrm{P}}-\varepsilon)n_b\right)\\
&\geq \prod_{b=1}^{B}\left(1-e^{-\frac{\delta^2 n_b}{18h^2(\lambda)B}}\right) \to 1,\ n\to\infty\,,
\end{aligned}$$

which completes the proof. □

The result given in (10.57) has the following interpretation: with probability exponentially fast tending to one if the total sample size $n = Bn_b$ converges to ∞ but B is fixed, the fraction of support vectors of the kernel based RLB estimator $f^{RLB,I}_{\mathrm{D}_n,\lambda,B}$ in a binary classification problem is essentially greater than the average of the Bayes risks for the bites. If we compare the rates of convergence in (10.57) and (10.58), we obtain the expected result that the probabilistic lower bound in both results is the same, but that the rate of convergence is better in (10.58) than in its counterpart in (10.57) for $B > 1$. This can be interpreted as the price we have to pay if we use RLB because of its computational advantages for huge data sets instead of $f_{\mathrm{D}_n,\lambda}$.

Let us now investigate conditions to guarantee that RLB estimators using SVMs are L-risk consistent; i.e.,

$$\mathcal{R}_{L,\mathrm{P}}\big(f^{RLB,I}_{\mathrm{D}_n,\lambda,B}\big) \to \mathcal{R}_{L,\mathrm{P}}$$

in probability for $n \to \infty$.

Theorem 10.36. *Let $f_{\mathrm{D},\lambda,b}$ be an L-risk consistent support vector machine with a convex loss function. Consider an RLB estimator $f_{\mathrm{D},\lambda,B}^{RLB,I}$ with weights $a_{B,b} \in (0,1)$, $\sum_{b=1}^{B} a_{B,b} = 1$, $B \geq 1$ fixed, and $\min_{1\leq b\leq B} n_b \to \infty$. Then $f_{\mathrm{D},\lambda,B}^{RLB,I}$ is L-risk consistent.*

Proof. The RLB estimator $f_{\mathrm{D},\lambda,B}^{RLB,I}$ is a convex combination of $f_{\mathrm{S}_b,\lambda}$, $b = 1,\ldots,B$, because $a_{B,b} \in (0,1)$ and $\sum_{b=1}^{B} a_{B,b} = 1$. Therefore,

$$
\begin{aligned}
0 &\leq \int L\bigl(Y, f_{\mathrm{D},\lambda,B}^{RLB,I}(X)\bigr)\, d\mathrm{P} - \mathcal{R}_{L,\mathrm{P}}^{*} \\
&= \int L\Bigl(Y, \sum_{b=1}^{B} a_{B,b} f_{\mathrm{S}_b,\lambda}(X)\Bigr)\, d\mathrm{P} - \mathcal{R}_{L,\mathrm{P}}^{*} \\
&\leq \int \sum_{b=1}^{B} a_{B,b}\, L\bigl(Y, f_{\mathrm{S}_b,\lambda}(X)\bigr)\, d\mathrm{P} - \mathcal{R}_{L,\mathrm{P}}^{*} \qquad (10.59) \\
&= \sum_{b=1}^{B} a_{B,b} \Bigl(\int L\bigl(Y, f_{\mathrm{S}_b,\lambda}(X)\bigr)\, d\mathrm{P} - \mathcal{R}_{L,\mathrm{P}}^{*} \Bigr), \qquad (10.60)
\end{aligned}
$$

which converges in probability to zero if $\min_{1\leq b\leq B} n_b \to \infty$, $n \to \infty$. Here we used the convexity of L in (10.59) and the L-risk consistency of $f_{\mathrm{D}_n,\lambda}$ in (10.60). □

From the no-free-lunch theorem (see Theorem 6.6), the proof given above cannot be modified in a simple way to cover the case where the number of bites $B = B(n)$ depends on the sample size because we have no uniform rate of consistency without restricting the class of probability measures.

Robustness Properties of RLB

Now we derive some results that show that certain robustness properties are inherited from the original estimator $f_{\mathrm{D}_n,\lambda}$ to the RLB estimator. Let us start with an investigation of the influence function of the RLB estimator $f_{\mathrm{P},\lambda,B}^{RLB,I}$.

Theorem 10.37 (Influence function of RLB). *Assume that the influence function of $f_{\mathrm{P},\lambda}$ exists for the distribution $\mathrm{P} \in \mathcal{M}_1$. Further assume that the weights satisfy $a_{B,b} \in [0,1]$ with $\sum_{b=1}^{B} a_{B,b} = 1$ and B fixed. Then the influence function of $f_{\mathrm{P},\lambda,B}^{RLB,I}$ using these weights exists and equals the influence function of $f_{\mathrm{P},\lambda}$.*

Proof. Let $z \in X \times Y$ and define $S_1(\mathrm{P}) := f_{\mathrm{P},\lambda,B}^{RLB,I}$ and $S_2(\mathrm{P}) := f_{\mathrm{P},\lambda}$, $\mathrm{P} \in \mathcal{M}_1$. It follows that

$$\begin{aligned}
\mathrm{IF}(z; S_1, \mathrm{P}) &= \lim_{\varepsilon \downarrow 0} \frac{f^{RLB,I}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda,B} - f^{RLB,I}_{\mathrm{P},\lambda,B}}{\varepsilon} \\
&= \lim_{\varepsilon \downarrow 0} \frac{\sum_{b=1}^{B} a_{B,b}\, f_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda} - \sum_{b=1}^{B} a_{B,b}\, f_{\mathrm{P},\lambda}}{\varepsilon} \\
&= \sum_{b=1}^{B} a_{B,b} \lim_{\varepsilon \downarrow 0} \frac{f_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda} - f_{\mathrm{P},\lambda}}{\varepsilon} \\
&= \mathrm{IF}(z; S_2, \mathrm{P})\,,
\end{aligned}$$

which gives the assertion. □

Hence, if $f_{\mathrm{P},\lambda}$ has a bounded influence function, the same is true for an RLB estimator of type I using a convex combination of weights as specified above. Recall that the existence and boundedness of the influence function of $f_{\mathrm{P},\lambda}$ were proven in Sections 10.3 and 10.4 under weak conditions on the loss function and on the kernel for classification problems and regression problems.

We have not considered breakdown points for $f_{\mathrm{P},\lambda}$ or $f_{\mathrm{D},\lambda}$. However, if we specify $\lambda = 0$ in combination with a linear kernel, then SVMs are just M-estimators in linear regression for which the breakdown points of M-estimators are well-known.

Theorem 10.38 (Finite-sample breakdown point of RLB). *Let $n_b \equiv n/B$, $b = 1, \ldots, B$. Assume that the finite-sample breakdown points $\varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda})$ of the estimators $f_{\mathrm{S}_b,\lambda}$ are all identical. Denote the finite-sample breakdown point of the estimator $\hat{\mu} = \hat{\mu}(f_{S_1,\lambda}, \ldots, f_{S_B,\lambda})$ in the aggregation step by $\varepsilon^*_B(\hat{\mu})$. Then the finite-sample breakdown point of an RLB estimator is given by*

$$\varepsilon^*_{RLB,n,B} = \varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda})\Big(\varepsilon^*_B(\hat{\mu}) + \frac{1}{B}\Big) + \frac{B}{n}\varepsilon^*_B(\hat{\mu})\,. \tag{10.61}$$

Proof. The minimum number of points needed to modify $f_{\mathrm{S}_b,\lambda}$ in bite S_b such that a breakdown occurs is given by $n_b\varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda}) + 1$, $b = 1, \ldots, B$. The RLB estimator breaks down if at least $B\varepsilon^*_B(\hat{\mu})+1$ of the estimators $f_{\mathrm{S}_1,\lambda}$, …, $f_{\mathrm{S}_B,\lambda}$ break down. This gives the assertion. □

Remark 10.39.

i) If B and $n_b = B/n$ are large, we obtain from (10.61) the lower bound

$$\varepsilon^*_{RLB,n,B} \geq \varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda})\,\varepsilon^*_B(\hat{\mu})\,. \tag{10.62}$$

ii) For the median, we obtain $\varepsilon^*_B(\hat{\mu}) = \frac{1}{2} - \frac{1}{B}$ if B is even and $\varepsilon^*_B(\hat{\mu}) = \frac{1}{2}$ if B is odd.

iii) For α-trimmed means, $\alpha \in (0, \frac{1}{2})$, we obtain $\varepsilon^*_B(\hat{\mu}) = \lfloor \alpha B \rfloor / B$.

iv) If the mean or any other estimator with $\varepsilon^*_B(\hat{\mu}) = 0$ is used in the aggregation step, RLB has a finite-sample breakdown point of $\varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda})/B \to 0$ if $B \to \infty$. ◁

Example 10.40 (Univariate location model). Consider the univariate location problem, where $x_i \equiv 1$ and $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, $n = 55$. Assume that all output values are different. The finite-sample breakdown point of the median is $\lfloor n/2 \rfloor / n = 0.49$. The mean has a finite-sample breakdown point of 0. Now let us investigate the robustness of the RLB approach with $B = 5$ and $n_b = 11$, $b = 1, \ldots, B$. *(i)* If the median is used as the location estimator in each bite and if the median is used in the aggregation step, the finite-sample breakdown point of the RLB estimator is $\varepsilon^*_{RLB,n,B} = 0.309$. This value is reasonably high but lower than the finite-sample breakdown point of the median for the whole data set, which is 0.49. Note that in a *fortunate* situation the impact of up to $(2 \cdot 11 + 5 \cdot 3)/55 = 0.672$ extremely large data points (say equal to $+\infty$) is still bounded for the RLB estimator in this setup: modify all data points in $B\varepsilon^*_B(\hat{\mu}) = 2$ bites and up to $n_b\, \varepsilon^*_{n_b}(f_{\mathrm{S}_b,\lambda}) = 5$ data points in the remaining $B(1 - \varepsilon^*_B(\hat{\mu})) = 3$ bites. This is no contradiction to (10.61) because the breakdown point measures the *worst-case* behavior. *(ii)* If the median is used as the location estimator in each bite and if the mean is used in the aggregation step, the finite-sample breakdown point of the RLB estimator is $\varepsilon^*_{RLB,n,B} = (1/B)\varepsilon^*_{n_b}(\hat{f}) = 0.09$. *(iii)* If the mean is used as the location estimator in each bite and also in the aggregation step, we of course obtain $\varepsilon^*_{RLB,n,B} = 0$. ◁

Determination of the Number of Bites

The number of bites obviously has some impact on the statistical behavior of the RLB estimator and also on the computation time and the necessary computer memory. An optimal choice of the number of bites B will in general depend on the unknown distribution P. But some general arguments are given on how to determine B in an appropriate manner.

One should take the sample size n, the computer resources (number of CPUs, RAM, hard disk), and the acceptable computation time into account. The quantity B should be much lower than n because otherwise there is not much hope of obtaining useful estimators from the bites. Further, B should depend on the dimensionality d of the input values. For example, a rule of thumb for *linear* regression is that n/d should be at least 5. Because the function f is completely unknown in non-parametric regression, assumptions on the complexity of f are crucial. The sample size n_b for each bite should converge to infinity if $n \to \infty$ to obtain consistency of RLB. The results from some numerical experiments not given here can be summarized as follows.

i) If B is too large, the computational overhead increases and the danger of bad fits increases because n_b is too small to provide reasonable estimators.

ii) A major decrease in computation time and memory saving is often already present if B is chosen in a way such that the numerical algorithms to fit each bite fit nicely into the computer (CPU, RAM, hard disk). Nowadays, robust estimators can often be computed for sample sizes up to $n_b = 10^4$ or $n_b = 10^5$. In this case, $B = \lfloor n/n_b \rfloor$ can be a reasonable choice.

iii) If distribution-free confidence intervals at the $(1-\alpha)$ level for the median of the predictions (i.e., $f^{RLB,II}_{\mathrm{D}_n,\lambda,B}(x) = \mathrm{median}_{1\le b\le B}\, f_{\mathrm{S}_b,\lambda}(x)$, $x \in X$) are needed, one should take into account that the actual confidence level of such confidence intervals based on order statistics can be conservative (i.e., higher than the specified level) for some pairs (r, s) of order statistics due to the discreteness of order statistics.

Numerical Example

Now we apply the RLB approach to kernel logistic regression (i.e., $L = L_{\text{c-logist}}$). All computations were done with the program `myKLR` (Rüping, 2003), which is an implementation of the algorithm proposed by Keerthi *et al.* (2005) to solve the dual problem. We choose KLR for two reasons. First, the computation of KLR needs much more time than SVMs based on the hinge loss because the latter solves a quadratic instead of a convex program in dual space and because for the hinge loss, the number of support vectors is usually much smaller than n. Therefore, the need for computational improvements is greater for $L_{\text{c-logist}}$ than for L_{hinge} and the potential gains of RLB can be more important. Second, the number of support vectors of KLR is approximately equal to n, which slows down the computation of predictions.

The simulated data sets contain n data points $(x_i, y_i) \in \mathbb{R}^8 \times \{-1, +1\}$ simulated in the following way. All eight components of $x_i = (x_{i,1}, \ldots, x_{i,8})$ are simulated independently from a uniform distribution on $(0, 1)$. The responses y_i are simulated independently from a logistic regression model according to $\mathrm{P}(Y_i = +1|X_i = x_i) = 1/\left(1 + e^{-f(x_i)}\right)$ and $\mathrm{P}(Y_i = -1|X_i = x_i) = 1 - \mathrm{P}(Y_i = +1|X_i = x_i)$. We set

$$f(x_i) = \sum_{j=1}^{8} x_{i,j} - x_{i,1}x_{i,2} - x_{i,2}x_{i,3} - x_{i,4}x_{i,5} - x_{i,1}x_{i,6}x_{i,7}\,.$$

The data points were saved as ASCII files, where $x_{i,j}$ is stored with four decimal places. The numerical results of fitting kernel logistic regression to such data sets are given in Table 10.4. It is obvious that in this situation RLB can save a lot of computation time. If the whole data set has $n = 10^5$ observations, approximately 10 hours is needed to compute KLR on a PC. If RLB with $B = 10$ bits are used, each with a subsample size of $n_b = 10^4$, one needs approximately 16 minutes if there is 1 GB of kernel cache available. This is a reduction by a factor of 38. If there are five CPUs available and each processor can use up to 200 MB kernel cache, RLB with $B = 10$ will need approximately 11 minutes, which is a reduction by a factor of 55.

When RLB was applied to a data set from a union of German insurance companies ($n \approx 4.5 \times 10^6$), the computation time decreased from months to days, see Christmann (2005) and Christmann *et al.* (2007) for details.

Table 10.4. Computation times for kernel logistic regression using `myKLR`.

Sample Size n	CPU Time	Used Cache in MB	Available Cache in MB
2000	4 sec	33	200
5000	25 sec	198	200
10000	5 min, 21 sec	200	200
10000	1 min, 33 sec	787	1000
20000	24 min, 11 sec	1000	1000
20000	14 min, 35 sec	1000	1000
100000	9 h, 56 min, 46 sec	1000	1000

10.6 Further Reading and Advanced Topics

Many different criteria have been proposed to define robustness or stability in a mathematical way, such as the minimax approach (Huber, 1964), qualitative robustness (Hampel, 1968, 1971), the sensitivity curve (Tukey, 1977), the approach based on least favorable local alternatives (Huber, 1981; Rieder, 1994), the approach based on influence functions (Hampel, 1974; Hampel *et al.*, 1986), the maxbias curve (Hampel *et al.*, 1986; Huber, 1964), the global concept of min-max bias robust estimation (He and Simpson, 1993; Martin *et al.*, 1989), the breakdown point (Donoho and Huber, 1983; Hampel, 1968), depth-based methods (Tukey, 1975), and configural polysampling (Morgenthaler and Tukey, 1991).

Section 10.2 described the main approaches of robust statistics. It is mainly based on Huber (1981), Hampel *et al.* (1986), Rousseeuw and Leroy (1987), Jurečková and Sen (1996), Davies and Gather (2004), and Maronna *et al.* (2006). For qualitative robustness, we refer also to Hampel (1968, 1971) and Cuevas (1988). The robustness results of SVMs for classification and regression treated in Sections 10.3 and 10.4 are based on Christmann (2002, 2004, 2005) and Christmann and Steinwart (2004, 2007). The robust learning from bites approach discussed in Section 10.5 was proposed by Christmann *et al.* (2007). Using Bouligand derivatives instead of Gâteaux derivatives, it was shown that SVMs based on Lipschitz-continuous loss functions have under weak assumptions a bounded Bouligand influence function provided that a bounded kernel is used; see Christmann and Van Messem (2008). Special cases are SVMs based on the ϵ-insensitive loss, Huber's loss, logistic loss for regression, and pinball loss for quantile regression. Debruyne *et al.* (2007) gave robustness results for a reweighted form of the SVM based on the least squares loss and Debruyne (2007) and proposed tools for model selection.

For a detailed description of outliers and approaches to detect, identify or test for outliers, we refer to Beckmann and Cook (1983), Barnett and Lewis (1994), Gather (1984, 1990), Davies and Gather (1993), Becker and Gather (1999), Gather and Pawlitschko (2004), Hampel *et al.* (1986),

and Christmann (1992). Steinwart *et al.* (2005) and Scott and Nowak (2006), among many others, used SVMs for anomaly detection and for learning minimum volume sets that are related to outlier regions proposed by Davies and Gather (1993).

Choice of a Metric

Neighborhoods around a probability distribution play an important role in robust statistics. Such neighborhoods are usually constructed by specifying a metric on the space of probability measures. However, the robustness properties of a statistical procedure can depend on the metric. The choice of a suitable metric is not always a simple task; see Hampel (1968), Huber (1981, Chapter 2), and Davies (1993, pp. 1851ff.) for nice treatments of this topic. For a recent controversial discussion of the question of which metrics are especially suitable for robust statistics, see Davies and Gather (2005, 2006) and Hampel (2005).

Sometimes one restricts attention to investigating robustness properties of statistical methods only in gross-error neighborhoods because such problems are often easier to solve. At first glance-neighborhoods defined via the norm of total variation seem to be an attractive alternative to gross-error neighborhoods. However, Hampel (1968, p. 43) showed the interesting fact that neighborhoods defined by the norm of total variation do not perform much better than gross-error neighborhoods. Consider two probability measures P and Q defined on the same measurable space, and let $\varepsilon \in (0,1)$. Then Q is an element of a neighborhood of radius ε around P with respect to the metric defined by the norm of total variation if and only if $\mathrm{Q} = (1-\varepsilon)\mathrm{P}+\varepsilon(\mathrm{P}+\mathrm{Q}_1-\mathrm{Q}_2)$, where $\mathrm{Q}_1 := \varepsilon^{-1}(\mathrm{Q} - \min\{\mathrm{P},\mathrm{Q}\})$ and $\mathrm{Q}_2 := \varepsilon^{-1}(\mathrm{P} - \min\{\mathrm{P},\mathrm{Q}\})$. Note that Q_1 and Q_2 are not necessarily probability measures but Q is a mixture of P and $\mathrm{P} + \mathrm{Q}_1 - \mathrm{Q}_2$ with mixing proportion ε.

Qualitative Robustness

Hampel (1968, 1971) proposed not only qualitative robustness but also the notion of Π-robustness. Informally, a sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is called Π-robust if it is qualitatively robust but also insensitive to "small" deviations from the assumption of independence. Hampel (1971, Theorem 3) proved that Π-robustness at some distribution P implies qualitative robustness but not vice versa. Boente *et al.* (1987) generalized Hampel's concept of Π-robustness for a sequence of estimators $(S_n)_{n\in\mathbb{N}}$ in Euclidean spaces to Polish spaces. Cuevas (1988) showed that qualitative robustness is incompatible with consistency of multivariate density estimates if the L_1-metric is used. Cuevas also investigated the case of qualitative robustness for certain stochastic processes with continuous trajectories on $[0,1]$ and proved that the simple mean function is not robust in this sense, but robust estimation is possible by generalizing M-estimators and α-trimmed means.

Robustness of SVMs

The boundedness of the sensitivity curve of SVMs for classification was essentially already established by Bousquet and Elisseeff (2002). The results given in Sections 10.3 and 10.4 and also some results for the more general case, where not only $f_{P,\lambda}$ but also an offset term $b_{P,\lambda}$ must be estimated, were derived by Christmann and Steinwart (2004, 2007, 2008). If attention is restricted to SVMs based on a Lipschitz-continuous loss function for regression and a bounded kernel, Christmann and Van Messem (2008) showed that $f_{P,\lambda}$ has even a bounded Bouligand influence function.

Robust Learning from Bites

The RLB approach proposed by Christmann (2005) and Christmann *et al.* (2007) has connections to the remedian proposed by Rousseeuw and Bassett (1990) for univariate location estimation, Rvote proposed by Breiman (1999a), DRvote with classification trees using majority voting (Chawla *et al.*, 2004), and subsampling (Politis *et al.*, 1999). Bootstrapping computer-intensive robust methods for huge data sets is often impossible due to computation time and memory limitations of the computer, but see Salibian-Barrera *et al.* (2008). RLB has some similarity to the algorithms FAST-LTS and FAST-MCD proposed by Rousseeuw and Van Driessen (1999, 2000) for robust estimation in linear regression or multivariate location and scatter models for large data sets. FAST-LTS and FAST-MCD split the data set into subsamples, optimize the objective function in each subsample, and use these solutions as starting values to optimize the objective function for the whole data set. This is in contrast to RLB, which aggregates estimation results from the bites to obtain robust confidence intervals.

Stability and Learnability

There is currently some interest in proving connections between stability, learnability, and predictivity for broad classes of ERM methods. We would like to cite Bousquet and Elisseeff (2002), Poggio *et al.* (2004), Mukherjee *et al.* (2006), and Elisseeff *et al.* (2005). Although different notions of stability are used in these papers, their notions of stability have a meaning similar to robustness. Several of these notions of stability only measure the impact of just *one* data point such that the connection to Tukey's sensitivity curve is obvious. Caponnetto and Rakhlin (2006) consider stability properties of empirical risk minimization over Donsker classes.

Robustness in Parametric and Non-parametric Models

There is a large body of literature on robust estimators and tests for outliers in *parametric* models, especially for linear regression, binary regression, and

multivariate location and scatter problems. From the viewpoint of robustness, the relationships between SVMs based on a convex loss function and classical M-estimation, say of Huber type, are strong. However, there are three important differences between the two approaches. First, classical M-estimation for linear regression or binary classification has no penalty term in its optimization problem (i.e., $\lambda = 0$). Second, SVMs often use non-linear kernels such as the Gaussian RBF kernel, which already can improve robustness properties (see, e.g., Theorems 10.10 and 10.18) whereas classical M-estimation uses only the unbounded linear kernel. Third, non-convex loss functions are sometimes used in classical M-estimation in linear models to improve robustness properties of such methods, although one is then often faced with the problem of non-uniqueness of the estimates and with algorithmic problems to find a global optimum and not only a local one.

Downweighting extreme points in the design space X is automatically done by SVMs using, for example, the Gaussian RBF kernel. Generalized M-estimators for regression are able to downweight not only data points extreme in the residual $y - f(x)$ but also extreme points x, so-called *leverage points*. For additional information on M-estimation and related methods in linear models, we refer to Huber (1964, 1981), Hampel *et al.* (1986), Fernholz and Morgenthaler (2005), Gather and Hilker (1997), Kent and Tyler (1991, 1996, 2001), and Maronna and Yohai (1981). Mendes and Tyler (1996) investigated constrained M-estimation for linear regression. Simpson *et al.* (1987) gave theoretical arguments in favor of M-estimators based on a twice Fréchet differentiable loss function. Rieder (1994) and Ruckdeschel and Rieder (2004) investigated robust asymptotic statistics based on shrinking neighborhoods. Rieder *et al.* (2008) proposed the radius-minimaxity concept.

Many robust alternatives to M-estimators were proposed in the literature. We did not describe such methods here because the connection between SVMs and M-estimation is much closer. However, we would like to mention the following alternatives to M-estimation in parametric models.

For L-estimators, which are linear combinations of order statistics, we refer to Serfling (1980) and Huber (1981). Regression quantiles are generalizations of L-estimators to the regression context and are treated, for example, by Koenker and Bassett (1978), Portnoy and Koenker (1997), and Koenker (2005). Schölkopf *et al.* (2000, p. 1216) and Takeuchi *et al.* (2006) investigated SVMs based on the Lipschitz-continuous pinball loss function to estimate quantile functions in a nonparametric way. Some results on L-risk consistency, rates of convergence, and various robustness properties of such SVMs were derived by Christmann and Steinwart (2007, 2008) and Steinwart and Christmann (2008).

Statistical properties of R-estimators were investigated, for example, by Serfling (1980), Coakley and Hettmansperger (1993), Hettmansperger *et al.* (1997), Ollila *et al.* (2002, 2003), and Hallin and Paindaveine (2004, 2005).

Of special interest in robust estimation in parametric models are methods with a positive or high-breakdown point to overcome the lack of a low

breakdown point of M-, L-, and R-estimators in high dimensions d. Probably the most important and most often used high-breakdown estimators for linear regression are the estimators least median of squares (LMS) and least trimmed squares (LTS), which were investigated in detail by Rousseeuw (1984, 1994, 1997a,b) and Rousseeuw and Leroy (1987). Both estimators were also used in many recent proposals to construct two-step estimators, which inherit from the starting estimators LMS or LTS a high-breakdown point and from the second step improved asymptotic behavior. LMS and LTS belong to the class of S-estimators and their generalizations. These estimators were investigated by Rousseeuw and Yohai (1984), Rousseeuw and Van Driessen (2000), Rousseeuw and van Zomeren (1990, 1991), Davies (1990, 1993, 1994), and Croux *et al.* (1994). Yohai (1987) proposed high-breakdown-point and high-efficiency robust estimates for linear regression; see also Yohai *et al.* (1991). Yohai and Zamar (1988) investigated high-breakdown point estimates of regression by means of the minimization of an efficient scale.

For recent advances of robust methods for online monitoring data we refer to Gather *et al.* (2002), Gather and Fried (2004), and Gather and Schettlinger (2007).

For robust estimation in generalized linear models, we refer to Pregibon (1982) for resistant estimation and Stefanski *et al.* (1986), Künsch *et al.* (1989), Morgenthaler (1992), Carroll and Pederson (1993), Bianco and Yohai (1996), and Cantoni and Ronchetti (2001) for M-estimation. Christmann (1992, 1994, 1998) proposed high-breakdown point estimators in generalized linear models. Rousseeuw and Christmann (2003) proposed robust estimates for logistic regression and other binary regression problems and solved the problem of non-existence of the classical non-robust maximum likelihood estimates and almost all robust estimates proposed earlier. The mathematical reason why many estimators do not have a solution for all possible data sets turned out to be the complete or quasi-complete separation of data points with positive weights; see Albert and Anderson (1984) and Santner and Duffy (1986).

For depth-related methods, see Tukey (1975) for the halfspace depth, Oja (1983) for Oja's median, Liu (1990, 1999) for the simplicial depth, Rousseeuw and Hubert (1999) for regression depth, Mizera (2002) for tangent depth, He and Wang (1997) for depth contours for multivariate data sets, and Zuo and Serfling (2000) and Mosler (2002) for various variants of statistical depth functions. For a numerical comparison between the support vector machine and the regression depth method proposed by Rousseeuw and Hubert (1999), see Christmann and Rousseeuw (2001) and Christmann *et al.* (2002).

Projection pursuit and its relation to robust estimation were investigated by Huber (1985, 1993) and Donoho and Johnstone (1989).

10.7 Summary

This chapter showed that support vector machines have—besides many other nice mathematical and algorithmic properties—good robustness properties if the loss function and the kernel are appropriately chosen. There is a natural connection between SVMs and M-estimation.

Robust statistics investigates how small violations of the model assumptions influence the results of the methods used. Many researchers still ignore deviations from ideal models because they hope that such deviations will not matter. However, J. W. Tukey, one of the pioneers of robust statistics, mentioned already in 1960 that: *"Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians"*; see Hampel *et al.* (1986, p. 21).

Small model violations are unavoidable in almost all practical applications because statistical models are usually simplifications of the true process that generated a data set and because typing errors, outliers, and many types of contamination occur in practice. In particular, this is true for data mining projects in which huge data sets with moderate data quality often must be analyzed; see Chapter 12.

It is therefore important to study the robustness properties of SVMs, although the model assumption of independent and identically distributed pairs of random variables without further restriction of the probability distribution is weak. Many approaches of robust statistics were developed for estimators of an unknown vector $\theta \in \mathbb{R}^d$ in parametric models. From a mathematical point of view, it is also interesting to investigate how these approaches can be used for SVMs where a function f in a (usually) infinite-dimensional Hilbert space H must be estimated.

A motivation of robust statistics was given in Section 10.1. Different approaches of robustness were described in Section 10.2, including topological neighborhoods in the space of probability distributions, qualitative robustness, influence functions with the related notions of gross error sensitivity, maxbias and sensitivity curve, and breakdown points.

Robustness properties of support vector machines for binary classification and regression were investigated in Sections 10.3 and 10.4. It was shown that the influence function of SVMs exists and is bounded if the loss function has a bounded first derivative and if the kernel is bounded and continuous, e.g., the Gaussian RBF kernel. This shows that SVMs based on a logistic loss function have nice robustness properties for classification and regression problems. In contrast, SVMs based on the least squares loss yield less robust results than SVMs based on the classical hinge loss or the ϵ-insensitive loss. Bounds for the maxbias, gross error sensitivity, and the sensitivity curve were derived; some of the bounds even turned out to be uniform. For the regression case with unbounded output space Y, it turned out that good robustness properties of SVMs are only available if the tail behavior of the probability distribution P and the growth type of the loss function L are appropriately

related to each other. Especially Lipschitz-continuous loss functions of growth type 1 offer good robustness properties. There are relationships between these results on robustness and stability results obtained by Poggio *et al.* (2004) and Mukherjee *et al.* (2006).

In Section 10.5, a simple but powerful subsampling strategy called robust learning from bites (RLB) was described to make SVMs usable for huge data sets (e.g., in data mining projects). RLB is designed for situations under which the original robust method cannot be applied due to excessive computation time or memory space problems. In these situations, RLB offers robust estimates and additionally robust confidence intervals. Although RLB estimators will in general not fulfill certain optimality criteria, the method has the advantages of scalability (the number of bites can be chosen according to the data set and the available hardware), performance (the computational steps for different bites can easily be distributed on several processors), robustness (RLB inherits robustness properties from the SVMs used by this strategy), and confidence intervals (no complex formulas are needed to obtain distribution-free (componentwise) confidence intervals for the estimates or the predictions).

10.8 Exercises

10.1. Mean and median ($\star$)
Prove that the mean and median are solutions of the optimization problems (10.5) and (10.6).

10.2. Property of Prohorov metric ($\star$)
Let $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1(Z)$. Define $\mathrm{P}_\delta := (1-\delta)\mathrm{P} + \delta\mathrm{Q}$ and $\mathrm{P}_\varepsilon := (1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}$ for $\delta, \varepsilon \in [0,1]$. Show that $d_{\mathrm{Pro}}(\mathrm{P}_\delta, \mathrm{P}_\varepsilon) \leq |\delta - \varepsilon|$ if $\delta \to \varepsilon$.

Hint: Dudley (2002).

10.3. Bounded Lipschitz metric ($\star\star\star$)
Prove Theorem A.4.22*iii)*.

Hint: The direction (b) $\Rightarrow$ (a) follows easily. One can use the Lagrange approach, the Kuhn and Tucker (1951) theorem, and Strassen's Theorem A.4.16 to prove the converse direction. See Huber (1981, pp. 30ff.) for details.

10.4. Lévy metric ($\star\star$)
The Lévy distance $d_{\mathrm{Lévy}}(\mathrm{P}_1, \mathrm{P}_2)$ of two probability measures P and Q on $\mathbb{R}$ is defined by

$$\inf\bigl\{\varepsilon > 0 : \mathrm{P}_1((-\infty, x-\varepsilon]) - \varepsilon \leq \mathrm{P}_2((-\infty, x]) \leq \mathrm{P}_1((-\infty, x+\varepsilon] + \varepsilon\bigr\}\,.$$

i) Show that $d_{\mathrm{Lévy}}$ is a metric that metricizes the weak* topology.

ii) Define $\mathrm{P}_\delta := (1-\delta)\mathrm{P} + \delta\mathrm{Q}$ and $\mathrm{P}_\varepsilon := (1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}$, $\delta, \varepsilon \in [0,1]$. Show that the Lévy metric shares with the Prohorov metric and the bounded Lipschitz metric the property $d_{\mathrm{Lévy}}(\mathrm{P}_\delta, \mathrm{P}_\varepsilon) = O(|\delta - \varepsilon|)$, $\delta \to \varepsilon$.

Hint: See Huber (1981, pp. 25ff.).

10.5. Kolmogorov metric and total variation distance ($\star\star$)

i) Prove that the Kolmogorov distance defined by

$$d_{\mathrm{Kol}}(\mathrm{P},\mathrm{Q}) := \sup_{x\in\mathbb{R}} |\mathrm{P}((-\infty,x]) - \mathrm{Q}((-\infty,x])|, \quad \mathrm{P},\mathrm{Q}\in\mathcal{M}_1(\mathbb{R}),$$

is a metric.

ii) Prove that the total variation distance defined by

$$d_{\mathrm{tv}}(\mathrm{P},\mathrm{Q}) := \sup_{A\in\mathcal{B}(\mathbb{R}^n)} |\mathrm{P}(A) - \mathrm{Q}(A)|, \quad \mathrm{P},\mathrm{Q}\in\mathcal{M}_1(\mathbb{R}^n),$$

is a metric.

iii) Clarify the connections between d_{Pro}, $d_{\mathrm{Lévy}}$, d_{Kol}, and d_{tv}.

iv) Do d_{tv} and d_{Kol} generate the weak* topology?

Hints: See Huber (1981, pp. 34ff.). *iii).* We have $d_{\mathrm{Lévy}}(\mathrm{P},\mathrm{Q}) \le d_{\mathrm{Pro}}(\mathrm{P},\mathrm{Q}) \le d_{\mathrm{tv}}(\mathrm{P},\mathrm{Q})$ and $d_{\mathrm{Lévy}}(\mathrm{P},\mathrm{Q}) \le d_{\mathrm{Kol}}(\mathrm{P},\mathrm{Q}) \le d_{\mathrm{tv}}(\mathrm{P},\mathrm{Q})$. *iv).* No.

10.6. Asymmetric logistic loss ($\star$)

Consider the asymmetric logistic loss $L : Y \times \mathbb{R} \to [0,\infty)$,

$$L(x,y,t) = \Big(\frac{c_2}{2} - c_1\Big)r - \frac{c_2}{2}\ln\Big(4\Lambda(r-c_3)\big(1-\Lambda(r-c_3)\big)\Big) + c_4\,,$$

where $r = y - t$, $0 < c_1 < c_2 < \infty$, $\Lambda(r) := 1/(1+e^{-r})$, $c_3 = -\Lambda^{-1}(c_1/c_2)$, and $c_4 = (c_2/2)\ln(4\frac{c_1}{c_2}(1-\frac{c_1}{c_2}))$.

i) Show that $(c_1,c_2) = (1,2)$ gives $L = L_{\text{r-logist}}$.

ii) Show that $L' = \partial_3 L$ and $L'' = \partial_{33} L$ are continuous and bounded.

iii) Is L a Nemitski loss function of some order $p \in (0,\infty)$?

Hint: iii). Yes.

10.7. Maxbias ($\star\star$)

Derive bounds for the maxbias$(\varepsilon; S, \mathrm{P})$ of SVMs over contamination neighborhoods $N_\varepsilon(\mathrm{P})$.

Hint: Use Theorems 10.25, 10.26, and 10.27.

10.8. Kernel-based LMS ($\star\star$)

Consider a non-parametric regression problem treated in Section 10.4. Define an estimator for $f \in H$ by $f^*_{\mathrm{P},\lambda} := \arg\min_{f\in H} \mathrm{Median}_{\mathrm{P}} L(Y, f(X)) + \lambda\|f\|_H^2$.

i) Explain why this estimator does *not* fit into the framework of SVMs for regression treated in Section 10.4.

ii) Show that this estimator corresponds to the least median of squares estimator for the special case of a linear kernel and $\lambda = 0$. Summarize the statistical properties of LMS from the literature.

Hint: See, e.g., Hampel (1975), Rousseeuw (1984), Rousseeuw and Leroy (1987), and Davies (1990, 1993).

10.9. Robustness properties of kernel-based LMS ($\star\star\star\star$)

Derive the robustness properties of $f^*_{\mathrm{P},\lambda}$ defined in Exercise (10.8).

Hint: Research.

11

Computational Aspects

Overview. *This chapter presents techniques to compute decision functions $f_{D,\lambda}$ of SVMs and mentions some software tools. In addition, we discuss how to choose suitable hyperparameters used by the regularization term, the kernel, and the loss.*

Prerequisites. *We need Chapter 2 on loss functions, Chapter 4 on kernels, Chapter 8 on SVMs for classification, and Chapter 9 on SVMs for regression problems. Some results from convex analysis, in particular from Section A.6.5 on convex programs and Section 6.5 on oracle inequalities for the purpose of parameter selection, are used.*

This chapter is concerned with the question of how to compute a decision function $f_{D,\lambda}$ of support vector machines for a given data set D. No attempt is made to describe in detail all relevant computational aspects regarding SVMs. From our point of view, there exists such a large body of literature on this topic that even the main research approaches published during the last decade would probably fill a textbook of their own. Here we only show some facets regarding computational aspects of SVMs. More precisely, we will concentrate in this chapter on addressing the following questions.

i) How can we compute the empirical SVM decision function $f_{D,\lambda}$?
ii) Are there algorithms to compute $f_{D,\lambda}$ that work well even for large sample sizes?
iii) Are there loss functions L such that the numerical problem of computing $f_{D,\lambda}$ can be solved especially fast?
iv) Are there loss functions such that $f_{D,\lambda}$ can efficiently be computed and have good robustness properties in the sense of Chapter 10?

After a short introduction, we show in Section 11.1 that empirical SVM solutions $f_{D,\lambda}$ are solutions of specific convex programs. SVMs based on the hinge loss and the logistic loss for classification purposes, ϵ-insensitive loss and least squares loss for regression, and pinball loss function for kernel-based quantile regression are treated as special cases. Some computational aspects of computing $f_{D,\lambda}$ efficiently for large sample sizes n are considered in Section 11.2, where special consideration is given to the hinge loss. SVMs depend in general not only on the loss function L and the kernel k but also on the determination of hyperparameters such as $\lambda > 0$, kernel parameters such as γ^2 for the Gaussian RBF kernel, or the parameter ϵ for the ϵ-insensitive loss

function. These hyperparameters can have a substantial impact on the quality of $f_{\mathrm{D},\lambda}$ and $\mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda})$ in applications. Section 11.3 lists some techniques to determine suitable choices of these hyperparameters. Section 11.4 mentions a few software tools available to compute empirical SVMs.

11.1 SVMs, Convex Programs, and Duality

In this section, it will be shown that the decision function $f_{\mathrm{D},\lambda}$ of SVMs for a given data set D with n data points is a solution of a certain *finite-dimensional* convex program. Throughout this section, we make the following assumptions if not otherwise stated.

Assumption 11.1 *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss function, H a reproducing kernel Hilbert space over a non-empty convex set X with positive definite kernel k and canonical feature map $\Phi(x) := k(\cdot, x)$, $x \in X$. Furthermore, let $\lambda > 0$ be the regularization parameter and $D = \{(x_i, y_i), i = 1,\ldots,n\} \subset X \times Y$ be a fixed data set with corresponding empirical measure $\mathrm{D} = \frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i,y_i)}$.*

We know by the representer theorem (see Theorem 5.5) that there exists a unique empirical SVM solution $f_{\mathrm{D},\lambda} \in H$ satisfying

$$\mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f_{\mathrm{D},\lambda}) = \inf_{f\in H} \mathcal{R}_{L,\mathrm{D}}(f) + \lambda\|f\|_H^2 \,. \tag{11.1}$$

In addition, there exists a vector $\bar{\alpha} = (\bar{\alpha}_1,\ldots,\bar{\alpha}_n) \in \mathbb{R}^n$, which in general is not uniquely determined, such that the empirical SVM solution has the representation

$$f_{\mathrm{D},\lambda} = \sum_{j=1}^{n} \bar{\alpha}_j \Phi(x_j) \in H_{|X'} \,, \tag{11.2}$$

where $H_{|X'} := \operatorname{span}\{\Phi(x_j) : j = 1,\ldots,n\}$. For notational convenience, let us denote for each $\alpha \in \mathbb{R}^n$ the corresponding function in $H_{|X'}$ by $w(\alpha)$,

$$w(\alpha) = \sum_{j=1}^{n} \alpha_j \Phi(x_j).$$

Note that we have

$$\|w(\alpha)\|_H^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j k(x_i,x_j) = \alpha^{\mathsf{T}} K \alpha, \tag{11.3}$$

where $K \in \mathbb{R}^{n\times n}$ is the symmetric kernel matrix (or Gram matrix) with coefficients $K_{i,j} := k(x_i,x_j)$ and α^{T} denotes the transpose of the vector α. Furthermore, we obtain

$$f_{\mathrm{D},\lambda}(x) = \langle w(\bar{\alpha}), \Phi(x) \rangle_H = \sum_{j=1}^{n} \bar{\alpha}_j k(x, x_j), \quad x \in X, \tag{11.4}$$

and

$$f_{\mathrm{D},\lambda}(x_i) = \bar{\alpha}_i^{\mathsf{T}} K e_i = e_i^{\mathsf{T}} K \bar{\alpha}_i, \quad i = 1, \ldots, n. \tag{11.5}$$

Note that $f_{\mathrm{D},\lambda} \in H_{|X'}$. Hence it is sufficient to optimize over $H_{|X'}$. Therefore, if we plug (11.3) and (11.4) into (11.1), we obtain

$$\begin{aligned} \mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f_{\mathrm{D},\lambda}) &= \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L\big(x_i, y_i, f(x_i)\big) + \lambda \|f\|_H^2 \\ &= \min_{w \in H_{|X'}} \frac{1}{n} \sum_{i=1}^{n} L\big(x_i, y_i, \langle w, \Phi(x_i) \rangle_H\big) + \lambda \|w\|^2_{H_{|X'}} \\ &= \min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} L\Big(x_i, y_i, \sum_{j=1}^{n} \alpha_j k(x_i, x_j)\Big) + \lambda\, \alpha^{\mathsf{T}} K \alpha\,. \end{aligned} \tag{11.6}$$

A finite sum of convex functions all defined on the same convex set is of course also a convex function. Therefore, the convexity of the loss function yields the convexity of the first term in (11.6). Furthermore, the kernel k is by Assumption 11.1 positive definite in the sense of Definition 4.15. This shows that the quadratic form $\|w(\alpha)\|_H^2 = \alpha^{\mathsf{T}} K \alpha$ is convex with respect to $\alpha \in \mathbb{R}^n$ whenever the symmetric matrix K is positive semi-definite; see (A.51). Combining these results, we conclude that the *decision function* $f_{\mathrm{D},\lambda}$ *is the solution of a finite-dimensional convex program* (see Definition A.6.22) if the Assumption 11.1 is valid. This is somewhat astonishing because the dimension of the reproducing kernel Hilbert space H is not assumed to be finite. Now recall that L is non-negative. Thus we can rewrite (11.6) in the form

$$\min_{\alpha,\, \xi \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \, \|w(\alpha)\|_H^2 \tag{11.7}$$

$$\text{s.t.} \quad \xi_i \geq L\big(x_i, y_i, \langle w(\alpha), \Phi(x_i) \rangle_H\big)\,, \quad i = 1, \ldots, n. \tag{11.8}$$

Hence many classical results about convex programs with constraints such as Lagrange multipliers, determination of saddle points, and algorithms to solve convex programs are applicable for empirical SVMs and will be used in the following considerations; see Section A.6.5 for details on convex programs.

Now we consider the convex programs for $f_{\mathrm{D},\lambda}$ corresponding to several loss functions often used for classification and regression purposes.

Example 11.2 (Margin-based loss). Assume that L is a margin-based loss function in the sense of Definition 2.24; i.e., $L(x, y, t) = \varphi(yt)$ for $y \in \{-1, +1\}$ and $t \in \mathbb{R}$. Hence the convex program (11.7) and (11.8) for $f_{\mathrm{D},\lambda}$ simplifies to

$$\min_{\alpha,\,\xi\in\mathbb{R}^n} \quad \frac{1}{n}\sum_{i=1}^n \xi_i + \lambda \left\|w(\alpha)\right\|_H^2 \tag{11.9}$$

$$\text{s.t.} \quad \xi_i \geq \varphi\big(y_i \left\langle w(\alpha), \Phi(x_i)\right\rangle_H\big), \quad i = 1,\ldots,n. \tag{11.10}$$

The decision function is $f_{\mathrm{D},\lambda} = \sum_{i=1}^n \bar{\alpha}_i \Phi(x_i)$, where $\bar{\alpha} = (\bar{\alpha}_1,\ldots,\bar{\alpha}_n)$ solves (11.9) and (11.10). ◁

Example 11.3 (Classification based on hinge loss). The hinge loss for classification is given by $L_{\mathrm{hinge}}(y,t) := \max\{0, 1-yt\}$ for $y \in \{-1,+1\}$ and $t \in \mathbb{R}$; see Example 2.27. The convex program for $f_{\mathrm{D},\lambda}$ is therefore given by

$$\min_{\alpha,\,\xi\in\mathbb{R}^n} \quad C\sum_{i=1}^n \xi_i + \frac{1}{2}\left\|w(\alpha)\right\|_H^2 \tag{11.11}$$

$$\text{s.t.} \quad \xi_i \geq 0,\ \ \xi_i \geq 1 - y_i \left\langle w(\alpha), \Phi(x_i)\right\rangle_H, \quad i = 1,\ldots,n, \tag{11.12}$$

where $C := 1/(2n\lambda)$ and $w(\alpha) = \sum_{j=1}^n \alpha_j y_j \Phi(x_j)$. The corresponding Lagrangian L^* is given by

$$C\sum_{i=1}^n \xi_i + \frac{1}{2}\left\|w(\alpha)\right\|_H^2 + \sum_{i=1}^n \alpha_i\big(1 - y_i \left\langle w(\alpha), \Phi(x_i)\right\rangle_H - \xi_i\big) - \sum_{i=1}^n \eta_i \xi_i\,,$$

where $\alpha := (\alpha_1,\ldots,\alpha_n) \in [0,\infty)^n$ and $\eta := (\eta_1,\ldots,\eta_n) \in [0,\infty)^n$. The corresponding dual program is found by differentiating L^* with respect to α and $\xi = (\xi_1,\ldots,\xi_n)$ imposing stationarity,

$$\nabla_\alpha L^*(\alpha,\xi) = w(\alpha) - \sum_{i=1}^n \alpha_i y_i \Phi(x_i) = 0, \tag{11.13}$$

$$\frac{\partial L^*}{\partial \xi_i} = C - \alpha_i - \eta_i = 0, \tag{11.14}$$

and resubstituting these relations into the primal problem. We obtain the dual program

$$\max_{\alpha\in[0,C]^n} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j k(x_i,x_j), \tag{11.15}$$

where the box constraint $\alpha \in [0,C]^n$ results from $\alpha_i \in [0,\infty)$, $\eta_i \in [0,\infty)$, and $\alpha_i = C - \eta_i$ due to (11.14). Therefore, $f_{\mathrm{D},\lambda}$ is the unique solution of even a *quadratic program with box constraints* if the hinge loss is used. We will see later on that this fact allows us to construct fast algorithms because quadratic problems can often be solved faster than general convex problems. For the case of an additional offset term b (i.e., we are computing $(f_{\mathrm{D},\lambda}, b_{\mathrm{D},\lambda}) \in H \times \mathbb{R}$) we have to replace $\langle w(\alpha), \Phi(x_i)\rangle_H$ by $\langle w(\alpha), \Phi(x_i)\rangle_H + b$ and add the additional constraint $\sum_{i=1}^n \alpha_i y_i = 0$ in (11.15); see Exercise 11.2. ◁

Example 11.4 (Classification based on logistic loss). The logistic loss for classification is given by $L_{\text{c-logist}}(y,t) = \varphi(yt) = \ln\bigl(1+\exp(-yt)\bigr)$, $y \in \{-1,+1\}$, $t \in \mathbb{R}$; see Example 2.29. Let us define the function $g : \mathbb{R} \to \mathbb{R}$, $g(r) := \varphi(-r) = \ln(1+e^r)$, $r \in \mathbb{R}$, and $C := 1/(2n\lambda)$. The empirical decision function based on this loss is thus given by

$$f_{\mathrm{D},\lambda} = \arg\min_{\alpha\in\mathbb{R}^n} C\sum_{i=1}^{n} g\bigl(-y_i \langle w(\alpha), \Phi(x_i)\rangle_H\bigr) + \frac{1}{2}\|w(\alpha)\|_H^2 . \tag{11.16}$$

Note that $w(\alpha) = \sum_{j=1}^n \alpha_j y_j \Phi(x_j)$. With $\xi_i := -\sum_{j=1}^n \alpha_j y_i y_j k(x_i,x_j)$ for $i,j \in \{1,\ldots,n\}$, it follows that

$$f_{\mathrm{D},\lambda} = \arg\min_{\alpha\in\mathbb{R}^n} C\sum_{i=1}^{n} g(\xi_i) + \frac{1}{2}\|w(\alpha)\|_H^2 \tag{11.17}$$

and $\|w(\alpha)\|_H^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i\alpha_j y_i y_j k(x_i,x_j)$. The Lagrangian for this problem is given by

$$L^*(\alpha,\xi) = C\sum_{i=1}^{n} g(\xi_i) + \frac{1}{2}\|w(\alpha)\|_H^2 - \sum_{i=1}^{n} \alpha_i\bigl(\xi_i + y_i \langle w(\alpha), \Phi(x_i)\rangle_H\bigr).$$

Note that $g'(r) = e^r/(1+e^r) = 1/\bigl(1+e^{-r}\bigr)$, $r \in \mathbb{R}$, equals the cumulative distribution function of the logistic distribution and that its inverse is the logit function $(g')^{-1}(u) = \ln(u/(1-u))$, $u \in (0,1)$. Computing the partial derivatives of the Lagrangian, we obtain the optimality conditions

$$\nabla_\alpha L^* = w(\alpha) - \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i) \;=\; 0, \tag{11.18}$$

$$\frac{\partial L^*}{\partial \xi_i} = Cg'(\xi_i) - \alpha_i \;=\; 0, \qquad\qquad i = 1,\ldots,n. \tag{11.19}$$

Define the function $G(u) = u\ln(u) + (1-u)\ln(1-u)$, $u \in (0,1)$. Hence $G'(u) = (g')^{-1}(u)$. It follows from (11.19) that $g'(\xi_i) = \alpha_i/C$ and $\xi_i = (g')^{-1}(\alpha_i/C) = G'(\alpha_i/C)$. We obtain the dual program for kernel-based logistic regression:

$$\min_{\alpha\in(0,C)^n} C\sum_{i=1}^{n} G\Bigl(\frac{\alpha_i}{C}\Bigr) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j k(x_i,x_j), \tag{11.20}$$

which is a finite-dimensional convex program; see also Exercise 11.3. ◁

Example 11.5 (Classification based on least squares loss). The least squares loss function for classification is given by $L_{\mathrm{LS}}(y,t) = (1-yt)^2 = (y-t)^2$, $y \in \{-1,+1\}$, $t \in \mathbb{R}$, considered in Example 2.26. The decision function $f_{\mathrm{D},\lambda}$ based on this loss function solves the convex programming problem

$$\begin{aligned} \min_{\alpha,\, r \in \mathbb{R}^n} \quad & \frac{C}{2} \sum_{i=1}^{n} r_i^2 \, + \frac{1}{2} \, \|w(\alpha)\|_H^2 \\ \text{s.t.} \quad & r_i = 1 - y_i \, \langle w(\alpha), \Phi(x_i) \rangle_H \,, \quad i = 1, \ldots, n, \end{aligned}$$

where $r = (r_1, \ldots, r_n)$ and $C := 1/(\lambda n)$. The Lagrangian is given by

$$L^*(\alpha, r) = \frac{C}{2} \sum_{i=1}^{n} r_i^2 \, + \frac{1}{2} \, \|w(\alpha)\|_H^2 + \sum_{i=1}^{n} \alpha_i \big(1 - y_i \, \langle w(\alpha), \Phi(x_i) \rangle_H - r_i\big).$$

After computing the partial derivatives of L^*, we get the following conditions for optimality:

$$\begin{aligned} w(\alpha) &= \sum_{j=1}^{n} \alpha_j y_j \Phi(x_j), \\ \alpha_i \quad &= C r_i \,, \qquad\qquad\qquad\qquad i = 1, \ldots, n, \\ r_i \quad &= 1 - y_i \, \langle w(\alpha), \Phi(x_i) \rangle_H \,, \quad i = 1, \ldots, n; \end{aligned} \tag{11.21}$$

see Exercise 11.4. This is a system of linear equations with respect to α. Note that sparseness of $f_{\mathrm{D},\lambda}$ is lost due to $\alpha_i = C r_i$ in (11.21). ◁

Example 11.6 (Distance-based loss). Assume that L is a distance-based loss function in the sense of Definition 2.32 (i.e., $L(x, y, t) = \psi(y - t)$, $y, t \in \mathbb{R}$) with $\psi : \mathbb{R} \to \mathbb{R}$ some suitable function. The convex program (11.7) and (11.8) for $f_{\mathrm{D},\lambda}$ simplifies to

$$\min_{\alpha,\, \xi \in \mathbb{R}^n} \quad \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \, \alpha^{\mathsf{T}} K \alpha \tag{11.22}$$

$$\text{s.t.} \quad \xi_i \geq \psi\Big(y_i - \sum_{j=1}^{n} \alpha_j k(x_i, x_j)\Big), \quad i = 1, \ldots, n. \tag{11.23}$$

Let us now consider for the sake of simplicity a convex distance-based loss function of the type

$$L(x, y, t) = \psi(y - t) = \tilde{\psi}\big(\max\big\{0, |y - t| - \epsilon\big\}\big),$$

where $\tilde{\psi} : [0, \infty) \to [0, \infty)$ and $\epsilon \geq 0$. We additionally assume that ψ has a continuous first derivative on $\mathbb{R} \backslash \{-\epsilon, +\epsilon\}$. We obtain

$$\psi(y - t) = \tilde{\psi}\big(\max\{0, y - t - \epsilon\}\big) + \tilde{\psi}\big(\max\{0, t - y - \epsilon\}\big), \quad y, t \in \mathbb{R}.$$

The convex program for $f_{\mathrm{D},\lambda}$ can then be rewritten as

$$\begin{aligned}\min_{\alpha,\,\xi^+,\,\xi^-\in\mathbb{R}^n}\quad & \frac{1}{n}\sum_{i=1}^{n}\bigl(\xi_i^+ + \xi_i^-\bigr) + \lambda\,\alpha^\mathsf{T} K\alpha \\ \text{s.t.}\quad & \xi_i^+ \geq \tilde{\psi}\Bigl(\max\Bigl\{0, y_i - \sum_{j=1}^{n}\alpha_j k(x_i,x_j) - \epsilon\Bigr\}\Bigr),\ i=1,\ldots,n, \\ & \xi_i^- \geq \tilde{\psi}\Bigl(\max\Bigl\{0, \sum_{j=1}^{n}\alpha_j k(x_i,x_j) - y_i - \epsilon\Bigr\}\Bigr),\ i=1,\ldots,n;\end{aligned}$$

see also Exercise 11.5. ◁

Example 11.7 (Regression based on ϵ-insensitive loss). Let $\epsilon > 0$. The ϵ-insensitive loss for regression uses $\psi(y-t) = \max\bigl\{0, |y-t| - \epsilon\bigr\}$ for $y, t \in \mathbb{R}$, considered in Example 2.42. This loss function is obviously an example of the distance-based symmetric loss functions considered in Example 11.6. The convex program for $f_{\mathrm{D},\lambda}$ can be rewritten as

$$\begin{aligned}\min_{\alpha,\,\xi^+,\,\xi^-\in\mathbb{R}^n}\quad & C\sum_{i=1}^{n}\bigl(\xi_i^+ + \xi_i^-\bigr) + \frac{1}{2}\,\alpha^\mathsf{T} K\alpha \\ \text{s.t.}\quad & \xi_i^+ \geq 0,\ \xi_i^+ \geq y_i - \sum_{j=1}^{n}\alpha_j k(x_i,x_j) - \epsilon,\ i=1,\ldots,n, \\ & \xi_i^- \geq 0,\ \xi_i^- \geq \sum_{j=1}^{n}\alpha_j k(x_i,x_j) - y_i - \epsilon,\ i=1,\ldots,n,\end{aligned}$$

where $C := 1/(2n\lambda)$. The Lagrangian is thus given by

$$\begin{aligned}L^*(\alpha,\xi^+,\xi^-) = C\sum_{i=1}^{n}\bigl(\xi_i^+ + \xi_i^-\bigr) \quad & + \frac{1}{2}\,\alpha^\mathsf{T} K\alpha \\ & + \sum_{i=1}^{n}\alpha_i^+\Bigl(y_i - \sum_{j=1}^{n}\alpha_j k(x_i,x_j) - \epsilon - \xi_i^+\Bigr) - \sum_{i=1}^{n}\eta_i^+\xi_i^+ \\ & + \sum_{i=1}^{n}\alpha_i^-\Bigl(\sum_{j=1}^{n}\alpha_j k(x_i,x_j) - y_i - \epsilon - \xi_i^-\Bigr) - \sum_{i=1}^{n}\eta_i^-\xi_i^-,\end{aligned}$$

where the Lagrange multipliers α_i^+, α_i^-, η_i^+, and η_i^-, $i=1,\ldots,n$, have to be non-negative. It follows from the saddle point condition of convex programs (see Theorem A.6.26) that the partial derivatives of L^* with respect to the primal variables $w(\alpha)$, ξ^+, and ξ^- have to vanish for optimality; i.e.,

$$\nabla_\alpha L^* = w(\alpha) - \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-)\Phi(x_i) = 0, \tag{11.24}$$

$$\frac{\partial L^*}{\partial \xi_i^+} = C - \alpha_i^+ - \eta_i^+ = 0\,, \qquad i = 1, \ldots, n,$$

$$\frac{\partial L^*}{\partial \xi_i^-} = C - \alpha_i^- - \eta_i^- = 0\,, \qquad i = 1, \ldots, n.$$

Note that we can easily eliminate η_i^+ and η_i^- from L^* because $\eta_i^+ = C - \alpha_i^+$ due to (11.25) and $\eta_i^- = C - \alpha_i^-$ due to (11.25), $i = 1, \ldots, n$. Substituting (11.24) and (11.25) into the formula for the Lagrangian L^* yields the dual program (see Exercise 11.6)

$$\max_{\alpha^+, \alpha^- \in \mathbb{R}^n} \; \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) y_i - \epsilon \sum_{i=1}^n (\alpha_i^+ + \alpha_i^-) - \frac{1}{2}(\alpha_i^+ - \alpha_i^-)^\mathsf{T} K (\alpha_i^+ - \alpha_i^-)$$
$$\text{s.t.} \qquad \alpha_i^+,\, \alpha_i^- \in [0, C], \quad i = 1, \ldots, n\,. \tag{11.25}$$

Due to (11.24), we thus obtain $f_{\mathrm{D},\lambda} = \sum_{i=1}^n \alpha_i \Phi(x_i)$ with $\alpha_i = \alpha_i^+ - \alpha_i^-$, $i = 1, \ldots, n$. The ϵ-insensitive loss and the hinge loss hence share the nice property that $f_{\mathrm{D},\lambda}$ is the unique solution of a *quadratic program with box constraints*.

For the case of an additional offset term b (i.e., we compute $\big(f_{\mathrm{D},\lambda}, b_{\mathrm{D},\lambda}\big) \in H \times \mathbb{R}$) we have to replace $\langle w(\alpha), \Phi(x_i)\rangle_H$ by $\langle w(\alpha), \Phi(x_i)\rangle_H + b$ and to add the constraint $\sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) = 0$ in (11.25) of the dual program. ◁

Example 11.8 (Regression based on least squares loss). The least squares loss function for regression is given by $L_{\mathrm{LS}}(y, t) = (y - t)^2$, $y, t \in \mathbb{R}$. The decision function $f_{\mathrm{D},\lambda}$ based on this loss function solves the convex programming problem

$$\min_{\alpha,\, r \in \mathbb{R}^n} \; \frac{C}{2} \sum_{i=1}^n r_i^2 \, + \frac{1}{2}\, \alpha^\mathsf{T} K \alpha$$
$$\text{s.t.} \quad r_i = y_i - \sum_{j=1}^n \alpha_j k(x_i, x_j)\,, \quad i = 1, \ldots, n,$$

where $r = (r_1, \ldots, r_n)$ and $C := 1/(\lambda n)$. The Lagrangian is given by

$$L^*(\alpha, r) = \frac{C}{2} \sum_{i=1}^n r_i^2 \, + \frac{1}{2}\, \alpha^\mathsf{T} K \alpha + \sum_{i=1}^n \alpha_i \Big(y_i - \sum_{j=1}^n \alpha_j k(x_i, x_j) - r_i \Big)\,.$$

After computing the partial derivatives of L^*, we get the following conditions for optimality:

$$\begin{aligned} w(\alpha) &= \sum_{i=1}^{n} \alpha_i \Phi(x_i), \\ \alpha_i &= C r_i\,, \qquad\qquad i = 1, \ldots, n, \\ r_i &= y_i - \langle w(\alpha), \Phi(x_i) \rangle_H\,, \quad i = 1, \ldots, n, \end{aligned}$$

see Exercise 11.7. This is a system of linear equations with respect to α. Note that one loses the sparseness property due to the conditions $\alpha_i = C r_i$ for $i = 1, \ldots, n$. ◁

Example 11.9 (Quantile regression based on the pinball loss). As our final example in this section, we consider the pinball loss function, which is suitable for kernel based quantile regression (see Example 2.43). This distance-based loss function uses $\psi_\tau(y-t) = (\tau-1)(y-t)$ for $y-t < 0$ and $\psi_\tau(y-t) = \tau(y-t)$ for $y - t \geq 0$, where $\tau \in (0,1)$ specifies the desired quantile level. The convex programming problem to determine $f_{\mathrm{D},\lambda}$ can be rewritten as

$$\min_{\alpha,\, \xi^+,\, \xi^- \in \mathbb{R}^n} \quad C \sum_{i=1}^{n} \bigl(\tau \xi_i^+ + (1-\tau)\xi_i^-\bigr) + \frac{1}{2}\, \alpha^{\mathsf{T}} K \alpha \tag{11.26}$$

$$\text{s.t.} \qquad \xi_i^+ \geq 0, \quad \xi_i^+ \geq y_i - \sum_{j=1}^{n} \alpha_j k(x_i, x_j),\ i = 1, \ldots, n, \tag{11.27}$$

$$\xi_i^- \geq 0, \quad \xi_i^- \geq \sum_{j=1}^{n} \alpha_j k(x_i, x_j) - y_i\,,\ i = 1, \ldots, n, \tag{11.28}$$

where $C = 1/(2n\lambda)$. Using the Lagrangian approach in a way similar to Example 11.7, we obtain the dual program

$$\max_{\alpha \in \mathbb{R}^n} \quad \alpha^{\mathsf{T}} y - \alpha^{\mathsf{T}} K \alpha \tag{11.29}$$

$$\text{s.t.} \quad C(\tau - 1) \leq \alpha_i \leq C\tau, \quad i = 1, \ldots, n\,, \tag{11.30}$$

where $y := (y_1, \ldots, y_n)$ and $C := 1/(2n\lambda)$ (see Exercise 11.8). Note that the box constraints in (11.30) are only symmetric around 0 for $\tau = \frac{1}{2}$; i.e., for median regression. It follows from (11.29) and (11.30) that $f_{\mathrm{D},\lambda}$ based on the pinball loss is the unique solution of even a *quadratic program with box constraints*.

For the case of an additional offset term b (i.e., we compute $\bigl(f_{\mathrm{D},\lambda}, b_{\mathrm{D},\lambda}\bigr) \in H \times \mathbb{R}$), we have to replace $\sum_{j=1}^{n} \alpha_j k(x_i, x_j)$ by $\sum_{j=1}^{n} \alpha_j k(x_i, x_j) + b$ and to add the constraint $\sum_{i=1}^{n} \alpha_i = 0$ in (11.30).

Suppose that we want to estimate $m \geq 2$ conditional quantile functions with quantile levels $0 < \tau_1 < \ldots < \tau_m < 1$ on the same data set D. Let us denote the corresponding empirical SVM solutions by $f_{\mathrm{D},\lambda,\tau_h}$, $h = 1, \ldots, m$. Then it can occur that for some $x \in X$ and some pair (h_1, h_2) with $1 \leq h_1 < h_2 \leq m$, the conditional quantile estimates are in reversed order; i.e.,

$$f_{\mathrm{D},\lambda,\tau_{h_1}}(x) > f_{\mathrm{D},\lambda,\tau_{h_2}}(x).$$

This undesired phenomenon is called the *crossing problem* and is not specific to SVMs based on the pinball loss but can also occur in classical parametric quantile regression. The reason why this phenomenon can occur is that the conditional quantile functions are independently estimated. One technique to overcome this problem is to fit all m conditional quantile functions simultaneously and to add constraints that enforce that the crossing problem cannot occur at ℓ points $\{x_j \in X : j = 1, \ldots, \ell\}$. Let us write the empirical SVM solution for the quantile level τ_h as $f_{\mathrm{D},\lambda,\tau_h}(x) = \langle w(\alpha_h), \Phi(x)\rangle_H$, where $w(\alpha_h) \in H_{|X'}$ for $j = 1, \ldots, m$, $x \in X$. The non-crossing constraints can be specified as linear constraints,

$$\langle w(\alpha_h), \Phi(x_j)\rangle_H \le \langle w(\alpha_{h+1}), \Phi(x_j)\rangle_H, \quad 1 \le h \le m-1,\ 1 \le j \le \ell,$$

in $H_{|X'}$. The primal optimization problem becomes

$$\begin{aligned}
\min_{w(\alpha_h),\,\xi_h^+,\,\xi_h^-,\,1\le h\le m-1} \quad & \sum_{h=1}^{m}\left(C\sum_{i=1}^{n}\Big(\tau_h\xi_{h,i}^+ + (1-\tau_h)\xi_{h,i}^-\Big) + \frac{1}{2}\|w(\alpha_h)\|_H^2\right) \\
\text{s.t.} \quad & \xi_{h,i}^+ - \xi_{h,i}^- = y_i - \langle w(\alpha_h), \Phi(x_i)\rangle_H, \\
& \qquad 1 \le h \le m,\ 1 \le i \le n, \\
& \langle w(\alpha_{h+1}), \Phi(x_j)\rangle_H - \langle w(\alpha_h), \Phi(x_j)\rangle_H \ge 0, \qquad (11.31) \\
& \qquad 1 \le h \le m-1,\ 1 \le j \le \ell.
\end{aligned}$$

Using the Lagrangian approach, we obtain the corresponding dual problem,

$$\begin{aligned}
\max_{\alpha_h,\,\beta_h\in\mathbb{R}^n,\,1\le h\le m-1} \quad & \sum_{h=1}^{m}\Big(\alpha_h^{\mathsf{T}} y - \frac{1}{2}\alpha_h^{\mathsf{T}} K \alpha_h - \alpha_h^{\mathsf{T}}\tilde{K}(\beta_{h-1} - \beta_h) \\
& \qquad - \frac{1}{2}(\beta_{h-1} - \beta_h)^{\mathsf{T}}\bar{K}(\beta_{h-1} - \beta_h)\Big) \\
\text{s.t.} \quad & C(\tau_h - 1) \le \alpha_{h,i} \le C\tau_h, \quad 1 \le h \le m,\ 1 \le i \le n, \\
& \beta_{h,j} \ge 0, \quad 1 \le h \le m,\ 1 \le j \le \ell,
\end{aligned}$$

where $\beta_{h,j}$ is the Lagrange multiplier from (11.31), $\tilde{K} \in \mathbb{R}^{n\times\ell}$ with entries $\tilde{K}_{i,j} = k(x_i, x_j)$, $\bar{K} \in \mathbb{R}^{\ell\times\ell}$ with entries $\bar{K}_{i,j} = k(x_i, x_j)$, and $\beta_h = (\beta_{h,1}, \ldots, \beta_{h,\ell}) \in \mathbb{R}^\ell$ for $h = 1, \ldots, m$. The empirical SVM solution for the conditional quantile function for quantile level τ_h is given by

$$f_{\mathrm{D},\lambda,\tau_h} = \sum_{i=1}^{n}\alpha_{h,i}\Phi(x_i) + \sum_{j=1}^{\ell}(\beta_{h-1,i} - \beta_{h,i})\Phi(x_j)\,. \qquad \triangleleft$$

11.2 Implementation Techniques

We saw in the previous section that SVM decision functions $f_{\mathrm{D},\lambda}$ are determined via the solution of convex programs. Hence classical results about

convex programs such as Lagrange multipliers, saddle points, and algorithms to solve convex programs can be used to compute $f_{\mathrm{D},\lambda}$. For details on convex programs and related topics, we refer to Section A.6.5.

There are several standard numerical methods to solve convex programs. The Nelder-Mead search (Nelder and Mead, 1965) is a versatile optimization algorithm that can even be used to minimize continuous but non-differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}$.[1] The Nelder-Mead algorithm does not compute or approximate gradients or Hessian matrices. It is based on the iterative construction of a simplex with $(n+1)$ points, and the function values at the points of this simplex are computed. Then the points of the simplex are iteratively changed by contraction, reflection, or expansion in an adaptive way. The Nelder-Mead search is in general not fast because many function values have to be evaluated before the simplex contracts to a point (i.e., before the algorithm converges) but it is numerically stable and easy to use.

The idea of gradient descent algorithms is to start with an initial vector $\alpha^{(0)}$ for α. Then a sequence of vectors $\alpha^{(\ell)}$, $\ell \in \mathbb{N}$, is iteratively computed such that $\alpha^{(\ell+1)}$ is evaluated on the basis of $\alpha^{(\ell)}$ and $\alpha^{(\ell+1)}$ is in the direction where the gradient of $g\big(\alpha^{(\ell)}\big)$ has steepest descent.

From a numerical point of view, algorithms such as Newton-Raphson are inefficient to solve convex programs for large sample sizes n that need to store the kernel matrix K into the RAM of the computer or use matrix inversions of K. The main reason is that K is an $n \times n$ matrix and in general not sparse. Hence, even if one takes the symmetry of K into account, one has to store approximately $n(n+1)/2$ coefficients. If 8 bytes are needed to store a single coefficient in double precision, one needs approximately $4n(n+1)$ bytes to store K on a computer. Table 11.1 clearly shows that the storage space needed to store K increases substantially if n increases. For large data sets, say with at least a million data points (x_i, y_i), it is not possible to store K in the RAM of current standard PCs or on a small cluster of workstations. Data sets with more than a million data points are now not unusual in bioinformatics or data mining projects (see also Chapter 12). This is one reason why subsampling strategies such as robust learning from bites (see Section 10.5) can be useful for data sets of this size.

If the sample size n is small to moderate, *interior point algorithms* belong to the most reliable and accurate optimization techniques. The main idea of interior point algorithms when used to compute an empirical SVM solution $f_{\mathrm{D},\lambda}$ is to solve the primal and the dual programs simultaneously. This is done by gradually enforcing the Karush-Kuhn-Tucker conditions to iteratively find a feasible solution. A vector α is called a *feasible solution* of a convex program if α is an element of the set over which the optimization is carried out and if α satisfies the constraints of the convex program. The *duality gap* between the objective functions of the primal and the dual programs is used to determine the quality of the current set of variables and to check whether the stopping

[1] There are variants of the Nelder-Mead algorithm that can deal with constraints.

Table 11.1. Relationship between sample size and space needed to store K.

Sample Size n	Storage Space $4n(n+1)$
100	40 KB
1000	4 MB
10000	400 MB
100000	40 GB
1000000	4 TB

criteria are fulfilled. For large-sized optimization problems, it is sometimes necessary to use approximations of interior point algorithms.

If the sample size n is rather large, *decomposition methods* are often helpful to compute $f_{\mathrm{D},\lambda}$. As explained before, the kernel matrix $K = \big(k(x_i, x_j)\big) \in \mathbb{R}^{n\times n}$ is often fully dense and may be too large to be stored in the RAM of a computer for large values of n. Decomposition methods are designed to handle this difficulty by breaking the optimization problem into smaller and manageable subproblems and solving these in an iterative manner. This technique to tackle the numerical problem is commonly referred to as *chunking* (Vapnik, 1982) or as *subset selection*. In other words, in contrast to many optimization methods for convex problems, where the whole vector $\alpha \in \mathbb{R}^n$ of the dual problem is iteratively updated in each step, decomposition methods modify only a subset of α in each iteration step. This subset, which is generally denoted as the *working set*

$$B := \{\alpha_j : j \in J \subset \{1, \ldots, n\}, |J| = q\},$$

leads to a relatively small subproblem to be minimized in each iteration step.

The idea of *sequential minimal optimization* proposed by Platt (1999) is to use the decomposition method in an extreme manner: the optimization step is done for only $q = 2$ points at each iteration step. At first sight, this approach might look too simple to be useful, but the opposite is true. SMO can be very effective for SVMs because the optimization problem for only two points can often be calculated analytically. This eliminates calling an iterative convex program optimizer at each iteration step, which can therefore save a lot of computation time. To describe this method, let us consider the convex problem

$$\min_{\alpha\in(0,C)^n} \quad C\sum_{i=1}^{n} G\Big(\frac{\alpha_i}{C}\Big) + \frac{1}{2}\alpha^{\mathsf{T}} K\alpha \tag{11.32}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \tag{11.33}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)$, G is some real-valued convex function, and $K = \big(k(x_i, x_j)\big) \in \mathbb{R}^{n\times n}$ denotes a positive definite kernel matrix. This problem

is typical for SVMs; see Example 11.4 and Exercise 11.3 on SVMs based on the logistic loss for classification problems. The basic algorithm for an SMO-type decomposition with $|B| = q = 2$ is then given by the following four steps.

Sequential Minimal Optimization Algorithm (SMO)

i) Find an initial feasible solution $\alpha^{(1)} \in \mathbb{R}^n$ fulfilling the constraints and set $\ell = 1$.

ii) If $\alpha^{(\ell)}$ solves the dual program with the desired numerical precision, then stop. Otherwise, find a working set $B^{(\ell)} \subset \{1, \ldots, n\}$ with $|B^{(\ell)}| = 2$.

iii) Define $A^{(\ell)} := \{\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n : \alpha_j = \alpha_j^{(\ell)} \text{ for } j \notin B^{(\ell)}\}$. Solve the following subproblem with respect to the two variables α_j, $j \in B^{(\ell)}$:

$$\begin{aligned} \min_{\alpha \in A^{(\ell)}} \quad & C \sum_{i=1}^{n} G\Big(\frac{\alpha_i}{C}\Big) + \frac{1}{2} \alpha^{\mathsf{T}} K \alpha \\ \text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0. \end{aligned}$$

This optimization problem is equivalent to

$$\begin{aligned} \min_{\alpha \in A^{(\ell)}} \quad & C \sum_{i \in B^{(\ell)}} G(\alpha_i/C) + \frac{1}{2} \sum_{i \in B^{(\ell)}} \sum_{j \in B^{(\ell)}} \alpha_i \alpha_j k(x_i, x_j) + c_*^{(\ell)} + \sum_{i \in B^{(\ell)}} \alpha_i c_i^{(\ell)} \\ \text{s.t.} \quad & \textstyle\sum_{i \in B^{(\ell)}} \alpha_i y_i = c_{**}^{(\ell)}, \end{aligned}$$

where

$$c_*^{(\ell)} = C \sum_{i \notin B^{(\ell)}} G(\alpha_i/C) + \frac{1}{2} \sum_{i \notin B^{(\ell)}} \sum_{j \notin B^{(\ell)}} \alpha_i \alpha_j k(x_i, x_j), \tag{11.34}$$

$$c_i^{(\ell)} = \sum_{j \notin B^{(\ell)}} \alpha_j k(x_i, x_j), \quad i \in B^{(\ell)}, \tag{11.35}$$

$$c_{**}^{(\ell)} = \sum_{j \notin B^{(\ell)}} \alpha_j y_j \,. \tag{11.36}$$

iv) Set $\alpha^{(\ell+1)}$ to be an optimal solution of the optimization problem given in step *iii)*. If the stopping rules of the algorithm are not yet met, increase ℓ to $\ell + 1$ and go to step *ii)*.

Of course, two problems are left open in this SMO-type decomposition algorithm: how to select the working sets and how to specify stopping rules. However, before we address these questions, let us first give an example for the SMO-type decomposition for the special case of the hinge loss. Let us further assume that there is a bias term $b \in \mathbb{R}$ to be estimated. The dual problem is thus

$$\max_{\alpha\in\mathbb{R}^n} \ \alpha^{\mathsf{T}}1 - \frac{1}{2}\alpha^{\mathsf{T}}\tilde{K}\alpha \tag{11.37}$$

$$\text{s.t.} \quad \alpha \in [0,C]^n \text{ and } \alpha^{\mathsf{T}}y = 0, \tag{11.38}$$

where C is the upper bound, $1 := (1,\ldots,1) \in \mathbb{R}^n$, $y = (y_1,\ldots,y_n) \in \{-1,+1\}^n$, and $\tilde{K} := \big(y_i y_j k(x_i,x_j)\big) \in \mathbb{R}^{n\times n}$ (see Exercise 11.2). Let $B \subset \{1,\ldots,n\}$ and denote by 1_B and 1_{B^c} vectors with $|B|$ and $|B^c|$ coefficients, respectively, all being equal to one. The first algorithm for an SMO-type decomposition is then given by the following algorithm.

Sequential Minimal Optimization Algorithm 1 (ALG1)

i) Find $\alpha^{(1)} \in \mathbb{R}^n$ as an initial feasible solution. Set $\ell = 1$.

ii) If $\alpha^{(\ell)}$ solves the dual program up to the desired numerical precision, stop. Otherwise, find a working set $B := \{i,j\} \subset \{1,\ldots,n\}$ with $|B| = 2$. Define the complement of B by $B^c = \{1,\ldots,n\}\backslash B$ and $\alpha_B^{(\ell)}$ and $\alpha_{B^c}^{(\ell)}$ being the subvectors of $\alpha^{(\ell)}$ with index sets B and B^c, respectively.

iii) Solve the following subproblem with the dual variable α_B, where we use obvious matrix notation:

$$\begin{aligned}
\max_{\alpha_B\in\mathbb{R}^2} \quad & \begin{bmatrix}\alpha_B^{\mathsf{T}} & (\alpha_{B^c}^{(\ell)})^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}1_B\\ 1_{B^c}\end{bmatrix} - \frac{1}{2}\begin{bmatrix}\alpha_B^{\mathsf{T}} & (\alpha_{B^c}^{(\ell)})^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}\tilde{K}_{B,B} & \tilde{K}_{B,B^c}\\ \tilde{K}_{B^c,B} & \tilde{K}_{B^c,B^c}\end{bmatrix}\begin{bmatrix}\alpha_B\\ \alpha_{B^c}^{(\ell)}\end{bmatrix} \\
&= \big(1_B - \tilde{K}_{B,B^c}\alpha_{B^c}^{(\ell)}\big)\alpha_B - \frac{1}{2}\alpha_B^{\mathsf{T}}\,\tilde{K}_{B,B}\,\alpha_B \ + \ c \\
&= \big(1_B - \tilde{K}_{B,B^c}\alpha_{B^c}^{(\ell)}\big)\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} - \frac{1}{2}[\alpha_i \ \ \alpha_j]\begin{bmatrix}\tilde{K}_{i,i} & \tilde{K}_{i,j}\\ \tilde{K}_{j,i} & \tilde{K}_{j,j}\end{bmatrix}\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} + c \\
\text{s.t.} \quad & \alpha_B \in [0,C]^2, \ \ \alpha_i y_i + \alpha_j y_j = -\sum\nolimits_{i\in B^c}\alpha_i^{(\ell)} y_i\,,
\end{aligned}$$

where c is some constant independent of α_B and

$$\begin{bmatrix}\tilde{K}_{B,B} & \tilde{K}_{B,B^c}\\ \tilde{K}_{B^c,B} & \tilde{K}_{B^c,B^c}\end{bmatrix}$$

is a permutation of K where rows and columns of K are permuted such that $\tilde{K}_{B,B} \in \mathbb{R}^{2\times 2}$ contains the elements of K with indexes $(i,j) \in B\times B$. The submatrices $\tilde{K}_{B,B^c}$, $\tilde{K}_{B^c,B}$, and $\tilde{K}_{B^c,B^c}$ are analogously constructed.

iv) Set $\alpha_B^{(\ell+1)}$ to be an optimal solution of the optimization problem given in step *iii)* and set $\alpha_{B^c}^{(\ell+1)} := \alpha_{B^c}^{(\ell)}$. If the stopping rules of the algorithm are not yet met, then increase ℓ to $\ell+1$ and go to step *ii)*.

Note that we denoted the working set by B instead of $B^{(\ell)}$ at every iteration step because misunderstandings are unlikely, although the working set changes from one iteration to another.

Now some results concerning the *working set selection* will be given. The decomposition method clearly has the advantage compared to some other

techniques to solve convex programs that this method can be used for data sets with large sample sizes n without storing the kernel matrix $K \in \mathbb{R}^{n\times n}$ or $\tilde{K} \in \mathbb{R}^{n\times n}$ into the RAM of the computer. However, since only q components of $\alpha \in \mathbb{R}^n$ with $q \ll n$ are updated per iteration, the decomposition method can suffer from slow convergence if n is large. In other words, even if any single iteration step can be done fast, the overall computation time can be long if many iteration steps are necessary. Hence it is essential to choose the working sets in a suitable way to reduce the number of iterations and the computation time. Some methods rely on a violation of the optimality condition for the subproblems. Several gradient-based methods belong to this class of methods. Recent research indicates that using a second-order approximation of the objective function to be optimized in the subproblems generally leads to faster overall convergence.

We will now describe an algorithm for SVMs based on the hinge loss. The dual program given in (11.37) and (11.38) is a quadratic program. Therefore, a second-order approximation

$$g : \mathbb{R}^n \to \mathbb{R}, \quad g(\alpha) := \frac{1}{2}\alpha^\mathsf{T}\tilde{K}\alpha - \alpha^\mathsf{T}1, \quad \alpha \in \mathbb{R}^n, \tag{11.39}$$

of the convex objective function to be minimized in the subproblems directly relates to a decrease of the objective function. Let us denote the gradient of $g(\alpha)$ by $\nabla g(\alpha) = \tilde{K}\alpha - 1$. To shorten the notation, we will often write $\nabla g(\alpha)_i$ instead of $\big(\nabla g(\alpha)\big)_i$, $i = 1, \ldots, n$.

One way to select the working set B is via a maximal violating pair that will formally be defined in Definition 11.10. Often we will write B instead of $B^{(\ell)}$ if misunderstandings are unlikely. The *maximal violating pair algorithm* (Keerthi *et al.*, 2001) can be described as follows.

Maximal Violating Pair Algorithm (ALG2)

i) Select two indexes $i, j \in \{1, \ldots, n\}$ such that

$$\begin{aligned} i &\in \arg\max\big\{-y_h \nabla g(\alpha^{(\ell)})_h : h \in I_{up}(\alpha^{(\ell)})\big\}, \\ j &\in \arg\min\big\{-y_h \nabla g(\alpha^{(\ell)})_h : h \in I_{low}(\alpha^{(\ell)})\big\}, \end{aligned}$$

where

$$\begin{aligned} I_{up}(\alpha) &:= \big\{h \in \{1, \ldots, n\} : \alpha_h > 0, y_h = -1 \text{ or } \alpha_h < C, y_h = +1\big\}, \\ I_{low}(\alpha) &:= \big\{h \in \{1, \ldots, n\} : \alpha_h > 0, y_h = +1 \text{ or } \alpha_h < C, y_h = -1\big\}. \end{aligned}$$

ii) Choose the working set $B := \{i, j\}$.

This technique to choose B is motivated by the Karush-Kuhn-Tucker conditions (see Section A.6.5): a vector $\alpha \in \mathbb{R}^n$ is a stationary point of the dual program (11.37) and (11.38) if and only if there exists a number $b \in \mathbb{R}$ and two vectors $\eta = (\eta_1, \ldots, \eta_n) \in [0, \infty)^n$ and $\mu = (\mu_1, \ldots, \mu_n) \in [0, \infty)^n$ such that

$$\begin{aligned} \nabla g(\alpha) + by &= \eta - \mu\,, \\ \eta_i \alpha_i &= 0, \\ \mu_i (C - \alpha_i) &= 0, \quad i = 1, \ldots, n. \end{aligned}$$

Note that we can rewrite this condition as

$$\begin{aligned} \nabla g(\alpha)_i + by_i &\geq 0, \text{ if } \alpha_i < C, \\ \nabla g(\alpha)_i + by_i &\leq 0, \text{ if } \alpha_i > 0, \end{aligned}$$

which is equivalent to

$$\begin{aligned} -y_i \nabla g(\alpha)_i &\leq b, \text{ if } i \in I_{up}(\alpha), \\ -y_i \nabla g(\alpha)_i &\geq b, \text{ if } i \in I_{low}(\alpha). \end{aligned}$$

Using these relations together with $Y = \{-1, +1\}$, we see that a feasible vector $a \in \mathbb{R}^n$ is a stationary point of (11.37) and (11.38) if and only if

$$m(\alpha) := \max_{i \in I_{up}(\alpha)} -y_i \nabla g(\alpha)_i \leq \min_{i \in I_{low}(\alpha)} -y_i \nabla g(\alpha)_i =: M(\alpha)\,. \tag{11.40}$$

Note that the maximum and the minimum in (11.40) are well-defined except for the case where $\sum_{i=1}^n y_i \in \{-n, +n\}$. For these special cases, the vector $(0, \ldots, 0) \in \mathbb{R}^n$ is the only feasible solution, and the iterative decomposition method already stops at the first step, provided we initialize with $(0, \ldots, 0)$.

Definition 11.10. *Consider a support vector machine based on the hinge loss with dual problem (11.37) and (11.38).*

i) A pair of indexes $\{i, j\} \subset \{1, \ldots, n\}$ with $i \neq j$ is called a ***violating pair*** *if $i \in I_{up}(\alpha)$, $j \in I_{low}(\alpha)$, and $y_i \nabla g(\alpha)_i < y_j \nabla g(\alpha)_j$.*

ii) A ***maximal violating pair*** *is a violating pair such that the difference $y_j \nabla g(\alpha)_j - y_i \nabla g(\alpha)_i$ is maximized over all violating pairs.*

A maximal violating pair is clearly a plausible choice of the working set B. Hush and Scovel (2003) showed for $\tilde{K}$ positive semi-definite that the sequence $\big(g(\alpha^{(\ell)})\big)_{\ell \in \mathbb{N}}$ strictly decreases for SMO-type methods if and only if the working set $B^{(\ell)}$ is a violating pair in each iteration. Unfortunately, having a violating pair even a strict decrease of $\big(g(\alpha^{(\ell)})\big)_{\ell \in \mathbb{N}}$ does not guarantee the convergence to a stationary point for $\ell \to \infty$. There exist counterexamples[2] if the working set B is a violating pair but not a maximal violating pair.

In the following we will show that the maximal violating pair is related to a *first-order approximation* of the objective function $g(\alpha)$ for an empirical SVM solution for classification based on the hinge loss having the dual problem (11.37) and (11.38). More precisely, the pair $\{i, j\}$ selected by the maximal violating pair algorithm satisfies

[2] See, e.g., Chen *et al.* (2006, pp. 895ff.).

$$\{i, j\} = \arg\min_B \mathrm{Sub}(B), \tag{11.41}$$

where the subproblem $\mathrm{Sub}(B)$ is defined by

$$\begin{aligned}
\mathrm{Sub}(B) := \min_{d_B} & \big(\nabla g(\alpha^{(\ell)})\big)_B^{\mathsf{T}} d_B && (11.42)\\
\text{s.t. } & d_B^{\mathsf{T}} y_B = 0, && (11.43)\\
& d_h \geq 0, \text{ if } \alpha_h^{(\ell)} = 0,\ h \in B, && (11.44)\\
& d_h \leq 0, \text{ if } \alpha_h^{(\ell)} = C,\ h \in B, && (11.45)\\
& d_h \in [-1, +1],\ h \in B, && (11.46)
\end{aligned}$$

the minimization of $\mathrm{Sub}(B)$ is over all subsets $B \subset \{1, \ldots, n\}$ with $|B| = 2$, and $\alpha_B^{(\ell)}, d_B, y_B \in \mathbb{R}^2$ are the subvectors of $\alpha^{(\ell)}, d, y \in \mathbb{R}^n$, respectively, where only coefficients with indexes in B are considered.

Now we will show that a maximal violating pair solves (11.41), provided that there exists at least one violating pair. Let us define $d := [d_B,\, 0_{B^c}] \in \mathbb{R}^n$. Then the objective function in (11.42) is obtained as a first-order approximation of $g(\alpha^{(\ell)} + d)$ because

$$g(\alpha^{(\ell)} + d) \;\approx\; g(\alpha^{(\ell)}) + \big(\nabla g(\alpha^{(\ell)})\big)^{\mathsf{T}} d \;=\; g(\alpha^{(\ell)}) + \big(\nabla g(\alpha^{(\ell)})\big)_B^{\mathsf{T}} d_B\,. \tag{11.47}$$

The constraint in (11.43) is from $(\alpha^{(\ell)} + d)^{\mathsf{T}} y = 0$ and $(\alpha^{(\ell)})^{\mathsf{T}} y = 0$. The condition $\alpha \in [0, C]^n$ leads to the inequalities (11.44) and (11.45), and (11.46) avoids that the value of the objective function approaches $-\infty$ because the function in (11.42) is linear. It is not necessary to consider all $n(n-1)/2$ possible subsets of $\{1, \ldots, n\}$ having two elements to find the optimal subset because the maximal violating pair algorithm solves (11.41) in $\mathcal{O}(n)$ steps, as can be seen as follows. For any set $\{i, j\} \subset \{1, \ldots, n\}$ with $i \neq j$, define $\hat{d}_i := y_i d_i$ and $\hat{d}_j := y_j d_j$ in (11.42)–(11.46). The objective function becomes

$$\big(-y_i \nabla g(\alpha^{(\ell)})_i + y_j \nabla g(\alpha^{(\ell)})_j\big)\hat{d}_j\,. \tag{11.48}$$

As $d_i = d_j = 0$ is feasible for (11.42)–(11.46), the minimum of (11.48) is less than or equal to zero. If $y_i \nabla g(\alpha^{(\ell)})_i < y_j \nabla g(\alpha^{(\ell)})_j$, then the term in (11.48) is negative if and only if $\hat{d}_j < 0$ and $\hat{d}_i > 0$ because $\hat{d}_i + \hat{d}_j = 0$. From the definition of $I_{low}(\alpha)$ and $I_{up}(\alpha)$ and (11.44) and (11.45), this corresponds to $i \in I_{up}(\alpha^{(\ell)})$ and $j \in I_{low}(\alpha^{(\ell)})$. Furthermore, the minimum occurs at $\hat{d}_i = -\hat{d}_j = 1$. Similar relations hold for $y_i \nabla g(\alpha^{(\ell)})_i > y_j \nabla g(\alpha^{(\ell)})_j$. Hence, solving (11.41) is essentially the same as solving

$$\begin{aligned}
&\min\Big\{\min\big\{0, y_i \nabla g(\alpha^{(\ell)})_i - y_j \nabla g(\alpha^{(\ell)})_j\big\} : i \in I_{up}(\alpha^{(\ell)}), j \in I_{low}(\alpha^{(\ell)})\Big\}\\
&= \min\Big\{0, -\max_{i \in I_{up}(\alpha^{(\ell)})} \big(-y_i \nabla g(\alpha^{(\ell)})_i\big) + \min_{i \in I_{low}(\alpha^{(\ell)})} \big(-y_i \nabla g(\alpha^{(\ell)})_i\big)\Big\}\,.
\end{aligned}$$

If we take (11.40) into account, we obtain that a maximal violating pair solves (11.41), provided that there exists at least one violating pair.

We continue with considering the dual program (11.37) and (11.38) for $f_{\mathrm{D},\lambda}$ based on the hinge loss. The objective function g in (11.39) is quadratic, and

$$\begin{aligned} g(\alpha^{(\ell)} + d) - g(\alpha^{(\ell)}) &= \big(\nabla g(\alpha^{(\ell)})\big)^{\mathsf{T}} d + \frac{1}{2} d^{\mathsf{T}} \nabla^2 g(\alpha^{(\ell)}) d \\ &= \big(\nabla g(\alpha^{(\ell)})\big)_B^{\mathsf{T}} d_B + \frac{1}{2} d_B^{\mathsf{T}} \big(\nabla^2 g(\alpha^{(\ell)})\big)_{B,B} d_B \end{aligned} \tag{11.49}$$

equals the reduction of the value of the objective function. Therefore, we obtain a selection method for the working set, taking the result of a second-order approximation into account if we replace the objective function in (11.42) by (11.49). We obtain the following *second-order approximation algorithm*

$$\{i, j\} = \arg\min_B \mathrm{Sub}(B), \tag{11.50}$$

where the subproblem $\mathrm{Sub}(B)$ is given by

$$\mathrm{Sub}(B) := \min_{d_B} \frac{1}{2} d_B^{\mathsf{T}} \big(\nabla^2 g(\alpha^{(\ell)})\big)_{B,B} d_B + \big(\nabla g(\alpha^{(\ell)})\big)_B^{\mathsf{T}} d_B \tag{11.51}$$

$$\text{s.t. } d_B^{\mathsf{T}} y_B = 0, \tag{11.52}$$

$$d_h \geq 0, \text{ if } \alpha_h^{(\ell)} = 0,\ h \in B, \tag{11.53}$$

$$d_h \leq 0, \text{ if } \alpha_h^{(\ell)} = C,\ h \in B, \tag{11.54}$$

the minimum is taken over all sets $B \subset \{1, \ldots, n\}$ with $|B| = 2$, and $\alpha_B^{(\ell)}, d_B, y_B \in \mathbb{R}^2$ are the subvectors of $\alpha^{(\ell)}, d, y \in \mathbb{R}^n$, respectively, where only coefficients with indexes in B are considered. The box constraints from (11.46) are removed because it will later be shown that the optimal value does not converge to $-\infty$. One hopes that (11.51)–(11.54) outperforms (11.42)–(11.46), but there seems to be no way that avoids considering all working sets having two elements to apply the second-order approximation algorithm. Hence the following heuristic method (Fan *et al.*, 2005) for an implementation of a second-order approximation is promising because not all possible working sets are considered.

Working Set Selection Algorithm (ALG3)

i) Select $i \in \arg\max_h \big\{-y_h \nabla g(\alpha^{(\ell)})_h : h \in I_{up}(\alpha^{(\ell)})\big\}$.
ii) Consider $\mathrm{Sub}(B)$ as defined in (11.51)–(11.54) and select

$$j \in \arg\min_h \big\{\mathrm{Sub}(\{i, h\}) : h \in I_{low}(\alpha^{(\ell)}), y_h \nabla g(\alpha^{(\ell)})_h > y_i \nabla g(\alpha^{(\ell)})_i\big\}. \tag{11.55}$$

iii) Choose $B := \{i, j\}$ as the working set.

If we use the same index i as in ALG2, we only have to check $\mathcal{O}(n)$ possible candidates for the working set B to choose the index j. An alternative is to choose $j \in \arg M(\alpha^{(\ell)})$ and search for the index i by a way similar to (11.55).

The following theorem shows that one can analytically solve (11.51)–(11.54) such that the working set selection method ALG3 does not cost much more than the algorithm ALG2. Recall that $K = \big(k(x_i, x_j)\big) \in \mathbb{R}^{n\times n}$ denotes the kernel matrix and that $\tilde{K} := \big(y_i y_j k(x_i, x_j)\big) \in \mathbb{R}^{n\times n}$. We denote the (i,j)-elements of these matrices by $K_{i,j}$ and $\tilde{K}_{i,j}$, respectively.

Theorem 11.11. *If L is the hinge loss, $B = \{i, j\}$ is a violating pair, and $b_{i,j} := K_{i,i} + K_{j,j} - 2K_{i,j} > 0$, then (11.51)–(11.54) have the optimal value*

$$-\frac{\big(y_j \nabla g(\alpha^{(\ell)})_j - y_i \nabla g(\alpha^{(\ell)})_i\big)^2}{2\, b_{i,j}}$$

of the objective function.

Proof. Define $\hat{d}_i := y_i d_i$ and $\hat{d}_j := y_j d_j$. From (11.52), we obtain $\hat{d}_i = -\hat{d}_j$, and the objective function in (11.51) becomes

$$\begin{aligned} &\frac{1}{2}\begin{bmatrix} d_i & d_j \end{bmatrix} \begin{bmatrix} \tilde{K}_{i,i} & \tilde{K}_{i,j} \\ \tilde{K}_{j,i} & \tilde{K}_{j,j} \end{bmatrix} \begin{bmatrix} d_i \\ d_j \end{bmatrix} + \begin{bmatrix} \nabla g(\alpha^{(\ell)})_i & \nabla g(\alpha^{(\ell)})_j \end{bmatrix} \begin{bmatrix} d_i \\ d_j \end{bmatrix} \\ &= \frac{1}{2}(K_{i,i} + K_{j,j} - 2K_{i,j})\hat{d}_j^2 + \big(y_j \nabla g(\alpha^{(\ell)})_j - y_i \nabla g(\alpha^{(\ell)})_i\big)\hat{d}_j\,. \end{aligned} \tag{11.56}$$

Since $K_{i,i} + K_{j,j} - 2K_{i,j} > 0$ and B is a violating pair, we define the positive constants

$$b_{i,j} := K_{i,i} + K_{j,j} - 2K_{i,j} \quad \text{and} \quad c_{i,j} := y_j \nabla g(\alpha^{(\ell)})_j - y_i \nabla g(\alpha^{(\ell)})_i\,. \tag{11.57}$$

Thus, the function in (11.56) has a minimum at

$$\hat{d}_j = -\hat{d}_i = -\frac{c_{i,j}}{b_{i,j}} < 0, \tag{11.58}$$

and the value of the objective function $\mathrm{Sub}(B)$ in (11.51) equals $-c_{i,j}^2/(2b_{i,j})$. Moreover, $\hat{d}_i$ and $\hat{d}_j$ fulfill the constraints (11.53) and (11.54). Indeed, we obtain for the case $j \in I_{low}(\alpha^{(\ell)})$ that $\alpha_j^{(\ell)} = 0$ implies $y_j = -1$ and hence $d_j = y_j \hat{d}_j > 0$. This condition is required by (11.53). The other cases can be treated in a similar way. Thus, $\hat{d}_i$ and $\hat{d}_j$ from (11.58) are optimal for (11.51)–(11.54). □

If the kernel matrix K is strictly positive definite, then the condition $K_{i,i} + K_{j,j} - 2K_{i,j} > 0$ of Theorem 11.11 is valid for any pair $\{i, j\}$ with $i \neq j$. Note that Theorem 11.11 enables us to write (11.55) as

$$j \in \arg\min_h \Big\{ -\frac{c_{i,h}^2}{b_{i,h}} \,:\, h \in I_{low}(\alpha^{(\ell)}),\, y_h \nabla g(\alpha^{(\ell)})_h > y_i \nabla g(\alpha^{(\ell)})_i \Big\},$$

where $b_{i,h}$ and $c_{i,h}$ are the constants defined in (11.57). If K is not strictly positive definite, $b_{i,j} = 0$ can occur. The following algorithm will address this situation.

Note that (11.42)–(11.46) and (11.51)–(11.54) are only used to select the working set B. Hence they do not have to fulfill the feasibility condition $\alpha_i^{(\ell)} + d_i \in [0, C]$ for all $i \in B$. However, feasibility must hold for the subproblem in step *iii)* of ALG1 used to compute $\alpha^{(\ell+1)}$ after the working set is determined.

Now we will consider a general working set algorithm (Chen *et al.*, 2006).

Working Set Selection Algorithm (ALG4)

i) Let $h^* : \mathbb{R} \to \mathbb{R}$ be a measurable function that is strictly increasing on the interval $[0, \infty)$ and fulfills $h^*(0) = 0$ and $h^*(z) \leq z$ for $z \in [0, \infty)$.

ii) Select $B = \{i, j\}$ as the working set if $i \in I_{up}(\alpha^{(\ell)})$, $j \in I_{low}(\alpha^{(\ell)})$, and

$$y_j \nabla g(\alpha^{(\ell)})_j - y_i \nabla g(\alpha^{(\ell)})_i \geq h^*\big(m(\alpha^{(\ell)}) - M(\alpha^{(\ell)})\big) > 0. \tag{11.59}$$

A special case of ALG4 is obtained by choosing $h(z) := \sigma z$, $z \in \mathbb{R}$, where $\sigma \in (0, 1]$ is fixed. In other words, ALG4 uses in this case a "constant-factor" violating pair as the working set.

Let us now consider the case where the kernel matrix K may not be strictly positive definite; e.g., $b_{i,j} = 0$ can happen for a linear kernel. The following algorithm (Chen *et al.*, 2006) for an SMO-type decomposition for the case of classification based on the hinge loss consists of four steps. The main idea of this approach is to slightly modify the subproblems if necessary (see (11.61)) to obtain an algorithm with nice properties.

Working Set Selection Algorithm (ALG5)

The steps *i)*, *ii)*, and *iv)* are the same as those in ALG1, but step *iii)* is replaced by the following:

iii′) Let τ be a small positive number being constant for all iteration steps, and define

$$\tilde{b}_{i,j} := \tilde{K}_{i,i} + \tilde{K}_{j,j} - 2 y_i y_j \tilde{K}_{i,j}\,. \tag{11.60}$$

If $\tilde{b}_{i,j} > 0$, then solve the subproblem from step *iii)* of the algorithm ALG1 and set $\alpha_B^{(\ell+1)}$ to be the optimal point of this subproblem. If $\tilde{b}_{i,j} \leq 0$, then solve the modified subproblem

$$\begin{aligned} \max_{(\alpha_i, \alpha_j) \in \mathbb{R}^2} \quad & \big(1_B - \tilde{K}_{B,B^c} \alpha_{B^c}^{(\ell)}\big)^{\mathsf{T}} \begin{bmatrix} \alpha_i \\ \alpha_j \end{bmatrix} - \frac{1}{2} [\alpha_i \;\; \alpha_j] \begin{bmatrix} \tilde{K}_{i,i} & \tilde{K}_{i,j} \\ \tilde{K}_{j,i} & \tilde{K}_{j,j} \end{bmatrix} \begin{bmatrix} \alpha_i \\ \alpha_j \end{bmatrix} \\ & - \frac{\tau - \tilde{b}_{i,j}}{4} \Big((\alpha_i - \alpha_i^{(\ell)})^2 + (\alpha_j - \alpha_j^{(\ell)})^2 \Big) \\ \text{s.t.} \quad & (\alpha_i, \alpha_j) \in [0, C]^2, \quad \alpha_i y_i + \alpha_j y_j = -\sum\nolimits_{i \in B^c} \alpha_i^{(\ell)} y_i\,, \end{aligned} \tag{11.61}$$

and set $\alpha_B^{(\ell+1)}$ to be the optimal point of this subproblem.

Let us first show how the subproblem from step *iii′)* of ALG2 can be solved. Using $d_i = -d_j$, we obtain

$$\frac{\tau - \tilde{b}_{i,j}}{4} \left\| \alpha_B - \alpha_B^{(\ell)} \right\|^2 = \frac{\tau - \tilde{b}_{i,j}}{2} d_j^2 \,. \tag{11.62}$$

Define

$$\tilde{b}^*_{i,j} := \begin{cases} \tilde{b}_{i,j} & \text{if } \tilde{b}_{i,j} > 0 \\ \tau & \text{otherwise,} \end{cases} \tag{11.63}$$

and

$$\tilde{c}^*_{i,j} := y_j \nabla g(\alpha^{(\ell)})_j - y_i \nabla g(\alpha^{(\ell)})_i > 0. \tag{11.64}$$

Hence (11.61) is essentially a strictly convex optimization problem of the form

$$\begin{aligned} \min_{d_j} \quad & \frac{1}{2} \tilde{b}^*_{i,j} d_j^2 + \tilde{c}^*_{i,j} d_j \\ \text{s.t.} \quad & c_{low} \le d_j \le c_{up} \,, \end{aligned} \tag{11.65}$$

where $\tilde{b}^*_{i,j}, \tilde{c}^*_{i,j} > 0$, $c_{low} < 0$, and $c_{up} \ge 0$. The optimum of the quadratic objective function is

$$\bar{d}_j \;=\; \max\bigl\{c_{low}, -\tilde{c}^*_{i,j}/\tilde{b}^*_{i,j}\bigr\} \;<\; 0. \tag{11.66}$$

Therefore, once $\tilde{b}^*_{i,j}$ is defined in (11.63), both subproblems can easily be solved no matter whether $\tilde{b}^*_{i,j}$ is positive or not.

Let us now check whether the value of the objective function actually decreases if ℓ increases. Some calculations using $\bar{d}_j < 0$, $\tilde{b}^*_{i,j}\bar{d}_j + \tilde{c}^*_{i,j} \ge 0$ due to (11.66) and $\|\alpha^{(\ell+1)} - \alpha^{(\ell)}\|_2^2 = 2\bar{d}_j^2$ yield that $g(\alpha) - g(\alpha^{(\ell)})$ equals

$$\frac{\bar{d}_j^2}{2}\bigl(\tilde{K}_{i,i} + \tilde{K}_{j,j} - 2y_i y_j \tilde{K}_{i,j}\bigr)^2 + \bigl(y_j \nabla g(\alpha^{(\ell)})_j - y_i \nabla g(\alpha^{(\ell)})_i\bigr) d_j \tag{11.67}$$

and

$$\begin{aligned} g(\alpha^{(\ell+1)}) - g(\alpha^{(\ell)}) = g(\bar{d}_j) = \frac{1}{2}\tilde{b}^*_{i,j}\bar{d}_j^2 + \tilde{c}^*_{i,j}\bar{d}_j \;=\; (\tilde{b}^*_{i,j}\bar{d}_j + \tilde{c}^*_{i,j})\bar{d}_j - \frac{\tilde{b}^*_{i,j}}{2}\,\bar{d}_j^2 \\ \le -\frac{\tilde{b}^*_{i,j}}{2}\,\bar{d}_j^2 \;=\; -\frac{\tilde{b}^*_{i,j}}{4}\left\|\alpha^{(\ell+1)} - \alpha^{(\ell)}\right\|_2^2, \quad \ell \in \mathbb{N}. \end{aligned}$$

We thus obtain the following result.

Lemma 11.12. *Suppose that L is the hinge loss and that the working set $B^{(\ell)}$ in algorithm ALG5 is a violating pair for all $\ell \in \mathbb{N}$, and let $\bigl(\alpha^{(\ell)}\bigr)_{\ell\in\mathbb{N}}$ be the sequence generated by this algorithm. Then*

$$g(\alpha^{(\ell+1)}) < g(\alpha^{(\ell)}) - \delta\left\|\alpha^{(\ell+1)} - \alpha^{(\ell)}\right\|_2^2, \quad \ell \in \mathbb{N}, \tag{11.68}$$

holds with

$$\delta := \frac{1}{4}\min\bigl\{\tau, \min\{\tilde{b}_{i,j} : \tilde{b}_{i,j} > 0\}\bigr\}. \tag{11.69}$$

Theorem 11.13. *Let $L = L_{\text{hinge}}$, and assume that $\tilde{K} \in \mathbb{R}^{n\times n}$ is positive semi-definite, the working set of algorithm ALG5 is a violating pair, and $\tau \leq \frac{2}{C}$. If $\tilde{b}_{i,j} = 0$, then the optimal solution of the subproblem in step iii′) from ALG5 is the same as the optimal solution of the subproblem in step iii) from ALG1.*

Proof. Suppose $\tilde{b}_{i,j} = 0$. From (11.67) and $\tilde{c}^*_{i,j} > 0$, it follows that the subproblem from step *iii)* has the optimum at $\bar{d}_j = c_{low}$. For the subproblem from step *iii′)* and the problem in (11.65), it follows from $\tilde{b}_{i,j} = 0$ that $\tilde{b}^*_{i,j} = \tau$. Since $\tilde{K}$ is positive definite by assumption, we obtain

$$0 = \tilde{b}_{i,j} = \tilde{K}_{i,i} + \tilde{K}_{j,j} - 2y_iy_j\tilde{K}_{i,j} = \|\Phi(x_i) - \Phi(x_j)\|^2_H\,,$$

which gives $\Phi(x_i) = \Phi(x_j)$. Hence $k(x_i, x_h) = k(x_j, x_h)$ for all $h = 1, \ldots, n$ implies

$$\begin{aligned} &y_j\nabla g(\alpha^{(\ell)})_j - y_i\nabla g(\alpha^{(\ell)})_i \\ &= \sum_{h=1}^{n} y_h k(x_j, x_h)\alpha_h^{(\ell)} - y_j - \sum_{h=1}^{n} y_h k(x_i, x_h)\alpha_h^{(\ell)} + y_i \;=\; y_i - y_j\,. \end{aligned}$$

Now $\{i, j\}$ is a violating pair. Hence $y_i - y_j > 0$ implies $\tilde{c}^*_{i,j} = y_i - y_j = 2$. As $\tau \leq \frac{2}{C}$ by assumption, (11.66) implies that $\bar{d}_j = c_{low}$ is the solution for the step *iii′)*. Hence both subproblems have the same optimal point. □

Before we can state a result concerning the asymptotic convergence of ALG5, we need the following result.

Lemma 11.14. *Let L be the hinge loss, and assume that the working set in each iteration of ALG5 is a violating pair. If a subsequence $\left(\alpha^{(\ell)}\right)_{\ell\in J}$, $J \subset \mathbb{N}$, converges to $\bar{\alpha}$, then for any given $s \in \mathbb{N}$, the sequence $\left(\alpha^{(\ell+s)}\right)_{\ell\in J}$ converges to $\bar{\alpha}$ as well.*

Proof. For the subsequence $\left(\alpha^{(\ell+1)}\right)_{\ell\in J}$ from $\left(\alpha^{(\ell)}\right)_{\ell\in\mathbb{N}}$ of Lemma 11.12 and the fact that the sequence $\left(g(\alpha^{(\ell)})\right)_{\ell\in\mathbb{N}}$ is bounded and decreasing, we obtain

$$\begin{aligned} \lim_{\ell\in J,\,\ell\to\infty} \|\alpha^{(\ell+1)} - \bar{\alpha}\| \;&\leq\; \lim_{\ell\in J,\,\ell\to\infty} \left(\|\alpha^{(\ell+1)} - \alpha^{(\ell)}\| + \|\alpha^{(\ell)} - \bar{\alpha}\|\right) \\ \leq \lim_{\ell\in J,\,\ell\to\infty} &\left(\delta^{-1/2}\left(g(\alpha^{(\ell)}) - g(\alpha^{(\ell+1)})\right)^{1/2} + \|\alpha^{(\ell)} - \bar{\alpha}\|\right) \;=\; 0. \end{aligned}$$

This yields $\alpha^{(\ell+1)} \to \bar{\alpha}$ for $\ell \to \infty$ and $\ell \in J$. By induction, we obtain the convergence $\alpha^{(\ell+s)} \to \bar{\alpha}$, $\ell \to \infty$, and $\ell \in J$ for any $s \in \mathbb{N}$. □

Note that the feasible region of (11.37) and (11.38) is compact due to Heine-Borel's Theorem A.2.4. Hence there exists a convergent subsequence of $\left(\alpha^{(\ell)}\right)$. The limit point of any convergent subsequence is a stationary point of (11.37) and (11.38), as the following result by Lin (2001) and Chen *et al.* (2006) shows.

Theorem 11.15. *Let L be the hinge loss and $\left(\alpha^{(\ell)}\right)_{\ell\in\mathbb{N}}$ be the infinite sequence generated by the SMO-type algorithm ALG5 using ALG4. Then any limit point of $\left(\alpha^{(\ell)}\right)_{\ell\in\mathbb{N}}$ is a stationary point of (11.37) and (11.38).*

Proof. Assume that $\bar{\alpha}$ is the limit point of a convergent subsequence $\left(\alpha^{(\ell)}\right)_{\ell\in J}$, $J \subset \mathbb{N}$. If $\bar{\alpha}$ is not a stationary point of (11.37) and (11.38), then it can not satisfy the optimality condition (11.40). Hence a maximal violating pair $(\bar{i}, \bar{j})$ exists such that

$$\bar{i} \in \arg\max_{i\in I_{up}(\alpha)} -y_i \nabla g(\alpha)_i\,, \quad \bar{j} \in \arg\min_{i\in I_{low}(\alpha)} -y_i \nabla g(\alpha)_i \tag{11.70}$$

and

$$\Delta := y_{\bar{j}} \nabla g(\bar{\alpha})_{\bar{j}} - y_{\bar{i}} \nabla g(\bar{\alpha})_{\bar{i}} \;>\; 0. \tag{11.71}$$

Let us further define the positive constant

$$\Delta' = \min\Big\{\Delta, \frac{1}{2}\min\big\{\big|y_s \nabla g(\bar{\alpha})_s - y_t \nabla g(\bar{\alpha})_t\big| : y_s \nabla g(\bar{\alpha})_s \neq y_t \nabla g(\bar{\alpha})_t\big\}\Big\}. \tag{11.72}$$

Lemma 11.14, the continuity of $\nabla g(\alpha)$, and $h^*(\Delta'/2) > 0$ imply that, for any given positive integer r, there exists an index $\bar{\ell} \in J$ such that, for all $\ell \in J$ with $\ell \geq \bar{\ell}$, the following relationships are valid (we will only give details for the derivation of (11.73) and (11.79), because the proofs to show the validity of (11.74)–(11.78) are somewhat similar):

$$y_{\bar{j}} \nabla g(\alpha^{(\ell+u)})_{\bar{j}} - y_{\bar{i}} \nabla g(\alpha^{(\ell+u)})_{\bar{i}} > \Delta' \qquad \text{if } u = 0, \ldots, r, \tag{11.73}$$

$$i \in I_{up}(\alpha^{(\ell)}), \ldots, i \in I_{up}(\alpha^{(\ell+r)}) \quad \text{if } i \in I_{up}(\bar{\alpha}), \tag{11.74}$$

$$i \in I_{low}(\alpha^{(\ell)}), \ldots, i \in I_{low}(\alpha^{(\ell+r)}) \qquad \text{if } i \in I_{low}(\bar{\alpha}), \tag{11.75}$$

$$\begin{gathered} y_j \nabla g(\alpha^{(\ell+u)})_j - \tfrac{\Delta'}{\sqrt{2}} > y_i \nabla g(\alpha^{(\ell+u)})_i \text{ for } u = 0, \ldots, r, \\ \text{if } y_j \nabla g(\bar{\alpha})_j > y_i \nabla g(\bar{\alpha})_i\,, \end{gathered} \tag{11.76}$$

$$\begin{gathered} \big|y_j \nabla g(\alpha^{(\ell+u)})_j - y_i \nabla g(\alpha^{(\ell+u)})_i\big| < h(\Delta') \text{ for } u = 0, \ldots, r, \\ \text{if } y_j \nabla g(\bar{\alpha})_j = y_i \nabla g(\bar{\alpha})_i\,, \end{gathered} \tag{11.77}$$

$$\begin{gathered} (\tau - \hat{b}_{i,j})\|\alpha^{(\ell+u+1)} - \alpha^{(\ell+u)}\| \leq \Delta' \text{ for } u = 0, \ldots, r-1, \text{ where} \\ \hat{b}_{i,j} := \min\{\tilde{K}_{i,i} + \tilde{K}_{j,j} - 2y_i y_j \tilde{K}_{i,j} : \tilde{K}_{i,i} + \tilde{K}_{j,j} - 2y_i y_j \tilde{K}_{i,j} < 0\}, \end{gathered} \tag{11.78}$$

$$\begin{gathered} i \notin I_{up}(\alpha^{(\ell+u+1)}) \text{ or } j \notin I_{low}(\alpha^{(\ell+u+1)}), \\ \text{if } y_j \nabla g(\bar{\alpha})_j > y_i \nabla g(\bar{\alpha})_i \text{ and } \{i, j\} \text{ is the working set at the} \\ (\ell+u)\text{-iteration for } u = 0, \ldots, r-1. \end{gathered} \tag{11.79}$$

We will now derive (11.73). Lemma 11.14 shows that the sequences $\left(\alpha^{(\ell+u)}\right)_{\ell\in J}$ for $u = 0, \ldots, r$ all converge to $\bar{\alpha}$. Now we use the continuity of $\nabla g(\alpha)$ and

(11.72) to obtain for any fixed $u \in \{0,\dots,r\}$ and the corresponding subsequence $\left(\alpha^{(\ell+u)}\right)_{\ell\in J}$ the existence of a constant $k_u \in \mathbb{N}$ such that (11.73) is satisfied for all values $k \in J$ with $k \geq k_u$. Define $\bar{\ell} := \max\{k_u : u \in \{0,\dots,r\}\}$. As r is finite, the validity of (11.73) follows.

Let us now consider (11.79). Similar to (11.40) for the problem (11.37) and (11.38), we obtain for the subproblem at the $(\ell+u)$-th iteration, if the dual subproblem (11.61) is considered and α_B is a stationary point, that

$$\begin{aligned}&\max_{t\in I_{up}(\alpha_B)} -y_t\left(\nabla g\left(\begin{bmatrix}\alpha_B\\ \alpha_{B^c}^{(\ell+u)}\end{bmatrix}\right)\right)_t - \frac{y_t(\tau-\tilde{b}_{i,j})}{2}\big(\alpha_t-\alpha_t^{(\ell)}\big)\\ \leq &\min_{t\in I_{low}(\alpha_B)} -y_t\left(\nabla g\left(\begin{bmatrix}\alpha_B\\ \alpha_{B^c}^{(\ell+u)}\end{bmatrix}\right)\right)_t - \frac{y_t(\tau-\tilde{b}_{i,j})}{2}\big(\alpha_t-\alpha_t^{(\ell)}\big)\,.\end{aligned}$$

Now $B=\{i,j\}$ and $\alpha_B^{(\ell+u+1)}$ is a stationary point of the subproblem that fulfills the inequality above. Suppose that

$$i \in I_{up}(\alpha^{(\ell+u+1)}) \quad \text{and} \quad j \in I_{low}(\alpha^{(\ell+u+1)}).$$

Then it follows from (11.78) and (11.62) that

$$\begin{aligned}&y_i\nabla g(\alpha^{(\ell+u+1)})_i\\ \geq\; &y_j\nabla g(\alpha^{(\ell+u+1)})_j - \frac{y_i(\tau-\tilde{b}_{i,j})}{2}\big(\alpha_i^{(\ell+u+1)}-\alpha_i^{(\ell+u)}\big)\\ &+\frac{y_j(\tau-\tilde{b}_{i,j})}{2}\big(\alpha_j^{(\ell+u+1)}-\alpha_j^{(\ell+u)}\big)\\ \geq\; &y_i\nabla g(\alpha^{(\ell)})_i - \frac{\tau-\tilde{b}_{i,j}}{\sqrt{2}}\left\|\alpha^{(\ell+u+1)}-\alpha^{(\ell+u)}\right\|\\ \geq\; &y_j\nabla g(\alpha^{(\ell)})_j - \frac{\Delta'}{\sqrt{2}}\,.\end{aligned} \tag{11.80}$$

However, this is a contradiction to (11.76) because $y_j\nabla g(\bar{\alpha})_j > y_i\nabla g(\bar{\alpha})_i$ implies (11.76) for $\alpha^{(\ell+u+1)}$. If $\tilde{b}_{i,j}>0$ and the subproblem of step *iii)* of the algorithm ALG1 is considered, then (11.80) has no term $\Delta'/\sqrt{2}>0$ which immediately gives the desired contradiction.

For notational convenience, let us now reorder the indexes of $\bar{\alpha}$ such that

$$y_1\nabla g(\bar{\alpha})_1 \geq \ldots \geq y_n\nabla g(\bar{\alpha})_n\,. \tag{11.81}$$

Further, we define

$$S_{up}(\ell) := \sum_{i\in I_{up}(\alpha^{(\ell)})} i \quad \text{and} \quad S_{low}(\ell) := \sum_{i\in I_{low}(\alpha^{(\ell)})} (n-i), \tag{11.82}$$

which gives

$$n \le S_{low}(\ell) + S_{up}(\ell) \le n(n-1). \tag{11.83}$$

Fix $u \in \{0, \ldots, r\}$. If the pair $\{i, j\}$ is selected at the $(\ell+u)$-th iteration, then it will be shown that

$$y_j \nabla g(\bar{\alpha})_j > y_i \nabla g(\bar{\alpha})_i \,. \tag{11.84}$$

Note that $y_j \nabla g(\bar{\alpha})_j < y_i \nabla g(\bar{\alpha})_i$ is impossible because $y_j \nabla g(\alpha^{(\ell+u)})_j < y_i \nabla g(\alpha^{(\ell+u)})_i$ from (11.76) then violates (11.59). Equality in (11.84) would give

$$\begin{aligned} & y_j \nabla g(\alpha^{(\ell+u)})_j - y_i \nabla g(\alpha^{(\ell+u)})_i & (11.85) \\ & < h^*(\Delta') < h^*\big(y_{\bar{j}} \nabla g(\alpha^{(\ell+u)})_{\bar{j}} - y_{\bar{i}} \nabla g(\alpha^{(\ell+u)})_{\bar{i}}\big) \\ & \le h^*\big(m(\alpha^{(\ell+u)}) - M(\alpha^{(\ell+u)})\big). & (11.86) \end{aligned}$$

Here we used that h^* is strictly increasing and (11.73) and (11.77) to obtain the first two inequalities, and the last inequality in (11.86) results from $\bar{i} \in I_{up}(\bar{\alpha})$, $\bar{j} \in I_{low}(\bar{\alpha})$, (11.74), and (11.75). However, (11.86) is a contradiction to (11.59), which shows that (11.84) is valid.

Now we will use a counting procedure that gives a contradiction to (11.70) and (11.71). Combining (11.84) and (11.79) shows that $i \notin I_{up}(\alpha^{(\ell+1)})$ or $j \notin I_{low}(\alpha^{(\ell+1)})$ if we consider the iteration step from ℓ to $\ell+1$. If $i \notin I_{up}(\alpha^{(\ell+1)})$, then (11.74) implies $i \notin I_{up}(\bar{\alpha})$ and hence $i \in I_{low}(\bar{\alpha})$. From (11.75) and the rule (11.59) to select the working set, it follows that $i \in I_{low}(\alpha^{(\ell)}) \cap I_{up}(\alpha^{(\ell)})$. Thus we have

$$i \in I_{low}(\alpha^{(\ell)}) \cap I_{up}(\alpha^{(\ell)}) \quad \text{and} \quad i \notin I_{up}(\alpha^{(\ell+1)}).$$

Note that $j - i \le -1$ due to (11.81). Therefore, we obtain with $j \in I_{low}(\alpha^{(\ell)})$ the inequalities

$$S_{up}(\ell+1) \le S_{up}(\ell) + j - i \le S_{up}(\ell) - 1 \;\text{ and }\; S_{low}(\ell+1) \le S_{low}(\ell). \tag{11.87}$$

In a similar manner, we get for the case $j \notin I_{low}(\alpha^{(\ell+1)})$ that

$$j \in I_{low}(\alpha^{(\ell)}) \cap I_{up}(\alpha^{(\ell)}) \quad \text{and} \quad j \notin I_{low}(\alpha^{(\ell+1)}).$$

For $i \in I_{up}(\alpha^{(\ell)})$, we have that

$$S_{up}(\ell+1) \le S_{up}(\ell) \text{ and } S_{low}(\ell+1) \le S_{low}(\ell) + (\ell - i) - (\ell - j) \le S_{low}(\ell) - 1. \tag{11.88}$$

The same arguments can be used to go from iteration step $(\ell+1)$ to $(\ell+2)$ because (11.79) can be used, as (11.84) holds for working sets selected during the iteration steps ℓ to $\ell + r$. Now (11.87) and (11.88) show that the term $S_{low}(\ell) + S_{up}(\ell)$ can be reduced to zero in $r := n(n-1)$ iterations, which gives the desired contradiction to (11.83). Hence, the assumptions (11.70) and (11.71) were wrong, which gives the assertion. □

The next result (Chen *et al.*, 2006) shows that the sequence $\big(\alpha^{(\ell)}\big)$ is even globally convergent under mild conditions.

Corollary 11.16. *Let L be the hinge loss function. If the matrix $\tilde{K}$ is strictly positive definite, then $\left(\alpha^{(\ell)}\right)_{\ell\in\mathbb{N}}$ globally converges to the unique maximum of the dual problem (11.37) and (11.38).*

Proof. Since $\tilde{K}$ is strictly positive definite, there exists a unique solution, say $\bar{\alpha}$, of the dual problem (11.37)-(11.38). Suppose that the sequence $\left(\alpha^{(\ell)}\right)_{\ell\in\mathbb{N}}$ does not globally converge to $\bar{\alpha}$. Then there exists a constant $\varepsilon > 0$ and an infinite subset $J \subset \mathbb{N}$ such that $\|\alpha^{(\ell)} - \bar{\alpha}\| \geq \varepsilon$ for all $\ell \in J$. Since $\{\alpha^{(\ell)} : \ell \in J\}$ is in a compact set, there exists a further subsequence that converges to some point, say α^*, with $\|\alpha^* - \bar{\alpha}\| \geq \varepsilon$. We know by Theorem 11.15 that α^* is an optimal solution of (11.37)-(11.38). This gives the desired contradiction because $\bar{\alpha}$ is the unique global maximum. □

The preceding corollary is quite useful for practical purposes, and we would like to give an example. Suppose that we consider a data set with $x_i \neq x_j$ for all $1 \leq i < j \leq n$. Further, let us assume that a Gaussian RBF kernel is used. Then the matrix $\tilde{K}$ is strictly positive definite and the sequence $\left(\alpha^{(\ell)}\right)_{\ell\in\mathbb{N}}$ converges globally to the unique maximum of the dual problem.

The following result (Chen *et al.*, 2006) shows that one can improve Theorem 11.15 if the matrix $\tilde{K}$ is positive definite.

Theorem 11.17. *Let L be the hinge loss, and $\tilde{K}$ be positive definite.*

i) If $\bar{\alpha} \neq \hat{\alpha}$ are any two optimal solutions of (11.37) and (11.38), then

$$y_i\big(\nabla g(\bar{\alpha})\big)_i = y_i\big(\nabla g(\hat{\alpha})\big)_i, \quad i = 1, \ldots, n, \tag{11.89}$$

and

$$m(\bar{\alpha}) = M(\bar{\alpha}) = m(\hat{\alpha}) = M(\hat{\alpha}). \tag{11.90}$$

ii) If there is an optimal solution $\bar{\alpha}$ satisfying $m(\bar{\alpha}) < M(\bar{\alpha})$, then $\bar{\alpha}$ is the unique optimal solution of (11.37) and (11.38).

iii) The following set is independent of any optimal solution $\bar{\alpha}$:

$$I := \{i \in \{1, \ldots, n\} : -y_i\big(\nabla g(\bar{\alpha})\big)_i > M(\bar{\alpha}) \;\text{ or }\; -y_i\big(\nabla g(\bar{\alpha})\big)_i < m(\bar{\alpha})\}. \tag{11.91}$$

Moreover, the problem (11.37) and (11.38) has a unique and bounded optimal solution at α_i, $i \in I$.

Proof. Since $\tilde{K} \in \mathbb{R}^{n\times n}$ is positive definite, the problem (11.37) and (11.38) is a convex programming problem and $\bar{\alpha}$ and $\hat{\alpha}$ are both global optima. Then

$$g(\bar{\alpha}) = g(\hat{\alpha}) = g(\delta\bar{\alpha} + (1-\delta)\hat{\alpha}), \qquad \text{for all } \delta \in [0,1],$$

implies

$$(\bar{\alpha} - \hat{\alpha})^{\mathsf{T}} \tilde{K} (\bar{\alpha} - \hat{\alpha}) = 0.$$

Now let us factorize $\tilde{K} = UU^{\mathsf{T}}$, which is possible because $\tilde{K}$ is positive definite. We obtain $\|U^{\mathsf{T}}(\bar{\alpha} - \hat{\alpha})\| = 0$ and hence $\tilde{K}\bar{\alpha} = \tilde{K}\hat{\alpha}$. From this we obtain (11.89).

To prove (11.90), we will show that

$$m(\hat{\alpha}) \geq M(\bar{\alpha}) \quad \text{and} \quad m(\bar{\alpha}) \geq M(\hat{\alpha}). \tag{11.92}$$

With the optimality conditions $M(\bar{\alpha}) \geq m(\bar{\alpha})$ and $M(\hat{\alpha}) \geq m(\hat{\alpha})$, the equalities in (11.90) hold. Due to symmetry, it is sufficient to prove the first case of (11.92). If it is false, then $m(\hat{\alpha}) < M(\bar{\alpha})$. We then investigate different cases by comparing $-y_i \nabla g(\bar{\alpha})_i$ with $M(\bar{\alpha})$ and $m(\hat{\alpha})$. If $m(\hat{\alpha}) < M(\bar{\alpha}) \leq -y_i \nabla g(\bar{\alpha})_i$, then $i \notin I_{up}(\hat{\alpha})$ and

$$\hat{\alpha}_i = \begin{cases} 0 & \text{if } y_i = -1 \\ C & \text{if } y_i = +1. \end{cases} \tag{11.93}$$

With $0 \leq \bar{\alpha}_i \leq C$, we obtain

$$y_i(\hat{\alpha}_i - \bar{\alpha}_i) \geq 0. \tag{11.94}$$

If $M(\bar{\alpha}) > m(\hat{\alpha}) \geq -y_i \nabla g(\bar{\alpha})_i$, then $i \notin I_{low}(\bar{\alpha})$ and

$$\bar{\alpha}_i = \begin{cases} C & \text{if } y_i = -1 \\ 0 & \text{if } y_i = +1, \end{cases} \tag{11.95}$$

and (11.94) still holds.

The other indexes are in the set

$$S := \big\{ i \,:\, m(\hat{\alpha}) < -y_i \nabla g(\hat{\alpha})_i = -y_i \nabla g(\bar{\alpha})_i < M(\bar{\alpha}) \big\}.$$

If $i \in S$, then $i \notin I_{up}(\hat{\alpha})$ and $i \notin I_{low}(\bar{\alpha})$. Hence (11.93) and (11.95) yield

$$y_i(\hat{\alpha}_i - \bar{\alpha}_i) = C, \tag{11.96}$$

and thus

$$0 = \sum_{i=1}^{n} y_i \hat{\alpha}_i - \sum_{i=1}^{n} y_i \bar{\alpha}_i = \sum_{i \notin S} y_i(\hat{\alpha}_i - \bar{\alpha}_i) + C|S|.$$

As $C > 0$ and (11.94) implies that each term in the sum above is non-negative, we obtain $|S| = 0$ and $\hat{\alpha}_i = \bar{\alpha}_i$ for all $i \notin S$. Thus, $\bar{\alpha} = \hat{\alpha}$. However, this is a contradiction to the assumption that $\bar{\alpha}$ and $\hat{\alpha}$ are different optimal solutions. Therefore, $m(\hat{\alpha}) < M(\bar{\alpha})$ is wrong and we obtain $m(\hat{\alpha}) \geq M(\bar{\alpha})$ in (11.92), which completes the proof of (11.90).

The second result of the theorem and the validity of the set I follow from (11.90). Moreover, the set I is independent of any optimal solution.

Assume that α is an optimal vector. If $i \in I$ and $m(\alpha) \leq M(\alpha) < -y_i \nabla g(\alpha)_i$, then $i \notin I_{up}(\alpha)$ and α_i is the same as $\hat{\alpha}_i$ in (11.93). Hence, the optimal coefficient α_i is unique and bounded. The case $m(\alpha) > -y_i \nabla g(\alpha)_i$ can be treated in a similar manner. □

Since in general the decomposition method approaches an optimum only after an infinite number of iteration steps, there is a need to specify *stopping criteria* to stop the iteration procedure after a finite number of steps. In general, it is not wise to specify in advance the number of steps to be carried out by the decomposition method because it is unknown how well the approximation of the optimum will be. In general, however, it can be useful to define an upper bound for the number of iteration steps or for the computation time and print an error message if the stopping criteria were not satisfied.

One possible stopping criterion is to specify in advance a small tolerance value, say $\varepsilon > 0$, and stop the iteration process if

$$m(\alpha^{(\ell)}) - M(\alpha^{(\ell)}) \leq \varepsilon.$$

This stopping condition is quite plausible and commonly used due to its closeness to the optimality condition (11.40). The following result (Chen *et al.*, 2006) shows that this stopping condition can actually be achieved in a finite number of iteration steps.

Theorem 11.18. *Let L be the hinge loss, $\tilde{K} \in \mathbb{R}^{n\times n}$ be positive definite, and suppose that the SMO-type decomposition method ALG5 using ALG4 generates an infinite sequence $\left(\alpha^{(\ell)}\right)_{\ell\in\mathbb{N}}$. Then*

$$\lim_{\ell\to\infty} m(\alpha^{(\ell)}) - M(\alpha^{(\ell)}) = 0. \tag{11.97}$$

Proof. Let us assume that the convergence in (11.97) is wrong. Then there exists an infinite set $\bar{J}$ and a constant $\Delta > 0$ such that

$$\left|m(\alpha^{(\ell)}) - M(\alpha^{(\ell)})\right| \geq \Delta, \quad k \in \bar{J}. \tag{11.98}$$

In the SMO-decomposition method we have $m(\alpha^{(\ell)}) > M(\alpha^{(\ell)})$ for all $\ell \in \mathbb{N}$, and hence (11.98) can be rewritten as

$$m(\alpha^{(\ell)}) - M(\alpha^{(\ell)}) \geq \Delta, \quad k \in \bar{J}. \tag{11.99}$$

In the set $\bar{J}$, there exists an infinite subset J such that

$$\lim_{\ell\in J,\, \ell\to\infty} \alpha^{(\ell)} = \bar{\alpha}.$$

Using the assumption that $\tilde{K}$ is positive definite, we obtain the global convergence of $\nabla g(\alpha^{(\ell)})$ by Theorem 11.17

$$\lim_{\ell\to\infty} \nabla g(\alpha^{(\ell)})_i = \nabla g(\bar{\alpha})_i \quad i = 1, \ldots, n. \tag{11.100}$$

Now we will use a counting approach similar to that of Theorem 11.15. First, rewrite (11.99) as

$$m(\alpha^{(\ell)}) \geq M(\alpha^{(\ell)}) + \Delta', \quad k \in \bar{J}, \tag{11.101}$$

where

$$\Delta' := \min\Big\{\Delta, \frac{1}{2}\min\big\{|y_s\nabla g(\bar{\alpha})_s - y_t\nabla g(\bar{\alpha})_t| : y_s\nabla g(\bar{\alpha})_s \neq y_t\nabla g(\bar{\alpha})_t\big\}\Big\} > 0. \tag{11.102}$$

We still require (11.73)–(11.79) but use (11.100) and the definition of Δ' in (11.102) to extend (11.76) and (11.77) for all $\ell \geq \bar{\ell}$ (i.e., not only for $\ell \in J$):

$$y_t\nabla g(\alpha^{(\ell)})_t < y_s\nabla g(\alpha^{(\ell)})_s \quad \text{if } y_t\nabla g(\bar{\alpha})_t < y_s\nabla g(\bar{\alpha})_s\,, \tag{11.103}$$

$$|y_s\nabla g(\alpha^{(\ell)})_s - y_t\nabla g(\alpha^{(\ell)})_t| > \Delta' \quad \text{if } y_t\nabla g(\bar{\alpha})_t \neq y_s\nabla g(\bar{\alpha})_s\,, \tag{11.104}$$

$$|y_s\nabla g(\alpha^{(\ell)})_s - y_t\nabla g(\alpha^{(\ell)})_t| < h^*(\Delta') \quad \text{if } y_t\nabla g(\bar{\alpha})_t = y_s\nabla g(\bar{\alpha})_s\,. \tag{11.105}$$

Then the proof follows Theorem 11.15 except (11.86), in which we need $m(\alpha^{(\ell+u)}) - M(\alpha^{(\ell+u)}) \geq \Delta'$ for all $u \in \{0, \ldots, r\}$. This condition does not follow from (11.101), which holds only for a subsequence. Therefore, our goal is to prove

$$m(\alpha^{(\ell)}) - M(\alpha^{(\ell)}) \geq \Delta', \quad \ell \geq \bar{\ell}. \tag{11.106}$$

Assume that there is a positive integer $\ell' \geq \bar{\ell}$ such that $m(\alpha^{(\ell')}) - M(\alpha^{(\ell')}) \in (0, \Delta')$ and that $\{i, j\}$ is the working set at this iteration step. Because $i \in I_{up}(\alpha^{(\ell')})$ and $j \in I_{low}(\alpha^{(\ell')})$ from the selection rule, we have

$$M(\alpha^{(\ell')}) \leq -y_j\nabla g(\alpha^{(\ell')})_j < -y_i\nabla g(\alpha^{(\ell')})_i \leq m(\alpha^{(\ell')}). \tag{11.107}$$

Using (11.104), we obtain that the set $\{i, j\}$ and indexes achieving $m(\alpha^{(\ell')})$ and $M(\alpha^{(\ell')})$ have the same value of $y_t\nabla g(\bar{\alpha})_t$ and are all from the set

$$\big\{t : y_t\nabla g(\bar{\alpha})_t = y_i\nabla g(\bar{\alpha})_i = y_j\nabla g(\bar{\alpha})_j\big\}. \tag{11.108}$$

Note that for elements not in this set, (11.103), (11.104), and (11.107) yield

$$\begin{aligned} &y_t\nabla g(\bar{\alpha})_t < y_i\nabla g(\bar{\alpha})_i \;\text{ implies} \\ &-y_t\nabla g(\alpha^{(\ell')})_t > -y_i\nabla g(\alpha^{(\ell')})_i + \Delta' > m(\alpha^{(\ell')}) \text{ and } t \notin I_{up}(\alpha^{(\ell')}). \end{aligned} \tag{11.109}$$

In a similar way, we obtain that

$$y_t\nabla g(\bar{\alpha})_t > y_i\nabla g(\bar{\alpha})_i \;\text{ implies }\; t \notin I_{low}(\alpha^{(\ell')}). \tag{11.110}$$

Because we have shown that the working set is from the set given in (11.108), other coefficients remain the same from iteration step ℓ' to $\ell' + 1$. Hence, indexes satisfying (11.109) and (11.110) fulfill $t \notin I_{up}(\alpha^{(\ell'+1)})$ and $t \notin I_{low}(\alpha^{(\ell'+1)})$, respectively. Furthermore, indexes in (11.109) have larger values of $-y_t\nabla g(\alpha^{(\ell'+1)})_t$ than others due to (11.103). Hence their values of $-y_t\nabla g(\alpha^{(\ell'+1)})_t$ are greater than $m(\alpha^{(\ell'+1)})$. Similarly, components in (11.110) are smaller than $M(\alpha^{(\ell'+1)})$. Using $m(\alpha^{(\ell'+1)}) > M(\alpha^{(\ell'+1)})$, we see that indexes that achieve $m(\alpha^{(\ell'+1)})$ and $M(\alpha^{(\ell'+1)})$ are again from the set in (11.108), and this is true for all $\ell \geq \ell'$. Now, by (11.105) and the conditions on h^*, we obtain

$$m(\alpha^{(\ell)}) - M(\alpha^{(\ell)}) < h^*(\Delta') \leq \Delta', \qquad \ell \geq \ell'.$$

This is the desired contradiction to (11.101), and thus (11.106) holds. □

One can argue that a main advantage of decomposition methods is to allow the computation of $f_{\mathrm{D},\lambda}$ even for large sample sizes n. However, the actual computation time and the number of necessary iteration steps until the stopping conditions are fulfilled can be quite large. Therefore, computational techniques to speed up the computation time or to decrease the number of iteration steps are desirable. Among such techniques, *shrinking* and *caching* have been shown to be successful for SVMs.

Shrinking is based on the idea that if an index $\alpha_i^{(\ell)}$ remains equal to 0 or to C for many iteration steps then it may stay at this value. The size of the optimization problem is reduced in shrinking algorithms without considering some bounded Lagrange multipliers. This has the advantage that the decomposition method then works on a smaller problem and hence a considerable reduction of CPU time is sometimes possible. In addition, less memory is used. Afterward, we have to add shrunken components back and must check whether an optimal solution of the original problem is obtained.

Besides shrinking, a caching strategy can also be helpful to speed up the computation of $f_{\mathrm{D},\lambda}$ for large sample sizes. Since $\tilde{K} \in \mathbb{R}^{n\times n}$ may then be too large to be stored into the RAM of the computer, the elements of $\tilde{K}$ are calculated when they are needed. The idea is to use the cache and the RAM of the computer (which allows relatively fast access to objects in it) to store recently used elements $\tilde{K}_{i,j}$. If in the final iterations only a small subset of columns of $\tilde{K}$ are actually needed and if the cache contains them, the computation of many kernel terms $k(x_i, x_j)$ becomes superfluous. Of course, this is especially interesting for SVMs having sparse solutions (i.e., if many Lagrange multipliers are equal to 0). This is often true for SVMs based on the hinge loss or on the ϵ-insensitive loss function.

The following result was shown by Chen *et al.* (2006).

Theorem 11.19. *Let L be the hinge loss and $\tilde{K} \in \mathbb{R}^{n\times n}$ be positive definite, and assume the SMO-type decomposition method ALG5 using ALG4. Let I be the set of indexes defined in (11.91).*

i) There exists an $\bar{\ell} \in \mathbb{N}$ such that, after $\ell > \bar{\ell}$ iteration steps, every Lagrange multiplier $\alpha_i^{(\ell)}$, $i \in I$, has reached the unique and bounded optimal solution. It remains the same in all subsequent iterations, and $i \in I$ is not *an element of the set*

$$\big\{t \in \{1,\ldots,n\} : M(\alpha^{(\ell)}) \leq -y_t \nabla g(\alpha^{(\ell)})_t \leq m(\alpha^{(\ell)})\big\}. \tag{11.111}$$

ii) If (i) has an optimal solution $\bar{\alpha}$ satisfying $m(\bar{\alpha}) < M(\bar{\alpha})$, then $\bar{\alpha}$ is the unique solution and the decomposition method reaches it in a finite *number of iterations.*

iii) If $\left(\alpha^{(\ell)}\right)_{\ell\in\mathbb{N}}$ *is an infinite sequence, then the following two limits exist and are equal:*

$$\lim_{\ell\to\infty} m(\alpha^{(\ell)}) = m(\bar{\alpha}) = \lim_{\ell\to\infty} M(\alpha^{(\ell)}) = M(\bar{\alpha}), \tag{11.112}$$

where $\bar{\alpha}$ *is any optimal solution.*

Proof. i). Suppose that the assertion is wrong. Then there exist an index $\bar{i} \in I$ and an infinite set $\hat{J} \subset \mathbb{N}$ such that

$$\alpha_{\bar{i}}^{(\ell)} \neq \hat{\alpha}_{\bar{i}}\,, \quad \ell \in \hat{J}, \tag{11.113}$$

where $\hat{\alpha}_{\bar{i}}$ is the i-th coefficient of the unique optimal solution according to Theorem 11.17. Using Theorem 11.15, there is a set $J \subset \hat{J}$ such that

$$\lim_{\ell\in J,\,\ell\to\infty} \alpha^{(\ell)} = \bar{\alpha} \tag{11.114}$$

is a stationary point. Furthermore, Theorem 11.17 implies that $\bar{\alpha}_{\bar{i}} = \hat{\alpha}_{\bar{i}}$ for $i \in I$; i.e., these coefficients are optimal and unique.

As $\bar{i} \in I$, let us first consider the case

$$M(\bar{\alpha}) < -y_{\bar{i}} \nabla g(\bar{\alpha})_{\bar{i}}\,. \tag{11.115}$$

In this situation, we have $\bar{i} \in I_{up}(\bar{\alpha})$, and (11.113) implies

$$\bar{i} \in I_{up}(\alpha^{(\ell)})\,, \quad \ell \in J. \tag{11.116}$$

For each index $j \in \arg M(\bar{\alpha})$, we have $j \in I_{low}(\bar{\alpha})$. It follows from (11.114) that there an integer $\bar{\ell} \in \mathbb{N}$ such that

$$j \in I_{low}(\alpha^{(\ell)})\,, \quad \ell \in J, \ell \geq \bar{\ell}. \tag{11.117}$$

Thus, (11.116) and (11.117) imply

$$m(\alpha^{(\ell)}) - M(\alpha^{(\ell)}) \geq y_j \nabla g(\alpha^{(\ell)})_j - y_{\bar{i}} \nabla g(\alpha^{(\ell)})_{\bar{i}}\,, \quad \ell \in J, \ell \geq \bar{\ell}. \tag{11.118}$$

Now, by (11.114), the continuity of $\nabla g(\alpha)$, and (11.97), and computing the limit on both sides of (11.118), we obtain

$$0 \geq y_j \nabla g(\bar{\alpha})_j - y_{\bar{i}} \nabla g(\bar{\alpha})_{\bar{i}} = -M(\bar{\alpha}) - y_{\bar{i}} \nabla g(\bar{\alpha})_{\bar{i}}\,.$$

However, this inequality violates the inequality in (11.115) which gives the desired contradiction. The proof for the case $m(\bar{\alpha}) > -y_{\bar{i}} \nabla g(\bar{\alpha})_{\bar{i}}$ is similar.

ii). Let us again assume that the assertion is false. Then $\left(\alpha^{(\ell)}\right)_{\ell\in\mathbb{N}}$ is an infinite sequence. It follows from Theorems 11.15 and 11.17 that $\bar{\alpha}$ is the unique optimal solution and $\alpha^{(\ell)}$ globally converges to $\bar{\alpha}$ if $\ell \to \infty$. Define the index sets

$$I_1 := \left\{ i \in \{1, \ldots, n\} : M(\bar{\alpha}) = -y_i \nabla g(\bar{\alpha})_i \right\},$$
$$I_2 := \left\{ i \in \{1, \ldots, n\} : m(\bar{\alpha}) = -y_i \nabla g(\bar{\alpha})_i \right\}.$$

Using part *i)* of the theorem, we see that $\arg m(\alpha^{(\ell)}) \subset I_1 \cup I_2$ and $\arg M(\alpha^{(\ell)}) \subset I_1 \cup I_2$ provided ℓ is sufficiently large. Now, by (11.97), the continuity of $\nabla g(\alpha)$, and the convergence $\lim_{\ell \to \infty} \alpha^{(\ell)} = \bar{\alpha}$, there exists an integer $\bar{\ell} \in \mathbb{N}$ such that for all $\ell \geq \bar{\ell}$

$$\arg m(\alpha^{(\ell)}) \cup \arg M(\alpha^{(\ell)}) \subset I_1 \quad \text{or} \quad \arg m(\alpha^{(\ell)}) \cup \arg M(\alpha^{(\ell)}) \subset I_2 \,. \tag{11.119}$$

Suppose that $\arg m(\alpha^{(\ell)}) \cup \arg M(\alpha^{(\ell)}) \subset I_1$ at the ℓ-th iteration. Then we can use the same argument as in (11.107) and (11.108) to obtain that the working set B is a subset of I_1. The decomposition method maintains feasibility, thus

$$\sum_{i \in B} y_i \alpha_i^{(\ell)} = \sum_{i \in B} y_i \alpha_i^{(\ell+1)} \,. \tag{11.120}$$

From $B \subset I_1$ and the assumption that $m(\bar{\alpha}) < M(\bar{\alpha})$, every $\bar{\alpha}_i$, $i \in B$, satisfies $i \notin I_{up}(\alpha)$. Hence $\bar{\alpha}_i = \hat{\alpha}_i = 0$, if $y_i = -1$ and $i \in B$, and $\bar{\alpha}_i = \hat{\alpha}_i = C$, if $y_i = +1$ and $i \in B$. If we combine this with (11.120), we obtain

$$\begin{aligned}
&\|\alpha^{(\ell+1)} - \bar{\alpha}\|_1 \\
&= \sum_{i \notin B} |\alpha_i^{(\ell+1)} - \bar{\alpha}_i| + \sum_{i \in B,\, y_i = +1} (C - \alpha_i^{(\ell+1)}) + \sum_{i \in B,\, y_i = -1} (\alpha_i^{(\ell+1)} - 0) \\
&= \sum_{i \notin B} |\alpha_i^{(\ell)} - \bar{\alpha}_i| + \sum_{i \in B,\, y_i = +1} (C - \alpha_i^{(\ell)}) + \sum_{i \in B,\, y_i = -1} (\alpha_i^{(\ell)} - 0) \\
&= \|\alpha^{(\ell)} - \bar{\alpha}\|_1 \,.
\end{aligned} \tag{11.121}$$

If $\arg m(\alpha^{(\ell)})$ and $\arg M(\alpha^{(\ell)})$ are both subsets of I_2, the equation (11.121) is still valid. Therefore, $0 \neq \|\alpha^{(\ell)} - \bar{\alpha}\|_1 = \|\alpha^{(\ell+r)} - \bar{\alpha}\|_1$, $r \in \mathbb{N}$, which gives the desired contradiction to the fact that $\left(\alpha^{(\ell)}\right)$ converges to $\bar{\alpha}$. Hence the decomposition method stops after a finite number of iteration steps.

iii). Since $\left(\alpha^{(\ell)}\right)_{\ell \in \mathbb{N}}$ is an infinite sequence, using the result of part *ii)* of the theorem, we see that the dual problem (11.37) and (11.38) has no optimal solution $\bar{\alpha}$ with the property $M(\bar{\alpha}) > m(\bar{\alpha})$. Using Theorem 11.17, this yields

$$M(\bar{\alpha}) = m(\bar{\alpha}) = -y_t \nabla g(\bar{\alpha})_t \,, \quad t \notin I. \tag{11.122}$$

Note that this result is valid for any optimal solution $\bar{\alpha}$. Now, part *i)* of the theorem guarantees the existence of $\bar{\ell} \in \mathbb{N}$ such that, for all $\ell \geq \bar{\ell}$, the index $i \in I$ is not an element of the set in (11.111). Therefore, the set in (11.111) is contained in the index set $I' := \{1, \ldots, n\} \backslash I$ and

$$\min_{i \in I'} -y_i \nabla g(\alpha^{(\ell)})_i \leq M(\alpha^{(\ell)}) < m(\alpha^{(\ell)}) \leq \max_{i \in I'} -y_i \nabla g(\alpha^{(\ell)})_i \,. \tag{11.123}$$

Although the sequence $\left(\alpha^{(\ell)}\right)_{\ell \in \mathbb{N}}$ may not be globally convergent, the sequences $\left(-y_i \nabla g(\alpha^{(\ell)})_i\right)_{\ell \in \mathbb{N}}$, $i = 1, \ldots, n$, are according to (11.89). The limits

of both sides of (11.123) are equal due to (11.122). Hence (11.112) follows and the assertion of part *iii)* is shown, which completes the proof. □

The preceding theorem shows, for SVMs based on the hinge loss, that the SMO-type decomposition method involves only indexes from I' in many iteration steps, which makes caching successful for this loss function. Recall that we know from Chapter 8 that the property of the hinge loss function being equal to zero in a whole interval implies that $f_{\mathrm{D},\lambda} = \sum_{i=1}^{n} \alpha_i \Phi(x_i)$ is usually sparse (i.e., many coefficients α_i are equal to 0) and this implies that caching can be effective. This theorem also illustrates two possible shrinking implementations for SVMs based on the hinge loss function. *(i)* Elements not in the set (11.111) are removed. This is done by the software `LIBSVM` (Chang and Lin, 2004). *(ii)* Any α_i that has stayed at the same bound for a certain number of iterations is removed. This strategy is implemented in $\texttt{SVM}^{light}$ (Joachims, 1999). We also refer to Section 11.4 for additional information regarding these software products. Caching and shrinking can probably offer such a big gain in computing $f_{\mathrm{D},\lambda}$ only for loss functions that allow a sparse representation of $f_{\mathrm{D},\lambda}$.

11.3 Determination of Hyperparameters

In this section, we consider some techniques for determining suitable combinations of the hyperparameters for SVMs. There exists a vast body of literature regarding the choice of hyperparameters for SVMs. Here we will only consider a few facets of how to choose such hyperparameters for classification and regression problems.

The quality of the estimator $\mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda})$ for the unknown risk $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ and the precision of predictions $f_{\mathrm{D},\lambda}(x)$ for the unknown values $f_{\mathrm{P},\lambda}(x)$ for unseen $x \in X$ critically depend not only on the data set D used for training purposes, the loss function, and the kernel but also on the choice of the hyperparameters such as the regularizing parameter $\lambda > 0$, kernel parameters, and parameters of the loss function. Examples are thus the value of γ for the Gaussian RBF kernel and ϵ used by the ϵ-insensitive loss function in regression. Unfortunately, choosing these hyperparameters in an optimal way usually requires computing $f_{\mathrm{D},\lambda}$ for many combinations of the hyperparameters. In other words, it is necessary to solve not just one convex problem but a series of them. This increases the computational effort for the use of SVMs in practice.

Let us first consider SVMs for regression based on the ϵ-insensitive loss function. There exists a linear relationship between the noise level of $\mathrm{P}(y|x)$ and the optimal value of ϵ for support vector regression using $L = L_{\epsilon\text{-insens}}$ (Smola *et al.*, 1998). Of course, P is unknown; otherwise we would probably not use $L_{\epsilon\text{-insens}}$, but, for example, the maximum likelihood loss $L\big(x, y, f(x)\big) = -\ln p\big(y - f(x)\big)$ if P has a density function p. There exists

a modification of the SVM based on $L_{\epsilon\text{-insens}}$ called *ν-support vector regression* exploiting this relationship. The idea is to modify (11.1) such that the hyperparameter ϵ becomes a variable of the optimization problem including a specific additional term in the primal objective function that attempts to minimize ϵ. For $L_{\epsilon\text{-insens}}$ and the case with an additional offset term $b \in \mathbb{R}$, the problem (11.1) is thus modified to

$$\inf_{f \in H,\, b \in \mathbb{R},\, \epsilon > 0} \mathbb{E}_{\mathrm{D}} L\big(Y, f(X) + b\big) + \lambda \|f\|_H^2 + \nu\epsilon \tag{11.124}$$

for some $\nu > 0$. Define $C = 1/(2n\lambda)$. Then we obtain the equivalent problem

$$\begin{aligned} \min_{\alpha,\, \xi^+,\, \xi^- \in \mathbb{R}^n,\, b \in \mathbb{R},\, \epsilon > 0} \quad & C \sum_{i=1}^n \big(\xi_i^+ + \xi_i^-\big) + \frac{1}{2}\|w(\alpha)\|_H^2 + Cn\nu\epsilon \\ \text{s.t.} \quad & \xi_i^+ \geq 0,\, \xi_i^+ \geq y_i - \langle w(\alpha), \Phi(x_i)\rangle_H - b - \epsilon, \\ & \xi_i^- \geq 0,\, \xi_i^- \geq \langle w(\alpha), \Phi(x_i)\rangle_H + b - y_i - \epsilon, \quad \forall i. \end{aligned}$$

The dual program becomes

$$\begin{aligned} \max_{\alpha^+,\, \alpha^- \in \mathbb{R}^n} \quad & \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) y_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) k(x_i, x_j) \\ \text{s.t.} \quad & \alpha_i^+,\, \alpha_i^- \in [0, C],\, \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) = 0,\, \sum_{i=1}^n (\alpha_i^+ + \alpha_i^-) \leq Cn\nu,\ \ \forall i. \end{aligned}$$

A pendant to ν-support vector regression exists for classification problems.

We will now consider the determination of a suitable combination of hyperparameters as an optimization problem and will summarize empirical results when we compare different numerical methods to solve this optimization problem. We will concentrate on classification and regression problems. A reasonable choice of the hyperparameters depends on the criteria used to measure their quality. One useful criterion is the *accuracy*. In classification problems, the accuracy is often measured by the empirical misclassification rate. One can also use the modification of TV-SVM described in Definition 8.20. In regression problems, the empirical L-risk or the empirical L-risk based on a suitable calibrated loss function are often used as accuracy criteria in regression problems.

Note that $f_{\mathrm{D},\lambda} \in H$ and the predictions $\hat{y} = f_{\mathrm{D},\lambda}(x) \in \mathbb{R}$ depend on the hyperparameters. As the derivative of both target functions on the hyperparameters is usually unknown, the optimal parameters have to be found numerically. The following six methods are often used to determine suitable hyperparameters and will be briefly described: random search, grid search, Nelder-Mead search, cross-validation, a heuristic search, and pattern search.

The simplest version of a *random search* can be described as follows. A random point of the parameter space is chosen, and the value of the objective function is evaluated. This is repeated N times, and the best point is taken

as the result (i.e., this point is considered as a suitable choice of the hyperparameters). Of course, the result of this search strongly depends on the chosen random points and on the number of random points. The random points for which the objective function is evaluated can be drawn, for example, from a multivariate normal distribution with the center of the search space as the mean.

Optimization by the *grid search* is also very simple. After the search space (i.e.; the set of all possible combinations of the hyperparameters) is specified, each search dimension is split into n_i parts. Often these splits are equidistant or geometrically distributed. The intersections of the splits—which form a (multi-)dimensional grid—are the trial points for which the objective function is evaluated. The best point is taken as the result. It is possible to use a two-stage grid search. The first grid covers a broad region of the space of possible hyperparameters, but this grid is relatively rough. The best point of the first grid is used as the center of a second and finer grid, and the best point of the second grid is taken as the result. Properties of grid searches are now relatively well investigated. The danger that the algorithm will only find a local optimum far away from the optimum is relatively small, provided the grid covers a broad region and that the grid is fine enough. Of course, searches with a fine grid for large data sets are very time-consuming, if at all possible.

The *Nelder-Mead algorithm* proposed by Nelder and Mead (1965) constructs a simplex of $m+1$ points for an m-dimensional optimization problem. There are variants of the Nelder-Mead algorithm that allow for constraints. For the determination of hyperparameters for SVMs, we typically have $1 \leq m \leq 4$ for classification and regression problems. The functional values are calculated for the vertices of the simplex, and the worst point is reflected through the opposite side of the simplex. If this trial point is best, the new simplex is expanded further out. If the function value is worse, then the second-worst point of the simplex is contracted. If no improvement at all is found, the simplex is shrunken toward the best point. The iteration terminates if the differences in the function values between the best and worst points are smaller than a pre-specified tolerance value. There is the danger that the algorithm will only find a local optimum.

Cross-validation is also a standard technique for finding a suitable set of hyperparameters, especially for small- to moderate-sized data sets. The data set is randomly divided into ℓ (e.g.; $\ell = 10$) disjoint subsets of equal size, and each subset is used once as a validation set, whereas the other $\ell - 1$ sets are put together to form a training set. In the simplest case, the average accuracy of the ℓ validation sets is used as an estimator for the accuracy of the method. The combination of the hyperparameters with the best performance is chosen. As Schölkopf and Smola (2002), among others, explain, there are some possible disadvantages regarding cross-validation, although it is quite often used in practice. One reason is the obvious danger of overfitting because the training data sets and the validation data sets are related to each other. Another point is that a suitable set of hyperparameters obtained for a data

set of size n may differ from a suitable set of hyperparameters obtained for subsets of this data set of size $(1-1/\ell)n$. Often, the smaller data set used for training purposes needs a slightly stronger regularization (e.g., a larger value of λ) and suitable parameters for the kernel and the loss function also may be slightly different.

Many *heuristic choices* have been proposed for the hyperparameters of SVMs. One approach was proposed by Cherkassky and Ma (2004). Their proposal is based on both theoretical considerations and empirical results. The following suggestions for the regularization parameter C, the width of the ϵ-insensitive loss, and the bandwidth parameter γ of the Gaussian RBF kernel are suited for the case where all input variables are scaled to the interval $[0,1]$. They can easily be adjusted to non-scaled data. Regarding the regularization parameter C, Cherkassky and Ma (2004) agree with the findings of Mattera and Haykin (1999) that C should be chosen according to the range of the values of the response variable in the training data. Since the range is not robust against outliers, Cherkassky and Ma (2004) propose $\epsilon := 3\sigma\sqrt{(\ln n)/n}$ and $C := \max\{|\bar{y} - 3\sigma_y|, |\bar{y} + 3\sigma_y|\}$, where $\bar{y}$ and σ_y denote the mean and the standard deviation of the responses y_i in the training data, respectively. Note that this choice of C does not result in a null sequence (λ_n) if $n \to \infty$. In practice, σ_y will be unknown and must be estimated. To accomplish this, Cherkassky and Ma (2004) proposed a nearest-neighbor regression where the number of neighbors is chosen between 3 and 7. The noise will then be estimated using the residuals of this regression. As Cherkassky and Ma (2004) base all their considerations on the RBF kernel, the kernel parameter γ must also be determined. It is chosen depending on the number of input variables of the regression problem, its dimension d, as $\gamma = \sqrt{2}c^{1/d}$, where c is a some constant between 0.1 and 0.5, for which good SVM performance can be achieved. This heuristic method has the advantage that the choice of the hyperparameters can be accessed directly from the data, which allows relativly fast computation. The authors give several numerical examples that show the power of their approach when used on artificial data. It seems to be unknown, however, whether their heuristic choice of (C, ϵ, γ) is always suitable when applied to real-life data.

Momma and Bennett (2002) proposed the *pattern search* algorithm as a directed search method to determine the hyperparameters for SVMs. It examines points in the parameter space that are arranged in a pattern around the actual optimal point. The pattern depends on the number of parameters in the SVM. For SVMs based on the hinge loss and a Gaussian RBF kernel using the logarithms of the parameter value, the pattern with four elements

$$M = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$$

can be used to construct a pattern in the parameter space (C, γ). For the three hyperparameters C, ϵ, and γ for an SVM based on the ϵ-insensitive loss and the Gaussian RBF kernel, this pattern can be expanded to

$$M^* = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}.$$

The columns of M and M^* describe the change applied to a given parameter vector $q = (C, \gamma)^\mathsf{T}$ or $q^* = (C, \epsilon, \gamma)^\mathsf{T}$. This means that only one parameter is changed at a time. The pattern search algorithm itself works as follows.

i) Initialization. Choose a start pattern center $q^{(0)}$ and compute the value of the function to be minimized $g(q^{(0)})$. Furthermore, choose a factor $\Delta^{(0)}$ that denotes the expansion of the pattern and τ, the expansion at which the algorithm should stop.

ii) Optimization step. Compute $q_i^{(k+1)} := q^{(k)} + \Delta^{(k)} m_i$ for all columns m_i of M and the corresponding $g(q_i^{(k+1)})$. If $\min g(q_i^{(k+1)}) < g(q^{(k)})$, set $q^{(k+1)} := \arg\min g(q_i^{(k+1)})$ and $\Delta^{(k+1)} := \Delta^{(k)}$. Otherwise, set $q^{(k+1)} := q^{(k)}$ and $\Delta^{(k+1)} := \Delta^{(k)}/2$ and proceed to the stopping rule.

iii) Stopping rule. If $\Delta^{(k)} < \tau$, stop the algorithm. Otherwise, perform another optimization step.

The algorithm searches the parameter space pattern-wise and memorizes the best hyperparameter combination it comes across. If the center of the pattern is optimal, the pattern will be made smaller, which corresponds to a finer grid search. If the pattern is small enough, the algorithm will stop. In principle, the pattern search works similar to a grid search, but it only makes calculations for a subset of the grid points. By choosing the direction of the steepest descent among pattern points, it will omit a lot of grid points, which may lead to unsatisfactory results when their respective parameter combinations are applied to the data. Furthermore, a more exhaustive search will be done automatically in the region of interest. This can lead to computational savings, but there is the danger that the algorithm will only find a local optimum.

Besides the accuracy, the *number of evaluations* (i.e., the number of combinations of the hyperparameters which are tested in order to find the best combination of the hyperparameters) is also important for practical purposes when several methods are compared.

To the best of our knowledge, there is currently no practical method known that chooses the hyperparameters of SVMs in an optimal manner for all data sets and is applicable for sample sizes of any size. Nevertheless, a few general results concerning how to find suitable hyperparameters for SVMs based on numerical research[3] for benchmark data sets and simulated data sets may be in order. Although for every fixed combination of hyperparameters we have a convex optimization problem to determine an empirical SVM solution, we are in general faced with a non-convex problem when we optimize over the hyperparameters. Typically, there is no single optimal choice of the hyperparameters but a connected region of close to optimal values. The change in

[3] See Christmann *et al.* (2005).

the level of the target function is sometimes approximately parallel to the input parameters, which seems to be one explanation why a pattern search often performs well. If computationally feasible, a fine grid covering a broad range of the parameter space or a two-stage grid often gives a suitable set of hyperparameters, but at a high computational cost. The Nelder-Mead search sometimes performs very poorly if the parameters for inflation or deflation of this algorithm are inappropriately chosen. Some practitioners do not care too much about possible disadvantages of cross-validation because this method often gives good results even if the resulting hyperparameters are not adjusted.

11.4 Software Packages

There exist well-established numerical packages (e.g., `NAG` and `IMSL`™) that can be used to solve convex or quadratic programs. An advantage of these software products is that they are in general numerically very stable. Some of these packages contain routines that are designed for large sparse systems, but this property is usually not needed to compute $f_{\mathrm{D},\lambda}$ as the kernel matrix K is dense (i.e., most coefficients $K_{i,j} = k(x_i, x_j)$ do not equal zero). One can argue that some commercial packages for general use have the disadvantages of a high price and a relatively large computation time for the determination of $f_{\mathrm{D},\lambda}$ because these programs are not specifically designed for SVMs.

Now we will mention a few implementations that were designed for solving the numerical problems of SVMs. A much longer list of programs to compute SVMs can be found, for example, on the websites `www.kernel-machines.org`[4] and `www.support-vector-machines.org`. Note that it is often helpful to scale all input variables to increase numerical stability, provided the implementation does not automatically scale the data.

`LIBSVM`

Chang and Lin (2004) developed a user-friendly library called `LIBSVM` for the computation of SVMs. This software is able to fit SVMs for classification, regression, and distribution estimation problems, is programmed in `C++` and `Java`, and belongs to the state-of-the-art software tools that are currently available for SVMs. In particular, the use of the hinge loss and the ϵ-insensitive loss is possible, as well as using ν-support vector machines. Updates are regularly available. This software is partially based on Fan *et al.* (2005). `LIBSVM` has the advantage that there are interfaces to several other software tools; e.g., to `R`, `MATLAB`®, `Python`, `Perl`, and the data mining software `Weka`. There exists a bundle of related programs called `LIBSVM Tools` and a graphical interface that is very suitable for demonstrating SVM classification and regression to

[4] In March 2008, there were around 45 software tools for SVMs mentioned on this website.

students. Additionally, there is a useful practical guide for SVM classification available written mainly for beginners.

SVM^{light}

Joachims (1999) developed SVM^{light}, which was one of the first implementations to make SVMs applicable in classification and regression problems for large data sets. It is written in C. In particular, one can use the hinge loss and the ϵ-insensitive loss function. Currently, SVM^{light} is one part of the software SVM^{struct} developed by the same author, which is a collection of SVM algorithms for predicting multivariate or structured outputs. It performs supervised learning by approximating a mapping from the input space X to the output space Y using labeled training examples $(x_1, y_1), \ldots, (x_n, y_n)$. Unlike regular SVMs, however, which consider only univariate predictions as in classification and regression, SVM^{struct} can predict complex objects y such as trees, sequences, or sets. Examples of problems with complex outputs are natural language parsing, sequence alignment in protein homology detection, and Markov models for part-of-speech tagging. The SVM^{struct} algorithm can also be used for linear-time training of binary and multi-class SVMs using a linear kernel. SVM^{struct} can be thought of as an API for implementing different kinds of complex prediction algorithms. SVM^{multiclass} is for multi-class classification problems, SVM^{map} has the goal of learning rankings, and SVM^{perf} is useful for learning a binary classification rule that directly optimizes the area under the receiver operating characteristic (ROC) curve or other criteria.

R

The statistical software package R (R Development Core Team, 2006) can be used to compute SVMs for classification and regression problems provided the function svm developed by D. Meyer from the add-on package e1071 is used. This function is based on LIBSVM and uses methods developed by Fan *et al.* (2005). Together with the graphical routines provided by R, this implementation for SVMs is from our point of view especially appropriate for small to moderate sized data sets, for running simulation studies for small data sets, and can easily be used by students. We also like to mention two other R packages. klarR developed at the Department of Statistics of the University of Dortmund contains an interface to SVM^{light} and the package svmpath developed by T. Hastie from the Stanford University can be used to compute the entire regularization path for an SVM based on the hinge for small data sets.

mySVM

Rüping (2000) developed the implementation mySVM for SVMs for classification, regression, and distribution estimation problems. This implementation

is based on SVM^{light}. The software can also be used for SVMs based on the pinball loss for quantile regression if the options epsilon=0, L+= $1-\tau$, and L-= τ are specified for the pinball loss function $L_{\tau\text{-pin}}$. There exists a Java implementation of mySVM designed to run inside of a database.

myKLR

Keerthi *et al.* (2005) developed a fast SMO-type algorithm for SVMs based on the logistic loss function for classification purposes. The algorithm uses many technical tricks and special properties of this particular loss function to solve the dual problem efficiently. The software myKLR is an implementation of this algorithm and was written by Rüping (2003). This implementation is much faster than quasi-Newton algorithms such as the Broyden-Fletcher-Goldfarb-Shanno algorithm with bound constraints applied to the primal problem for large data sets. Nevertheless, myKLR needs considerably more computation time than comparable SMO algorithms for SVMs based on the hinge loss. This is due to the fact that the empirical solution of SVMs based on the logistic loss is not sparse and that a convex and not (only) a quadratic optimization problem as for the case of the hinge loss must be solved.

LS-SVMlab

LS-SVMlab is a toolbox for SVMs based on the least squares loss function and uses methods described in the textbook by Suykens *et al.* (2002). This software is written in MATLAB® and C and contains besides implementations to solve SVMs in classification and regression problems routines for kernel-based principal component analysis.

11.5 Further Reading and Advanced Topics

Section 11.1, which showed that empirical SVM decision functions $f_{\mathrm{D},\lambda}$ are solutions of special convex or even quadratic programs with constraints, is mainly based on Schölkopf and Smola (2002), Cristianini and Shawe-Taylor (2000), and Smola and Schölkopf (2004). More details on kernel logistic regression can be found in Keerthi *et al.* (2005). We refer to Schölkopf *et al.* (2000) and Smola and Schölkopf (2004) for additional information regarding ν-support vector regression and related topics and to the textbook by Suykens *et al.* (2002) for SVMs based on the least squares loss. For quantile regression, we refer to Koenker and Bassett (1978), He (1997), Koenker (2005), and Takeuchi *et al.* (2006). For problems with monotonicity constraints, we refer to Takeuchi *et al.* (2006).

Section 11.2, on implementation techniques to compute SVMs, is mainly based on Keerthi *et al.* (2001), Fan *et al.* (2005), and Chen *et al.* (2006).

These papers also investigate generalizations of the algorithms given here. Chen *et al.* (2006) also offer results that show that we "only" have linear convergence for decomposition methods based on the algorithm WSS2 for SVMs based on the hinge loss. Many of these results are valid for SVMs based on the ϵ-insensitive loss function and for one-class SVMs, too. We conjecture that this is also true for SVMs based on the pinball loss, but as far as we know, this has not yet been proven. For additional details on decomposition methods, we refer to Osuna *et al.* (1997), Joachims (1999), and Platt (1999). Some improvements for Platt's SMO algorithm for SVM classifiers were proposed by Keerthi *et al.* (2001). The optimization problem (11.41) related to the maximal violating pair algorithm was probably first considered by Joachims (1999). List and Simon (2004) give a general convergence theorem for the decomposition method. Hush and Scovel (2003) propose a polynomial-time decomposition algorithm for SVMs and prove necessary and sufficient conditions for stepwise improvement of their algorithm. For general polynomial time decomposition algorithms, we refer to List and Simon (2007). As far as we know, it is not yet known whether existing SVM algorithms satisfy the conditions, but the authors also provide an algorithm that fulfills the conditions. Let $c(\tilde{K})$ denote the maximum of the norms of the (2×2) submatrices determined by restricting $\tilde{K}$ from the dual program to two indices. If the constant $C = 1/(2n\lambda)$ satisfies $\sqrt{1/2} \le C \le nc(\tilde{K})$, then this algorithm for the computation of an empirical SVM solution based on the hinge loss needs at most $4c(\tilde{K})C^2n^4/\varepsilon$ iterations with a guaranteed precision of ε. For a formal analysis of stopping criteria of decomposition methods for SVMs, we refer also to Lin (2002a), Chen *et al.* (2006), and List *et al.* (2007).

For leave-one-out estimates, we refer to Schölkopf and Smola (2002, Chapter 12), Joachims (2002), and Mukherjee *et al.* (2006). Seeger (2007) proposed cross-validation optimization for large-scale hierarchical classification kernel methods, and the kernel hyperparameters are chosen automatically by maximizing the cross-validation log likelihood in a gradient-based way. Keerthi *et al.* (2007) proposed an efficient method for gradient-based adaptation of the hyperparameters. Davies *et al.* (2008) discussed general nonparametric regression as an example of model choice.

There is a large and rapidly increasing body of literature on implementation techniques for SVMs. Much more information than in Section 11.2 can be found for example in Schölkopf and Smola (2002, Chapter 10) and Cristianini and Shawe-Taylor (2000, Chapter 7). Keerthi *et al.* (2005) proposed a fast dual algorithm for SVMs based on the logistic loss for classification based on an SMO decomposition. This algorithm is implemented in the software `myKLR` (Rüping, 2003). An overview of SVM solvers is given by Bottou and Lin (2006). Joachims (1999), Osuna and Girosi (1999), Platt (1999), Huang *et al.* (2006), and Bottou *et al.* (2007) describe techniques especially designed for making SVMs applicable for large data sets. Joachims (2002) considers fast algorithms for SVMs in the context of text classification. Smola and Schölkopf (2004) describe methods for the numerical computation

of SVMs, with special emphasis on regression problems. For data sets with millions of data points, a subsampling strategy such as robust learning from bites may be useful, too.

Most literature on computational aspects of SVMs currently concentrates on solving the dual optimization problem, but there is increasing interest also in algorithms that solve the primal problem of SVMs. Mangasarian (2002) proposed a finite Newton method for classificaton purposes. Keerthi and DeCoste (2005) proposed an algorithm to solve the primal problem of linear SVMs based on the least squares loss function; see also Suykens *et al.* (2002) for such SVMs. Joachims (2006) developed an algorithm and software to train linear SVMs in *linear* time. Chapelle (2007) argued that the primal problem can often be solved efficiently both for linear and nonlinear SVMs. This also offers the opportunity to investigate new families of algorithms for large-scale SVMs.

Corresponding to the large number of implementation techniques that were proposed for the numerical computation of $f_{\mathrm{D},\lambda}$, there exist many software implementations. A longer list of implementations than the one we gave for the computation of SVMs and related methods can again be found on the websites `www.kernel-machines.org` and `www.support-vector-machines.org`.

If the number of input variables d is very large, *feature selection* can be helpful to increase the precision and to decrease the computational burden of SVMs. Many researchers proposed feature selection methods or compared such methods. A general framework for feature selection is described by Schölkopf and Smola (2002, Chapter 14). We refer to Guyon *et al.* (2002) for recursive feature elimination in the context of gene selection for cancer classification using SVMs. Krishnapuram *et al.* (2004) considered joint feature selection and classifier design in the context of gene expression analysis, and Hochreiter and Obermayer (2004) applied SVMs in the context of gene selection for microarray data. Neumann *et al.* (2005) considered combined SVM-based feature selection and pattern recognition. Their approach is based on additional regularization and embedded nonlinear feature selection and uses difference of convex functions programming from the general framework of non-convex continuous optimization. Cai *et al.* (2007) compared several feature selection and classification algorithms to identify malicious executables and found SVM classifiers to be superior in terms of good prediction accuracy, short training time, and low danger of overfitting. Song *et al.* (2007) investigated supervised feature selection via dependence estimation.

11.6 Summary

The empirical SVM decision function $f_{\mathrm{D},\lambda}$ is defined as the solution of a minimization problem over an infinite-dimensional reproducing kernel Hilbert space H that can have an infinite dimension. Nevertheless, $f_{\mathrm{D},\lambda}$ can be evaluated numerically by solving a finite-dimensional convex program.

Many classical numerical algorithms for solving such convex programs are not well-suited to compute $f_{D,\lambda}$ for large sample sizes n. However, there are algorithms to compute $f_{D,\lambda}$ efficiently even for large values of n. Some of these algorithms are based on sequential minimal optimization (SMO). One main advantage of such algorithms is that it is not necessary to store the $(n \times n)$ matrix $K = (k(x_j, x_i))$ in the memory of the computer.

There are loss functions L such that the numerical problem of computing $f_{D,\lambda}$ can be solved relatively quickly. Among those loss functions are the hinge loss and the least squares loss for classification, the ϵ-insensitive loss function and the least squares loss function for regression, and the pinball loss function for kernel-based quantile regression.

There exist loss functions such that not only $f_{D,\lambda}$ can be computed relatively quickly but it also has good robustness properties in the sense of Chapter 10. Examples are the hinge loss, the ϵ-insensitive loss, and the pinball loss. The least squares loss is not Lipschitz-continuous and yields usually non-robust estimates.

It is not always suitable to use a loss function fulfilling the above-mentioned properties of fast computation and robustness. One counterexample is the hinge loss, which does not allow estimation of the conditional probabilities $P(Y|x)$. In contrast, it is possible to estimate these conditional probabilities based on the Lipschitz-continuous logistic loss function for classification problems. The empirical SVM decision function $f_{D,\lambda}$ based on this loss function offers good robustness properties if used in combination with a bounded universal kernel (e.g., the Gaussian RBF kernel) but has the disadvantage of a substantially higher computation time for large sample sizes compared with the hinge loss.

The SVM decision function $f_{D,\lambda}$ and the corresponding empirical risk $\mathcal{R}_{L,D}(f_{D,\lambda})$ depend critically on hyperparameters such as λ and parameters used by the loss function and the kernel. Currently, there seems to be no easy and computationally fast way to determine these hyperparameters for all data sets in an optimal manner, although different computationally intensive methods can offer a suitable choice.

11.7 Exercises

11.1. Numerical exercise ($\star$)
Compute $f_{D,\lambda}$ and make plots similar to those in Figure 10.11 for the daily milk consumption data set given in Table 10.1 using the ϵ-insensitive loss function and a Gaussian RBF kernel and a polynomial kernel. Use different values of the hyperparameters and study their effect.

Hint: Use, for example, one of the software products `LIBSVM`, $\texttt{SVM}^{light}$, or `mySVM`, or the function `svm` of the `R`-package `e1071`.

11.2. SVM based on hinge loss ($\star$)
Derive the Lagrangian and the dual problem for the computation of $f_{D,\lambda}$

based on the hinge loss function. Furthermore, consider an SVM based on L_{hinge} and the classification problem with an additional offset term $b \in \mathbb{R}$. Derive the primal convex program, the Lagrangian L^*, and the dual program for the computation of $(f_{\text{D},\lambda}, b_{\text{D},\lambda})$.

Hint: Schölkopf and Smola (2002).

11.3. SVM based on logistic classification loss $(\star)$
Derive the Lagrangian and the dual problem for the computation of $f_{\text{D},\lambda}$ based on the logistic loss for classification. Furthermore, consider an SVM based on $L_{\text{c-logist}}$ and the classification problem with an additional offset term $b \in \mathbb{R}$. Derive the primal convex program, the Lagrangian L^*, and the dual program for the computation of $(f_{\text{D},\lambda}, b_{\text{D},\lambda})$.

Hint: Keerthi *et al.* (2005).

11.4. SVM based on least squares loss for classification$(\star)$
Work out the details for Example 11.5. Compute the Lagrangian and its partial derivatives. Show that $f_{\text{D},\lambda}$ is the solution of a set of linear equations. Repeat the calculations for the case of an additional offset term $b \in \mathbb{R}$.

Hint: Suykens *et al.* (2002).

11.5. SVM based on distance-based loss $(\star)$
Work out the details for Example 11.6. Derive L^* and the dual program.

Hint: Smola and Schölkopf (1998) and Schölkopf and Smola (2002).

11.6. SVM based on ϵ-insensitive loss $(\star)$
Derive the Lagrangian and the dual problem for the computation of $f_{\text{D},\lambda}$ based on the ϵ-insensitive loss function. Furthermore, consider an SVM based on $L_{\epsilon\text{-insens}}$ and the regression problem with an additional offset term $b \in \mathbb{R}$. Derive the primal convex program, the Lagrangian L^*, and the dual program for the computation of $(f_{\text{D},\lambda}, b_{\text{D},\lambda})$.

Hint: Smola and Schölkopf (1998).

11.7. SVM based on least squares loss for regression $(\star)$
Consider a regression problem with $X = \mathbb{R}^d$ and $Y = \mathbb{R}$. Derive the primal program, the Lagrangian L^*, and the dual program for the computation of $f_{\text{D},\lambda}$ based on $L_{\text{LS}}(y,t) = (y-t)^2$, $y, t \in \mathbb{R}$. Explain why $f_{\text{D},\lambda}$ can be computed relatively quickly even for large sample sizes n. Furthermore, consider an SVM based on this loss function and a regression problem with an additional offset term $b \in \mathbb{R}$. Derive the primal convex program, the Lagrangian L^*, and the dual program for the computation of $(f_{\text{D},\lambda}, b_{\text{D},\lambda})$.

Hint: Suykens *et al.* (2002).

11.8. SVMs based on the pinball loss for quantile regression $(\star)$
Work out the details for Example 11.9. Compute also the Lagrangian and the dual program.

Hint: Schölkopf and Smola (2002) and Takeuchi *et al.* (2006).

12

Data Mining

Overview. *Support vector machines are often used in data mining. In this case, SVMs are only one part of a complex process. This chapter describes a general data mining strategy and explains which role SVMs can play within this process. Some competitors of SVMs and software tools for data mining are briefly described.*

Prerequisites. *Basic knowledge on SVMs from Chapter 1.*

Support vector machines and other kernel-based techniques treated in this monograph have two main areas of application: risk minimization in machine learning and data mining. In this chapter, we describe the data mining process and explain the role of SVMs in this process.

The goal in standard areas of statistical machine learning is to minimize the empirical risk. Here it is not of primary importance to extract knowledge about the internal structure of the data set as long as the empirical risk is minimized. A typical example is the automatic classification of incoming emails as "no spam" or "spam". Other examples are detection of credit card fraud and the automatic recognition of hand-written digits. SVMs are often successfully applied in these cases, although the data analyst usually has no or only vague prior information about the probability distribution P that generated the data set. In the former chapters of this monograph, we gave a theoretical foundation of why SVMs based on appropriate choices of the loss function and the kernel and using suitable hyperparameters are able to learn, which makes SVMs especially valuable for these standard areas.

Another field where SVMs and related kernel methods are successfully applied is data mining projects, where these methods are one—but only one—cornerstone of the whole process. The main goal of data mining projects is generally to extract and model formerly unknown information contained usually in large and complex data sets. It seems unrealistic to hope that the application of SVMs can really be successful in data mining without some knowledge about the general data mining process. Therefore, this chapter gives a short overview of data mining and describes the main phases in such projects.

In Section 12.1, we give a definition of data mining and explain why data mining is important. In Section 12.2, we describe a general data mining strategy called CRISP-DM, which was proposed by Chapman *et al.* (2000). CRISP-DM is the abbreviation of CRoss-Industry Standard Process for Data

Mining. In Section 12.3, we explain the role of SVMs as one part in the whole data mining process. Section 12.4 mentions a few software tools for data mining. Section 12.5 contains information about further literature, and Section 12.6 gives a summary of this chapter.

12.1 Introduction

Hand *et al.* (2001, p. 1) define data mining in the following way.

> *Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.*

This definition contains several keywords. Data mining deals with *observational data sets*, which means that the data set is generally not collected only for the purpose of the data mining project. Furthermore, observational data are generally not random samples but often quite large and a number of cases between 10^5 and 10^7 is not unusual. The owner of such a data set sometimes has prior information about the data or the data-generating process such that the goal is to *extract new information* from the data. The kind of information desired grossly differs between data mining projects. Examples are a model with low prediction error, the identification, of high-risk subgroups or the detection of dependencies between attributes. The result of a data mining project is not only the extraction of new information but also to make the result applicable in practice. The information obtained should therefore be summarized in a way that offers high interpretability both from a business and a mathematical point of view. Thus, one can argue that data mining is more than the application of modeling techniques either from parametric statistics, semi-parametric statistics, or non-parametric statistical machine learning theory to large data sets. Some examples of data mining projects are:

- customer relationship management (CRM): customer acquisition, customer assessment, and customer churn analysis
- eCommerce: prediction of sales and detection of associations between customers and products
- text mining and web mining
- credit risk scoring: banking
- insurance tariffs and identification of high-risk subgroups
- analysis of gene expression data: microarray experiments

In the next section, we describe one particular strategy for data mining.

12.2 CRISP-DM Strategy

The CRISP-DM project (Chapman *et al.*, 2000) has developed an industry- and tool-neutral data mining process model. CRISP-DM is the abbreviation of CRoss-Industry Standard Process for Data Mining and was developed by DaimlerChrysler AG (Germany), Teradata being a subdivision of NCR Systems Engineering Copenhagen (USA and Denmark), OHRA Verzekeringen en Bank Groep B.V. (The Netherlands), and the statistical software company SPSS®(USA). The project was partially funded by the European Commission under the ESPRIT program. Starting from the knowledge discovery processes used in industry today and responding directly to user requirements, this project defined and validated a data mining process that is applicable in diverse business sectors. The goal of CRISP-DM is to make large data mining projects faster, cheaper, more reliable, and more manageable.

The CRISP-DM strategy consists of the six main phases shown in Figure 12.1 and described in the rest of this section.

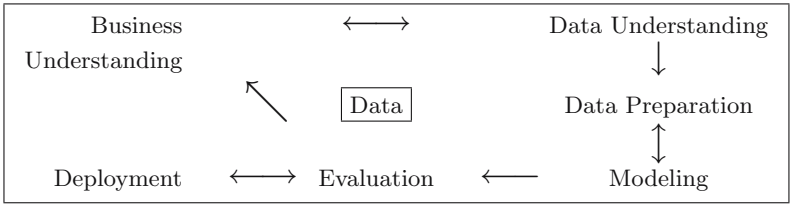


Fig. 12.1. Main phases of data mining according to the CRISP-DM strategy.

Phase 1: Business Understanding

The aim in this phase is the determination of the objectives from a *business perspective*. Often the customer has several competing objectives and constraints that must be balanced. The analyst has the goal of uncovering important but unknown factors that can influence the final outcome. A description is usually given of any solution currently in use for the problem together with a list of their advantages and disadvantages. Further, the *business success criteria* for a successful or at least useful outcome to the project are determined from an applied point of view. Some examples for such criteria are the following:

- Estimate the probability of an event (e.g., a genomic defect) given a list of inputs.
- Determine an insurance tariff that minimizes the risk (i.e., the expectation of the losses).

- Give useful insights as to which customers produce the most expensive claims for an insurance company.
- Improve the response rate in a direct marketing campaign by at least 5 percent.
- Identify the subgroup of patients having a certain type of cancer who will have the highest benefit from a new drug.

At this early stage of a data mining project, one assesses the current situation by considering all resources, constraints, and assumptions that should be considered in the determination of the data analysis goal and by the project plan. One should take into account the available or necessary personnel, type of data files, computing resources, available data mining software tools, and other relevant software. All requirements of the project, including scheduled date of completion, comprehensibility, and quality or precision of the results are listed. Of course, one should make sure that access to the data files is allowed and possible from a technical point of view. All the project-specific assumptions should be listed, no matter whether they can be checked during the data mining process or not. Assumptions are made for the precision of the estimates. The constraints made on the project are listed (e.g., lack of resources to carry out some tasks within the timescale or legal or ethical constraints). An additional aspect is the determination of lower bounds on the required sample size such that the conclusions can be made with a desired precision. All assumptions on external factors such as competitive products or technical advances should be listed. A decision should be made whether the final model should be interpretable in business terminology or not because this can easily influence the choice of the modeling technique. The results of SVMs for example are often harder to interpret from a business point of view than for parametric models such as generalized linear models; see Table 12.2. Additionally, the starting point and the endpoint of the project are listed together with possible risks depending on the size of the data or the data quality.

A cost versus benefit analysis for the data mining project is prepared that compares the cost of the data mining project with the potential benefit to the business. Of course, a data mining project will in general only be done if the potential benefit dominates the cost.

Then the *data mining success criteria* are determined in statistical terms. All business questions are translated into *data mining goals*. For example, a direct-marketing campaign requires segmentation of customers in order to decide who to approach in the campaign, and one should also specify the size of the segments. During this phase, one specifies the type of data mining problem; e.g., classification (see Chapters 5 and 8) or regression (see Chapter 9). Additional criteria for model assessment are specified such as model accuracy, performance, and complexity. Furthermore, benchmarks for the evaluation criteria are determined. An example is the level of the predictive accuracy.

Then a *project plan* is made that lists all stages of the data mining project. The project plan should include duration, required resources, inputs, outputs, and dependencies. Dependencies between the time schedule and risks are described. Recommendations for actions are also given for the case where risks appear. The project plan also contains a list of people, specifying who is responsible for which steps. From our point of view, it is essential for a successful data mining project to take the following rule of thumb into account during the construction of the project plan:

50%–70% of the time and effort for the data preparation phase;
15%–25% for the data understanding phase;
10%–20% for the business understanding phase, modeling, and evaluation;
5%–10% for the deployment phase.

Phase 2: Data Understanding

The first task in this phase is the *collection of initial data* listed in the project resources. Therefore, a list of necessary data sets or databases and their types is constructed. Further, the software tools and the methods to acquire them are listed. If problems are encountered, they should also be listed.

A *data description report* is made that describes the main properties of the data, including

- quantity of data (d : number of variables or attributes, n : number of cases or records);
- format of the data;
- coding, percentage, and patterns of missing values;
- identifier variables needed for merging data from different databases or tables;
- time period when the data were collected.

Then a *data exploration report* is made that describes the distribution of the key attribute(s); e.g., the main response variables (or target attribute) of a prediction problem. A list of possible values and a contingency table are given for categorical variables. For continuous variables, some descriptive statistics are listed; e.g., minimum and maximum values, mean and standard deviation, or their robust pendants, such as median and median absolute deviation (MAD). Furthermore, low-dimensional relationships and dependencies between pairs or a small number of attributes are computed taking the scales (nominal, ordinal, continuous) of the attributes into account. This report also describes properties of interesting subpopulations for further examination; e.g., stratification by gender, age, or geographical region.

Additionally, the *data quality report* lists the results of the data quality verification. It also mentions possible solutions for the case of quality problems. Such solutions often depend on deep knowledge of the business and of the data itself. Many authors argue that the data quality is essential for success in data mining projects; see, e.g., Hipp *et al.* (2001).

Phase 3: Data Preparation

The goal of this phase is to obtain clean data sets or databases that can be used in the modeling phase. First, *data selection* is done by deciding which attributes will be included or excluded in the next phases. The selection covers the selection of attributes (columns) and the choice of cases (rows). Possible criteria for the decisions are the relevance to the data mining goals, the percentage of missing values, and the data quality.

Then a *data cleaning report* is made that describes the actions to increase the data quality. The report describes which actions were taken to overcome the data quality problems. If missing values are replaced by imputation methods or other strategies (see Rubin, 1987), the report should describe which methods were used and how many data points were modified.

The *construction of the clean data sets* includes data preparation operations such as correction of typing errors and the transformation of existing attributes (by using logarithms, square roots, Box-Cox transformations, indicator variables, etc.) and by defining derived attributes. As an example, we mention the definition of the body mass index (BMI) for adults, which is calculated by the following metric formula:

$$\text{BMI} = \frac{\text{height (in meters)}}{\text{weight}^2\text{(in kilograms)}} .$$

These BMI values are usually classified into groups such as "underweight": BMI below 18.5; "normal": BMI between 18.5 and 24.9; etc. There exist modified formulas to compute the BMI for children and teens, taking gender and age into account. The goals of derived attributes such as the body mass index are twofold: a reduction of dimensionality and ease of interpretation.

The next step of the data preparation phase is the *integration of data*. This is generally needed because data from multiple tables or data sets have to be merged together or new cases must be created about the same object or person. Sometimes merged data also cover aggregations that are operations to summarize information from multiple cases or tables. As an example, we mention a data mining project for analyzing insurance data. Assume that there are three tables partially due to data security:

- Table A, with a unique index variable, say ID, and personal and demographic information about the customers;
- Table B, containing the ID variable, the number of claims, the claim amount for the current year, and possible explanatory variables (inputs);
- Table C, containing the ID variable and the claim history covering the last decade.

Here the ID variable is needed to merge the data belonging to a single customer together into one large table. A useful aggregation step in this example is the construction of new variables for the sum of years without a claim, the total

number of claims per year, and the average claim amount per year for each customer.

The final step in this phase is *formatting the data*, which is generally required by the modeling tool and to increase the readability of the results. It is often useful to format the values of categorical attributes such that the preferred reference class is the first class or the last class. It can be helpful to convert text attributes into uppercase after trimming blanks.

Phase 4: Modeling

First, one has to *select the modeling technique(s)*, taking the data mining goals, the properties of the data, and the plausibility of model assumptions into account. One advantage of kernel methods including SVMs is that these non-parametric methods only need rather weak assumptions in comparison with parametric methods. Support vector machines are of course only one class of modeling techniques used in data mining projects, and we would like to mention three strong competitors: generalized linear models, generalized additive models, and tree-based methods. These methods are implemented in several data mining tools.

Generalized Linear Models

A generalized linear model (GLIM) is a parametric regression model having three components: a stochastic component, a linear predictor, and a link function; see Nelder and Wedderburn (1972), McCullagh and Nelder (1989), and Fahrmeir and Kaufmann (1985). The *stochastic component* assumes that the $(d+1)$-dimensional random variables (X_i, Y_i), $i = 1, \ldots, n$, are stochastically independent and that the conditional distribution of Y_i given $X_i = x_i$ is an element of an exponential family (see (A.17)). The *linear predictor* describes the assumption that the vector x influences Y only via the linear combination $\eta = x^\mathsf{T}\theta$. The bijective *link function* g connects the linear predictor to the conditional expectation $\mu(\theta) := \mathbb{E}_{\mathrm{P}_\theta}(Y|x)$ via $g(\mu(\theta)) = \eta$. GLIMs make implicitly a hard additional assumption, namely that there is a functional relationship specified by some fixed variance function $v : \mathbb{R} \to [0, \infty)$ between the conditional expectation and the conditional variance of Y given x. Table 12.1 lists the conditional distribution of Y given x, the link function, and the variance function of special GLIMs. Note that the variance function is a polynomial in these cases. This simple relationship between expectation and variance of the response variable can be grossly violated in practice, see Christmann (2005) for a data set from insurance companies.

The classical estimator of $\theta \in \mathbb{R}^d$ is the maximum likelihood (ML) estimator, which has nice properties (consistency in probability or almost sure, convergence rate of $n^{-1/2}$, asymptotic efficiency, and asymptotic normality) if the assumptions of the generalized linear model are satisfied; see Fahrmeir and Kaufmann (1985, 1986). These asymptotic properties of the ML estimator

Table 12.1. Important special cases of GLIMs. Here Λ and Φ denote the cumulative distribution functions of the standard logistic and standard Gaussian distributions, respectively.

Model	Distribution of $Y\|x$	Link Function	Variance Function
Linear regression	Normal	identity: $\eta = \mu$	$v(\mu) = 1$
Logistic regression	Binomial	logit: $\eta = \Lambda^{-1}(\mu)$	$v(\mu) = \mu(1-\mu)$
Probit regression	Binomial	probit: $\eta = \Phi^{-1}(\mu)$	$v(\mu) = \mu(1-\mu)$
Poisson regression	Poisson	$\eta = \ln(\mu)$	$v(\mu) = \mu$
Gamma regression	Gamma	$\eta = 1/\mu$	$v(\mu) = \mu^2$
Gamma regression	Gamma	$\eta = \ln(\mu)$	$v(\mu) = \mu^2$
Inverse Gaussian regression	Inverse Gaussian	$\eta = \mu^{-2}$	$v(\mu) = \mu^3$
Negative Binomial regression	Negative Binomial	$\eta = \ln(\mu)$	$v(\mu) = \mu + k\mu^2$

allow the construction of asymptotically optimal hypothesis tests and confidence regions for β. However, this estimator can have two serious drawbacks. It may not exist for some data sets; see Albert and Anderson (1984) and Santner and Duffy (1986). Furthermore, the maximum likelihood estimator is non-robust in several special cases, including linear regression and logistic regression. In a rather informal way, one can describe robust methods by the property that small violations of the model assumptions should have only a small and bounded impact on the result; see Chapter 10 for details. Robust alternatives to the maximum likelihood estimator were proposed for example by Rousseeuw (1984) and Rousseeuw and Yohai (1984) for linear regression, and Künsch *et al.* (1989) and Christmann (1994, 1998) for generalized linear models; see Chapter 10.

Generalized Additive Models

A generalized additive model (GAM) is a semi-parametric regression model having three components: a stochastic component, an additive predictor, and a link function; see Hastie and Tibshirani (1990). In contrast to the linear predictor $\eta = x^{\mathsf{T}}\beta$ used by GLIMs, a GAM allows the explanatory variables $x = (x_1, \ldots, x_\ell, \ldots, x_d) \in \mathbb{R}^d$ to influence the conditional distribution $Y|x$ via the more flexible way

$$\eta = \alpha + (x_1, \ldots, x_\ell)^{\mathsf{T}}\beta + \sum_{j=\ell+1}^{d} f_j(x_j)\,, \tag{12.1}$$

where the intercept term $\alpha \in \mathbb{R}$, the slope parameters $\beta \in \mathbb{R}^\ell$, and the functions $f_j : \mathbb{R} \to \mathbb{R}$, $j = \ell+1, \ldots, d$, $0 \le \ell \le d$, are unknown and must be estimated from the data. Linear regression methods, including parametric splines

and Fourier fits or univariate regression smoothers such as local polynomial regression, are used to estimate the unknown functions f_j, $j = \ell + 1, \ldots, d$. The backfitting algorithm is often used to fit the unknown functions in an iterative manner; see Hastie *et al.* (2001, p. 260). Second-order interactions (i.e., $f_{j_1,j_2}(x_{j_1}, x_{j_2})$ with $\ell < j_1 < j_2 \leq d$) can be included by using surface smoothers, but such smoothers increase the computational effort. It can be quite hard to model higher-order interactions using such smoothers in a nonparametric way with generalized additive models due to the curse of high dimensionality. SVMs can be useful here to fit the non-parametric part of GAMs, including higher-order interactions.

Tree-Based Methods

Finally, we mention tree-based methods as important alternatives to support vector machines; see Breiman *et al.* (1984) and Hastie *et al.* (2001). Special cases are classification trees and regression trees. The basic principle of tree-based methods is quite different from those of GLIMs and GAMs. Trees partition the space X defined by the input variables in a recursive manner to maximize a score of class purity, which means that the majority of the data points in each cell of the partition belong to one class for the case of classification trees. The construction of the tree begins with the whole data set (i.e., all data points are in the same cell). A tree-based method then computes the best split or partition consisting of ℓ cells of the whole data set such that the data points in each cell are significantly more homogeneous than data points of different cells. The cells are called *nodes*. A binary splitting rule ($\ell = 2$) is often preferred to a multi-way splitting rule ($\ell > 2$), which may fragment the data too quickly, leaving insufficient data at the next stage. Then the procedure is repeated: each node is again split into ℓ subnodes. This recursive technique is repeated until a user-defined stopping rule is satisfied. Finally, a tree constructed in this way is usually truncated to the most important splits and the corresponding nodes. This step is called *pruning*. Tree-based methods differ with respect to the splitting, stopping, and pruning rules. Trees often perform well for low-dimensional data sets and the computation time is often short, but trees can have problems in modeling complex high-dimensional dependency structures.

In phase 4 of CRISP-DM, one has to decide which modeling techniques are appropriate for the project. This decision is often non-trivial, because many aspects should be taken into account. A—partially subjective—comparison of advantages and disadvantages of GLIMs, GAMs, trees, and SVMs is given in Table 12.2. Data quality can play a major role in data mining projects, as was explained above. Furthermore, poor data quality makes it hard to justify hard parametric model assumptions, and large databases often contain a small amount of typing errors, different types of outliers, and data points measured in an imprecise manner. Hence, knowledge from the first three data mining

phases is also useful to decide whether robustness aspects are important for the specific data mining project.

Table 12.2. Properties of several modeling techniques (mod.: moderate; ML: maximum likelihood estimation; robust: robust estimation).

Characteristic	GLIM (ML/robust)	GAM	Trees	SVMs
Type of model	parametric	semi-parametric	non-parametric	non-parametric
Predictive power	bad/mod.	mod.	bad	good
Ability to extract linear combinations of features	good	good	bad	good
Ability to detect complex dependencies	bad	mod.	good	good
Natural handling of data of mixed type*	mod.	mod.	good	mod.
Handling of missing values	bad	bad	good	mod.
Interpretability	good	good	mod.	low–mod.
Dependency on hyper-parameters	no/yes	yes	yes	yes
Robustness w.r.t. outliers in inputs	bad/good	bad	mod.–good	mod.–good
Robustness w.r.t. outliers in outputs	bad/good	bad	mod.–good	mod.–good
Insensitive to monotone transformations of inputs	bad	mod.	mod.–good	mod.
Computation time	good	mod.	mod.	mod.
Computational scalability (large n)	good	mod.	mod.	mod.
Availability in data mining tools	good	mod.	good	mod. (increasing)

* Nominal, ordinal, continuous.

Generating a test design is the next step in phase 4 of CRISP-DM. A description of the intended plan should be given to ensure that the criteria to measure the empirical risk or the goodness of the predictions are fair and that the data set was not overfitted. A common practice in data mining projects is to split the data set randomly into three parts called the training data set, validation data set, and test data set. The training data set is usually modeled several times with the same modeling technique but with different sets of so-called hyperparameters. As an example, we mention the support vector

regression: the hyperparameters are $(\epsilon, \gamma, \lambda) \in (0, \infty)^3$ if the ϵ-insensitive loss function $L_{\epsilon\text{-insens}}$, the Gaussian RBF kernel $k_{\text{RBF}}(x, x') = \exp(-\gamma^{-2}\|x - x'\|_2^2)$, $x, x' \in \mathbb{R}^d$, and the regularizing constant λ are used; see Chapters 4 and 9. The validation data set is used for fine-tuning purposes. The models learned from the training data set are applied to the validation data set to obtain an optimal setting of the hyperparameters yielding the best properties of the modeling technique for the validation data set. The test data set is necessary to obtain a fair measure of how well the modeling technique actually works for *unseen* data points never used before.

Additionally, the procedure should be described as how to divide the whole data into these disjoint parts (e.g., by simple random sampling without replacement or proportional stratified random sampling based on a certain stratification attribute). Proportional stratified random sampling involves dividing the whole data set into homogeneous and disjoint subgroups defined by the values of the stratification attribute and then taking a simple random sample in each subgroup. Which attribute is appropriate for stratification purposes depends on the data mining project. A common rule in stratified random sampling is that data points in the same stratum should be more similar to each other with respect to the response variable than data points of different strata. Often a stratification by gender, age group, geographical region, or a combination of these variables is useful.

There are several reasons why a stratified sampling may be preferable over simple random sampling; see Cochran (1977) or Levy and Lemeshow (1999). It assures that results can be obtained not only for the overall data set but also for interesting subgroups of the data set. This can be quite important, for example, to assure that the training, validation, and the test data sets all contain a reasonable number of people for interesting minority groups. If one data mining goal is to obtain new knowledge about subgroups, this may be the only way to effectively assure that this goal can be achieved. Finally, if one of the subgroups is extremely small, one can use different sampling fractions within the different strata to randomly oversample the small group, which results in non-proportional stratified random sampling. It is of course necessary in this situation to weight the within-strata estimates depending on the sampling fraction whenever overall population estimates are needed.

Cross-validation can be an alternative to the splitting approach described above especially for data sets of only moderate size.

In the *model building step*, the modeling methods are run on the prepared and carefully cleaned data sets to obtain the following outputs:

- The best choice of the *hyperparameters* that was detected for the validation data set is listed. This is done for each modeling method that depends on such parameters.
- The *fitted models* are given using these hyperparameters together with predictions $\hat{y}_i$ for all data points in the test data set.

- A *model description* of the fitted models is also given. The description contains, for example, the risk evaluated for the test data set and a corresponding list of significant explanatory variables.

The final step within this CRISP-DM phase is the *model assessment*. The statistician has to interpret the models, taking into account the data mining success criteria, domain knowledge, and the desired test design. A ranking of the models is useful according to the formerly specified evaluation criteria such as accuracy or generality of the model if different models were fitted.

It is not unusual for the hyperparameters to be revised several times due to fine-tuning of the model. This results in an iteration of the model building step and the model assessment step until no essential improvement can be made. It is recommended to document all such revisions and assessments for future phases and for future data mining projects of similar type.

Phase 5: Evaluation

This phase starts with the *evaluation of the results*. The statistician generally discusses the results with business analysts and domain experts to ensure that not only the narrower technical success criteria treated in phase 4 are (hopefully) met but also the business success criteria. One goal is to determine whether there exists a business-relevant reason why the "optimal" model obtained during the former phase is deficient. Furthermore, the model can be evaluated to check whether the test application meets all budgets or time constraints.

The *review process* summarizes the whole data mining process so far and describes the unexpected findings, missed activities, and overlooked potential risk factors and lists actions that should be repeated.

Finally a decision is made as how to proceed. Are the data mining results sufficiently worthwhile to move to the deployment phase, or should the project stop?

Phase 6: Deployment

If the data mining results were strong enough to justify deployment into the business, a *deployment plan* is made. It describes a strategy for how the relevant findings of the data mining project can become part of the business solution used in practice. A *monitoring plan* and a *maintenance plan* can help to avoid unnecessarily long periods of wrong or suboptimal usage of the data mining results. Of course, *final reports* and a *final presentation* are made at the end of the data mining project. A final *project review* is also useful to document experience that can be important for further data mining projects, such as pitfalls, misleading or encouraging approaches, and problems with certain software tools.

12.3 Role of SVMs in Data Mining

It is obvious from the description of the typical phases in the data mining process given in Section 12.2 that SVMs and related kernel methods are of interest mostly in the modeling phase (i.e., in phase 4 of CRISP-DM). We refer to Chapter 8 for classification, and Chapter 9 for regression problems. However, there is ongoing research on how to use SVMs in other phases, too.

In the data preparation phase (phase 3), kernel feature extraction by kernel PCA can be quite useful to reduce the dimensionality of the inputs and to construct derived attributes. We refer to Schölkopf and Smola (2002), Shawe-Taylor and Cristianini (2004), and the references cited in these books for details on kernel PCA.

Pelckmans *et al.* (2005) proposed kernel-based classifiers using a worst-case analysis of a finite set of observations, including missing values of the inputs. The approach is based on a component-wise SVM and an empirical measure of maximal variation of the components to bound the influence of the components that cannot be evaluated due to missing values.

12.4 Software Tools for Data Mining

There are many commercial and non-commercial software tools for data mining. No attempt is made to mention all of them.

`CART`® and `TreeNet`®

The commercial software tools `CART`® and `TreeNet` are distributed by Salford Systems. `CART`® is a well-known classification and regression tree package. `TreeNet`® is based on boosted decision trees. It is an accurate model builder and can also serve as a powerful initial data exploration tool. Both software tools are more specialized to only a few methods than the other software products listed below but have proven to give good results in several competitions on knowledge discovery and data mining.

`IBM`® `DB2 Intelligent Miner for Data`

The `IBM`® `DB2 Intelligent Miner` is based on a client-server architecture. The server executes mining and processing functions and can host mining results. The client is powered with administrative and visualization tools. Hence, the client can be used to visually build a data mining operation, execute it on the server, and have the results returned for visualization and further analysis. In addition, the application programming interface (API) provides `C++` classes and methods as well as `C` structures and functions for application programmers. This allows the user to define and use new methods. The Intelligent Miner can read data stored by `Oracle`, `SAS`®, and `SPSS`® and contains scalable algorithms.

SAS® Enterprise Miner™

The SAS Enterprise Miner is one of the leading commercial software tools for data mining. SAS propagates the so-called SEMMA strategy which is similar to the CRISP-DM strategy described in Section 12.2. SEMMA is the abbreviation of Sample, Explore, Modify, Model, and Assess. The SAS® Enterprise Miner™ has a user-friendly graphical user interface (GUI) and contains tools for all necessary steps such as data pre-processing, sampling, classification, regression, trees, clustering, association rules, visualization, assessment, and scoring. Additionally, this software allows user-defined models and ensemble techniques, making it a flexible data mining tool. The SAS® Enterprise Miner™ runs stably and is capable of handling huge data sets efficiently because it is delivered as a distributed client-server system. To our knowledge there (currently) exists only a non-productive version of an SVM implementation in the SAS Enterprise Miner™.

SPSS® Clementine

The commercial data mining tool SPSS®Clementine has a structure similar to the SAS®Enterprise Miner™ and has a user-friendly GUI. New versions of Clementine also use a client-server architecture. However, Clementine is— from our point of view— less flexible and offers fewer analytical methods than SAS® Enterprise Miner™.

Weka

Weka is a collection of machine learning algorithms for data mining tasks; see Witten and Frank (2005). The algorithms can either be applied directly to a data set or called from user-defined code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License. There exists an interface from the statistical software system R to Weka.

12.5 Further Reading and Advanced Topics

There are a lot of well-written textbooks on data mining. Besides the CRISP-DM approach developed by Chapman *et al.* (2000), we would like to mention the following books for a detailed description of data mining. Hand *et al.* (2001) treat data mining from a broad view covering all fundamental topics. The textbook by Berry and Linoff (1997) is aimed especially for business users, and Mitchell (1997) and Witten and Frank (2005) emphasize the machine learning viewpoint of data mining. An overview and a description of modeling techniques from a more statistical point of view is given by Hastie *et al.* (2001).

McCullagh and Nelder (1989), Cox and Snell (1989), and Agresti (1996) treat generalized linear models, maximum likelihood estimation and hypothesis testing in detail. Robust parameter estimation is treated, for example, by Hampel *et al.* (1986), Rousseeuw (1984), Künsch *et al.* (1989), Bickel *et al.* (1993), Christmann (1994, 1998), and Bianco and Yohai (1996) for the case of a generalized linear model.

Cochran (1977) and Levy and Lemeshow (1999) investigate sampling techniques. Methods for the determination of sample sizes for simple random sampling, stratified random sampling, and other sampling plans are treated in detail by Desu and Raghavarao (1990) and Lemeshow *et al.* (1990).

Methods dealing with missing values are treated by Rubin (1987) and Little and Rubin (1987). The statistical software package SAS® contains the procedures PROC MI and PROC MIANALYZE, which can be used to detect patterns of missing values and for the imputation for such values. These procedures are based on the EM algorithm or on the Markov Chain Monte Carlo (MCMC) method. The statistical software package SPSS® offers the module "missing value analysis" for the same task.

12.6 Summary

This chapter briefly described data mining, the main phases of data mining projects, and the role of kernel methods treated in this book within the whole data mining process. Data mining is the analysis of usually large observational data sets to detect and model unsuspected relationships and summarize the data in ways that are both understandable and useful to the data owner; see Hand *et al.* (2001, p. 1). The CRISP-DM project (Chapman *et al.*, 2000) has developed an industry- and tool-neutral data mining process strategy. This data mining strategy consists of six main phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Support vector machines treated in this monograph are mainly used in the modeling phase, but there is ongoing research to use SVMs in other phases also.

12.7 Exercises

12.1. CRISP-DM versus SEMMA ($\star$)
Compare the data mining strategies CRISP-DM proposed by Chapman *et al.* (2000) and SEMMA used by the SAS® Enterprise Miner™.

12.2. Importance of data preparation and data quality ($\star$)
Explain why data preparation and methods to improve the data quality can be quite time-consuming. Explain why these steps are important even for non-parametric methods that make only minor model assumptions.

Appendix

In this appendix, we summarize several results and notions from different mathematical disciplines that are used in the book. We decided to present these results in a form that is useful for the purposes of the book, and hence some of these results can actually be formulated in a more general form.

A.1 Basic Equations, Inequalities, and Functions

In this section, we recall various conceptionally simple facts from mathematics that do not fit into the more focused sections that follow.

Let us begin by recalling that for $n \geq 0$ the n-th **Hermite polynomial** is defined by

$$h_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}, \qquad x \in \mathbb{R}. \tag{A.1}$$

It is simple to check that $h_n(-x) = (-1)^n h_n(x)$ for all all $n \geq 0$, $x \in \mathbb{R}$. Furthermore, the Hermite polynomials are "orthogonal" in the sense of

$$\int_{-\infty}^{\infty} h_n(x) h_m(x)\, e^{-x^2}\, dx = 2^n n! \sqrt{\pi}\ \delta_{n,m}\,, \tag{A.2}$$

where $\delta_{n,m}$ is the Kronecker symbol, i.e., $\delta_{n,m} = 1$ if $n = m$ and $\delta_{n,m} = 0$ otherwise. Moreover, one can show that they form an orthogonal basis of the Hilbert space of measurable functions $f : X \to \mathbb{R}$ satisfying

$$\int_{-\infty}^{\infty} |f(x)|^2\ e^{-x^2}\, dx < \infty\,.$$

We refer to p. 91–93 of Courant and Hilbert (1953) for these calculations and to Section A.5.2 for Hilbert spaces and their bases.

Let us now recall that for $a \in \mathbb{R}$ the **incomplete gamma function** is defined by

$$\Gamma(a,x) := \int_x^\infty e^{-t}\, t^{a-1} dt\,, \qquad\qquad x \geq 0.$$

Note that $\Gamma(a,x) < \infty$ for all $x > 0$, and for $a > 0$ the **gamma function** $\Gamma(a) := \Gamma(a,0)$ is also finite. Moreover, for $\sigma, r > 0$ and $d \geq 1$, we have

$$\Big(\frac{1}{2\sigma^2\pi}\Big)^{d/2} \int_{\|x\|_2 \geq r} e^{-\frac{\|y\|_2^2}{2\sigma^2}} = \frac{1}{\Gamma(d/2)} \Gamma\Big(\frac{d}{2}, \frac{r^2}{2\sigma^2}\Big)\,, \tag{A.3}$$

where $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$. Consequently, $\Gamma(a,x)$ can be used to compute tails for multivariate normal distributions. The following lemma collects some useful estimates of the incomplete gamma function.

Lemma A.1.1 (Properties of the incomplete gamma function). *For all $a \in \mathbb{R}$ and $x > 0$, we have $\Gamma(a+1,x) = a\Gamma(a,x) + e^{-x}x^a$, and for $n \in \mathbb{N}$ this yields $\Gamma(n+1) = n!$. Moreover, for $a > 0$ and $x \geq a$, the following estimates hold:*

$$a\Gamma(a-1,x) \leq \Gamma(a,x) \tag{A.4}$$

$$\min\{1,a\}\, e^{-x} x^{a-1} \leq \Gamma(a,x) \leq \max\{1,a\}\, e^{-x} x^{a-1}\,. \tag{A.5}$$

Finally, for $x \geq 0$ and $a, b > 0$, we have $\Gamma(a,x) \leq \Gamma(a+b) x^{-b}$.

Proof. For all $a \in \mathbb{R}$ and $x > 0$, integration by parts yields

$$\Gamma(a+1,x) = \int_x^\infty e^{-t} t^a dt = -e^{-t}t^a\Big|_x^\infty + a\int_x^\infty e^{-t} t^{a-1} dt = x^a e^{-x} + a\Gamma(a,x).$$

Moreover, for $a > 0$ and $x \geq a$, we have

$$a\Gamma(a-1,x) = a\int_x^\infty e^{-t} t^{a-2} dt \leq \int_x^\infty e^{-t} t^{a-1} dt = \Gamma(a,x)\,,$$

i.e., we have shown (A.4). Let us now show (A.5) for $a \geq 1$. In this case, the left inequality follows from

$$e^{-x}x^{a-1} = \int_x^\infty e^{-t} x^{a-1} dt \leq \int_x^\infty e^{-t} t^{a-1} dt = \Gamma(a,x),$$

while the right inequality follows from

$$\Gamma(a,x) = a\Gamma(a,x) - (a-1)\Gamma(a,x) \leq a\Gamma(a,x) - a(a-1)\Gamma(a-1,x) = ae^{-x}x^{a-1}.$$

The case $0 < a < 1$ can be shown analogously. Finally, for $t \geq x$, we have $1 \leq x^{-b}t^b$, and hence we find $\Gamma(a,x) \leq x^{-b}\Gamma(a+b,x) \leq \Gamma(a+b)x^{-b}$. □

The next lemma recalls the multinomial formula that generalizes the binomial formula to more than two addends.

Lemma A.1.2 (Multinomial formula). *For $n \in \mathbb{N}$ and $z_1, \dots, z_n \in \mathbb{C}$, we have*

$$(z_1 + \cdots + z_n)^n = \sum_{\substack{j_1,\dots,j_d \geq 0 \\ j_1+\cdots+j_d=n}} n! \prod_{i=1}^{d} \frac{z_i^{j_i}}{j_i!} \,.$$

The following lemma shows how to "interpolate" two upper bounds of a non-negative sequence.

Lemma A.1.3. *Let $(a_n) \subset [0,\infty)$ be a sequence for which there exist constants $r, t \in (0,\infty)$ and $c_r, c_t \in (0,\infty)$ such that $r \leq t$ and both*

$$a_n \leq c_r \, n^{-1/r} \qquad \text{and} \qquad a_n \leq c_t \, n^{-1/t}$$

for all $n \geq 1$. Then, for all $s \in [r,t]$ and all $n \geq 1$, we have

$$a_n \leq c_r^{\frac{r(t-s)}{s(t-r)}} \, c_t^{\frac{t(s-r)}{s(t-r)}} \, n^{-1/s} \,.$$

Proof. Let us write $a := (c_r/c_t)^{\frac{rt}{t-r}}$. For $n \geq a$, we then have

$$a_n \leq c_r \, n^{-\frac{1}{r}} = c_r \, n^{\frac{1}{s}-\frac{1}{r}} n^{-\frac{1}{s}} \leq c_r \, a^{\frac{1}{s}-\frac{1}{r}} n^{-\frac{1}{s}} = c_r^{\frac{r(t-s)}{s(t-r)}} \, c_t^{\frac{t(s-r)}{s(t-r)}} \, n^{-\frac{1}{s}} \,.$$

Analogously, for $n \leq a$, we obtain

$$a_n \leq c_t \, n^{-\frac{1}{t}} = c_t \, n^{\frac{1}{s}-\frac{1}{t}} n^{-\frac{1}{s}} \leq c_t \, a^{\frac{1}{s}-\frac{1}{t}} n^{-\frac{1}{s}} = c_r^{\frac{r(t-s)}{s(t-r)}} \, c_t^{\frac{t(s-r)}{s(t-r)}} \, n^{-\frac{1}{s}} \,. \qquad \square$$

The next lemma describes a simple yet powerful technique to achieve convergence to zero in certain situations.

Lemma A.1.4 (Selection lemma). *Let $F : (0,\infty) \times \mathbb{N} \to [0,\infty)$ be a function such that $\lim_{n\to\infty} F(\lambda, n) = 0$ for all $\lambda > 0$. Then there exists a decreasing sequence $(\lambda_n) \subset (0,1]$ such that $\lim_{n\to\infty} \lambda_n = 0$ and $\lim_{n\to\infty} F(\lambda_n, n) = 0$.*

Proof. For $k \geq 1$, there exists an $n_k > 1$ such that for all $n \geq n_k$ we have

$$F(k^{-1}, n) < k^{-1} \,. \tag{A.6}$$

We may assume without loss of generality that $n_k < n_{k+1}$ for all $k \geq 1$. For $n \geq 1$, we write

$$\lambda_n := \begin{cases} 1 & \text{if } 1 \leq n < n_1 \\ k^{-1} & \text{if } n_k \leq n < n_{k+1} \,. \end{cases}$$

Obviously, the sequence (λ_n) is decreasing. Now let $\varepsilon > 0$. Then there exists an integer $k \geq 1$ with $k^{-1} \leq \varepsilon$. Let us fix an $n \geq n_k$. Then there exists an $i \geq k$ with $n_i \leq n < n_{i+1}$, and consequently we have $\lambda_n = i^{-1}$. This gives

$$\lambda_n = i^{-1} \leq k^{-1} \leq \varepsilon \,,$$

and since (A.6) together with $n_i \leq n$ yields $F(i^{-1}, n) \leq i^{-1}$, we also find

$$F(\lambda_n, n) = F(i^{-1}, n) \leq i^{-1} \leq \varepsilon \,. \qquad \square$$

The following lemmas compute extrema for various functions. Since these calculations are often used in the book we decided to collect them in the appendix. The proof of the first lemma is elementary calculus and hence omitted.

Lemma A.1.5. *Let $c_1, c_2 > 0$ and $a, b > 0$. Then we have*

$$\min_{t>0} c_1 t^a + c_2 t^{-b} = \frac{a+b}{b} \left(\frac{b}{a}\right)^{\frac{a}{a+b}} c_1^{\frac{b}{a+b}} c_2^{\frac{a}{a+b}} ,$$

and the minimum is attained at $t^* := (\frac{bc_2}{ac_1})^{1/(a+b)}$.

Lemma A.1.6. *For $\alpha, \beta, \gamma > 0$ and $p > 0$, $q \geq 0$, there exists a constant $c_1 > 0$ such that for all $x > 0$ we have*

$$\min_{s,t\in[0,\infty)} \left(s^\alpha t^{-\gamma} + t^\beta + x s^{-p} t^{-q}\right) = c_1 \, x^{\frac{\alpha\beta}{\alpha\beta+\beta p+\gamma p+\alpha q}} .$$

Moreover, there exist constants $c_2 > 0$ and $c_3 > 0$ independent of x such that the minimum above is attained at

$$\begin{aligned} s^* &:= c_2 \, x^{\frac{\beta+\gamma}{\alpha\beta+\beta p+\gamma p+\alpha q}} , \\ t^* &= c_3 \, x^{\frac{\alpha}{\alpha\beta+\beta p+\gamma p+\alpha q}} . \end{aligned}$$

In particular, there exists another constant $c_4 > 0$ independent of x such that

$$\min_{s,t\in[0,1]} \left(s^\alpha t^{-\gamma} + t^\beta + x s^{-p} t^{-q}\right) \leq c_4 \, x^{\frac{\alpha\beta}{\alpha\beta+\beta p+\gamma p+\alpha q}} , \qquad x \in (0,1].$$

Proof. First minimize the function $s \mapsto s^\alpha t^{-\gamma} + t^\beta + x s^{-p} t^{-q}$ with the help of Lemma A.1.5. This gives the minimizer $s^* := (p/\alpha)^{1/(\alpha+p)} x^{1/(\alpha+p)} t^{(\gamma-q)/(\alpha+p)}$ and the expression

$$\min_{\substack{s>0\\ t>0}} \left(s^\alpha t^{-\gamma} + t^\beta + x s^{-p} t^{-q}\right) = \min_{t>0} \left(t^\beta + c_{\alpha,p} \, x^{\frac{\alpha}{\alpha+p}} t^{-\frac{\gamma p+\alpha q}{\alpha+p}}\right),$$

where $c_{\alpha,p} = \frac{\alpha+p}{p} (\frac{p}{\alpha})^{\frac{\alpha}{\alpha+p}}$. Another application of Lemma A.1.5 then yields the first three assertions. The fourth assertion follows from inserting the values s^*/c_2 and t^*/c_3 into the objective function. □

Lemma A.1.7. *For all $\alpha, \beta, \gamma \in (0,1]$ and $r \geq 1$, there exists a constant c such that, for*

$$\rho := \min\left\{ \frac{r\beta}{1+(r-1)\beta}, \frac{\gamma\beta}{\beta+\alpha\gamma} \right\}$$

and all $t \in (0,1]$, we have

$$c\, t^\rho \;\leq\; \min_{s\in(0,1]} \left(s^\beta + t^\gamma s^{-\alpha\gamma} + t s^{(\beta-1)/r}\right) \;\leq\; 3\, t^\rho .$$

Proof. Let us first consider the case $\frac{r\beta}{1+(r-1)\beta} < \frac{\gamma\beta}{\beta+\alpha\gamma}$. In this case, we have $\beta \in (0,1)$, and, by Lemma A.1.5, we thus obtain

$$\min_{s\in(0,1]}\Big(s^\beta + t^\gamma s^{-\alpha\gamma} + ts^{(\beta-1)/r}\Big) \geq \min_{s>0}\Big(s^\beta + ts^{(\beta-1)/r}\Big) \geq c\, t^{\frac{\beta}{\beta+(1-\beta)/r}} = c\, t^\rho$$

for a suitable constant $c > 0$. Moreover, for $s^* := t^{\frac{r}{1+(r-1)\beta}}$, we obtain

$$\begin{aligned}\min_{s\in(0,1]}\Big(s^\beta + t^\gamma s^{-\alpha\gamma} + ts^{(\beta-1)/r}\Big) &\leq (s^*)^\beta + t^\gamma (s^*)^{-\alpha\gamma} + t(s^*)^{(\beta-1)/r} \\ &= 2t^{\frac{r\beta}{1+(r-1)\beta}} + t^{\frac{r\beta}{1+(r-1)\beta}}\, t^{\frac{\gamma+\beta(r-1)\gamma-r\beta-r\alpha\gamma}{1+(r-1)\beta}},\end{aligned}$$

and using that $\frac{r\beta}{1+(r-1)\beta} < \frac{\gamma\beta}{\beta+\alpha\gamma}$ is equivalent to $\gamma+\beta(r-1)\gamma-r\beta-r\alpha\gamma > 0$, we then find the assertion.

Let us now consider the opposite case $\frac{r\beta}{1+(r-1)\beta} \geq \frac{\gamma\beta}{\beta+\alpha\gamma}$. Then Lemma A.1.5 yields

$$\min_{s\in(0,1]}\Big(s^\beta + t^\gamma s^{-\alpha\gamma} + ts^{(\beta-1)/r}\Big) \geq \min_{s>0}\Big(s^\beta + t^\gamma s^{-\alpha\gamma}\Big) \geq c\, t^{\frac{\beta\gamma}{\beta+\alpha\gamma}} = c\, t^\rho\,.$$

Moreover, for $s^* := t^{\frac{\gamma}{\beta+\alpha\gamma}}$, we obtain

$$\begin{aligned}\min_{s\in(0,1]}\Big(s^\beta + t^\gamma s^{-\alpha\gamma} + ts^{(\beta-1)/r}\Big) &\leq (s^*)^\beta + t^\gamma (s^*)^{-\alpha\gamma} + t(s^*)^{(\beta-1)/r} \\ &= 2t^{\frac{\beta\gamma}{\beta+\alpha\gamma}} + t^{\frac{\beta\gamma}{\beta+\alpha\gamma}}\, t^{\frac{r\beta+r\alpha\gamma-\gamma-(r-1)\beta\gamma}{r(\beta+\alpha\gamma)}},\end{aligned}$$

and using that $\frac{r\beta}{1+(r-1)\beta} \geq \frac{\gamma\beta}{\beta+\alpha\gamma}$ is equivalent to $r\beta+r\alpha\gamma-\gamma-(r-1)\beta\gamma \geq 0$, we then find the assertion. □

A.2 Topology

In various chapters of the book, we need results about metric spaces, or more generally topological spaces. Roughly speaking, both topological spaces and metric spaces describe "neighborhoods" of their points, so basic concepts, such as convergence and continuity, can be defined. The following definitions and facts can be found in various textbooks on topology or functional analysis, such as Kelley (1955), Willard (1970), or Rudin (1976).

Definition A.2.1. *Given a set X, a subset τ of the power set 2^X of X is called a* ***topology*** *on X if it satisfies:*

i) $\emptyset \in \tau$, $X \in \tau$.
ii) If $O_1 \in \tau$ and $O_2 \in \tau$, then $O_1 \cap O_2 \in \tau$.
iii) If I is any index set and $O_i \in \tau$ for all $i \in I$, then $\bigcup_{i\in I} O_i \in \tau$.

In this case, the pair (X,τ) is called a ***topological space****. Moreover, each $O \in \tau$ is called an* ***open set****.*

Let X be any non-empty set. Then $\tau := \{\emptyset, X\}$ is the smallest topology on X and is called the **indiscrete topology** on X. Obviously, every topology on X contains the indiscrete topology. Moreover, the power set 2^X of X is the largest topology on X and is called the **discrete topology** on X. Again, it is obvious that every topology on X is contained in the discrete topology. Given a topological space (X, τ) and a subset $A \subset X$, the set

$$\tau_A := \{O \cap A : O \in \tau\}$$

defines a topology on A, called the **trace** or **relative topology** of τon A.

A topological space is said to be a **Hausdorff space** if for any pair of distinct points $x, y \in X$, there exist $O_x, O_y \in \tau$ such that $x \in O_x$, $y \in O_y$, and $O_x \cap O_y = \emptyset$. For our purposes, the most important examples of topological Hausdorff spaces are **metric spaces**, which we introduce now.

Definition A.2.2. *A map $d : X \times X \to [0, \infty)$ on a non-empty set X is called a **metric** if it satisfies:*

i) $d(x, y) = 0$ *if and only if* $x = y$.
ii) $d(x, y) = d(y, x)$ *for all* $x, y \in X$.
iii) $d(x, y) \leq d(x, z) + d(z, y)$ *for all* $x, y, z \in X$.

*In this case, the pair (X, d) is called a **metric space**, and if d is clear from the context, we usually omit it by calling X a metric space. Moreover, if only ii) and iii) together with $d(x, x) = 0$ for all $x \in X$ are satisfied, then d is called a **pseudo-metric***

An example of a metric space is the **Euclidean space** $\mathbb{R}^d$, $d \in \mathbb{N}$, with the **Euclidean distance**

$$d(x, y) = ||x - y||_2 := \left(\sum_{i=1}^{d} |x_i - y_i|^2 \right)^{1/2}, \qquad x, y \in \mathbb{R}^d .$$

More generally, every normed space (see Section A.5) is a metric space. Moreover, the **discrete metric** d on a set $X \neq \emptyset$ is defined by $d(x, x) := 0$ and $d(x, y) := 1$ if $x \neq y$. Finally, given a (pseudo-)metric space (X, d), the pair $(A, d_{|A \times A})$ is a (pseudo-)metric space for all non-empty $A \subset X$.

For a pseudo-metric space (X, d), we define the **open ball** with radius $\varepsilon > 0$ and center $x \in X$ by

$$B_d(x, \varepsilon) := \{y \in X : d(x, y) < \varepsilon\}, \tag{A.7}$$

and the corresponding **closed ball** is defined by replacing "$<$" with "$\leq$" in (A.7). Moreover, we call a subset $O \subset X$ **open** if for all $x \in O$ there exists an $\varepsilon > 0$ such that $B_\varepsilon(x) \subset O$. Using the triangle inequality, it is easy to see that every open ball is an open subset. Moreover, the set τ of open subsets of (X, d), i.e.,

$$\tau := \big\{ O \subset X : O \text{ is an open subset of } X \big\},$$

is a topology on X that is called the topology of the pseudo-metric d. Note that this topology is Hausdorff if d is a metric.

A **basis** of a topology τ is any subset τ_1 of τ such that every open set is a union of sets in τ_1. For example, it is easy to check that the set of open balls of a pseudo-metric space is a basis of its topology. In particular, the open balls in $\mathbb{R}^d$ form a basis of the topology of its metric.

Let us now introduce some more concepts for topological spaces.

Definition A.2.3 (Closed and compact subsets). *Let (X, τ) be a topological space. We say that $A \subset X$ is:*

i) ***closed*** *if its complement $X \setminus A$ is open, i.e., $X \setminus A \in \tau$.*
ii) ***compact*** *if, for every family $(O_i)_{i \in I}$ of open sets with $A \subset \bigcup_{i \in I} O_i$, there exist finitely many indexes $i_1, \dots, i_n \in I$ with $A \subset \bigcup_{j=1}^{n} O_{i_j}$.*

Note that open and closed are *not* mutually exclusive concepts. Indeed, X and $\emptyset$ are open *and* closed subsets with respect to every topology on X. Moreover, every compact subset of a Hausdorff space is closed, but in general the converse is not true, as for example $\mathbb{R}$ as a subset of $\mathbb{R}$ shows. However, every closed subset of a compact subset is again compact. Finally, the following theorem (see, e.g., Rudin, 1976, Theorem 2.41) characterizes compact subsets of $\mathbb{R}^d$. Note, that it is false in *every infinite-dimensional* normed space.

Theorem A.2.4. *A set $A \subset \mathbb{R}^d$ is compact if and only if A is closed and bounded.*

Given a topological space (X, τ) and a subset $A \subset X$, we define the **interior** of A by

$$\mathring{A} := \bigcup_{O \subset A \text{ open}} O$$

and the **closure** of A by

$$\overline{A} := \bigcap_{A \subset C \text{ closed}} C\,.$$

It is elementary to see that $\mathring{A}$ is open and $\overline{A}$ is closed. Moreover, A is called **dense** if $\overline{A} = X$. Finally, (X, τ) is called **separable** if there exists a countable and dense subset of X.

Definition A.2.5. *Let (X_1, τ_1) and (X_2, τ_2) be topological spaces and $x_0 \in X$. A map $f : X_1 \to X_2$ is called* ***continuous at*** *x_0 if for all $O_2 \in \tau_2$ with $f(x_0) \in O_2$ there exists an $O_1 \in \tau_1$ such that $x_0 \in O_1$ and $f(O_1) \subset O_2$. Moreover, f is called* ***continuous*** *if f is continuous at every $x \in X$.*

It is easy to see that the function $f : X_1 \to X_2$ is continuous if and only if $f^{-1}(O) \in \tau_1$ for all $O \in \tau_2$. Morover, using complements, it is not hard to see that the continuity is also equivalent to the condition that $f^{-1}(A)$ is closed

for all closed subsets $A \subset X_2$. In general, similar statements are *not* true if one considers the images instead of the pre-images. However, every continuous image of a **compact** set is compact, and consequently, if X_1 is compact and X_2 is Hausdorff, then $f(A)$ is compact (and thus closed) for all closed subsets A of X_1.

Let (X, τ) be a topological space and $f : X \to \mathbb{R}$ be a function. Then the **support** of f is defined by

$$\operatorname{supp} f \ := \ \overline{\{x \in X : f(x) \neq 0\}}\,.$$

Given two topological spaces (X_1, τ_1) and (X_2, τ_2) the **product topology** $\tau_1 \otimes \tau_2$ is the smallest topology on $X_1 \times X_2$ ensuring that both projections $\pi_i : X_1 \times X_2 \to X_i$, $i = 1, 2$, are continuous.

For real-valued functions, we often need the following, weaker concept of continuity.

Definition A.2.6. *Let X be a topological space and $f : X \to \mathbb{R} \cup \{\infty\}$. Then f is called **lower semi-continuous (l.s.c.)** if the level sets $\{x \in X : f(x) \leq a\}$ are closed for all $a \in \mathbb{R}$.*

The following lemma immediately follows from the definition and the fact that arbitrary intersections of closed sets are closed.

Lemma A.2.7 (Supremum of l.s.c. functions). *Let X be a topological space, $I \neq \emptyset$, and $(f_i)_{i \in I}$ be a family of l.s.c. functions $f_i : X \to \mathbb{R}$. Then $f : X \to \mathbb{R} \cup \{\infty\}$ defined by $f(x) := \sup_{i \in I} f_i(x)$ is lower semi-continuous.*

The lower semi-continuity is often used in minimization problems. In this regard, the following elementary lemma that allows us to "ignore" the behavior of l.s.c. functions outside a closed set is of special importance.

Lemma A.2.8. *Let X be a topological space, $A \subset X$ be a closed subset, and $f : A \to \mathbb{R}$ be a function that is lower semi-continuous with respect to the trace topology on A. Then $\bar{f} : X \to \mathbb{R} \cup \{\infty\}$, defined by $\bar{f}(x) := f(x)$ if $x \in A$ and $\bar{f}(x) := \infty$ otherwise, is lower semi-continuous.*

In the rest of this section, we present some further concepts and results for metric spaces. Let us begin by recalling that a sequence (x_n) in a metric space (X, d) is said to **converge** if there exists an element $x \in X$ such that for all $\varepsilon > 0$ there exists an $n_0 \geq 1$ such that for all $n \geq n_0$ we have $d(x, x_n) \leq \varepsilon$. In this case, the **limit** x is unique and denoted by $\lim_{n \to \infty} x_n$. We sometimes also write $x_n \to x$ if (x_n) converges to x and call a sequence $(a_n) \subset \mathbb{R}$ with $a_n \to 0$ a **null sequence**. Moreover, (x_n) is called a **Cauchy sequence** if for every $\varepsilon > 0$ there is an $n_0 \geq 1$ such that $d(x_n, x_m) < \varepsilon$ for all $n \geq n_0$ and $m \geq n_0$. Obviously, every convergent sequence is a Cauchy sequence, but in general the converse is false. This leads to the following definition.

Definition A.2.9. *A metric space is called* ***complete*** *if every Cauchy sequence converges.*

In metric spaces, many topological concepts introduced earlier can be characterized by sequences. The following lemma illustrates this.

Lemma A.2.10. *Let (X, d) and (X', d') be metric spaces, $f : X \to X'$ and $g : X \to \mathbb{R}$ be maps, and $A \subset X$. Then the following statements are true:*

i) A is closed if and only if, for all sequences $(x_n) \subset A$ that are convergent to some $x \in X$, we have $x \in A$.
ii) A is compact if and only if every sequence $(x_n) \subset A$ has a subsequence that converges to some $x \in A$.
iii) f is continuous at $x \in X$ if and only if, for all sequences $(x_n) \subset X$ satisfying $\lim_{n\to\infty} x_n = x$, we have $\lim_{n\to\infty} f(x_n) = f(x)$.
iv) g is l.s.c. if and only if $g(x_0) \leq \liminf_{x\to x_0} g(x)$ for all $x_0 \in X$.

Let us finally introduce two classes of topological spaces that are very useful when considering measures on topological spaces.

Definition A.2.11. *A topological space (X, τ) is called a* ***Polish space*** *if τ has a countable basis and there exists a complete metric defining τ.*

The Euclidean spaces $\mathbb{R}^d$ are Polish. Moreover, both compact Hausdorff spaces with countable basis and complete separable metric spaces are Polish. In addition, the topological product of two Polish spaces is Polish. Finally, all open and closed subsets of a Polish space become Polish when equipped with the trace topology.

To introduce the second type of topological spaces (X, τ), we say that a subset $V \subset X$ is a **neighborhood** of a point $x \in X$ if there exists an $O \in \tau$ such that $x \in O \subset V$. Note that the neighborhood V need not be open.

Definition A.2.12. *A topological space (X, τ) is called* ***locally compact*** *if for every $x \in X$ there exists a compact neighborhood of x.*

Every compact topological space is locally compact, and the Euclidean spaces are the standard examples of locally compact but not compact spaces.

A.3 Measure and Integration Theory

Measure and integration theory is used throughout this book, and hence this section provides some necessary background from this discipline.

A.3.1 Some Basic Facts

In this subsection, we briefly recall elementary notions and results from measure and integration theory. A more detailed treatment can be found in the books by Bauer (2001) and Dudley (2002).

Definition A.3.1. *Given a non-empty set X, a subset $\mathcal{A}$ of the power set 2^X of X is called a* ***σ-algebra*** *on X if it satisfies:*

i) $X \in \mathcal{A}$.
ii) $A^c := X \backslash A \in \mathcal{A}$ *for all* $A \in \mathcal{A}$.
iii) $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$ *for all sequences* $(A_n)_{n\in\mathbb{N}}$ *of sets in* $\mathcal{A}$.

In this case, $(X, \mathcal{A})$ is called a ***measurable space*** *and the elements of $\mathcal{A}$ are called* ***measurable sets****. Finally, if $\mathcal{A}$ is clear from the context or its specific form is irrelevant, we often omit it by calling X a measurable space.*

Obviously, both $\mathcal{A} := \{\emptyset, X\}$ and $\mathcal{A} := 2^X$ are σ-algebras called the **indiscrete** and **discrete** σ-algebra, respectively. If $\mathcal{A}$ is a σ-algebra on X and $\tilde{X} \subset X$, then $\tilde{X} \cap \mathcal{A} := \{\tilde{X} \cap A : A \in \mathcal{A}\}$ is a σ-algebra on $\tilde{X}$ called the **trace σ-algebra** of $\mathcal{A}$ in $\tilde{X}$.

It is easy to show that the intersection $\bigcap_{i\in I} \mathcal{A}_i$ of arbitrary σ-algebras $\mathcal{A}_i$ on X is again a σ-algebra on X. For any $C \subset 2^X$, there hence exists a smallest σ-algebra containing C, i.e., a σ-algebra $\sigma(C)$ such that both $C \subset \sigma(C)$ and $\sigma(C) \subset \mathcal{A}$ for all σ-algebras $\mathcal{A}$ on X with $C \subset \mathcal{A}$. Obviously, $\sigma(C)$ is uniquely determined, and we say that $\sigma(C)$ is the σ-algebra **generated** by C.

Let us give a few examples of generated σ-algebras. To this end, we first consider a sequence $(X_n, \mathcal{A}_n)_{n\in\mathbb{N}}$ of measurable spaces. Then the **product σ-algebra** $\bigotimes_{n\in\mathbb{N}} \mathcal{A}_n$ on the product space $\mathsf{X}_{n\in\mathbb{N}} X_n$ is the σ-algebra that is generated by all cylinder sets $A_n \times \mathsf{X}_{m\neq n} X_m$, $A_n \in \mathcal{A}_n$, $n \in \mathbb{N}$. Moreover, a topological (X, τ) is endowed with its **Borel σ-algebra** $\mathcal{B}(\tau) := \mathcal{B}(X) := \sigma(\tau)$, and the elements of $\mathcal{B}(\tau)$ are called **Borel sets**. Note that if X_n, $n \geq 1$, are *separable* metric spaces, we have $\mathcal{B}(\mathsf{X}_{n\in\mathbb{N}} X_n) = \bigotimes_{n\in\mathbb{N}} \mathcal{B}(X_n)$.

Given measurable spaces $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ and a map $f : X_1 \to X_2$, it is easy to show that $f^{-1}\mathcal{A}_2 := \{f^{-1}(A) : A \in A_2\}$ is a σ-algebra on X_1. Moreover, f is said to be $(\mathcal{A}_1, \mathcal{A}_2)$-**measurable**, or simply measurable, if $f^{-1}\mathcal{A}_2 \subset \mathcal{A}_1$.

Given a sequence $(X_n, \mathcal{A}_n)_{n\in\mathbb{N}}$ of measurable spaces, the coordinate projections $\pi_m : \mathsf{X}_{n\in\mathbb{N}} X_n \to X_m$ are $(\bigotimes_{n\in\mathbb{N}} \mathcal{A}_n, \mathcal{A}_m)$-measurable. In addition, if f is a continuous map between two topological spaces (X_1, τ_1) and (X_2, τ_2), then f is $(\mathcal{B}(\tau_1), \mathcal{B}(\tau_2))$-measurable. Moreover, the composition $f := f_1 \circ f_2$ of arbitrary measurable functions f_1 and f_2 is measurable. Combining these results, one sees that sums and products of measurable functions are measurable. In particular, **simple functions**, i.e., functions of the form $f := \sum_{i=1}^n c_i \mathbf{1}_{A_i}$, where $c_i \in \mathbb{R}$ and $A_i \subset X$ are measurable if and only if $A_i \in \mathcal{A}$ for all $i = 1, \ldots, n$. Furthermore, if $(f_n)_{n\in\mathbb{N}}$ is a sequence of measurable functions from $(X, \mathcal{A})$ to $\bar{\mathbb{R}} := [-\infty, \infty]$, then $\sup_{n\in\mathbb{N}} f_n$, $\inf_{n\in\mathbb{N}} f_n$, $\limsup_{n\to\infty} f_n$, and

$\liminf_{n\to\infty} f_n$ are also measurable. In addition, one can show that for any measurable function $f : X \to [0,\infty]$ there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of simple non-negative measurable functions with $f_n \uparrow f$ pointwise, i.e., $f_n(x) \to f(x)$ for all $x \in X$ and $f_n(x) \le f_{n+1}(x)$ for all $x \in X$ and $n \ge 1$. Finally, if f is bounded, then we can actually pick an increasing sequence (f_n) such that the convergence is uniform, i.e., $\|f - f_n\|_\infty \to 0$.

Definition A.3.2. *Given a measurable space $(X, \mathcal{A})$, we say that a function $\mu : \mathcal{A} \to [-\infty, +\infty]$ is a* ***signed measure*** *if both $\mu(\emptyset) = 0$ and*

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n)$$

for every sequence $(A_n)_{n\in\mathbb{N}}$ of mutually disjoint sets in $\mathcal{A}$.

A signed measure μ is a ***measure*** *if $\mu(A) \ge 0$ for all $A \in \mathcal{A}$, and a* ***finite measure*** *if in addition $\mu(X) < \infty$. Moreover, a measure μ with $\mu(X) = 1$ is called a* ***probability measure*** *or a* ***distribution****. Finally, a measure is called σ-finite if $X = \bigcup_{n\in\mathbb{N}} A_n$ for some sets $A_n \in \mathcal{A}$ satisfying $\mu(A_n) < \infty$, $n \in \mathbb{N}$.*

The triple $(X, \mathcal{A}, \mu)$ is called a **(σ-finite, finite) measure space** or a **probability space** if $(X, \mathcal{A})$ is a measurable space and μ is a (σ-finite, finite) measure or a probability measure on $\mathcal{A}$, respectively. In the latter case, we often use the letter P instead of μ.

Examples of σ-finite measures are the **counting measure** μ on $(\mathbb{Z}, 2^{\mathbb{Z}})$, where $\mu(A)$ equals the number of points in $A \in 2^{\mathbb{Z}}$, and the **Lebesgue measure** on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, which is specified by the requirement

$$\mu\big(\{x \in \mathbb{R}^d : a_i < x_i \le b_i, i = 1, \dots, d\}\big) = \prod_{i=1}^{d} (b_i - a_i)\,,$$

for all $a_i < b_i$, $i = 1, \dots, d$. In other words, for bounded rectangles, the Lebesgue measure equals the ordinary volume. Finally, given a measurable space $(X, \mathcal{A})$ and an $x \in X$, the **Dirac measure** δ_x is defined by $\delta_x(A) := 1$ if $x \in A$ and $\delta_x(A) := 0$ otherwise.

Given a measure space $(X, \mathcal{A}, \mu)$, we say that $N \in \mathcal{A}$ is a μ-**zero set** or μ-**null set** if $\mu(N) = 0$. Moreover, we say that a property $\mathsf{P}(\mathsf{x})$ holds μ-**almost surely** or μ-**almost everywhere** if $\mu(\{x \in X : \mathsf{P}(\mathsf{x}) \text{ false}\}) = 0$. For example, a sequence (f_n) of measurable functions converges μ-almost surely to a measurable function f if $\mu\{x \in X : f_n(x) \text{ does not converge to } f(x)\} = 0$.

Let us now consider a probability space $(X, \mathcal{A}, \mathrm{P})$. Then in general the subsets of P-zero sets are *not* P-zero sets since they may not be measurable. However, the following lemma shows that we can always add such sets to $\mathcal{A}$.

Lemma A.3.3. *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space. Then*

$$\mathcal{A}_{\mathrm{P}} := \big\{A \cup B : A \in \mathcal{A},\, \exists N \in \mathcal{A} \text{ with } \mathrm{P}(N) = 0 \text{ and } B \subset N\big\}$$

is a σ-algebra, called the P***-completion*** *of $\mathcal{A}$, and $\hat{\mathrm{P}} : \mathcal{A}_{\mathrm{P}} \to [0,1]$ defined by $\hat{\mathrm{P}}(A \cup B) := \mathrm{P}(A)$, $A \cup B \in \mathcal{A}_{\mathrm{P}}$, is a probability measure with $\hat{\mathrm{P}}_{|\mathcal{A}} = \mathrm{P}$. We call $\tilde{\mathrm{P}}$ the* ***extension*** *of P to $\mathcal{A}_{\mathrm{P}}$.*

Given a measurable space $(X, \mathcal{A})$ and a probability measure $\mathrm{P} : \mathcal{A} \to [0,1]$, we say that $(X, \mathcal{A})$ is P**-complete** if $\mathcal{A} = \mathcal{A}_{\mathrm{P}}$. Moreover, the σ-algebra

$$\hat{\mathcal{A}} := \bigcap_{\mathrm{P}:\mathcal{A}\to[0,1]} \mathcal{A}_{\mathrm{P}}\,,$$

where P runs over all probability measures on $\mathcal{A}$, is called the **universal completion** of $\mathcal{A}$. In addition, $(X, \mathcal{A})$ is called a **complete measurable space** if $\mathcal{A} = \hat{\mathcal{A}}$. Note that we always have $\mathcal{A} \subset \hat{\mathcal{A}} \subset \mathcal{A}_{\mathrm{P}}$, and consequently $\mathcal{A}$ is complete if it is P-complete for some P. Moreover, using standard arguments, one can easily show that for every $\mathcal{A}_{\mathrm{P}}$-measurable function $f : X \to \mathbb{R}$ there exists an $\mathcal{A}$-measurable function $\bar{f} : X \to \mathbb{R}$ and a set $N \in \mathcal{A}$ with $\mathrm{P}(N) = 0$ and

$$\big\{x \in X : \bar{f}(x) \neq f(x)\big\} \subset N\,. \tag{A.8}$$

In other words, $\mathcal{A}_{\mathrm{P}}$-measurable functions only differ from $\mathcal{A}$-measurable functions on $\mathcal{A}_{\mathrm{P}}$-sets of measure zero.

Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space. A sequence (f_n) of measurable functions $f_n : X \to \mathbb{R}$ is said to **converge** to a measurable function $f : X \to \mathbb{R}$ **in probability** P if for all $\varepsilon > 0$ and $\delta > 0$ there exists an $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$, we have $\mathrm{P}(\{x \in X : |f(x) - f_n(x)| \geq \delta\}) \leq \varepsilon$. Moreover, as mentioned above, the sequence **converges** P**-almost surely** if $f_n(x) \to f(x)$ for P-almost all $x \in X$. It is easy to show that the latter convergence implies convergence in probability P. Conversely, if (f_n) converges to f in probability P then every subsequence of (f_n) has a subsequence that converges P-almost surely to f. Finally, Markov's inequality shows that convergence with respect to $\|\cdot\|_{L_p(\mathrm{P})}$ for some $p > 0$, implies convergence in probability P.

Let us now recall the definition of the **integral** for measurable $f : X \to \mathbb{R}$, where $(X, \mathcal{A}, \mu)$ is some measure space. To this end, we begin with simple non-negative measurable functions $f = \sum_{i=1}^n c_i \mathbf{1}_{A_i}$, for which we define the integral by

$$\int_X f\, d\mu := \sum_{i=1}^n c_i \mu(A_i)\,,$$

where as usual in measure theory we define $0 \cdot \infty := 0$. It is a simple routine to show that the integral above is independent of the representation of f. Now let $f : X \to [0, \infty]$ be a non-negative measurable function. Then, as we have mentioned earlier, there exists a sequence (f_n) of simple non-negative measurable functions with $f_n \uparrow f$ pointwise. We define

$$\int_X f\, d\mu := \sup \int_X f_n\, d\mu\,,$$

where again this definition turns out to be independent of the chosen sequence (f_n). Finally, a measurable function $f : X \to [-\infty, \infty]$ is said to be **μ-integrable** if

$$\int_X |f| \, d\mu < \infty \, .$$

Note that every $f : X \to [-\infty, \infty]$ can be written as $f = f^+ - f^-$, where $f^+ := \max(f, 0)$ and $f^- := \max(-f, 0)$. For a μ-integrable function f, we thus define its integral by the difference of the integrals of f^+ and f^-. We write $\mathcal{L}_1(\mu)$ for the set of all μ-integrable functions. It turns out that the integral is linear on $\mathcal{L}_1(\mu)$, and, in addition, it is monotone, i.e. the integral of non-negative functions is non-negative. Moreover, it enjoys the following limit theorems.

Theorem A.3.4 (Fatou's lemma). *Let $(X, \mathcal{A}, \mu)$ be a measure space and $f_n : X \to [0, \infty]$, $n \geq 1$, be measurable functions. Then we have*

$$\int_X \liminf_{n\to\infty} f_n \, d\mu \leq \liminf_{n\to\infty} \int_X f_n \, d\mu \, .$$

Roughly speaking, Fatou's lemma shows that the integral is l.s.c. with respect to almost sure convergence. The next theorem states that the integral is continuous with respect to monotone convergence.

Theorem A.3.5 (Monotone convergence, Beppo Levi). *Let $(X, \mathcal{A}, \mu)$ be a measure space and $f_n : X \to [0, \infty]$, $n \geq 1$, be measurable functions with $f_n \leq f_{n+1}$ for all $n \geq 1$. Then we have*

$$\int_X \sup_{n\in\mathbb{N}} f_n \, d\mu = \sup_{n\in\mathbb{N}} \int_X f_n \, d\mu \, .$$

The last convergence theorem shows the continuity of the integral with respect to almost sure convergence provided that the functions of the sequence considered have an integrable envelope function.

Theorem A.3.6 (Dominated convergence, Lebesgue). *Let $(X, \mathcal{A}, \mu)$ be a measure space and $f_n : X \to [-\infty, \infty]$, $n \geq 1$, be measurable functions that converge μ-almost surely to an $f : X \to [-\infty, \infty]$. If there exists a $g \in \mathcal{L}_1(\mu)$ such that $|f_n| \leq g$ for all $n \geq 1$, then $f \in \mathcal{L}_1(\mu)$ and*

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X \lim_{n\to\infty} f_n \, d\mu = \int_X f \, d\mu \, .$$

With the help of Lebesgue's theorem we can now describe a simple condition that ensures that we can interchange the order of differentiation and integration.

Corollary A.3.7. *Let $(X, \mathcal{A}, \mu)$ be a measure space, $I \subset \mathbb{R}$ be an open interval, and $f : X \times I \to \mathbb{R}$ be a function such that $f(x, \cdot) : I \to \mathbb{R}$ is differentiable for all $x \in X$ and $f(\cdot, t) \in \mathcal{L}_1(\mu)$ for all $t \in I$. Assume that there exists a $g \in \mathcal{L}_1(\mu)$ such that $|\frac{\partial f}{\partial t}(x,t)| \leq g(x)$ for all $(x,t) \in X \times I$. Then the function $t \mapsto \int_X f(x,t)\mu(dx)$ is differentiable and we have*

$$\frac{\partial}{\partial t} \int_X f(x,t)\, d\mu(x) = \int_X \frac{\partial f}{\partial t} f(x,t)\, d\mu(x)\,.$$

The following theorem, which can be found, for example, on p. 98 of the book by Ash and Doléans-Dade (2000), shows that almost sure convergence implies uniform convergence up to some "small" set.

Theorem A.3.8 (Egorov's theorem). *Let $(X, \mathcal{A}, \mu)$ be a finite measure space. Furthermore, let (f_n) be a sequence of measurable functions $f_n : X \to \mathbb{R}$ that μ-almost surely converges to a measurable function $f : X \to \mathbb{R}$. Then, for all $\varepsilon > 0$, there exists a measurable set $A \subset X$ such that $\mu(X \backslash A) \leq \varepsilon$ and*

$$\lim_{n\to\infty} \|(f_n - f)_{|A}\|_\infty = 0\,.$$

If $(X, \mathcal{A}, \mu)$ is a measure space and $f : X \to \mathbb{R}$ is a non-negative and measurable function, then

$$\nu(A) := \int_A f d\mu\,, \qquad\qquad A \in \mathcal{A},$$

defines a new measure over $(X, \mathcal{A})$ and we write $d\nu = f d\mu$ or $f = \frac{d\nu}{d\mu}$. Such a function f is called a **Radon-Nikodym derivative** of ν with respect to μ, and if ν is even a probability measure, the function f is called a **probability density** of ν with respect to μ.

Now let μ and ν be arbitrary measures on $(X, \mathcal{A})$. Then ν is called **absolutely continuous** with respect to μ if $\mu(A) = 0$ implies $\nu(A) = 0$, $A \in \mathcal{A}$. The next theorem shows that in this case ν has a density with respect to μ.

Theorem A.3.9 (Radon-Nikodym theorem). *Let μ and ν be σ-finite measures on some measurable space $(X, \mathcal{A})$. Then there exists a measurable function $f = \frac{d\nu}{d\mu}$ if and only if ν is absolutely continuous with respect to μ.*

Given two σ-finite measure spaces $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{C}, \nu)$ one can show that there exists a uniquely determined measure $\mu \otimes \nu$ on $\mathcal{A} \otimes \mathcal{C}$ such that

$$\mu \otimes \nu(A \times C) = \mu(A)\, \nu(C)\,, \qquad\qquad A \in \mathcal{A},\, C \in \mathcal{C}.$$

The following result, whose proof can be found, for example, on p. 137 of Dudley (2002), shows how to integrate with respect to the **product measure** $\mu \otimes \nu$.

Theorem A.3.10 (Tonelli-Fubini). *Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{C}, \nu)$ be σ-finite measure spaces and $f : X \times Y \to [0, \infty]$ be measurable. Then we have*

$$\int_{X\times Y} f \, d(\mu \otimes \nu) = \int_Y \int_X f(x, y) \, d\mu(x) \, d\nu(y) = \int_X \int_Y f(x, y) \, d\nu(y) \, d\mu(x) \,,$$

where the inner integrals are measurable in the remaining argument. In addition, the same formula holds for $f \in \mathcal{L}_1(\mu \otimes \nu)$, but in this case the inner integrals are only almost surely defined.

Of course, products of more than two measures can be defined by induction, and it is straightforward to see that a suitable modification of Theorem A.3.10 still holds for such *finite* products of measures. Moreover, one can show that, for probability measures, even countable products can be defined. We refer to p. 113ff of Ash and Doléans-Dade (2000) for details.

The following lemma, which can be found, for example, on p. 141 of the book by Bauer (2001), provides a very handy formula for computing expectations by tail bounds.

Lemma A.3.11. *Let $(X, \mathcal{A}, \mu)$ be a finite measure space and $f : X \to [0, \infty)$ be a measurable function. Furthermore, let $\varphi : [0, \infty) \to [0, \infty)$ be a continuous function that is continuously differentiable on $(0, \infty)$ and satisfies $\varphi(0) = 0$. Then we have*

$$\int_X \varphi \circ f d\mu \;=\; \int_0^\infty \varphi'(t) \, \mu\big(f \geq t\big) \, dt \,.$$

The following definition will be used to describe measures that can be cut in arbitrary pieces.

Definition A.3.12. *Let $(X, \mathcal{A}, \mu)$ be a measure space. An $A \in \mathcal{A}$ is called an **atom** if for all measurable $B \subset A$ we either have $\mu(A \backslash B) = 0$ or $\mu(B) = 0$. Moreover, $(X, \mathcal{A}, \mu)$ is called **atom-free** if no $A \in \mathcal{A}$ is an atom.*

Obviously, the Lebesgue measure on subsets of $\mathbb{R}^d$ is atom free, and in addition every measure that has a density with respect to some atom-free measure is itself atom-free. On the other hand, for a measure μ on $\mathbb{N}$, each singleton $\{n\}$ with $\mu(\{n\}) > 0$ is an atom. The following theorem shows that atom-free probability measures can be cut arbitrarily.

Theorem A.3.13 (Lyapunov). *Let $(X, \mathcal{A}, \mu)$ be an atom-free probability space and let $f_1, \ldots, f_n \in \mathcal{L}_1(\mu)$. Then*

$$\left\{ \left(\int_X \mathbf{1}_A f_1 d\mu \,, \, \ldots \,, \, \int_X \mathbf{1}_A f_n d\mu \right) : A \in \mathcal{A} \right\}$$

is a compact and convex set in $\mathbb{R}^n$. In particular, $\{\mu(A) : A \in \mathcal{A}\} = [0, 1]$.

For any signed measure μ on a measurable space $(X, \mathcal{A})$ there exists a $B_\mu \in \mathcal{A}$ such that for all $A \in \mathcal{A}$ we have $\mu^+(A) := \mu(A \cap B_\mu) \geq 0$ and $\mu^-(A) := -\mu(A \setminus B_\mu) \geq 0$. It is easy to check that μ^+ and μ^- are measures. Moreover, at least one of these measures is finite and we have $\mu = \mu^+ - \mu^-$. In addition, μ^+ and μ^- are singular, i.e. there exists a $C \in \mathcal{A}$ with $\mu^+(C) = \mu^-(X \setminus C) = 0$. These properties uniquely determine μ^+, μ^-, and B_μ, and (μ^+, μ^-, B_μ) is called the **Hahn-Jordan decomposition** of μ. Moreover, the measure $|\mu| := \mu^+ + \mu^-$ is called the **total variation measure** for μ. For a function $g : X \to \mathbb{R}$ that is μ^+- and μ^--integrable, we define

$$\int_X g \, d\mu := \int_X g \, d\mu^+ - \int_X g \, d\mu^- .$$

Finally, the Radon-Nikodym Theorem A.3.9 can be extended to finite signed measures μ, see, e.g., Corollary 5.6.2 by Dudley (2002).

A.3.2 Measures on Topological Spaces

In this subsection, we briefly recall σ-algebras and corresponding measures that are defined by a topology. To this end let us recall from the previous subsection that for topological spaces (X, τ) the Borel σ-algebra was defined by $\mathcal{B}(X) := \sigma(\tau)$. In the following, we call a measure $\mu : \mathcal{B}(X) \to [0, \infty]$ a **Borel measure**. It is often useful if one can approximate Borel measures by their behavior on sets associated to the underlying topology. This idea is formally expressed in the following definition.

Definition A.3.14. *Let (X, τ) be a topological space and μ be a Borel measure on X. Then μ is called **regular** if for each $A \in \mathcal{B}(X)$ we have **outer regularity**, i.e.,*

$$\mu(A) = \inf\{\mu(O) : A \subset O,\ O \text{ open}\}, \tag{A.9}$$

*and **inner regularity**, i.e.,*

$$\mu(A) = \sup\{\mu(C) : C \subset A,\ C \text{ compact}\}. \tag{A.10}$$

The next theorem shows that regular measures naturally occur on benign topological spaces. Its proof can be found on p. 225 of Dudley (2002).

Theorem A.3.15 (Ulam's theorem). *Every finite Borel measure on a Polish space is regular.*

Let us now assume that μ is a Borel measure on a locally compact space (X, τ). Furthermore, assume that μ is inner regular, i.e., it satisfies (A.10), and that for all $x \in X$ there exists an $O \in \tau$ such that $x \in O$ and $\mu(O) < \infty$. In the literature such measures are called **Radon measures**. One can show that for Radon measures μ there exists a largest open μ-zero set G, i.e., the union

G of *all* open μ-zero sets is again a μ-zero set. In this case, the **support** of μ is defined to be $\operatorname{supp}\mu := X\backslash G$, and μ is called **strictly positive** if $\operatorname{supp}\mu = X$. Note that the same construction can be made for Polish spaces.

Polish spaces also play an important role when "disintegrating" probability measures on product spaces, as the following result (see, e.g., Section 10.2 of Dudley, 2002) shows.

Lemma A.3.16. *Let $(X, \mathcal{A})$ be a measurable space and Y be a Polish space with Borel σ-algebra $\mathcal{B}(Y)$. Furthermore, let P be a probability measure on $\mathcal{A} \otimes \mathcal{B}(Y)$. Then there exists a map $\mathrm{P}(\,\cdot\,|\,\cdot\,) : \mathcal{B}(Y) \times X \to [0,1]$ such that*

i) $\mathrm{P}(\,\cdot\,|x)$ is a probability measure on $\mathcal{B}(Y)$ for all $x \in X$.
ii) $x \mapsto \mathrm{P}(B|x)$ is measurable for all $B \in \mathcal{B}(Y)$.
iii) For all $A \in \mathcal{A}$, $B \in \mathcal{B}(Y)$, we have

$$\mathrm{P}(A \times B) \;=\; \int_A \mathrm{P}(B|x)\, d\mathrm{P}_X(x)\,. \tag{A.11}$$

*The map $\mathrm{P}(\,\cdot\,|\,\cdot\,)$ is called a **regular conditional probability** or **regular conditional distribution** of P.*

Finally, the following lemma is often useful to guarantee the measurability of functions defined on a product space. Its proof can be found on p. 70 of Castaing and Valadier (1977).

Lemma A.3.17 (Carathéodory). *Let $(X, \mathcal{A})$ be a measurable space, Z be a Polish space equipped with its Borel σ-algebra, and $h : X \times Z \to \mathbb{R}$ be a map. Then h is measurable if the following two conditions are satisfied:*

i) $h(x, \cdot\,) : Z \to \mathbb{R}$ is continuous for all $x \in X$.
ii) $h(\,\cdot\,, z) : X \to \mathbb{R}$ is measurable for all $z \in Z$.

A.3.3 Aumann's Measurable Selection Principle

In various parts of the book, we need measurable selections. In this subsection, we will present a general result on the existence of measurable selections that goes back to Aumann.

We begin with some basic preparations. To this end, let X, Y, and Z be measurable spaces, where for notational convenience we omit the corresponding σ-algebras. Moreover, let $h : X \times Z \to Y$ and $A \subset Y$ be measurable, and

$$\begin{aligned} F : X &\to 2^Z \\ x &\mapsto \big\{z \in Z : h(x,z) \in A\big\}\,, \end{aligned} \tag{A.12}$$

where 2^Z denotes the set of all subsets of Z. Note that, given a measurable $Z_0 \subset Z$, the constant function $F : X \to 2^Z$ defined by $F(x) := Z_0$ is of the form (A.12) since $Z_0 = \{z \in Z : \mathbf{1}_{X \times Z_0}(x,z) \in \{1\}\}$. Furthermore, we write

$$\begin{aligned}\operatorname{Dom} F &:= \big\{x \in X : F(x) \neq \emptyset\big\}\,,\\ \operatorname{Gr} F &:= \big\{(x,z) \in X \times Z : z \in F(x)\big\}\,,\\ F^{-1}(B) &:= \big\{x \in X : F(x) \cap B \neq \emptyset\big\} \qquad\qquad B \subset Z.\end{aligned}$$

Then we have $\operatorname{Dom} F = F^{-1}(Z)$ and $\operatorname{Gr} F = \{(x,z) \in X \times Z : h(x,z) \in A\}$, and consequently $\operatorname{Gr} F$ is measurable. Furthermore, if $\pi_X : X \times Z \to X$ denotes the projection onto X, then we have

$$\begin{aligned}\operatorname{Dom} F &= \big\{x \in X : \exists z \in Z \text{ with } h(x,z) \in A\big\}\\ &= \pi_X\Big(\big\{(x,z) \in X \times Z : h(x,z) \in A\big\}\Big)\\ &= \pi_X\big(\operatorname{Gr} F\big)\,.\end{aligned}$$

Now the following result provides a sufficient condition under which $\operatorname{Dom} F$ is measurable and F admits measurable selections.

Lemma A.3.18 (Aumann's measurable selection principle). *Let $(X, \mathcal{A})$ be a complete measurable space, Z be a Polish space equipped with its Borel σ-algebra, and Y be a measurable space. Furthermore, let $h : X \times Z \to Y$ be a measurable map, $A \subset Y$ be measurable, and $F : X \to 2^Z$ be defined by (A.12). Then the following statements are true:*

i) $\operatorname{Dom} F$ *is measurable.*

ii) There exists a sequence of measurable functions $f_n : X \to Z$ such that for all $x \in \operatorname{Dom} F$ the set $\{f_n(x) : n \in \mathbb{N}\}$ is dense in $F(x)$.

iii) Let $\varphi : X \times Z \to [0,\infty]$ be measurable and $\psi : X \to [0,\infty]$ be defined by

$$\psi(x) := \inf_{z \in F(x)} \varphi(x,z)\,, \qquad\qquad x \in X. \tag{A.13}$$

Then ψ is measurable. Furthermore, for all $n \geq 1$, there exists a measurable $f_n : X \to Z$ such that for all $x \in \operatorname{Dom} F$ we have $f_n(x) \in F(x)$ and $\varphi(x, f_n(x)) \leq \psi(x) + 1/n$, and consequently

$$\psi(x) = \inf_{n \in \mathbb{N}} \varphi\big(x, f_n(x)\big)\,, \qquad\qquad x \in \operatorname{Dom} F\,.$$

In addition, if the infimum in (A.13) is attained for all $x \in \operatorname{Dom} F$, then there exists a measurable function $f^* : X \to Z$ with $\psi(x) = \varphi(x, f^*(x))$ for all $x \in \operatorname{Dom} F$.

Since the complete proof is out of the scope of this book, we only show how the above lemma follows from Aumann's original result. We closely follow the presentation of Castaing and Valadier (1977).

Proof. i). Let us first recall that the projection theorem (see Theorem III.23 on p. 75 of Castaing and Valadier, 1977) ensures $\pi_X(B) \in \mathcal{A}$ for all $B \in \mathcal{A} \otimes \mathcal{B}(Z)$. Now the assertion follows directly from $\operatorname{Dom} F = \pi_X(\operatorname{Gr} F)$.

ii). Let $\mathcal{A} \cap \operatorname{Dom} F$ denote the trace σ-algebra of $\mathcal{A}$ in $\operatorname{Dom} F$. Then it is easy to see that $\mathcal{A} \cap \operatorname{Dom} F$ is a complete σ-algebra. Furthermore, $F_{|\operatorname{Dom} F} : \operatorname{Dom} F \to 2^Z$ obviously maps $\operatorname{Dom} F$ to non-empty subsets of Z. In addition, we have

$$\operatorname{Gr}(F_{|\operatorname{Dom} F}) = \big\{(x,z) \in \operatorname{Dom} F \times Z : z \in F(x)\big\} = \operatorname{Gr} F \cap \big(\operatorname{Dom} F \times Z\big)\,,$$

and hence $\operatorname{Gr}(F_{|\operatorname{Dom} F})$ is measurable by part *i)*. Now Aumann's selection theorem (see Theorem III.22 on p. 74 of Castaing and Valadier, 1977) gives a sequence (f_n) of $(\mathcal{A} \cap \operatorname{Dom} F)$-measurable functions $f_n : \operatorname{Dom} F \to Z$ such that the set $\{f_n(x) : n \in \mathbb{N}\}$ is dense in $F(x)$ for all $x \in \operatorname{Dom} F$. Extending these functions to measurable functions $\tilde{f}_n : X \to Z$ gives the assertion.

iii) The measurability of ψ follows from (Castaing and Valadier, 1977, Lemma III.39 on p. 86). Moreover, on the measurable set $\{x \in X : \psi(x) = \infty\}$, there is nothing to prove, and hence we may restrict our considerations to $\operatorname{Dom} F \cap \{x \in X : \psi(x) < \infty\}$ equipped with the trace σ-algebra of $\mathcal{A}$. Then the existence of f_n is shown on p. 87 of Castaing and Valadier (1977). Finally, the existence of the measurable function $f^* : X \to Z$ is shown on p. 86 of the same book. □

A.4 Probability Theory and Statistics

A.4.1 Some Basic Facts

In this section, we briefly collect the basic notions from probability theory. To this end, let $(X, \mathcal{A})$ be a measurable space. Recall that a measure P on $\mathcal{A}$ is called a **probability measure** or **distribution** if $\mathrm{P}(X) = 1$, and the triple $(X, \mathcal{A}, \mathrm{P})$ is called **probability space**. Moreover, the σ-algebra $\mathcal{A}$ is typically known from the context, and in this case we usually do not mention it explicitly. Moreover, for a function $f \in L_1(\mathrm{P})$, we often use one of the notations

$$\mathbb{E}_{\mathrm{P}} f := \mathbb{E} f := \mathbb{E}_{x \sim \mathrm{P}} f(x) := \int_X f d\mathrm{P}$$

and call the value of the integral above the **expectation** of f. If $f \in L_2(\mathrm{P})$, we often write

$$\operatorname{Var}_{\mathrm{P}} f := \mathbb{E}_{\mathrm{P}} (f - \mathbb{E}_{\mathrm{P}} f)^2$$

and call the value of this integral the **variance** of f. Let us now introduce some fundamental objects from probability theory.

Definition A.4.1. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. A mapping $\xi : X \to Y$ is called a **random variable** if ξ is $(\mathcal{A}, \mathcal{B})$-measurable.*

If there is a probability measure P on the σ-algebra $\mathcal{A}$ above, then the random variable induces a probability measure on $\mathcal{B}$. This is made precise in the following definition.

Definition A.4.2. *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space, $(Y, \mathcal{B})$ be a measurable space, and $\xi : X \to Y$ be a random variable. Then*

$$\mathrm{P}_\xi(B) := \mathrm{P}(\xi \in B) := \mathrm{P}(\xi^{-1}(B))\,, \qquad B \in \mathcal{B},$$

defines a probability measure on $\mathcal{B}$. This measure is called the ***image measure*** *of P under ξ and often also called the* ***distribution*** *of ξ.*

We are typically interested in integrals for functions of the form $f \circ \xi$. The following theorem relates such integrals with the distribution of ξ.

Theorem A.4.3 (Transformation formula). *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space, $(Y, \mathcal{B})$ be a measurable space, and $\xi : X \to Y$ be a random variable. Furthermore, let E be a separable Banach space and $f : Y \to E$ be a measurable function. Then we have $f \circ \xi \in L_1(\mathrm{P})$ if and only if $f \in L_1(\mathrm{P}_\xi)$, and in this case we have the identity $\mathbb{E}_\mathrm{P} f \circ \xi \;=\; \mathbb{E}_{\mathrm{P}_\xi} f$.*

Let us now assume that we have two random variables $\xi : X \to Y$ and $\xi' : X' \to Y$ defined on the probability spaces $(X, \mathcal{A}, \mathrm{P})$ and $(X', \mathcal{A}', \mathrm{P}')$, respectively. We say that ξ and ξ' are **equal in distribution** or **identically distributed** if $\mathrm{P}_\xi = \mathrm{P}'_{\xi'}$. In this case, the transformation formula yields

$$\mathbb{E}_\mathrm{P} f \circ \xi \;=\; \mathbb{E}_{\mathrm{P}_\xi} f \;=\; \mathbb{E}_{\mathrm{P}'_{\xi'}} f \;=\; \mathbb{E}_{\mathrm{P}'} f \circ \xi'\,;$$

i.e., the expectations of $f \circ \xi$ and $f \circ \xi'$ coincide if they exist. Consequently, if we are only interested in statements involving expectations of the form $\mathbb{E}_\mathrm{P} f \circ \xi$, we can exchange ξ by an equally distributed random variable ξ'. Remarkably, there are important situations introduced below in which we can find such an alternative ξ', which is easier to analyze than the original random variable ξ.

Given measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ and a random variable $\xi : X \to Y$, the set

$$\sigma(\xi) := \bigl\{\xi^{-1}(B) : B \in \mathcal{B}\bigr\}$$

is called the σ-algebra **induced** or **generated** by ξ. Obviously, $\sigma(\xi)$ is a sub-σ-algebra of $\mathcal{A}$.

Let us now recall independence for σ-algebras and random variables. To this end, let $(X, \mathcal{A}, \mathrm{P})$ be a probability space, I be an index set, and $(A_i)_{i \in I}$ be a family of sets with $A_i \in \mathcal{A}$ for all $i \in I$. Then the events A_i, $i \in I$, are called **(stochastically) independent** if for all distinct indexes $i_1, \ldots, i_n \in I$ and all $n \in \mathbb{N}$ we have

$$\mathrm{P}\Bigl(\bigcap_{j=1}^{n} A_{i_j}\Bigr) \;=\; \prod_{i=1}^{n} \mathrm{P}(A_{i_j})\,. \tag{A.14}$$

Now let $(\mathcal{A}_i)_{i \in I}$ be a family of σ-algebras with $\mathcal{A}_i \subset \mathcal{A}$ for all $i \in I$. Then the σ-algebra's $\mathcal{A}_i$, $i \in I$, are called independent if all families $(A_i)_{i \in I}$ of events with $A_i \in \mathcal{A}_i$, $i \in I$, are independent. Finally, let $(Y_i, \mathcal{B}_i)$, $i \in I$, be

measurable spaces. Then the random variables $\xi_i : X \to Y_i$, $i \in I$, are called independent if their induced σ-algebras $\sigma(\xi_i)$, $i \in I$, are independent. In this case, it is easy to check that, for a given partition $(I_j)_{j\in J}$ of I, the grouped random variables $\eta_j := (\xi_i)_{i\in I_j}$, $j \in J$, are again independent. Moreover, for random variables $f_i : Y_i \to Y_i'$ mapping into measurable spaces $(Y_i', \mathcal{B}_i')$, $i \in I$, it immediately follows from the definition that the random variables $f_i \circ \xi_i$, $i \in I$, are independent if the (ξ_i) are independent.

The following lemma gives a useful characterization of finite sequences of independent random variables.

Lemma A.4.4. *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space, $(Y_i, \mathcal{B}_i)$, $i = 1, \ldots, n$, be measurable spaces, and $\xi_i : X \to Y_i$, $i = 1, \ldots, n$, be random variables. Then $\xi_1, \ldots, \xi_n$ are independent if and only if the image measure of the random variable $\xi := (\xi_1, \ldots, \xi_n) : X \to Y_1 \times \cdots \times Y_n$ is given by $\mathrm{P}_\xi = \mathrm{P}_{\xi_1} \otimes \cdots \otimes \mathrm{P}_{\xi_n}$.*

The preceding lemma in particular shows that the projections $\pi_i : Y_1 \times \cdots \times Y_n \to Y_i$, $i = 1, \ldots, n$, are independent with respect to every product measure $\mathrm{P} = \mathrm{P}_1 \otimes \cdots \otimes \mathrm{P}_n$ on $Y_1 \times \cdots \times Y_n$. Moreover, if $\mathrm{P} = \mathrm{P}_{(\xi_1,\ldots,\xi_n)}$, these projections have the same joint distribution as the finite sequence $\xi_1, \ldots, \xi_n$, and hence they can serve as a canonical representation of the independent random variables $\xi_1, \ldots, \xi_n$ whenever we are only interested in properties that can be expressed in terms of P_ξ.

Now let $(X, \mathcal{A}, \mathrm{P})$ be a probability space and $\xi_i : X \to \mathbb{R}$, $i = 1, \ldots, n$, be independent $\mathbb{R}$-valued random variables. Then it is easy to derive from (A.14) that

$$\mathbb{E} \prod_{i=1}^{n} \xi_i = \prod_{i=1}^{n} \mathbb{E}\xi_i \,. \tag{A.15}$$

Moreover, a similar result holds for Hilbert space valued independent random variables. Namely, if H is a separable Hilbert space and $\xi_1, \xi_2 : X \to H$ are independent random variables, then we have

$$\mathbb{E}_\mathrm{P}\langle \xi_1, \xi_2 \rangle_H \;=\; \langle \mathbb{E}_\mathrm{P}\xi_1, \mathbb{E}_\mathrm{P}\xi_2 \rangle_H \,. \tag{A.16}$$

A generalization to longer products is possible, but since we do not need such generalizations, we omit the details.

Theorem A.4.5. *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space, $\xi : X \to \mathbb{R}$ be a random variable that is P-integrable or non-negative, and $\mathcal{B} \subset \mathcal{A}$ be a σ-algebra. Then there exists a $\mathcal{B}$-measurable function $\eta : X \to \mathbb{R}$ such that $\mathbb{E}_\mathrm{P}\mathbf{1}_B\eta = \mathbb{E}_\mathrm{P}\mathbf{1}_B\xi$ for all $B \in \mathcal{B}$. Any two such functions coincide P-almost surely. Finally, η is called the **conditional expectation** of ξ and is denoted by $\mathbb{E}_\mathrm{P}(\xi \,|\, \mathcal{B})$ or $\mathbb{E}(\xi \,|\, \mathcal{B})$.*

Note that in general we do *not* have $\mathbb{E}(\xi \,|\, \mathcal{B}) = \xi$ since $\mathbb{E}(\xi \,|\, \mathcal{B})$ is required to be $\mathcal{B}$-measurable, whereas ξ is only assumed to be $\mathcal{A}$-measurable. The following lemma collects some useful properties of conditional expectations.

Lemma A.4.6. *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space, $\xi, \eta : X \to \mathbb{R}$ be random variables, $\mathcal{B}, \mathcal{C} \subset \mathcal{A}$ be σ-algebras, and $\psi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then the following statements are true provided that the occurring (conditional) expectations exist:*

i) $\mathbb{E}(\cdot \,|\, \mathcal{B}) : L_1(\mathrm{P}) \to L_1(\mathrm{P})$ *is a linear projection onto the space* $L_1(\mathcal{B}, \mathrm{P})$ *of all* $\mathcal{B}$*-measurable and* P*-integrable functions. Moreover, it is a 1-projection; i.e., it satisfies* $\mathbb{E}|\mathbb{E}(\xi \,|\, \mathcal{B})| \leq \mathbb{E}|\xi|$.
ii) $\xi \geq 0$ *implies* $\mathbb{E}(\xi \,|\, \mathcal{B}) \geq 0$.
iii) $\mathbb{E}(\eta\xi \,|\, \mathcal{B}) = \eta\, \mathbb{E}(\xi \,|\, \mathcal{B})$ *if* η *is* $\mathcal{B}$*-measurable.*
iv) $\mathbb{E}\big(\eta \cdot \mathbb{E}(\xi \,|\, \mathcal{B})\big) = \mathbb{E}\big(\mathbb{E}(\eta \,|\, \mathcal{B}) \cdot \xi\big) = \mathbb{E}\big(\mathbb{E}(\eta \,|\, \mathcal{B}) \cdot \mathbb{E}(\xi \,|\, \mathcal{B})\big)$.
v) $\mathbb{E}\big(\mathbb{E}(\xi \,|\, \mathcal{B}) \big| \mathcal{C}\big) = \mathbb{E}(\xi \,|\, \mathcal{C})$ *if* $\mathcal{C} \subset \mathcal{B}$.
vi) $\mathbb{E}(\xi \,|\, \mathcal{B}) = \mathbb{E}\xi$ *if* $\sigma(\xi)$ *and* $\mathcal{B}$ *are independent.*
vii) ***Jensen's inequality:*** $\psi\big(\mathbb{E}(\xi \,|\, \mathcal{B})\big) \leq \mathbb{E}\big(\psi(\xi) \,|\, \mathcal{B}\big)$.

We now recall the definition of exponential families that is used in Chapter 12. Let $q \in \mathbb{N}$ and $\Theta \subset \mathbb{R}^q$. A set of probability distributions $\mathcal{P} := \{\mathrm{P}_\theta : \theta \in \Theta\}$ on some measurable space $(X, \mathcal{A})$ is called **dominated** by a σ-finite measure ν if all P_θ are absolutely continuous with respect to ν. Hence, the Radon-Nikodym Theorem A.3.9 yields the existence of ν-densities $\frac{d\mathrm{P}_\theta}{d\nu}$. A dominated set of probability distributions $\mathcal{P} := \{\mathrm{P}_\theta : \theta \in \Theta\}$ is called a ***q*-parameter exponential family** in η and T with parameter θ if all $\mathrm{P}_\theta \in \mathcal{P}$ have ν-densities of the form

$$\frac{d\mathrm{P}_\theta}{d\nu}(x) = c(\theta)\, \exp\big(\langle \eta(\theta), T(x) \rangle\big), \qquad x \in X, \tag{A.17}$$

where *(i)* 1 and $\eta_j : \Theta \to \mathbb{R}$ for $j = 1, \ldots, q$ are linearly independent real-valued functions, *(ii)* 1, $T_j : \Theta \to \mathbb{R}$ for $j = 1, \ldots, q$ are on the complement of every ν-null set linearly independent Borel measurable functions, and *(iii)* $\langle \eta(\theta), T(x) \rangle := \sum_{j=1}^q \eta_j(\theta) T_j(x)$. The set of all $\eta \in \mathbb{R}^q$ with $0 < \int_X \exp\big(\langle \eta, T(x) \rangle\big)\, d\nu(x) < \infty$ is called canonical parameter space. All moments of order $p \in \mathbb{N}$ of the distribution of $(T_1, \ldots, T_q)$ exist. We refer to Witting (1985) and Lehmann and Romano (2005) for further details.

A.4.2 Some Limit Theorems

In this section, we collect some useful limit theorems for sequences of random variables and empirical measures.

The following result is often useful to show almost sure convergence. Recall that $\limsup A_n = \bigcap_{n \in \mathbb{N}} \bigcup_{i \geq n} A_i$ for a sequence of sets A_n, $n \in \mathbb{N}$.

Theorem A.4.7 (Borel-Cantelli lemma). *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space and $(A_n)_{n \in \mathbb{N}}$ a sequence of sets $A_n \in \mathcal{A}$. Then the following statements are true:*

i) If $\sum_{n \in \mathbb{N}} \mathrm{P}(A_n) < \infty$, *then* $\mathrm{P}(\limsup A_n) = 0$.

ii) If $\sum_{n\in\mathbb{N}} \mathrm{P}(A_n) = \infty$ *and if* $A_1, A_2, \ldots$ *are stochastically independent, then* $\mathrm{P}(\limsup A_n) = 1$.

Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space. A sequence $(X_n)_{n\in\mathbb{N}}$ of real random variables on X is said to satisfy the **strong law of large numbers** if and only if $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges P-almost surely to some constant c; i.e., $\mathrm{P}\big(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = c\big) = 1$. A sequence $(X_n)_{n\in\mathbb{N}}$ of real random variables on X is said to satisfy the **weak law of large numbers** if and only if $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to some constant c in probability P; i.e.; there exists some constant c such that, for all $\varepsilon > 0$, $\lim_{n\to\infty} \mathrm{P}\big(\big|\frac{1}{n}\sum_{i=1}^{n} X_i - c\big| > \varepsilon\big) = 0$.

Theorem A.4.8 (Weak law of large numbers).

i) If $(X_n)_{n\in\mathbb{N}}$ *is a sequence of real random variables on* $(X, \mathcal{A})$ *with* $\mathbb{E}\, X_i = 0$, $\mathbb{E}\, X_i^2 = 1$, *and* $\mathbb{E}\, X_i X_j = 0$ *for all* $i \neq j$, *then the weak law of large numbers holds for* $(X_n)_{n\in\mathbb{N}}$.

ii) If $(X_n)_{n\in\mathbb{N}}$ *is a sequence of independent, identically distributed real random variables on* $(X, \mathcal{A})$ *with* $\mathbb{E}\, X_i = \mu \in \mathbb{R}$ *and* $\mathrm{Var} X_i = \sigma^2 \in (0, \infty)$, *then the weak law of large numbers holds for* $\big(\frac{X_n - \mu}{\sigma}\big)_{n\in\mathbb{N}}$ *and* $\frac{1}{n}\sum_{i=1}^{n} X_n$ *converges to* $\mathbb{E}\, X_1$ *in probability.*

Proof. See Dudley (2002, pp. 261ff.). Note that the hypothesis in part *i)* means that the random variables are orthonormal in the Hilbert space $L_2(\mathrm{P})$. □

Theorem A.4.9 (Strong law of large numbers). *If* $(X_n)_{n\in\mathbb{N}}$ *is a sequence of independent, identically distributed real random variables on* $(X, \mathcal{A})$ *with* $\mathbb{E}\,|X_1| < \infty$, *then the strong law of large numbers holds for* $(X_n)_{n\in\mathbb{N}}$ *and* $\frac{1}{n}\sum_{i=1}^{n} X_i$ *converges to* $\mathbb{E}\, X_1$ *with probability one. If* $\mathbb{E}\,|X_1| = \infty$, *then almost surely* $\frac{1}{n}\sum_{i=1}^{n} X_i$ *does not converge to any finite limit.*

Proof. See Dudley (2002, p. 263). □

Theorem A.4.10 (Central limit theorem). *Assume that for each fixed* $n \in \mathbb{N}$ *the real-valued random variables* $X_{n,j}$, $j = 1, \ldots, k(n)$, *are independent with* $\mathbb{E}X_{n,j} = 0$, $\sigma_{n,j}^2 := \mathbb{E}X_{n,j}^2$, *and* $\sum_{1\le j\le k(n)} \sigma_{n,j}^2 = 1$. *Define*

$$S_n := \sum_{j=1}^{k(n)} X_{n,j} \quad \text{and} \quad L_{n,j}(\varepsilon) := \int_{|x|>\varepsilon} x^2 d\mathrm{P}_{X_{n,j}}(x)\,, \quad \varepsilon > 0.$$

If $\lim_{n\to\infty} \sum_{1\le j\le k(n)} L_{n,j}(\varepsilon) = 0$ *for all* $\varepsilon > 0$, *then* $d_{\mathrm{Pro}}\big(\mathrm{P}_{S_n}, N(0,1)\big) \to 0$ *for* $n \to \infty$, *where* d_{Pro} *denotes the Prohorov metric from Definition 10.1.*

Proof. See Dudley (2002, p. 316). □

Theorem A.4.11 (Varadarajan). *Let* (X, d) *be a separable metric space and* P *be any Borel probability measure on* X. *Then the empirical measures* P_n *converge to* P *almost surely; i.e.,* $\mathrm{P}\big(\{\omega : \mathrm{P}_n(\cdot)(\omega) \to \mathrm{P}\}\big) = 1$.

Proof. See Dudley (2002, p. 399). □

Theorem A.4.12 (Glivenko-Cantelli theorem). *Let* P *be any probability measure on* $(\mathbb{R}, \mathcal{B})$ *with distribution function* $F(x) = \mathrm{P}((-\infty, x])$, $x \in \mathbb{R}$. *Define the empirical distribution function of* P_n *by* $F_n(x)(\omega) := \mathrm{P}_n((-\infty, x])(\omega)$. *Then almost surely* $F_n(\cdot)(\omega) \to F$ *uniformly on* $\mathbb{R}$ *as* $n \to \infty$.

Proof. See Dudley (2002, p. 400). □

A.4.3 The Weak* Topology and Its Metrization

We now mention some results useful for Chapter 10 on robustness. To this end, let us now assume that Z is a Polish space. We write $\mathcal{M}_1(Z)$, or simply $\mathcal{M}_1$, for the set of all Borel probability measures on Z. Furthermore, recall that Ulam's Theorem A.3.15 shows that every finite Borel measure μ on a Polish space is regular in the sense that the value of $\mu(A)$, $A \in \mathcal{B}(Z)$, can be approximated by compact sets from below and by open sets from above.

Theorem A.4.13. *Let* (Z, τ) *be a Polish space with Borel* σ*-algebra* $\mathcal{B}(Z)$. *The following statements are equivalent:*

i) *A sequence of probability distributions* P_n, $n \in \mathbb{N}$, *converges weakly to a probability distribution* P *if* $n \to \infty$.
ii) $\liminf \mathrm{P}_n(A) \geq \mathrm{P}(A)$ *for all open sets* $A \in \mathcal{B}(Z)$.
iii) $\limsup \mathrm{P}_n(A) \leq \mathrm{P}(A)$ *for all closed sets* $A \in \mathcal{B}(Z)$.
iv) $\lim \mathrm{P}_n(A) = \mathrm{P}(A)$ *for all Borel sets* $A \in \mathcal{B}(Z)$ *with* $\mathrm{P}(\partial A) = 0$, *where* ∂A *denotes the boundary of* A.

Proof. See Huber (1981, p. 22). □

Definition A.4.14. *Suppose that* (Z, τ) *is a Polish space. A set* $\mathcal{P} \subset \mathcal{M}_1(Z)$ *is called* ***tight*** *if for every* $\varepsilon > 0$ *there exists a compact set* $K \subset Z$ *such that* $\mathrm{P}(K) \geq 1 - \varepsilon$ *for all* $\mathrm{P} \in \mathcal{P}$. *A set* $\mathcal{P} \subset \mathcal{M}_1(Z)$ *is called* ***relatively compact*** *if every sequence of elements of* $\mathcal{P}$ *contains a weakly convergent subsequence.*

Theorem A.4.15 (Prohorov's theorem). *Suppose that* (Z, τ) *is a Polish space. Then* $\mathcal{P} \subset \mathcal{M}_1(Z)$ *is tight if and only if* $\mathcal{P}$ *is relatively compact.*

Proof. See Prohorov (1956). □

Theorem A.4.16 (Strassen's theorem). *Let* (Z, τ) *be a Polish space with complete metric* d_Z, P_1 *and* P_2 *be probability measures on* Z, *and denote the closed* δ*-neighborhood of a set* A *by* $A^\delta := \{x \in Z : \inf_{y \in A} d_Z(x, y) \leq \delta\}$. *Then the following two statements are equivalent:*

(i) $\mathrm{P}_1(A) \leq \mathrm{P}_2(A^\delta) + \varepsilon$ *for all* $A \in \mathcal{B}(Z)$.
(ii) *There are (dependent)* Z*-valued random variables* Z_i *such that* P_i *is the probability distribution of* Z_i, $i = 1, 2$, *and* $\mathrm{P}(d_Z(Z_1, Z_2) \leq \delta) \geq 1 - \varepsilon$.

Proof. See Strassen (1965, pp. 436ff). □

Definition A.4.17. *Let (Z, τ) be a Polish space. The* ***weak* topology*** *in the set of probability measures $\mathcal{M}_1(Z)$ is the weakest topology such that for every continuous $f : X \to \mathbb{R}$ the map $g_f : \mathcal{M}_1(Z) \to \mathbb{R}$, $g_f(\mathrm{P}) = \mathbb{E}_\mathrm{P} f$ is continuous.*

Theorem A.4.18. *Let (Z, τ) be a Polish space. Then the set $\mathcal{M}_1(Z)$ of all probability measures on Z endowed with the weak* topology is a Polish space.*

Proof. See Huber (1981, p. 29). □

Denote by $\mathcal{M}_1'$ the set of finite signed measures (see Definition A.3.2) on (Z, τ). Let h be the restriction to $\mathcal{M}_1$ of a linear functional on $\mathcal{M}_1'$. Then a linear functional h is *weakly continuous* on $\mathcal{M}_1$ if and only if it has the representation $h(\mu) = \int f \, d\mu$ for some bounded and continuous function f.

Now we will consider certain metrics on $\mathcal{M}_1$ that describe the weak* topology and that are useful for robust statistics. This list of metrics is, however, by no means complete. We begin with the Prohorov metric introduced in Definition 10.1.

Theorem A.4.19. *The Prohorov metric d_Pro defines a metric on $\mathcal{M}_1(Z)$.*

Proof. See Prohorov (1956) or Dudley (2002, p. 394). □

Theorem A.4.20. *Let (Z, τ) be a Polish space. The Prohorov metric metrizes the weak* topology in $\mathcal{M}_1(Z)$.*

Proof. See Prohorov (1956) or Dudley (2002, p. 395). □

Definition A.4.21. *Let $\mathcal{M}_1$ be the set of probability measures on Z, where (Z, τ) is a Polish space. Consider a complete metric d_Z on Z that is bounded by one.*[1] *The* ***bounded Lipschitz metric*** *is defined by $d_\mathrm{bL}(\mathrm{P}_1, \mathrm{P}_2) := \sup_f |\mathbb{E}_{\mathrm{P}_1} f - \mathbb{E}_{\mathrm{P}_2} f|$, where the supremum is taken over all functions f satisfying the Lipschitz condition $|f(x) - f(y)| \le d_Z(x, y)$.*

This metric has some nice properties summarized in the next theorem. Note that part *iii)* can be viewed as an analogue to Theorem A.4.16.

Theorem A.4.22. *Let (Z, τ) be a Polish space.*

i) The bounded Lipschitz metric d_bL is a metric.
ii) If $\mathrm{P}_1, \mathrm{P}_2 \in \mathcal{M}_1(Z)$, then $d^2_\mathrm{Pro}(\mathrm{P}_1, \mathrm{P}_2) \le d_\mathrm{bL}(\mathrm{P}_1, \mathrm{P}_2) \le 2 d_\mathrm{Pro}(\mathrm{P}_1, \mathrm{P}_2)$.
iii) The following two statements are equivalent:
  (a) $d_\mathrm{bL}(\mathrm{P}_1, \mathrm{P}_2) \le \varepsilon$, where $\mathrm{P}_1, \mathrm{P}_2 \in \mathcal{M}_1(Z)$.
  (b) There are (dependent) Z-valued random variables Z_i with probability distributions P_i, $i = 1, 2$, and $\mathbb{E}\, d_Z(Z_1, Z_2) \le \varepsilon$.

[1] Otherwise replace $d_Z(x, y)$ by $d^*_Z(x, y) := d_Z(x, y)/(1 + d_Z(x, y))$.

Proof. See Huber (1981, pp. 29ff). □

Definition A.4.23. *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space, (Z, τ) a Polish space with complete metric d_Z, and Z_1, Z_2 measurable functions from X to Z. The* ***Ky Fan metric*** *d_{KyFan} is defined by*

$$d_{\mathrm{KyFan}}(Z_1, Z_2) = \inf\{\varepsilon \geq 0 : \mathrm{P}(d_Z(Z_1, Z_2) > \varepsilon) \leq \varepsilon\}.$$

Theorem A.4.24. *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space, (Z, τ) a Polish space with complete metric d_Z, and $L_0(X, Z)$ the space of equivalence classes of measurable functions from X to Z. Then d_{KyFan} metrizes convergence in probability such that $\lim_{n\to\infty} d_{\mathrm{KyFan}}(Z_n, Z_0) = 0$ if and only if, for all $\varepsilon > 0$, $\lim_{n\to\infty} \mathrm{P}\big(d_Z(Z_n, Z_0) > \varepsilon\big) = 0$.*

Proof. See Dudley (2002, p. 289). □

The next result gives a relationship between the Prohorov metric and the Ky Fan metric. We denote the probability measure of Z_1 by P_{Z_1}.

Theorem A.4.25. *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space, (Z, τ) a Polish space with complete metric d_Z, and Z_1, Z_2 measurable functions from X to Z. Then $d_{\mathrm{Pro}}(\mathrm{P}_{Z_1}, \mathrm{P}_{Z_2}) \leq d_{\mathrm{KyFan}}(Z_1, Z_2)$.*

Proof. See Dudley (2002, p. 397). □

The next two theorems provide useful necessary and sufficient conditions for qualitative robustness in the sense of Definition 10.2.

Theorem A.4.26. *Let (Z, τ_Z), (W, τ_W) be Polish spaces and $Z_1, \ldots, Z_n$ independent and identically distributed Z-valued random functions. Furthermore, let $(S_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions where $S_n : Z^n \to W$, $S_n(Z_1, \ldots, Z_n) \in W$. Assume:*

i) The sequence $(S_n)_{n\in\mathbb{N}}$ is qualitatively robust at $\mathrm{P} \in \mathcal{P} \subset \mathcal{M}_1(Z)$.
ii) The probability measure $\mathrm{Q} \in N(\mathrm{P}) \subset \mathcal{P}$, where $N(\mathrm{P})$ is a neighborhood of $\mathrm{P} \in \mathcal{M}_1(Z)$ in the relative topology of $\mathcal{P}$.
iii) There exists a measurable function $S_\infty : N(\mathrm{P}) \to W$ such that

$$\lim_{n\to\infty} d_{\mathrm{KyFan}}\big(S_n(Z_1, \ldots, Z_n), S_\infty(\mathrm{Q})\big) = 0, \quad \forall\, \mathrm{Q} \in N(\mathrm{P}).$$

Then, S_∞ is continuous at P.

Proof. See Hampel (1968) and Cuevas (1988). □

Theorem A.4.27. *Let (Z, τ_Z), (W, τ_W) be Polish spaces and $Z_1, \ldots, Z_n$ independent and identically distributed Z-valued random functions. Furthermore, let $(S_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions, where $S_n : Z^n \to W$, $S_n(Z_1, \ldots, Z_n) \in W$, such that there exists a measurable function $S : \mathcal{M}_1(Z) \to W$ with $S_n(Z_1, \ldots, Z_n) = S(\mathrm{P}_n)$, where $\mathrm{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$. If S is continuous on $\mathcal{M}_1(Z)$, then the sequence $(S_n)_{n\in\mathbb{N}}$ is qualitatively robust at P for all $\mathrm{P} \in \mathcal{M}_1(Z)$.*

Proof. See Hampel (1968) and Cuevas (1988). □

A.5 Functional Analysis

In this section, we provide various concepts and results from functional analysis. We strongly advise the reader unfamiliar with this subject to consult additional textbooks such as Rudin (1973), Conway (1990), Riesz and Nagy (1990), Dudley (2002), Lax (2002), and the unfortunately not translated Werner (1995) for a more thorough introduction. Most of the results and concepts we will present can be found in any of these textbooks, and thus we omit providing a reference for them. However, some more advanced results are more difficult to find, and thus we do provide corresponding references.

A.5.1 Essentials on Banach Spaces and Linear Operators

In this subsection, we introduce Banach spaces, bounded linear operators, and related notions. In addition, we present various results for these concepts. Examples of important Banach spaces can be found in Section A.5.5.

Throughout this and the following subsections, $\mathbb{K}$ stands for either $\mathbb{R}$ or $\mathbb{C}$ if not stated otherwise. Let us now begin by recalling the definition of vector and Banach spaces.

Definition A.5.1. *A $\mathbb{K}$-**vector space** is a triple $(E, +, \cdot)$, where E is a nonempty set and $+ : E \times E \to E$ and $\cdot : \mathbb{K} \times E \to E$ are maps satisfying:*

i) $(x + y) + z = x + (y + z)$ *for all* $x, y, z \in E$.
ii) $x + y = y + x$ *for all* $x, y \in E$.
iii) There exists an element $0 \in E$ *with* $x + 0 = x$ *for all* $x \in E$.
iv) For all $x \in X$ *there exists an element* $-x \in E$ *with* $x + (-x) = 0$.
v) $(\alpha\beta) \cdot x = \alpha \cdot (\beta \cdot x)$ *for all* $\alpha, \beta \in \mathbb{K}$, $x \in E$.
vi) $1 \cdot x = x$ *for all* $x \in E$.
vii) $(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x$ *for all* $\alpha, \beta \in \mathbb{K}$ *and* $x \in E$.
viii) $\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y$ *for all* $\alpha \in \mathbb{K}$ *and* $x, y \in E$.

*Moreover, the $\cdot$ denoting the **scalar multiplication** is usually omitted.*

A subset $F \subset E$ is called a **(linear) subspace** of E if for all $x, y \in F$ and $\alpha \in \mathbb{R}$ we have $x + y \in F$ and $\alpha x \in F$. The **linear span** of a subset $A \subset E$ is the smallest linear subspace of E containing A. We write $\operatorname{span} A$ for this space. A family $(x_i)_{i \in I} \subset E$ is called **linearly independent** if for all $n \geq 1$, $i_1 \ldots, i_n \in I$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{K}$ with

$$\alpha_1 x_{i_1} + \cdots + \alpha_n x_{i_n} = 0\,,$$

we have $\alpha_1 = \cdots = \alpha_n = 0$. Moreover, the family is an **algebraic basis** if it is linearly independent and $\operatorname{span}\{x_i : i \in I\} = E$. In this case, the **dimension** is defined by

$$\dim E := |I|\,.$$

Definition A.5.2. *Let E be a $\mathbb{K}$-vector space. A map $\|\cdot\| : E \to [0,\infty)$ is called a* ***quasi-norm*** *if there exists a constant $c \in [1,\infty)$ such that:*

i) $\|x\| = 0$ *if and only if* $x = 0$.
ii) $\|\alpha x\| = |\alpha|\,\|x\|$ *for all* $\alpha \in \mathbb{K}$ *and* $x \in X$.
iii) $\|x + \tilde{x}\| \le c\,(\|x\| + \|\tilde{x}\|)$ *for all* $x, \tilde{x} \in X$.

In this case the pair $(E, \|\cdot\|)$ is called a ***quasi-normed space****, and if $c = 1$ we simply speak of a* ***norm*** *and a* ***normed space****, respectively.*

Given a norm $\|\cdot\|$ on E, the map $(x,\tilde{x}) \mapsto \|x - \tilde{x}\|$ is a metric on E. If this metric is complete, i.e., every Cauchy sequence with respect to $\|\cdot - \cdot\|$ has a limit in E, then $(E, \|\cdot\|)$ is called a **Banach space**. Analogously we define quasi-Banach spaces, though we note that, in general, quasi-norms do not define metrics but only translation-invariant topologies.

A routine exercise shows that the addition and the scalar multiplication are continuous in normed spaces. Furthermore, the norm itself is also continuous by the so-called **inverse triangle inequality** $|\,\|x\| - \|\tilde{x}\|\,| \le \|x - \tilde{x}\|$, which follows from the **triangle inequality** $\|x + \tilde{x}\| \le \|x\| + \|\tilde{x}\|$.

If no confusion can arise, we often write E rather than $(E, \|\cdot\|)$. Furthermore, in order to distinguish between different (quasi)-norms, we often use the notation $\|\cdot\|_E$ for the (quasi)-norm of the (quasi)-normed space E. We further write $B_E := \{x \in E : \|x\| \le 1\}$ for the **closed unit ball** and $\mathring{B}_E := \{x \in E : \|x\| < 1\}$ for the **open unit ball**. A set $A \subset E$ is called **bounded** if $A \subset cB_E$ for some $c \in [0,\infty)$. Moreover, A is called convex if for all $x, \tilde{x} \in A$ and all $\lambda \in [0,1]$ we have $\lambda x + (1-\lambda)\tilde{x} \in A$. Finally, the **convex hull** $\operatorname{co} A$ of an arbitrary $A \subset E$ is the smallest convex set containing A.

Besides Banach spaces, the most fundamental concept of functional analysis is that of continuous linear operators. In order to introduce this concept, let us recall that given two $\mathbb{K}$-vector spaces E and F, a map $S : E \to F$ is called a **(linear) operator** if it satisfies $S(\alpha x) = \alpha Sx$ and $S(x+\tilde{x}) = Sx + S\tilde{x}$ for all $\alpha \in \mathbb{K}$ and $x, \tilde{x} \in E$, where we used the standard notational convention $Sx := S(x)$. By the linearity, it is not hard to see that a linear operator is continuous if and only if it is continuous at 0. Moreover, the following easy yet important characterization holds.

Lemma A.5.3 (Bounded operator). *Let E and F be normed spaces and $S : E \to F$ be a linear operator. Then the following statements are equivalent:*

i) S is continuous.
ii) S is bounded, i.e., the image SB_E of B_E under S is bounded.
iii) There exists a constant $c \in [0,\infty)$ such that for all $x \in E$ we have

$$\|Sx\|_F \le c\|x\|_E \tag{A.18}$$

In this case, $\|S\| := \sup_{x \in B_E} \|Sx\|_F$ is the smallest $c \ge 0$ satisfying (A.18).

We often write $\mathcal{L}(E,F)$ for the space of all bounded operators mapping from E to F. Note that $S \mapsto \|S\|$ defined in Lemma A.5.3 is a norm on $\mathcal{L}(E,F)$, usually called the **operator norm**, and this norm is complete if F is complete. We sometimes use the notation $\|S : E \to F\|$ in order to distinguish between different norms, and we often write $\mathcal{L}(E) := \mathcal{L}(E,E)$. In addition, the **rank** of an operator $S \in \mathcal{S}(E,F)$ is defined to be the dimension of its image, i.e., $\operatorname{rank} S := \dim S(E)$. Finally, the next result usually makes it easier to check whether a linear operator is bounded.

Theorem A.5.4 (Closed graph theorem). *Let E and F be Banach spaces and $S : E \to F$ be a linear operator. Then S is bounded if and only if it has a closed graph, i.e., for all sequences $(x_n) \subset E$ for which there exist an $x \in E$ and a $y \in F$ satisfying $x_n \to x$ and $Sx_n \to y$, we have $Sx = y$.*

If a bounded linear operator $S : E \to F$ satisfies $\|Sx\|_F = \|x\|_E$ for all $x \in E$, then S is called an **isometric embedding** and E is said to be **isometrically embedded** into F. It is obvious that in this case S is injective. If S, in addition, is surjective, then S is called an **isometric isomorphism** and E and F are said to be **isometrically isomorphic**. Finally, S is called a **metric surjection** if $S\mathring{B}_E = \mathring{B}_F$, where $\mathring{B}_E$ and $\mathring{B}_F$ denote the open unit balls of E and F, respectively. Obviously, metric surjections are surjective, and injective metric surjections are isometric isomorphisms.

Given a normed space E, there exists a Banach space $\tilde{E}$ and an isometric embedding $I : E \to \tilde{E}$ such that $I(E)$ is dense in $\tilde{E}$. It turns out that the space $\tilde{E}$ is uniquely (up to isometric isomorpy) determined by this property, and hence $\tilde{E}$ is called the **completion** of E.

A bounded linear operator $S : E \to F$ is said to be **compact** if $\overline{SB_E}$ is a compact subset in F. We write $\mathcal{K}(E,F)$ for the set of all compact linear operators. Note that every bounded linear operator with $\operatorname{rank} S < \infty$ is compact, but the converse is in general not true. Moreover, an isometric embedding is compact if and only if its domain is finite-dimensional. Finally, the following classical theorem, which we only state in a simplified version, holds.

Theorem A.5.5 (Fredholm alternative). *Let E be a Banach space and $S \in \mathcal{L}(E)$ be compact. Then $\mathrm{id}_E + S$ is surjective if and only if it is injective.*

An important special case of linear operators are the bounded linear **functionals**, i.e., the elements of the **dual space** $E' := \mathcal{L}(E,\mathbb{R})$. Note that, by the completeness of $\mathbb{R}$, dual spaces are always Banach spaces. In addition, the **Hahn-Banach theorem** ensures $\{0\} \subsetneq E'$ whenever $\{0\} \subsetneq E$, i.e., on every non-trivial normed space, there exists a non-trivial bounded linear functional. Given a normed space E and elements $x \in E$ and $x' \in E'$, we often write the evaluation of x' at x as a **dual pairing**, i.e., we write $\langle x', x\rangle_{E',E} := x'(x)$. Moreover, we omit the subscript E', E whenever is it clear from the context.

Every Banach space E is isometrically embedded into its **bi-dual space** $E'' := (E')'$ via the canonical embedding $\iota_E : E \to E''$ that is defined by

$x \mapsto (x' \mapsto \langle x', x\rangle_{E',E})$. The space E is called **reflexive** if this embedding is surjective, i.e., if E is isometrically isomorphic to E''. Finite dimensional spaces as well as Hilbert spaces and the spaces $L_p(\mu)$, $p \in (1,\infty)$ (see the following subsections for definitions) are reflexive. Finally, the smallest topology on E' for which the maps $x' \mapsto \langle x', x\rangle_{E',E}$ are continuous on E' for all $x \in E$ is called the **weak* topology**.

Now let E and F be two normed spaces and $S : E \to F$ be a bounded linear operator. Then the **adjoint operator** $S' : F' \to E'$ is defined by

$$\langle S'y', x\rangle_{E',E} := \langle y', Sx\rangle_{F',F} \qquad \text{for all } x \in E,\ y' \in F'. \tag{A.19}$$

The adjoint operator is again a bounded linear operator with $\|S'\| = \|S\|$. Moreover, by Schauder's theorem, S is compact if and only if S' is compact. In addition, S has a dense image, i.e., $\overline{S(E)} = F$, if and only if S' is injective, see, e.g., Conway (1990), p. 168, Proposition 1.8. Finally, the **bi-adjoint operator** $S'' : E'' \to F''$ satisfies

$$S''x = Sx\,, \qquad x \in E, \tag{A.20}$$

where E and F are interpreted as subspaces of E'' and F'' via the canonical embeddings ι_E and ι_F.

The smallest topology on E for which all $x' \in E'$ are continuous is called the **weak topology**. Consequently, a sequence $(x_n) \subset E$ is called **weakly convergent** if there exists an $x \in E$ with $\lim_{n\to\infty}\langle x', x_n\rangle = \langle x', x\rangle$ for all $x' \in E'$. Note that norm-convergent sequences are weakly convergent, but in general the converse is not true. Moreover, the following theorem (see Theorem II.3.28 of Dunford and Schwartz, 1988) states that every bounded subset of a reflexive space is **sequentially weakly compact**. Interestingly, the analogous statement for *norm* convergence is in general false.

Theorem A.5.6. *Let E be a reflexive Banach space and $A \subset E$ be a bounded subset. Then, for all sequences $(x_n) \subset A$, there exists a subsequence $(x_{n_k}) \subset A$ and an $x \in E$ such that $\lim_{k\to\infty}\langle x', x_{n_k}\rangle \to \langle x', x\rangle$ for all $x' \in E'$.*

For a sequence $(x_n) \subset E$ converging weakly to some $x \in X$, we have $|\langle x', x\rangle| = \lim_{n\to\infty}|\langle x', x_n\rangle| \le \liminf_{n\to\infty}\|x_n\|$ for all $x' \in B_{E'}$. In addition, the canonical embedding $\iota_E : E \to E''$ is isometric and hence we have $\|x\| = \sup_{x'\in B_{E'}}|\langle x', x\rangle|$. This leads to

$$\|x\| \le \liminf_{n\to\infty}\|x_n\|\,. \tag{A.21}$$

An **algebra** $\mathcal{A}$ is a vector space equipped with an additional associative and commutative multiplication $\cdot : \mathcal{A}\times\mathcal{A} \to \mathcal{A}$ such that

$$\begin{aligned} x\cdot(y+z) &= x\cdot y + x\cdot z\,,\\ \lambda(x\cdot y) &= (\lambda x)\cdot y\,, \end{aligned}$$

holds for all $x, y, z \in \mathcal{A}$ and $\lambda \in \mathbb{R}$. A classical example of an algebra is the space $C(X)$ of all continuous functions $f : X \to \mathbb{R}$ on the compact metric space (X, d) endowed with the usual supremum norm $\|\,.\,\|_\infty$. The following approximation theorem of Stone-Weierstraß (see, e.g., Corollary 4.3.5 of Pedersen, 1988, or Theorem 18.5 of Brown and Pearcy, 1977) states that certain subalgebras of $C(X)$ are dense.

Theorem A.5.7 (Stone-Weierstraß). *Let (X, d) be a compact metric space and $\mathcal{A} \subset C(X)$ be an algebra. Then $\mathcal{A}$ is dense in $C(X)$ if both $\mathcal{A}$ does not vanish, i.e., for all $x \in X$, there exists an $f \in \mathcal{A}$ with $f(x) \neq 0$, and $\mathcal{A}$ separates points, i.e., for all $x, y \in X$ with $x \neq y$, there exists an $f \in \mathcal{A}$ with $f(x) \neq f(y)$.*

A.5.2 Hilbert Spaces

One of the most important examples of Banach spaces are Hilbert spaces, which in some sense are natural generalizations of the finite-dimensional Euclidian spaces. This subsection reviews some properties of these spaces. Let us begin with the following basic definition.

Definition A.5.8. *A map $\langle \cdot\,, \cdot \rangle : H \times H \to \mathbb{K}$ on a $\mathbb{K}$-vector space H is called an **inner product** if it satisfies:*

i) $\langle x_1 + x_2, x\rangle = \langle x_1, x\rangle + \langle x_2, x\rangle$ *for all* $x_1, x_2, x \in H$.
ii) $\langle \alpha x, x'\rangle = \alpha\langle x, x'\rangle$ *for all* $\alpha \in \mathbb{K}$ *and* $x, x' \in H$.
iii) $\langle x, x'\rangle = \overline{\langle x', x\rangle}$ *for all* $x, x' \in H$.
iv) $\langle x, x\rangle \geq 0$ *for all* $x \in H$, *and* $\langle x, x\rangle = 0$ *if and only if* $x = 0$.

Note that in the case $\mathbb{K} = \mathbb{R}$ condition *iii)* reduces to $\langle x, x'\rangle = \langle x', x\rangle$. In this case, an inner product is thus linear in *both* arguments.

If $\langle \cdot\,, \cdot \rangle$ is an inner product on H, the pair $(H, \langle \cdot\,, \cdot \rangle)$ is called a **pre-Hilbert space**. In order to distinguish between different inner products, we sometimes write $\langle \cdot\,, \cdot \rangle_H$. Moreover, if the inner product is clear from the context, we often call H a pre-Hilbert space.

It is well-known that for inner products the **Cauchy-Schwarz inequality**

$$\left|\langle x, x'\rangle\right|^2 \leq \langle x, x\rangle \, \langle x', x'\rangle\,, \qquad x, x' \in H,$$

holds. This inequality plays a central role in many considerations on inner products. For example, it can be used to show that

$$\|x\|_H := \sqrt{\langle x, x\rangle}\,, \qquad x \in H,$$

defines a norm on H. If this norm is complete, the pair $(H, \langle \cdot\,, \cdot \rangle)$ is called a **Hilbert space**. The following lemma shows that we can retrieve the inner product from this norm. In particular, it shows, that the completion of a pre-Hilbert space is a Hilbert space.

Lemma A.5.9 (Polarization and parallelogram identity). *Let us fix a* $\mathbb{K}$*-pre-Hilbert space* $(H, \langle \cdot\,, \cdot \rangle)$*. Then the following equations are true for all* $x, x' \in H$*:*

$$4\langle x, x' \rangle = \|x + x'\|_H^2 - \|x - x'\|_H^2 \qquad \text{if } \mathbb{K} = \mathbb{R},$$
$$4\langle x, x' \rangle = \|x + x'\|_H^2 - \|x - x'\|_H^2 + i\|x + ix'\|_H^2 - i\|x - ix'\|_H^2 \qquad \text{if } \mathbb{K} = \mathbb{C}.$$

In addition, we have $\|x+x'\|_H^2 + \|x-x'\|_H^2 = 2\|x\|_H^2 + 2\|x'\|_H^2$ *for all* $x, x' \in H$*.*

Given two Hilbert spaces H_1 and H_2, their **sum** $H_1 \oplus H_2$ is the Hilbert space that consists of all pairs $(x_1, x_2) \in H_1 \times H_2$, and whose norm is defined by $\|(x_1, x_2)\|_{H_1 \oplus H_2}^2 := \|x_1\|_{H_1}^2 + \|x_2\|_{H_2}^2$. In order to define the **tensor product** $H_1 \otimes H_2$ of H_1 and H_2 we need to recall that, given a vector space E, a map $f : H_1 \times H_2 \to E$ is called **bilinear** if $f(x_1, \cdot\,) : H_2 \to E$ and $f(\,\cdot\,, x_2) : H_1 \to E$ are linear maps for all $x_1 \in H_1$ and $x_2 \in H_2$. Given two Hilbert spaces H_1 and H_2, or more generally, two vector spaces, one can show that there exists a vector space E and a bilinear map $\pi : H_1 \times H_2 \to E$ such that for all vector spaces F and all bilinear maps $f : H_1 \times H_2 \to F$ there exists exactly one linear map $\varphi : E \to F$ such that $f = \varphi \circ \pi$. We write $x_1 \otimes x_2 := \pi(x_1, x_2)$, $(x_1, x_2) \in H_1 \times H_2$. It turns out that the space E is uniquely determined up to isomorphy and $E = \operatorname{span}\{x_1 \otimes x_2 : (x_1, x_2) \in H_1 \times H_2\}$. This justifies the notation $H_1 \otimes H_2 := E$ for the tensor product of H_1 and H_2. Moreover, if H_1 and H_2 are Hilbert spaces, it is straightforward to show that there exists a unique inner product $\langle \cdot\,, \cdot \rangle_{H_1 \otimes H_2}$ on $H_1 \otimes H_2$ whose corresponding norm $\|\cdot\|_{H_1 \otimes H_2}$ satisfies

$$\|x_1 \otimes x_2\|_{H_1 \otimes H_2} = \|x_1\|_{H_1} \cdot \|x_2\|_{H_2}\,, \qquad (x_1, x_2) \in H_1 \times H_2.$$

In general, this norm is *not* complete, and hence we write $H_1 \hat{\otimes} H_2$ for the corresponding completion of $H_1 \otimes H_2$.

In Euclidian spaces, *orthogonal* elements can be used to find nice algebraic bases, namely orthonormal bases. In the following, we recall how these concepts can be generalized to arbitrary Hilbert spaces.

Definition A.5.10. *Let* H *be a pre-Hilbert space. Then we call* $x, x' \in H$ ***orthogonal*** *if they satisfy* $\langle x, x' \rangle_H = 0$*. In this case, we write* $x \perp x'$*. In addition, we call two subsets* $A, B \subset H$ *orthogonal if* $x \perp x'$ *for all* $x \in A$ *and* $x' \in B$*. Finally, the* ***orthogonal complement*** *of a subset* $A \subset H$ *is defined by*

$$A^\perp := \left\{ x' \in H : x \perp x' \text{ for all } x \in A \right\}.$$

Given a closed subspace H_1 of a Hilbert space H, every element $x \in H$ can be represented by a sum $x = x_1 + x_2$, where $x_1 \in H_1$ and $x_2 \in H_1^\perp$ are *uniquely* determined elements. The bounded linear operator $P_{H_1} : H \to H$ defined by $P_{H_1} x := x_1$ is called the **orthogonal projection** onto H_1. It is a projection, i.e., it satisfies $P_{H_1}^2 = P_{H_1}$, and it is also known that $\|P_{H_1}\| = 1$. Finally, note that from these properties we can deduce the formula $H = H_1 \oplus H_1^\perp$.

Before we give a definition of an orthogonal basis, we need to recall the notion of **unconditionally convergent series.** To this end, let I be an *arbitrary* index set, E be a Banach space, and $x_i \in E$, $i \in I$, be arbitrary elements. Given an $x \in E$, we say that the sum $\sum_{i\in I} x_i$ converges unconditionally to x if $I_0 := \{i : x_i \neq 0\}$ is at most countable and for every enumeration $\{i_1, i_2, \dots\}$ of I_0 the equation $\sum_{i=1}^{\infty} x_i = x$ holds. In this case, the element x is, of course, uniquely determined, and hence the notation $x = \sum_{i\in I} x_i$ is justified.

A family $(e_i)_{i\in I}$ in a Hilbert space H is called an **orthonormal system (ONS)** if for all $i, j \in I$ we have

$$\langle e_i, e_j\rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

If, in addition, the closure of $\operatorname{span}\{e_i : i \in I\}$ equals H, the family $(e_i)_{i\in I}$ is called an **orthonormal basis (ONB)**. One can show that every ONS can be extended to an ONB, i.e., given an ONS $(e_i)_{i\in I}$, there exists a set J with $I \subset J$ and an ONB $(\tilde{e}_i)_{i\in J}$ such that $\tilde{e}_i = e_i$ for all $i \in I$. In particular, every Hilbert space has an ONB. Moreover, it turns out that a Hilbert space H has a *countable* ONB if and only if H is separable.

Given an ONS $(e_i)_{i\in I}$ of a Hilbert space H and an element $x \in H$, the **Fourier coefficients** $(\langle x, e_i\rangle)_{i\in I}$ satisfy **Bessel's inequality**

$$\sum_{i\in I} \big|\langle x, e_i\rangle\big|^2 \leq \|x\|_H^2 \,,$$

where the sum converges unconditionally. Moreover, denoting the orthogonal projection of H onto the closure of $\operatorname{span}\{e_i : i \in I\}$ by P, we have

$$Px = \sum_{i\in I} \langle x, e_i\rangle x \,, \qquad\qquad x \in H,$$

where the sum converges unconditionally. In fact, the Fourier coefficients are the only coefficients $(a_i)_{i\in I}$ for which $Px = \sum_{i\in I} a_i e_i$. The following proposition shows that for ONBs even more can be said.

Lemma A.5.11 (Parseval's identity). *Let H be a Hilbert space and $(e_i)_{i\in I}$ be an ONS. Then $(e_i)_{i\in I}$ is an ONB if and only if for all $x \in H$ we have*

$$\|x\|_H^2 = \sum_{i\in I} \big|\langle x, e_i\rangle\big|^2 .$$

Moreover, in this case, we have

$$x = \sum_{i\in I} \langle x, e_i\rangle\, e_i \,, \qquad\qquad x \in H. \tag{A.22}$$

Finally, an ONS is an ONB if and only if (A.22) holds.

Let us now consider the dual H' of a Hilbert space H. To this end, note that given an $x \in H$ the map $\langle \cdot, x\rangle : H \to \mathbb{K}$ defined by $y \mapsto \langle y, x\rangle$ is a bounded linear functional on H, i.e., an element of H'. The following theorem states that *all* bounded linear functionals on H are of this form. As a consequence of this theorem, we easily see that every Hilbert space is reflexive.

Theorem A.5.12 (Fréchet-Riesz representation). *Let H be a $\mathbb{K}$-Hilbert space and H' its dual. Then the map $\iota : H \to H'$ defined by $\iota x := \langle \cdot, x\rangle$ for all $x \in H$ is isometric and surjective. Moreover, in the case $\mathbb{K} = \mathbb{R}$ it is even an isometric isomorphism.*

Let us now consider bounded linear operators $S : H_1 \to H_2$ acting between $\mathbb{R}$-Hilbert spaces H_1 and H_2. To this end, let $S' : H_2' \to H_1'$ be the adjoint of S and $\iota_i : H_i \to H_i'$ be the isometric isomorphism of the Fréchet-Riesz representation. Then we call the operator $S^* := \iota_1^{-1} S' \iota_2 : H_2 \to H_1$ the **adjoint** of S in the Hilbert space sense, or simply the adjoint if no confusion can arise. It is easy to check that this operator is *characterized* by the relation

$$\langle Sx, y\rangle_{H_2} = \langle x, S^* y\rangle_{H_1}, \qquad x \in H_1,\, y \in H_2.$$

Moreover, $^* : \mathcal{L}(H_1, H_2) \to \mathcal{L}(H_2, H_1)$ is an isometric isomorphism that in addition is self-inverse, i.e., $S^{**} = S$ for all $S \in \mathcal{L}(H_1, H_2)$. Furthermore, we have $(RS)^* = S^* R^*$ for all $S \in \mathcal{L}(H_1, H_2)$ and $R \in \mathcal{L}(H_2, H_3)$. An operator $T \in \mathcal{L}(H)$ is called **self-adjoint** if $T^* = T$, and it is called **positive** if $\langle Tx, x\rangle \geq 0$. Moreover, if the latter inequality is strict for all $x \neq 0$, we say that T is **strictly positive**. Given an $S \in \mathcal{L}(H_1, H_2)$, it is elementary to see that S^*S and SS^* are self-adjoint and positive. Furthermore, if S is injective, then S^*S is strictly positive, and conversely, if S^* is injective, then SS^* is strictly positive. Finally, every orthogonal projection is self-adjoint and positive.

Our next goal is to show that *compact* self-adjoint operators can be diagonalized similarly to symmetric matrices. To this end, recall that a $\lambda \in \mathbb{R}$ is called an **eigenvalue** of $T \in \mathcal{L}(H)$ if there exists an $x \neq 0$ such that $Tx = \lambda x$. Every such x is called an **eigenvector** of T and λ. Note that for every eigenvector there exists a unique eigenvalue, but the converse is obviously not true. For an eigenvalue λ, we thus write $E(\lambda) := \ker(\lambda \operatorname{id}_H - T)$ for the corresponding **eigenspace** and call $\dim E(\lambda)$ the **geometric multiplicity** of λ. It is easy to see that for $\lambda_1 \neq \lambda_2$ the eigenspaces $E(\lambda_1)$ and $E(\lambda_2)$ are orthogonal whenever T is *self-adjoint*. Moreover, for (strictly) positive operators, every eigenvalue is obviously (strictly) positive.

For an operator $S \in \mathcal{L}(H_1, H_2)$, we now consider the positive and self-adjoint operators $T_1 := S^*S : H_1 \to H_1$ and $T_2 := SS^* : H_2 \to H_2$. To this end, we write $E_1(\lambda) := \ker(\lambda \operatorname{id}_H - T_1)$ and $E_2(\lambda) := \ker(\lambda \operatorname{id}_H - T_2)$ for $\lambda \neq 0$. Let us assume that $\lambda \neq 0$ is an eigenvalue of T_1. Then it is easy to check that $S : E_1(\lambda) \to E_2(\lambda)$ is well-defined and injective and consequently λ is an eigenvalue of T_2. In addition, we have $(S^*S)_{|E_1(\lambda)} = \lambda \operatorname{id}_{E_1(\lambda)}$. Moreover,

by symmetry, we can obviously interchange the roles of T_1 and T_2, and hence we see that S^*S and SS^* have exactly the same *non-zero* eigenvalues with the same geometric multiplicities.

In the following, we wish to consider the eigenvalues of compact self-adjoint operators. In order to deal with the cases of finitely many and infinitely many eigenvalues *simultaneously*, we say that a family $(\lambda_i)_{i\in I} \subset \mathbb{R}$ converges to zero if either $I = \{1, 2, \ldots, n\}$ or $I = \mathbb{N}$ and $\lim_{i\to\infty} \lambda_i = 0$.

Theorem A.5.13 (Spectral theorem). *Let H be an $\mathbb{R}$-Hilbert space and $T \in \mathcal{L}(H)$ be compact and self-adjoint. Then there exist an at most countable ONS $(e_i)_{i\in I}$ and a family $(\lambda_i(T))_{i\in I}$ converging to 0 such that $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$ and*

$$Tx = \sum_{i\in I} \lambda_i(T)\langle x, e_i\rangle e_i, \qquad x \in H. \tag{A.23}$$

Moreover, $\{\lambda_i(T) : i \in I\}$ is the set of non-zero eigenvalues of T.

For compact, *positive*, and self-adjoint $T \in \mathcal{L}(H)$, the eigenvalues in (A.23) are non-negative and hence we can define operators $T^r : H \to H$, $r \geq 0$, by

$$T^r x = \sum_{i\in I} \lambda_i^r \langle x, e_i\rangle e_i, \qquad x \in H. \tag{A.24}$$

It is easy to check that $T^{r+s} = T^r T^s$ for $r, s \geq 0$, $T^1 = T$, and $T^2 = TT$. For this reason, T^r is called a **fractional power** of T. Moreover, in the case $r = 1/2$, we have $T^{1/2}T^{1/2} = T$, and thus we call $T^{1/2}$ the **square root** of T and write $\sqrt{T} := T^{1/2}$. Obviously, we have $\lambda_i(T^r) = \lambda_i^r(T)$, $i \in I$, $r \geq 0$.

Let us now consider a compact $S \in \mathcal{L}(H_1, H_2)$. Then $S^*S : H_1 \to H_1$ is compact, positive, and self-adjoint, and hence it enjoys a representation of the form (A.23) with non-negative eigenvalues. We write

$$s_i(S) := \begin{cases} \sqrt{\lambda_i(S^*S)} = \lambda_i(\sqrt{S^*S}) & \text{if } i \in I \\ 0 & \text{if } i \in \mathbb{N}\setminus I \end{cases} \tag{A.25}$$

for the **singular numbers** of S. Recall that S^*S and SS^* have exactly the same non-zero eigenvalues with the same geometric multiplicities and hence we find $s_i(S^*) = s_i(S)$ for all $i \geq 1$. Finally, if $T \in \mathcal{L}(H)$ is compact, positive, and self-adjoint, we have $s_i(T) = \sqrt{\lambda_i(T^*T)} = \sqrt{\lambda_i(T^2)} = \lambda_i(T)$, $i \in I$.

So far, we have seen that the singular numbers of compact operators converge to zero. Let us finally refine this analysis by considering operators whose singular numbers converge with a certain speed. To this end, let $T \in \mathcal{L}(H)$ be a compact operator. We say that T is **nuclear** or of **trace class** if

$$\|T\|_{\mathsf{nuc}} := \sum_{i=1}^{\infty} s_i(T) < \infty.$$

Using the representations above, it is then easy to check that a *self-adjoint* and compact $T \in \mathcal{L}(H)$ is nuclear if and only if $\sum_{i\in I} |\lambda_i(T)| < \infty$, and in this case the latter sum equals $\|T\|_{\mathsf{nuc}}$.

It turns out that besides summable singular numbers, square summable singular numbers are of particular interest. To this end, we say that an operator $S \in \mathcal{L}(H_1, H_2)$ is **Hilbert-Schmidt** if

$$\|S\|_{\mathsf{HS}} := \left(\sum_{j\in J} \|Se_j\|_{H_2}^2\right)^{1/2} < \infty\,, \tag{A.26}$$

where $(e_j)_{j\in J}$ is an *arbitrary* ONB of H_1. One can show that Hilbert-Schmidt operators are compact and that the **Hilbert-Schmidt norm** $\|S\|_{\mathsf{HS}}$ is *independent* of the choice of the ONB. As a matter of fact, we have

$$\|S\|_{\mathsf{HS}} = \left(\sum_{i=1}^{\infty} s_i^2(S)\right)^{1/2}, \tag{A.27}$$

and using $s_i^2(S) = \lambda_i(S^*S) = s_i(S^*S)$ we hence see $\|S\|_{\mathsf{HS}}^2 = \|S^*S\|_{\mathsf{nuc}}$. Consequently, S is a Hilbert-Schmidt operator if and only if S^*S is nuclear. Moreover, one can show in a similar fashion that $\|S\|_{\mathsf{HS}} = \|S^*\|_{\mathsf{HS}}$. If $S \in \mathcal{L}(H_1, H_2)$ has finite rank, i.e., $\operatorname{rank} S < \infty$, then it is automatically Hilbert-Schmidt, and for orthogonal projections $P : H \to H$ with $\operatorname{rank} P < \infty$, the definition of the Hilbert-Schmidt norm immediately yields $\|P\|_{\mathsf{HS}}^2 = \operatorname{rank} P$. Finally, given a Hilbert space H, the set $\mathsf{HS}(H)$ of all Hilbert-Schmidt operators acting on H becomes a Hilbert space itself if we equip it with the Hilbert-Schmidt norm $\|\cdot\|_{\mathsf{HS}}$. In this case, the corresponding inner product is given by

$$\langle T_1, T_2\rangle_{\mathsf{HS}(H)} = \sum_{j\in J} \langle T_1 e_j, T_2 e_j\rangle_H\,, \qquad T_1, T_2 \in \mathsf{HS}(H)\,, \tag{A.28}$$

where $(e_j)_{j\in J}$ is an *arbitrary* ONB of H. Finally, $\mathsf{HS}(H)$ is separable if and only if H is separable.

Let us now consider another interesting property of the singular numbers. To this end, let $S \in \mathcal{L}(E, F)$ be an operator acting between arbitrary Banach spaces E and F. For $i \geq 1$, its i-th **approximation number** is defined by

$$a_i(S) := \inf\bigl\{\|S - A\| : A \in \mathcal{L}(E, F) \text{ with } \operatorname{rank} A < i\bigr\}\,. \tag{A.29}$$

Obviously, $(a_i(S))_{i\geq 1}$ is decreasing, and if $\operatorname{rank} S < \infty$, we also have $a_i(S) = 0$ for all $i > \operatorname{rank} S$. Moreover, if F is a Hilbert space H, then it suffices (see, e.g., Proposition 2.4.5 of Carl and Stephani, 1990) to consider operators of the form $A = PS$ in (A.29), where $P \in \mathcal{L}(H)$ is an orthogonal projection with $\operatorname{rank} P < i$. Moreover, by diagonalization (see, e.g., Section 2.11 of Pietsch, 1987), one can show that $s_i(S) = a_i(S)$ for all compact $S \in \mathcal{L}(H_1, H_2)$ acting between Hilbert spaces and all $i \geq 1$. Consequently, we have $\lambda_i(T) = a_i(T)$ for all compact, self-adjoint, and positive $T \in \mathcal{L}(H)$ and all $i \in I$.

A.5.3 The Calculus in Normed Spaces

In one-dimensional calculus, differentiability of a map f at a point means that f can be well-approximated in terms of a linear map. This idea is used in the following general definition.

Definition A.5.14. *Let E and F be normed spaces, $U \subset E$ and $V \subset F$ be open sets, and $G : U \to V$ be a map. We say that G is **Gâteaux differentiable** at $x_0 \in U$ if there exists a bounded linear operator $A : E \to F$ such that*

$$\lim_{\substack{t \to 0 \\ t \neq 0}} \frac{\|G(x_0 + tx) - G(x_0) - tAx\|_F}{t} = 0\,, \qquad x \in E.$$

*In this case, A is called the **derivative** of G at x_0, and since A is uniquely determined, we write*

$$G'(x_0) := \frac{\partial G}{\partial E}(x_0) := A\,.$$

*Moreover, we say G that **Fréchet differentiable** at x_0 if A actually satisfies*

$$\lim_{\substack{x \to 0 \\ x \neq 0}} \frac{\|G(x_0 + x) - G(x_0) - Ax\|_F}{\|x\|_E} = 0\,.$$

Furthermore, we say that G is (Gâteaux, Fréchet) differentiable if it is (Gâteaux, Fréchet) differentiable at every $x_0 \in U$.

*Finally, G is said to be **continuously differentiable** if it is Fréchet differentiable and the derivative $G' : U \to \mathcal{L}(E, F)$ is continuous.*

It is obvious that Fréchet differentiability implies Gâteaux differentiability, but in general the converse is not true. However, if $E = F = \mathbb{R}$, both notions actually fall together and coincide with the "standard" notion of differentiability. It is also obvious that every bounded linear operator $S : E \to F$ is Fréchet differentiable with derivative

$$S'(x_0) = S\,, \qquad x_0 \in E.$$

Furthermore, if a function $G : E \to \mathbb{R}$ has a local extremum at $x_0 \in E$ and is also Gâteaux differentiable at x_0, then we have $G'(x_0) = 0$. The latter fact can be shown as in the one-dimensional case.

The following lemma presents the calculus for the concepts of differentiability. For a proof, we refer to Chapter 4 of Zeidler (1986).

Lemma A.5.15 (Calculus in normed spaces). *Let $E, F, \tilde{F}$ be normed spaces, $U \subset E$, $V \subset F$, and $\tilde{V} \subset \tilde{F}$ be open sets, $G_1, G_2 : U \to V$, $\tilde{G} : V \to \tilde{V}$ be maps, and $\alpha_1, \alpha_2 \in \mathbb{R}$. Then the following statements are true:*

i) If G_1 and G_2 are (Gâteaux, Fréchet) differentiable at some $x_0 \in U$, then $\alpha_1 G_1 + \alpha_2 G_2$ is (Gâteaux, Fréchet) differentiable at x_0 and we have

$$(\alpha_1 G_1 + \alpha_2 G_2)'(x_0) = \alpha_1 G_1'(x_0) + \alpha_2 G_2'(x_0).$$

ii) If G_1 is (Gâteaux, Fréchet) differentiable at some $x_0 \in U$ and $\tilde{G}$ is Fréchet differentiable at $G_1(x_0)$, then $\tilde{G} \circ G_1$ is (Gâteaux, Fréchet) differentiable at $x_0 \in E$ and we have

$$\big(\tilde{G} \circ G_1\big)'(x_0) = \tilde{G}'\big(G_1(x_0)\big) \circ G_1'(x_0)\,.$$

The following theorem provides a useful tool for establishing Fréchet differentiability of functions defined on product spaces. Its proof can be found, for example, in Akerkar (1999, Theorem 2.6 on p. 37).

Theorem A.5.16 (Partial Fréchet differentiability). *Let E_1, E_2, and F be Banach spaces, $U_1 \subset E_1$ and $U_2 \subset E_2$ be open subsets, and $G : U_1 \times U_2 \to F$ be a continuous map. Then G is continuously differentiable if and only if G is partially Fréchet differentiable and the partial derivatives $\frac{\partial G}{\partial E_1}$ and $\frac{\partial G}{\partial E_2}$ are continuous. In this case, the derivative of G at $(x_1, x_2) \in U_1 \times U_2$ is given by*

$$G'(x_1,x_2)(y_1,y_2) = \frac{\partial G}{\partial E_1}(x_1,x_2)y_1 + \frac{\partial G}{\partial E_2}(x_1,x_2)y_2\,, \qquad (y_1,y_2) \in E_1 \times E_2\,.$$

Finally, we need the following (simplified) version of the implicit function theorem, whose proof can be found in Chapter 4 of Zeidler (1986) and in Chapter 4 of Akerkar (1999).

Theorem A.5.17 (Implicit function theorem). *Let E and F be Banach spaces and $G : E \times F \to F$ be a continuously differentiable map. Suppose that we have a pair $(x_0, y_0) \in E \times F$ such that $G(x_0, y_0) = 0$ and $\frac{\partial G}{\partial F}(x_0, y_0)$ is an invertible operator. Then there exists a $\delta > 0$ and a continuously differentiable map $f : x_0 + \delta B_E \to y_0 + \delta B_F$ such that for all $x \in x_0 + \delta B_E$, $y \in y_0 + \delta B_F$, we have*

$$G(x,y) = 0 \qquad \text{if and only if} \qquad y = f(x)\,.$$

Moreover, the derivative of f is given by

$$f'(x) = -\left(\frac{\partial G}{\partial F}\big(x, f(x)\big)\right)^{-1} \frac{\partial G}{\partial E}\big(x, f(x)\big)\,.$$

A.5.4 Banach Space Valued Integration

In this subsection, we briefly present the very basics of Banach space valued integration. For a thorough treatment and proofs of the results mentioned below, we refer to Chapter II of the book by Diestel and Uhl (1977) and to Chapter 1 of the book by Dinculeanu (2000).

Let us begin by introducing a suitable concept of measurability. To this end, let $(\Omega, \mathcal{A})$ be a measurable space and E be a Banach space. A function $f : \Omega \to E$ is called a **measurable step function** if there exists $x_1, \ldots, x_n \in E$ and $A_1, \ldots, A_n \in \mathcal{A}$ with

$$f = \sum_{i=1}^{n} \mathbf{1}_{A_i} x_i \,. \tag{A.30}$$

Moreover, we say that $f : \Omega \to E$ is an **E-valued measurable function** if there exists a sequence (f_n) of measurable step functions $f_n : \Omega \to E$ such that $\lim_{n\to\infty} \|f(\omega) - f_n(\omega)\|_E = 0$ holds for all $\omega \in \Omega$. The following lemma (see Theorem 8 on p. 5 in the book by Dinculeanu, 2000) relates this measurability notion to standard measurability.

Lemma A.5.18. *Let E be a Banach space, $(\Omega, \mathcal{A})$ be a measurable space, and $f : \Omega \to E$. Then the following statements are equivalent:*

i) f is an E-valued measurable function.
ii) $f(\Omega)$ is separable and $f^{-1}(B)$ is measurable for all Borel sets $B \subset E$.

In almost all situations, we deal with separable Banach spaces E. It is thus good to remember that in such cases the preceding lemma shows that both notions of measurability coincide.

It is not hard to see that the E-valued measurability is preserved by standard operations such as addition, multiplication, and limits. Moreover, if $f : \Omega \to E$ is an E-valued measurable function and $S : E \to F$ is a bounded linear operator, then it is easy to see that $S \circ f : \Omega \to F$ is an F-valued measurable function. In particular, the functions $\langle x', f\rangle : \Omega \to \mathbb{R}$ are measurable for all $x' \in E'$. The following theorem, which can be found, for example, on p. 9 of the book by Dinculeanu (2000) and on p. 42 of the book by Diestel and Uhl (1977), gives the converse implication, provided that E is separable.

Theorem A.5.19 (Petti's measurability theorem). *Let E be a Banach space and $(\Omega, \mathcal{A})$ be a measurable space. Then $f : \Omega \to E$ is an E-valued measurable function if and only if the following two conditions are satisfied:*

i) f is weakly measurable, i.e., $\langle x', f\rangle : \Omega \to \mathbb{R}$ is measurable for all $x' \in E'$.
ii) $f(\Omega)$ is a separable subset of E.

In order to illustrate the utility of the preceding theorem, let us assume that (Z, d) is a *separable* metric space (equipped with the Borel σ-algebra) and that $f : Z \to E$ is a continuous function. Then f is obviously weakly measurable, and by the separability of Z and the continuity of f, the image $f(Z)$ is separable. Consequently, f is an E-valued measurable function.

Given a measurable step function $f : \Omega \to E$ with representation (A.30) and a σ-finite measure μ on Ω, we define the **integral** of f by

$$\int_\Omega f \, d\mu := \sum_{i=1}^{n} \mu(A_i)\, x_i \,.$$

It is a simple exercise to check that this definition is independent of the representation (A.30), and another such exercise shows that the following definition is correct.

Definition A.5.20. *Let E be a Banach space and $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space. An E-valued measurable function $f : \Omega \to E$ is called* ***Bochner μ-integrable*** *if there exists a sequence (f_n) of E-valued measurable step functions $f_n : \Omega \to E$ such that*

$$\lim_{n\to\infty} \int_\Omega \|f_n - f\|_E \, d\mu = 0 \,.$$

In this case, the limit

$$\int_\Omega f \, d\mu := \lim_{n\to\infty} \int_\Omega f_n \, d\mu$$

exists and is called the ***Bochner integral*** *of f. Finally, if μ is a probability measure, we sometimes write $\mathbb{E}_\mu f$ for this integral.*

It can be easily shown that the Bochner integral is linear. Moreover, an E-valued measurable function $f : \Omega \to E$ is Bochner μ-integrable if and only if $\omega \mapsto \|f(\omega)\|_E$ is μ-integrable, and in this case we also have

$$\Big\| \int_\Omega f \, d\mu \Big\|_E \;\le\; \int_\Omega \|f\|_E \, d\mu \,. \tag{A.31}$$

In particular, if $S : E \to F$ is a bounded linear operator and $f : \Omega \to E$ is Bochner μ-integrable, then $S \circ f : \Omega \to F$ is Bochner μ-integrable. Moreover, in this case, the integral commutes with S, i.e., we have

$$S\Big(\int_\Omega f \, d\mu\Big) \;=\; \int_\Omega Sf \, d\mu \,. \tag{A.32}$$

In addition, a straightforward application of the scalar dominated convergence theorem yields (see, e.g., Theorem 3 on p. 45 in Diestel and Uhl, 1977).

Theorem A.5.21 (Dominated convergence theorem). *Let E be a Banach space, $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space, and (f_n) be a sequence of Bochner μ-integrable functions $f_n : \Omega \to E$. If $\lim_{n\to\infty} f_n(\omega) = f(\omega)$ for μ-almost all $\omega \in \Omega$ and if there exists a μ-integrable function $g : \Omega \to \mathbb{R}$ with $\|f_n\| \le g$, then f is Bochner μ-integrable and*

$$\lim_{n\to\infty} \int_\Omega f_n \, d\mu = \int_\Omega f \, d\mu \,.$$

Finally, the next result (see, e.g., Corollary 8 on p. 48 in Diestel and Uhl, 1977) shows that the Bochner integral is in some sense a convex combination.

Theorem A.5.22. *Let $(\Omega, \mathcal{A}, \mu)$ be a finite measure space, E be a Banach space, and $f : \Omega \to E$ be Bochner μ-integrable. Then, for each $A \in \mathcal{A}$ with $\mu(A) > 0$, we have*

$$\frac{1}{\mu(A)} \int_A f d\mu \;\in\; \overline{\mathrm{co}}(f(A)).$$

A.5.5 Some Important Banach Spaces

In this subsection, we will briefly introduce some important Banach spaces frequently used in this book.

Let us begin by defining $\|f\|_\infty := \sup_{x\in X} |f(x)|$ for an arbitrary function $f : X \to \mathbb{R}$. Then the set $B(X) := \{f : X \to \mathbb{R} \,|\, \|f\|_\infty < \infty\}$ equipped with $\|\cdot\|_\infty$ is a Banach space.

Given a measurable space $(X, \mathcal{A})$, we write $\mathcal{L}_0(X)$ for the set of all real-valued measurable functions on X, i.e., $\mathcal{L}_0(X) := \{f : X \to \mathbb{R} \,|\, f \text{ measurable}\}$. Obviously, this set is a vector space when considering the usual addition and multiplication. Moreover, the subspace $\mathcal{L}_\infty(X) := \{f \in \mathcal{L}_0(X) : \|f\|_\infty < \infty\}$ of all bounded measurable functions on X becomes a Banach space when equipped with the norm $\|\cdot\|_\infty$. Let us now assume that we have a measure μ on $\mathcal{A}$. For $p \in (0, \infty)$ and $f \in \mathcal{L}_0(X)$, we then write

$$\|f\|_{\mathcal{L}_p(\mu)} := \left(\int_X |f|^p d\mu\right)^{1/p}.$$

To treat the case $p = \infty$, we say that $N \in \mathcal{A}$ is a local μ-zero set, if $\mu(N \cap A) = 0$ for all $A \in \mathcal{A}$ with $\mu(A) < \infty$. Now we define

$$\begin{aligned}\|f\|_{\mathcal{L}_\infty(\mu)} &:= \operatorname*{ess\text{-}sup}_{x\in X} |f(x)| \\ &:= \inf\bigl\{a \geq 0 : \{x \in X : |f(x)| > a\} \text{ is a local } \mu\text{-zero set}\bigr\}.\end{aligned}$$

In both cases, the sets[2]

$$\mathcal{L}_p(\mu) := \{f \in \mathcal{L}_0(X) : \|f\|_{\mathcal{L}_p(\mu)} < \infty\}$$

are vector spaces of functions, and for $p \in [1, \infty]$ the map $\|\cdot\|_{\mathcal{L}_p(\mu)}$ enjoys all properties of a norm on $\mathcal{L}_p(\mu)$ besides definiteness, i.e., in general $\|f\|_{\mathcal{L}_p(\mu)} = 0$ does *not* imply $f = 0$. To address the latter, we say that $f, f' \in \mathcal{L}_p(\mu)$ are equivalent, written as $f \sim f'$, if $\|f - f'\|_{\mathcal{L}_p(\mu)} = 0$. In other words, $f \sim f'$ if and only if $f(x) = f'(x)$ for μ-almost all $x \in X$. Now consider the set of equivalence classes

$$L_p(\mu) := \bigl\{[f]_\sim : f \in \mathcal{L}_p(\mu)\bigr\}, \tag{A.33}$$

where $[f]_\sim := \{f' \in \mathcal{L}_p(\mu) : f \sim f'\}$ denotes the equivalence class of f. It is straightforward to show that $L_p(\mu)$ becomes a vector space with addition and scalar multiplication defined by $[f]_\sim + [g]_\sim := [f + g]_\sim$ and $\alpha[f]_\sim := [\alpha f]_\sim$, $f, g \in \mathcal{L}_p(\mu)$, $\alpha \in \mathbb{R}$. Moreover, $\|[f]_\sim\|_{L_p(\mu)} := \|f\|_{\mathcal{L}_p(\mu)}$ defines a complete norm on $L_p(\mu)$ for $p \in [1, \infty]$, i.e., $(L_p(\mu), \|\cdot\|_{L_p(\mu)})$ is a Banach

[2] Note that in Section A.3.1 we allowed $\mathcal{L}_1(\mu)$ to contain $[-\infty, \infty]$-valued functions, whereas here we only consider $\mathbb{R}$-valued functions. However, it is easy to show that a μ-integrable function is μ-almost surely finite, and hence this notational conflict can be ignored in essentially all situations.

space. Furthermore, for $p \in (0,1)$, $\|\cdot\|_{L_p(\mu)}$ is a complete quasi-norm, which, however, is almost never a norm. In addition, $L_p(\mu)$ is a Hilbert space if and only if $p = 2$. It is common practice to identify the **Lebesgue spaces** $L_p(\mu)$ and $\mathcal{L}_p(\mu)$, and in many situations such an identification can actually be made rigorous. However, one should always be aware that strictly speaking $L_p(\mu)$ does *not* consist of functions. In particular, evaluations $f(x)$ of elements $f \in L_p(\mu)$ are usually not defined, whereas for $f \in \mathcal{L}_p(\mu)$ such evaluations always make perfect sense.

We often use the standard abbreviations $\|\cdot\|_p := \|\cdot\|_{\mathcal{L}_p(\mu)}$ and $\|\cdot\|_p := \|\cdot\|_{L_p(\mu)}$, respectively. In addition, we usually write $\mathcal{L}_p(X) := \mathcal{L}_p(\mu)$ and $L_p(X) := L_p(\mu)$ if $X \subset \mathbb{R}^d$ and μ is the Lebesgue measure on X. Furthermore, if μ is the counting measure on some set X, we write $\ell_p(X)$ instead of $\mathcal{L}_p(\mu)$. Note that for counting measures the equivalence classes $[\cdot]_\sim$ are singletons, and hence it does not make sense to distinguish between $\mathcal{L}_p(\mu)$ and $L_p(\mu)$. Moreover, for the particular cases $X = \mathbb{N}$ and $X = \{1, \ldots, d\}$, we write ℓ_p and ℓ_p^d, respectively. In particular, ℓ_2^d denotes the d-dimensional Euclidean space.

Given a $p \in [1,\infty]$, there exists a unique $p' \in [1,\infty]$, called the **conjugate exponent**, such that $1/p + 1/p' = 1$. For $f \in \mathcal{L}_p(\mu)$ and $g \in \mathcal{L}_{p'}(\mu)$, **Hölder's inequality** then states $fg \in \mathcal{L}_1(\mu)$ and $\|fg\|_1 \leq \|f\|_p \, \|g\|_{p'}$. Moreover, one can show that the map $\iota : L_{p'}(\mu) \to (L_p(\mu))'$ that maps every $g \in L_{p'}(\mu)$ to the bounded linear functional $\iota g : L_p(\mu) \to \mathbb{R}$ defined by

$$\iota g(f) := \int_X f g \, d\mu \,, \qquad f \in L_p(\mu) \,, \tag{A.34}$$

is isometric, and for $p \in [1,\infty)$ it is even an isometric isomorphism. Informally speaking, the latter justifies the identification of $(L_p(\mu))'$ with $L_{p'}(\mu)$ for $p \in [1,\infty)$. Let us end this discussion with two other inequalities. For the first, we refer to Proposition 6.6.20 by Pedersen (1988), and for the second, we refer to Theorem 2.38 by Adams and Fournier (2003).

Theorem A.5.23 (Young's inequality). *Let $p \in [1,\infty]$, $f \in \mathcal{L}_1(\mathbb{R}^d)$, and $g \in \mathcal{L}_p(\mathbb{R}^d)$. Then*

$$f * g\,(x) := \int_{\mathbb{R}^d} f(x-y) g(y) dy$$

*exists for Lebesgue-almost all $x \in \mathbb{R}^d$. Moreover, the **convolution** $f * g$ of f and g is contained in $L_p(\mathbb{R}^d)$ and satisfies*

$$\|f * g\|_p \;\leq\; \|f\|_1 \, \|g\|_p \,.$$

Lemma A.5.24 (Clarkson's inequality). *Let $(X, \mathcal{A}, \mu)$ be a measure space and $p \in [2,\infty)$. Then for all $f, g \in \mathcal{L}_p(\mu)$, we have*

$$\left\| \frac{f+g}{2} \right\|_p^p + \left\| \frac{f-g}{2} \right\|_p^p \leq \frac{\|f\|_p^p + \|g\|_p^p}{2} \,.$$

Let us now introduce a few more spaces defined by measures. To this end, let $(X, \mathcal{A}, \mu)$ be a σ-finite measure space. For $f \in \mathcal{L}_0(X)$ and $0 < p < \infty$, we write

$$\|f\|_{\mathcal{L}_{p,\infty}(\mu)} := \inf\Big\{c > 0 : \mu\big(\{x \in X : |f(x)| \geq t\}\big) \;\leq\; (c/t)^p \text{ for all } t > 0\Big\}$$

and $\mathcal{L}_{p,\infty}(\mu) := \{f \in \mathcal{L}_0(X) : \|f\|_{\mathcal{L}_{p,\infty}(\mu)} < \infty\}$. One can show that $\|\cdot\|_{\mathcal{L}_{p,\infty}(\mu)}$ is a quasi-norm on $\mathcal{L}_{p,\infty}(\mu)$. Moreover, a simple modification of Markov's inequality yields $\|f\|_{\mathcal{L}_{p,\infty}(\mu)} \leq \|f\|_{\mathcal{L}_p(\mu)}$ for all $f \in \mathcal{L}_p(\mu)$. Conversely, one can show that if μ is a *finite* measure, then for all sufficiently small $\varepsilon > 0$ there exists a constant $c_{p,\varepsilon} \in [0, \infty)$ such that $\|f\|_{\mathcal{L}_{p-\varepsilon}(\mu)} \leq c_{p,\varepsilon}\|f\|_{\mathcal{L}_{p,\infty}(\mu)}$ for all $f \in \mathcal{L}_{p,\infty}(\mu)$. Finally, note that $\mathcal{L}_{p,\infty}(\mu)$ is a so-called **Lorentz space**, though usually these spaces are introduced in a slightly different (see, e.g., Bennett and Sharpley, 1988) yet equivalent way. We decided to use the definition above since it provides an obvious way to quantify tail bounds.

Let us finally introduce the space $\mathcal{L}_p(\mu)$ for $p = 0$, where for simplicity we additionally assume that μ is a *finite* measure. In this case, we simply define $\mathcal{L}_0(\mu) := \mathcal{L}_0(X)$, i.e., $\mathcal{L}_0(\mu)$ consists of all measurable functions on X. Furthermore, we define a *metric* on $\mathcal{L}_0(\mu)$ by

$$d_\mu(f, g) := \int_X \min\{1, |f - g|\} d\mu\,.$$

It is easy to show that this metric defines the **convergence in measure** μ. Moreover, addition and scalar multiplication in $\mathcal{L}_0(\mu)$ are continuous with respect to d_μ, and d_μ is translation-invariant in the sense of $d_\mu(f+h, g+h) = d_\mu(f, g)$, $f, g, h \in \mathcal{L}_0(\mu)$. This motivates the intuitive notation $\|f\|_{\mathcal{L}_0(\mu)} := d_\mu(f, 0)$, which implies $\|f - g\|_{\mathcal{L}_0(\mu)} = d_\mu(f - g, 0) = d_\mu(f, g)$.

Let us now introduce some **spaces of continuous functions**. To this end, we fix a topological Hausdorff space (X, τ) and write

$$C(X) := \big\{f : X \to \mathbb{R} \,|\, f \text{ is continuous}\big\}\,.$$

Furthermore, we write $C_b(X) := \{f \in C(X) : f \text{ is bounded}\}$ for the set of bounded continuous functions and

$$C_c(X) := \big\{f \in C(X) : \operatorname{supp} f \text{ is compact}\big\}$$

for the set of continuous functions with compact support. For later use, we note that the pair $(C_b(X), \|\cdot\|_\infty)$ is a Banach space, and if X is compact we further have $C(X) = C_b(X) = C_c(X)$. The next result, which is Theorem 29.14 in the book by Bauer (2001), shows that in many cases $C_c(X)$ is dense in the Lebesgue spaces.

Theorem A.5.25. *If X is a locally compact space, then $C_c(X)$ is dense in $\mathcal{L}_p(\mu)$ for all regular Borel measures μ on X and all $p \in [1, \infty)$.*

To introduce **spaces of differentiable functions**, we use for a given multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ the standard notation $|\alpha| := \alpha_1 + \cdots + \alpha_d$. Moreover, we write $\partial^\alpha := \partial_1^{\alpha_1} \ldots \partial_d^{\alpha_d}$, where $\partial_i^{\alpha_i}$ denotes the α_i-th partial derivative operator in direction i, i.e., $\partial_i^{\alpha_i} f := \partial^{\alpha_i} f / \partial^{\alpha_i} x_i$ and arbitrary reorderings of the partial derivative operators are allowed. Note that for the multi-index $\alpha = 0$ we have $\partial^\alpha f = f$. Now, for $m \in \mathbb{N}_0$ and a non-empty *open* $X \subset \mathbb{R}^d$, the set

$$C^m(X) := \left\{ f : X \to \mathbb{R} \,\middle|\, \partial^\alpha f \in C(X) \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \leq m \right\}$$

of m-times continuously differentiable functions is obviously a vector space. Furthermore, its subspace

$$C_b^m(X) := \left\{ f \in C^m(X) : \|\partial^\alpha f\|_\infty < \infty \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \leq m \right\}$$

equipped with $\|f\|_{C_b^m(X)} := \max_{|\alpha| \leq m} \|\partial^\alpha f\|_\infty$ becomes a Banach space. Note that $C_b^0(X) = C_b(X)$ and $\|\cdot\|_{C_b^0(X)} = \|\cdot\|_\infty$. We further write $C^\infty(X) := \bigcap_{m \geq 0} C^m(X)$. Moreover, $C_b^\infty(X)$ denotes the subspace of $\bigcap_{m \geq 0} C_b^m(X)$ that consists of the functions f satisfying $\|f\|_{C_b^\infty(X)} := \max_{|\alpha| < \infty} \|\partial^\alpha f\|_\infty < \infty$. Note that $\|\cdot\|_{C_b^\infty(X)}$ is a complete norm on $C_b^\infty(X)$. Let us now assume that $X \subset \mathbb{R}^d$ is a *bounded* open subset. We write

$$C^m(\overline{X}) := \left\{ f \in C^m(X) : \forall \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \leq m \, \exists g \in C(\overline{X}) \text{ s.t. } g_{|X} = \partial^\alpha f \right\}$$

and equip this subspace of $C_b^m(X)$ with $\|\cdot\|_{C_b^m(X)}$. It is not hard to see that $C^m(\overline{X})$ is the set of m-times continuously differentiable functions whose partial derivatives up to the order m are bounded and uniformly continuous. Moreover, for $f \in C^m(\overline{X})$, there exists a *unique* function $\hat{f} \in C(\overline{X})$ such that $\hat{f}_{|X} = f$, and it is easy to check that $\|f\|_\infty = \|\hat{f}\|_\infty$.

Our next goal is to introduce a generalization of the (partial) derivative. To this end, we define the set of **test functions** over an open non-empty subset $X \subset \mathbb{R}^d$ by

$$\mathcal{D}(X) := \left\{ \varphi \in C^\infty(X) : \operatorname{supp} \varphi \text{ is compact} \right\}.$$

Let us now fix a multi-index $\alpha \in \mathbb{N}_0^d$ and an $f \in L_2(X)$. We say that f is **weakly α-differentiable** if there exists a $g \in L_2(X)$ such that

$$\langle g, \varphi \rangle_{L_2(X)} = (-1)^{|\alpha|} \langle f, \partial^\alpha \varphi \rangle_{L_2(X)}$$

for all $\varphi \in \mathcal{D}(X)$. It can be shown that in this case g is uniquely determined, which motivates the notation $\partial^{(\alpha)} f := g$. Moreover, we have $\partial^{(\alpha)} f = \partial^\alpha f$ for all $f \in C^m(\overline{X})$, where, for an open interval X, the latter is an immediate consequence of the integration-by-parts formula. For $m \geq 0$, we now write

$$W^m(X) := \left\{ f \in L_2(X) : \partial^{(\alpha)} f \text{ exists for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha| \leq m \right\}$$

for the space of all m-times weakly differentiable functions. Moreover, for $f \in W^m(X)$, we define the so-called **Sobolev norm** by

$$\|f\|_{W^m(X)} := \left(\sum_{|\alpha| \le m} \|\partial^{(\alpha)} f\|^2_{L_2(X)} \right)^{1/2} .$$

The pair $(W^m(X), \|\cdot\|_{W^m(X)})$ is called **Sobolev space** of order m. It is well-known that Sobolev spaces are Hilbert spaces and that $C^m(X) \cap W^m(X)$ is a dense subspace of $W^m(X)$, where the latter can be found, for example, on p. 67 of Adams and Fournier (2003) and p. 54 of Ziemer (1989). To discuss some further properties of these spaces, we assume for simplicity that X is an open Euclidean ball in $\mathbb{R}^d$. In this case, $C^m(\overline{X}) \cap W^m(X)$ is a dense subspace of $W^m(X)$, see Theorem 3.22 of Adams and Fournier (2003). Moreover, **Sobolev's embedding theorem**, see Theorem 4.12 of Adams and Fournier (2003) for this and related results, states that for $j \ge 0$ and $m > d/2$ there exists a constant $c > 0$ such that for every $f \in W^{m+j}(X)$ there exists a $g \in C^j_b(X)$ such that $f = g$ almost everywhere and

$$\|g\|_{C^j_b(X)} \le c \, \|f\|_{W^{m+j}(X)} .$$

In other words, $W^{m+j}(X)$ can be identified with a subspace of $C^j_b(X)$ whenever $m > d/2$. In particular, $W^m(X)$ "consists" of continuous functions for such m. Given an $f \in W^m(\mathbb{R}^d)$, one can easily show that the restriction $f_{|X}$ to X satisfies $f_{|X} \in W^m(X)$ with $\|f_{|X}\|_{W^m(X)} \le \|f\|_{W^m(\mathbb{R}^d)}$. Interestingly, for Euclidean balls X, a form of inverse inequality also holds. More precisely, the **Calderón-Stein extension theorem** shows that there exists a linear operator $\tilde{} : W^0(X) \to W^0(\mathbb{R}^d)$ such that

$$\tilde{f} = f \quad \text{almost everywhere on } X$$

for all $f \in W^0(X)$ and $\tilde{f} \in W^m(\mathbb{R}^d)$ with

$$\|\tilde{f}\|_{W^m(\mathbb{R}^d)} \le c_m(X) \|f\|_{W^m(X)}$$

for all $f \in W^m(X)$, where $c_m(X) > 0$ is a constant depending only on m and X. We refer to p. 154 of Adams and Fournier (2003) and Chapter 6 of Stein (1970) for a proof. Note that an immediate consequence of this theorem is that

$$\|f\| := \inf\left\{ \|g\|_{W^m(\mathbb{R}^d)} : g \in W^m(\mathbb{R}^d) \text{ with } g_{|X} = f \right\}, \tag{A.35}$$

$f \in W^m(X)$, defines an equivalent norm on $W^m(X)$. Finally, we refer to the books by Adams and Fournier (2003) and Ziemer (1989) for more information on Sobolev spaces.

A.5.6 Entropy Numbers

In this subsection, we present some properties of (dyadic) entropy numbers introduced in Section 6.3. Let us begin by recalling their definition.

Definition A.5.26. *Let (T,d) be a metric space and $n \geq 1$ be an integer. Then the n-th **(dyadic) entropy number** of (T,d) is defined by*

$$e_n(T,d) \;:=\; \inf\left\{\varepsilon > 0 : \exists\, t_1,\ldots,t_{2^{n-1}} \in T \text{ such that } T \subset \bigcup_{i=1}^{2^{n-1}} B_d(t_i,\varepsilon)\right\},$$

where we use the convention $\inf\emptyset := \infty$.

Moreover, if (T,d) is a subspace of a normed space $(E,\|\cdot\|)$ we write $e_n(T,\|\cdot\|) := e_n(T,E) := e_n(T,d)$.

Finally, if $S : E \to F$ is a bounded, linear operator between the normed spaces E and F, we write $e_n(S) := e_n(SB_E,\|\cdot\|_F)$.

Entropy numbers have been extensively studied in the literature. For a gentle introduction on their basics, we refer to Chapter 1 of Carl and Stephani (1990), from which the following properties are taken if not stated otherwise. To present these properties, we assume that we have a metric space (T,d). Then the entropy numbers are obviously **monotone**, i.e., $e_n(T,d) \geq e_{n+1}(T,d)$ for all $n \geq 1$. Now let us fix a subset $A \subset T$. Then A equipped with the trace metric $d_A := d_{|A\times A}$ of d is a metric space and hence the n-th entropy number of (A,d_A) is given by

$$e_n(A,d_A) \;=\; \inf\left\{\varepsilon > 0 : \exists\, t_1,\ldots,t_{2^{n-1}} \in A \text{ such that } A \subset \bigcup_{i=1}^{2^{n-1}} B_d(t_i,\varepsilon)\right\}.$$

However, one could also consider the quantity

$$\tilde{e}_n(A,d) \;:=\; \inf\left\{\varepsilon > 0 : \exists\, t_1,\ldots,t_{2^{n-1}} \in T \text{ such that } A \subset \bigcup_{i=1}^{2^{n-1}} B_d(t_i,\varepsilon)\,,\right\}$$

which allows the ε-net to be taken from T instead of A. Fortunately, both quantities are closely related; namely we have

$$\tilde{e}_n(A,d) \;\leq\; e_n(A,d_A) \;\leq\; 2\tilde{e}_n(A,d)\,, \tag{A.36}$$

where the first inequality follows from $A \subset T$ and the second inequality can be derived from (1.1.3) and (1.1.4) in the book by Carl and Stephani (1990).

Now let E and F be normed spaces and $S_1 : E \to F$ and $S_2 : E \to F$ be bounded linear operators. Then the dyadic entropy numbers are **additive** in the sense of

$$e_{m+n-1}(S_1+S_2) \;\leq\; e_m(S_1) + e_n(S_2)\,, \qquad n,m \geq 1\,. \tag{A.37}$$

Moreover, if Z is another normed space and $R : E \to Z$ and $S : Z \to F$ are bounded linear operators, then the dyadic entropy numbers are also **multiplicative** in the sense of

$$e_{m+n-1}(SR) \;\le\; e_m(S)e_n(R)\,, \qquad\qquad n, m \ge 1\,. \tag{A.38}$$

In particular, we have $e_n(SR) \le \|S\|e_n(R)$ and $e_n(SR) \le \|R\|e_n(S)$ for all $n \ge 1$, where we used the equation

$$e_1(S) = \|S\|\,. \tag{A.39}$$

Let us now assume that we have Banach spaces $\tilde{E}$, E, F, and $\tilde{F}$, a bounded linear operator $S : E \to F$, a metric surjection $Q : \tilde{E} \to E$, and an isometric embedding $I : F \to \tilde{F}$. Then the (dyadic) entropy numbers are **surjective** and **injective**, i.e., for $n \ge 1$, we have

$$e_n(SQ) = e_n(S) \qquad\qquad \text{and} \qquad\qquad e_n(IS) \le e_n(S) \le 2e_n(IS), \tag{A.40}$$

respectively. In addition, if I happens to be bijective, i.e., to be an isometric isomorphism, we actually have $e_n(IS) = e_n(S)$.

An operator $S : E \to F$ between $\mathbb{R}$-Banach spaces E and F is of finite rank, i.e., $d := \operatorname{rank} S < \infty$, if and only if there exists a constant $C > 0$ such that

$$C2^{-(n-1)/d} \;\le\; e_n(S) \;\le\; 4\|S\|\,2^{-(n-1)/d}\,, \qquad\qquad n \ge 1. \tag{A.41}$$

In particular, there exists *no* operator $S \ne 0$ whose dyadic entropy numbers decrease faster than 2^{-n}.

Entropy numbers are closely related to the approximation numbers introduced in (A.29). Namely, Carl's inequality states that for all $0 < p \le \infty$ and $0 < q < \infty$ there exists a constant $c_{p,q} > 0$ such that

$$\sum_{i=1}^{m} i^{q/p-1} e_i^q(S) \le c_{p,q} \sum_{i=1}^{m} i^{q/p-1} a_i^q(S) \tag{A.42}$$

for all bounded operators $S : E \to F$ acting between Banach spaces and all $m \ge 1$. For a proof, we refer to Theorem 3.1.2 of Carl and Stephani (1990). Moreover, Lemma 1.5.1 by Carl and Stephani (1990) shows that (A.42) is equivalent to

$$\sum_{i=1}^{m} 2^{iq/p} e_{2^i}^q(S) \le \tilde{c}_{p,q} \sum_{i=1}^{m} 2^{iq/p} a_{2^i}^q(S)\,, \tag{A.43}$$

where $\tilde{c}_{p,q}$ is another constant independent of S, E, F, and m. Finally, Carl and Stephani (1990) show on p. 120 that, for Hilbert spaces H and $T \in \mathcal{K}(H)$, we have the following strong inverse of the inequalities above:

$$a_i(T) \le 2e_i(T)\,, \qquad\qquad i \ge 1\,. \tag{A.44}$$

Let us finally collect some well-established bounds on entropy numbers for certain embeddings. To this end, let us fix a bounded, convex, and open subset $X \subset \mathbb{R}^d$ and an integer $m \geq 0$. Then Kolmogorov and Tikhomirov (1961), see also Theorem 2.7.1 of van der Vaart and Wellner (1996) for the upper bound, showed that there exist constants $\tilde{c}_{m,d}(X), c_{m,d}(X) > 0$ such that

$$\tilde{c}_{m,d}(X)\, n^{-m/d} \leq e_n\bigl(\mathrm{id} : C_b^m(X) \to \ell_\infty(X)\bigr) \leq c_{m,d}(X)\, n^{-m/d} \qquad \text{(A.45)}$$

for all $n \geq 1$. Since $C^m(\overline{X})$ is a subspace of $C_b^m(X)$, the multiplicativity (A.38) shows that the upper bound also holds for $\mathrm{id} : C^m(\overline{X}) \to \ell_\infty(X)$. Moreover, $\mathrm{id} : C^0(\overline{X}) \to \ell_\infty(X)$ is an isometric embedding and hence (A.40) yields

$$e_n\bigl(\mathrm{id} : C^m(\overline{X}) \to C^0(\overline{X})\bigr) \leq 2c_{m,d}(X)\, n^{-m/d}\,, \qquad n \geq 1. \qquad \text{(A.46)}$$

In order to have a closer look at the constant $c_{m,d}(X)$, we fix an $r > 0$ and write $\tau_r f(x) := f(rx)$ for functions $f : rX \to \mathbb{R}$ and $x \in X$. It is straightforward to check that $\|\tau_r f\|_{C_b^m(X)} \leq r^m \|f\|_{C_b^m(rX)}$ for all $f \in C_b^m(rX)$ and $r \geq 1$. In addition, we obviously have $\|\tau_{1/r} f\|_\infty = \|f\|_\infty$ for all $f \in \ell_\infty(X)$ and $r > 0$. Moreover, we have the commutative diagram

$$\begin{array}{ccc} C_b^m(rX) & \xrightarrow{\ \mathrm{id}\ } & \ell_\infty(rX) \\ {\scriptstyle \tau_r}\downarrow & & \uparrow{\scriptstyle \tau_{1/r}} \\ C_b^m(X) & \xrightarrow[\ \mathrm{id}\ ]{} & \ell_\infty(X) \end{array}$$

which yields $e_n(\mathrm{id} : C_b^m(rX) \to \ell_\infty(rX)) \leq r^m e_n(\mathrm{id} : C_b^m(X) \to \ell_\infty(X))$ for all $r \geq 1$ and $n \geq 1$ by the multiplicativity (A.38). Consequently, for $r \geq 1$, the constant $c_{m,d}(rX)$ can be assumed to be of the form

$$c_{m,d}(rX) = r^m c_{m,d}(X)\,. \qquad \text{(A.47)}$$

Our next goal is to present bounds similar to (A.45) for Sobolev spaces. To this end, we assume that X is an open Euclidean ball in $\mathbb{R}^d$. Then the first bound, originally proved by Birman and Solomyak (1967), states that for all $m \geq 0$ there exist constants $\tilde{c}_m(X), c_m(X) > 0$ such that

$$\tilde{c}_m(X)\, n^{-m/d} \leq e_n\bigl(\mathrm{id} : W^m(X) \to L_2(X)\bigr) \leq c_m(X)\, n^{-m/d} \qquad \text{(A.48)}$$

for all $n \geq 1$. Moreover, Birman and Solomyak also proved that for $m > d/2$ there exist (different) constants $\tilde{c}_m(X), c_m(X) > 0$ such that

$$\tilde{c}_m(X)\, n^{-m/d} \leq e_n\bigl(\mathrm{id} : W^m(X) \to L_\infty(X)\bigr) \leq c_m(X)\, n^{-m/d} \qquad \text{(A.49)}$$

for all $n \geq 1$. These bounds are special cases of more general results on embeddings thoroughly presented by Edmunds and Triebel (1996). Since the

original paper by Birman and Solomyak is in Russian, we now briefly describe how the bounds above can be recovered from the general theory. To this end, we first mention that on p. 24f, Edmunds and Triebel introduce two scales of quasi-Banach spaces, denoted by $B^s_{p,q}(\mathbb{R}^d)$ and $F^s_{p,q}(\mathbb{R}^d)$, where $s \in \mathbb{R}$, $q \in (0,\infty]$, and $p \in (0,\infty]$ for the B-scale but only $p \in (0,\infty)$ for the F-scale. It turns out that $B^m_{2,2}(\mathbb{R}^d) = F^m_{2,2}(\mathbb{R}^d) = W^m(\mathbb{R}^d)$ in the sense of isomorphisms for all $m \geq 0$, see p. 44 and p. 25. Moreover, they also show on p. 44 that $B^0_{\infty,1}(\mathbb{R}^d)$ is continuously embedded into $L_\infty(\mathbb{R}^d)$. Let us write $A^s_{p,q}(\mathbb{R}^d)$ if we mean either $B^s_{p,q}(\mathbb{R}^d)$ or $F^s_{p,q}(\mathbb{R}^d)$. For an open Euclidean ball $X \subset \mathbb{R}^d$, Edmunds and Triebel then define $A^s_{p,q}(X) := \{g_{|X} : g \in A^s_{p,q}(\mathbb{R}^d)\}$ and

$$\|f\|_{A^s_{p,q}(X)} := \inf\bigl\{\|g\|_{A^s_{p,q}(\mathbb{R}^d)} : g \in A^s_{p,q}(\mathbb{R}^d) \text{ with } g_{|X} = f\bigr\}\,.$$

Using (A.35), we conclude that $B^m_{2,2}(X) = F^m_{2,2}(X) = W^m(X)$ in the sense of isomorphisms for all $m \geq 0$, as well as $B^0_{\infty,1}(X) \subset L_\infty(X)$ in the sense of a continuous embedding. Let us now fix $s_2 < s_1$ and $p_1, p_2, q_1, q_2 \in (0,\infty]$ such that

$$s_1 - s_2 - d\bigl(p_1^{-1} - p_2^{-1}\bigr)_+ > 0\,,$$

where for the F-scale, we addionally assume $p_1, p_2 < \infty$. Then Theorem 2 on p. 118 of Edmunds and Triebel (1996) shows that there exist constants $\tilde{c}, c > 0$ such that

$$\tilde{c}\, n^{-(s_1-s_2)/d} \leq e_n\bigl(\mathrm{id} : A^{s_1}_{p_1,q_1}(X) \to A^{s_2}_{p_2,q_2}(X)\bigr) \leq c\, n^{-(s_1-s_2)/d} \qquad \text{(A.50)}$$

for all $n \geq 1$. From this, (A.48) follows by taking $s_1 := m$, $s_2 := 0$, $p_i := q_i := 2$. In addition, the lower bound of (A.49) follows from (A.48) since $L_\infty(X)$ is continuously embedded into $L_2(X)$. Finally, the upper bound can be derived from (A.50) by considering the B-scale and $s_1 := m$, $p_1 := q_1 := 2$, $s_2 := 0$, $p_2 := \infty$, and $q_2 := 1$.

A.6 Convex Analysis

In this section, we discuss some properties of convex functions. To this end, recall that a subset $A \subset E$ of a Banach space E is called **convex** if, for all $x_1, x_2 \in A$ and all $\alpha \in [0,1]$, we have $\alpha x_1 + (1-\alpha)x_2 \in A$. In this case, a function $f : A \to \mathbb{R} \cup \{\infty\}$ is called **convex** if, for all $x_1, x_2 \in A$ and all $\alpha \in [0,1]$, we have

$$f\bigl(\alpha x_1 + (1-\alpha)x_2\bigr) \;\leq\; \alpha f(x_1) + (1-\alpha) f(x_2)\,.$$

In addition, f is called **concave** if $-f$ is convex.

In the following subsection, we recall some continuity properties of convex functions, and in Subsection A.6.2 we review the subdifferential calculus for convex functions. Then, in Subsection A.6.3, we discuss some stronger notions of convexity and their relations to each other. In Subsection A.6.4, we present some important properties of the Fenchel-Legendre bi-conjugate operation.

A.6.1 Basic Properties of Convex Functions

In this section, we provide some elementary facts on convex functions. We begin with the following result, which can be found, for example, on p. 27 of Rockafellar (1970).

Lemma A.6.1 (Hessian matrix of convex functions). *Let $O \subset \mathbb{R}^n$ be an open convex set and $g : O \to \mathbb{R}$ be a twice continuously differentiable function. Then g is convex if and only if its Hessian matrix $Q_x = (q_{i,j}(x))_{i,j}$ defined by*

$$q_{i,j}(x) := \frac{\partial^2 g}{\partial x_i \, \partial x_j}(x_1, \ldots, x_n)$$

is positive semi-definite for every $x \in O$.

An immediate consequence of this result is the following example for a convex function from $\mathbb{R}^n$ to $\mathbb{R}$. Let $K \in \mathbb{R}^{n \times n}$ be a symmetric matrix, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. The quadratic function

$$g(a) := a^T K a + a^T b + c, \quad a \in \mathbb{R}^n, \tag{A.51}$$

is convex on $\mathbb{R}^n$ if and only if K is positive semi-definite, i.e., if $a^T K a \geq 0$ for every $a \in \mathbb{R}^n$.

We now give a continuity result for convex functions.

Lemma A.6.2 (Continuity of convex functions). *Let $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be a convex function and* $\operatorname{Dom} f := \{t \in \mathbb{R} : f(t) < \infty\}$. *Then we have:*

i) f is continuous at all $t \in \operatorname{Int} \operatorname{Dom} f$.
ii) If f is lower semi-continuous, then $f_{|\overline{\operatorname{Dom} f}}$ is continuous.

Proof. See Theorem 2.35 and Corollary 2.37 of Rockafellar and Wets (1998) or Theorem 2.1.6 of Zălinescu (2002). □

Note that the preceding lemma is not true in general Banach spaces. Indeed, every infinite-dimensional Banach space E has a *linear*, and thus convex, functional $x' : E \to \mathbb{R}$ that is nowhere continuous.

The following result is shown in, e.g., Theorem 2.1.3 of Zălinescu (2002).

Lemma A.6.3 (Supremum of convex functions). *Let E be a Banach space, $I \neq \emptyset$ and $(f_i)_{i \in I}$ be a family of convex functions $f_i : E \to \mathbb{R}$. Then $f : E \to \mathbb{R} \cup \{\infty\}$ defined by $f(x) := \sup_{i \in I} f_i(x)$, $x \in E$, is convex.*

Since affine linear functions on $\mathbb{R}$ are continuous and convex, we obtain the following result by combining the preceding lemmas with Lemma A.2.7.

Lemma A.6.4. *Let $I \neq \emptyset$ and $(f_i)_{i \in I}$ be a family of affine linear functions $f_i : \mathbb{R} \to \mathbb{R}$. Then $f(x) := \sup_{i \in I} f_i(x)$ defines a convex and l.s.c. function $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ whose restriction $f_{|\overline{\operatorname{Dom} f}}$ is continuous.*

Given two Banach spaces E and F and a subset $A \subset E$, we call a function $f : A \to F$ **Lipschitz continuous** if there exists a constant $c \geq 0$ such that $\|f(x) - f(x')\|_F \leq c\|x - x'\|_E$ for all $x, x' \in A$. In this case, the smallest such constant c is denoted by $|f|_1$. Moreover, a function $f : \mathbb{R} \to \mathbb{R}$ is called **locally Lipschitz continuous** if for all $t > 0$ the restriction $f_{|[-t,t]}$ of f to the interval $[-t, t]$ is Lipschitz continuous.

Lemma A.6.5 (Local Lipschitz continuity and convexity). *Every convex $f : \mathbb{R} \to \mathbb{R}$ is locally Lipschitz continuous, and for $t > 0$ we have*

$$\left|f_{|[-t,t]}\right|_1 \;\leq\; \frac{2}{t}\,\|f_{|[-2t,2t]}\|_\infty\,.$$

If in addition $f(0) = 0$, then $s \mapsto \frac{f(s)}{s}$ is increasing on $(0, \infty)$ and we have

$$\|f_{|[-t,t]}\|_\infty \;\leq\; t \cdot \left|f_{|[-t,t]}\right|_1\,, \qquad\qquad t > 0.$$

Proof. The first inequality follows from the proof of Proposition 1.6 of Phelps (1993). In addition, for $0 < s < s'$ and $\alpha := s/s' \in [0, 1]$, we have $f(s) = f(\alpha s' + (1 - a)0) \leq \alpha f(s')$, which shows the monotonicity assertion. Finally, the second inequality follows from $|f(r)| = |f(r) - f(0)| \leq \left|f_{|[-r,r]}\right|_1 \cdot r$. □

Let us now recall the fundamental theorem of calculus for Lipschitz continuous functions. Note that it actually holds for *absolutely continuous* functions; however, we do not need this full generality, and hence we omit the details.

Theorem A.6.6 (Fundamental theorem of calculus). *Let $f : [a, b] \to \mathbb{R}$ be a Lipschitz continuous function. Then f is differentiable at Lebesgue-almost all $t \in [a, b]$. Furthermore, the almost everywhere defined derivative f' of f is Lebesgue integrable and satisfies*

$$f(x) = f(a) + \int_a^x f'(t)dt\,, \qquad\qquad x \in [a, b]. \tag{A.52}$$

Proof. The proof of this classical result from Lebesgue's integration theory can be found in many textbooks on real analysis. Here we only mention the Theorems 26–28 in Chapter X of Graves (1956) and the Theorems 271, 269, and 274 by Kestelman (1960). □

Note that for restrictions $f_{|[a,b]}$ of *convex* functions $f : \mathbb{R} \to \mathbb{R}$, the Lipschitz condition in the preceding theorem is automatically satisfied by Lemma A.6.5. Consequently, convex functions on $\mathbb{R}$ are almost everywhere differentiable and satisfy (A.52).

Theorem A.6.6 can be used to establish a convexity test and a formula for computing the (local) Lipschitz constants. These results are provided by the following two lemmas.

Lemma A.6.7 (Convexity test). *Let $f : [a,b] \to \mathbb{R}$ be a Lipschitz continuous function and $N \subset [a,b]$ be a Lebesgue null set such that f is differentiable at all $t \in [a,b] \setminus N$. Assume that, for all $s,t \in [a,b]\setminus N$ with $s \le t$, we have $f'(s) \le f'(t)$. Then f is convex.*

Proof. Let us fix $x_1, x_2 \in [a,b]$ with $x_1 < x_2$ and a real number $\lambda \in (0,1)$. We define $x := \lambda x_1 + (1-\lambda)x_2$. With the help of Theorem A.6.6, we then obtain

$$\frac{f(x)-f(x_1)}{x-x_1} = \frac{1}{x-x_1}\int_{x_1}^{x} f'(t)dt \;\le\; \sup_{t\in[x_1,x]\setminus N} f'(t)$$

and

$$\frac{f(x_2)-f(x)}{x_2-x} = \frac{1}{x_2-x}\int_{x}^{x_2} f'(t)dt \ge \inf_{t\in[x,x_2]\setminus N} f'(t)\,.$$

Combining both inequalities, by our monotonicity assumption on f', we then easily find the assertion. □

Lemma A.6.8 (Computation of Lipschitz constant). *Let $f : [a,b] \to \mathbb{R}$ be a Lipschitz continuous function and $N \subset [a,b]$ be a Lebesgue null set such that f is differentiable at all $t \in [a,b]\setminus N$. Then we have*

$$|f|_1 = \sup_{t\in[a,b]\setminus N} |f'(t)|\,.$$

Proof. With the help of Theorem A.6.6, we obtain

$$\begin{aligned} |f|_1 = \sup_{\substack{x_1,x_2\in[a,b]\\ x_1\neq x_2}} \left|\frac{f(x_2)-f(x_1)}{x_2-x_1}\right| &\le \sup_{\substack{x_1,x_2\in[a,b]\\ x_1<x_2}} \frac{1}{x_2-x_1}\int_{x_1}^{x_2} |f'(t)|dt \\ &\le \sup_{t\in[a,b]\setminus N} |f'(t)|\,. \end{aligned}$$

Conversely, for $t \in [a,b]\setminus N$, an easy estimate shows

$$|f'(t)| = \left|\lim_{\substack{s\to t\\ s\neq t}} \frac{f(s)-f(t)}{s-t}\right| \le \sup_{\substack{s\in[a,b]\\ s\neq t}} \left|\frac{f(s)-f(t)}{s-t}\right| = |f|_1\,. \qquad \square$$

Again note that, for a convex function $f : \mathbb{R} \to \mathbb{R}$, the Lipschitz continuity of its restrictions $f_{|[-a,a]}$, $a > 0$, is guaranteed by Lemma A.6.5. Consequently, Lemma A.6.8 provides a simple way to calculate the local Lipschitz constants for convex f.

The following theorem provides a simple-to-use criterion for the existence of a global minimizer.

Theorem A.6.9 (Existence of minimizers). *Let E be a reflexive Banach space and $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex and lower semi-continuous map.*

If there exists an $M > 0$ such that $\{x \in E : f(x) \le M\}$ is non-empty and bounded, then f has a global minimum, i.e., there exists an $x_0 \in E$ with

$$f(x_0) \le f(x)\,, \qquad x \in E.$$

Moreover, if f is strictly convex, then x_0 is the only element minimizing f.

Proof. Proposition 6 on p. 75 of Ekeland and Turnbull (1983) shows the existence, and the uniqueness is a consequence of the strict convexity. □

A.6.2 Subdifferential Calculus for Convex Functions

In this subsection, we collect some important properties of subdifferentials. Throughout this subsection, E and F denote $\mathbb{R}$-Banach spaces. Let us begin by recalling the definition of subdifferentials.

Definition A.6.10. *Let E be a Banach space, $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex function, and $w \in E$ with $f(w) < \infty$. Then the **subdifferential** of f at w is defined by*

$$\partial f(w) := \left\{ w' \in E' : \langle w', v - w \rangle \le f(v) - f(w) \text{ for all } v \in E \right\}.$$

We begin with a proposition that provides some elementary facts of the subdifferential (see Phelps, 1993, Proposition 1.11).

Proposition A.6.11. *Let $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex function and $w \in E$ such that $f(w) < \infty$. If f is continuous at w, then the subdifferential $\partial f(w)$ is a non-empty, convex, and weak*-compact subset of E'. In addition, if $c \ge 0$ and $\delta > 0$ are constants satisfying*

$$\left| f(v) - f(w) \right| \;\le\; c\,\|v - w\|\,, \qquad v \in w + \delta B_E\,,$$

then we have $\|w'\| \le c$ for all $w' \in \partial f(w)$.

The following proposition shows the extent to which the known rules of calculus carry over to subdifferentials. For the proofs of these rules, we refer to, e.g., Castaing and Valadier (1977), Ekeland and Turnbull (1983), Phelps (1993), and Zălinescu (2002).

Proposition A.6.12 (Subdifferential calculus). *Let $f, g : E \to \mathbb{R} \cup \{\infty\}$ be convex functions, $\lambda \ge 0$, and $A : F \to E$ be a bounded linear operator. Then the following rules are true:*

Homogeneity. *For all $w \in E$ with $f(x) < \infty$, we have $\partial(\lambda f)(w) = \lambda \partial f(w)$.*
Additivity. *If there exists a $w_0 \in E$ at which f is continuous, then, for all $w \in E$ satisfying both $f(w) < \infty$ and $g(w) < \infty$, we have*

$$\partial(f + g)(w) = \partial f(w) + \partial g(w)\,.$$

Chain rule. *If there exists an $v_0 \in F$ such that f is finite and continuous at Av_0, then, for all $v \in F$ satisfying $f(Av) < \infty$, we have*

$$\partial(f \circ A)(v) \;=\; A'\partial f(Av),$$

where $A' : E' \to F'$ denotes the adjoint operator of A.

Minima. *The function f has a global minimum at $w \in E$ if and only if $0 \in \partial f(w)$.*

Differentiability. *If f is finite and continuous at $w \in E$, then f is Gâteaux differentiable at w if and only if $\partial f(w)$ is a singleton, and in this case we have $\partial f(w) = \{f'(w)\}$.*

Monotonicity. *If f is finite and continuous at all $w \in E$, then ∂f is a* ***monotone operator****, i.e., for all $v, w \in E$ and $v' \in \partial f(v)$, $w' \in \partial f(w)$, we have*

$$\langle v' - w', v - w\rangle \geq 0\,.$$

The following proposition shows how the subdifferential of a function defined by an integral can be computed.

Proposition A.6.13. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss function, P be a distribution on $X \times Y$, and $p \in [1, \infty)$. We define $R : L_p(\mathrm{P}) \to [0, \infty]$ by*

$$R(f) := \int_{X\times Y} L\big(x, y, f(x, y)\big)\, d\mathrm{P}(x, y)\,, \qquad\qquad f \in L_p(\mathrm{P}).$$

If we have $R(f) < \infty$ for all $f \in L_p(\mathrm{P})$, then, for all $f \in L_p(\mathrm{P})$, we have

$$\partial R(f) = \Big\{h \in L_{p'}(\mathrm{P}) : h(x, y) \in \partial L\big(x, y, f(x, y)\big) \text{ for } \mathrm{P}\text{-almost all } (x, y)\Big\}\,,$$

where $\partial L(x, y, t)$ denotes the subdifferential of $L(x, y, \cdot\,)$ at the point t.

Proof. Since L is convex and finite, it is a continuous loss. Consequently, it is a normal convex integrand by Proposition 2C of Rockafellar (1976). Then Corollary 3E of Rockafellar (1976) gives the assertion. □

The next proposition shows that the subdifferential is in some sense semi-continuous (for a proof see, e.g., Proposition 2.5 of Phelps, 1993).

Proposition A.6.14. *If $f : \mathbb{R}^d \to \mathbb{R}$ is continuous and convex, then the subdifferential map is upper semi-continuous in the sense that, for all $w \in \mathbb{R}^d$ and $\varepsilon > 0$, there exists a $\delta > 0$ with*

$$\partial f(w + \delta B_{\mathbb{R}^d}) \;\subset\; \partial f(w) + \varepsilon B_{\mathbb{R}^d}\,.$$

Here we used the notation $\partial f(M) := \bigcup_{v\in M} \partial f(v)$ for a set $M \subset \mathbb{R}^d$.

Let us now present a simple description of the subdifferential for functions defined on $\mathbb{R}$. To this end, recall (see, e.g., Lemma 1.2 of Phelps, 1993) that, for a convex function $f : \mathbb{R} \to \mathbb{R}$, both the left and the right derivatives

$$d^- f(x) := \lim_{t \searrow 0} \frac{f(x) - f(x-t)}{t}\,, \qquad d^+ f(x) := \lim_{t \searrow 0} \frac{f(x+t) - f(x)}{t}\,,$$

exist for all $x \in \mathbb{R}$. This leads to the following lemma.

Lemma A.6.15. *For every convex function $f : \mathbb{R} \to \mathbb{R}$ and all $x \in \mathbb{R}$, we have*

$$\partial f(x) = \big[d^- f(x), d^+ f(x)\big].$$

Proof. By Theorem 3.1.8 of Borwein and Lewis (2000), we find $d^- f(x) = \min \partial f(x)$ and $d^+ f(x) = \max \partial f(x)$. Now the assertion follows from the fact that $\partial f(x)$ is a convex set by Proposition A.6.11. □

The preceding lemma can be used to establish the next lemma, which sometimes makes it possible to substitute a convex function by a *Lipschitz continuous* convex function.

Lemma A.6.16. *Let $f : \mathbb{R} \to [0, \infty)$ be a convex function and $a, b \in \mathbb{R}$ be real numbers with $a < b$. Then there exists a convex, Lipschitz continuous function $\tilde{f} : \mathbb{R} \to [0, \infty)$ with the following properties:*

$$\tilde{f}_{|[a,b]} = f_{|[a,b]}\,, \tag{A.53}$$

$$|\tilde{f}|_1 = \big|f_{|[a,b]}\big|_1\,, \tag{A.54}$$

$$\partial \tilde{f}(x) \subset \partial f(x)\,, \qquad x \in [a,b]. \tag{A.55}$$

Proof. Let us first assume that $d^+ f(a) \le 0$ and $d^- f(b) \ge 0$. Then we define $\tilde{f}$ by

$$\tilde{f}(x) := \begin{cases} d^+ f(a) \cdot (x-a) + f(a) & \text{if } x < a \\ f(x) & \text{if } x \in [a,b] \\ d^- f(b) \cdot (x-b) + f(b) & \text{if } x > b\,. \end{cases}$$

It is obvious that $\tilde{f}$ is a non-negative function satisfying (A.53). In order to show the other assertions, recall that by Lemma A.6.5 and Theorem A.6.6 there exists a Lebesgue null set $N \subset [a,b]$ such that $f_{|[a,b]}$ is differentiable at all $x \in [a,b] \setminus N$. Consequently, $\tilde{f}$ is differentiable at all $x \in [a,b] \setminus N$, and its derivative at these points is given by $\tilde{f}'(x) = f'(x)$. Moreover, for $x \le a$, we clearly have $\tilde{f}'(x) = d^+ f(a)$ and, analogously, for $x \ge b$, we have $\tilde{f}'(x) = d^- f(b)$. In addition, Lemma A.6.15 together with the monotonicity of ∂f yields

$$d^+ f(a) \le f'(x_1) \le f'(x_2) \le d^- f(b) \tag{A.56}$$

for all $x_1, x_2 \in [a,b] \setminus N$ with $x_1 \le x_2$. Using Lemma A.6.7, we then see that $\tilde{f}$ is convex. Moreover, from (A.56), we can also conclude that

$$\sup_{x\in[a,b]\setminus N} |f'(x)| \le \max\bigl\{\bigl|d^+f(a)\bigr|, \bigl|d^-f(b)\bigr|\bigr\} \tag{A.57}$$

To show the converse inequality, we fix an $\varepsilon > 0$. By Lemma A.6.14, there then exists a $\delta > 0$ such that

$$\partial f\bigl(b + [-\delta,\delta]\bigr) \;\subset\; \partial f(b) + [-\varepsilon,\varepsilon] = \bigl[d^-f(b) - \varepsilon, d^+f(b) + \varepsilon\bigr]\,,$$

where for the last equality we used Lemma A.6.15. Moreover, since N is a null set there exists a $x \in [b-\delta, b] \setminus N$, and since for this x we have $f'(x) \in \partial f(b + [-\delta,\delta])$, we find $f'(x) \ge d^-f(b) - \varepsilon$. This together with (A.56) shows

$$\lim_{\substack{x\to b\\ x\in[a,b]\setminus N}} f'(x) = d^-f(b)\,.$$

Since we can make an analogous consideration for $d^+f(a)$, we find that we actually have equality in (A.57). Using Lemma A.6.8, we then find

$$|\tilde f|_1 = \max\Bigl\{\bigl|d^+f(a)\bigr|, \bigl|d^-f(b)\bigr|, \sup_{x\in[a,b]\setminus N} |f'(x)|\Bigr\} = \sup_{x\in[a,b]\setminus N} |f'(x)| = \bigl|f_{|[a,b]}\bigr|_1\,.$$

Finally, (A.55) is obvious for $x \in (a,b)$, and for, say, $x = b$, it follows from $\tilde f'(b) = d^-f(b) \in \partial f(b)$.

Let us now consider the case $d^+f(a) \le 0$ and $d^-f(b) < 0$. If f has no minimum, we modify f on $(-\infty, a)$ analogously to the first case and keep it unchanged on $[a,\infty)$. Repeating the reasoning above, then gives the assertion. Moreover, if f has a minimum at, say, b', we modify f on $(-\infty, a)$ analogously to the first case and keep it unchanged on $[a, b']$. Moreover, for $x \in [b', \infty)$, we define $\tilde f(x) := f(b')$. Again, repeating the arguments of the first case gives the assertion. The remaining cases can be treated in a symmetric way. □

A.6.3 Some Further Notions of Convexity

In this section, we introduce some stronger notions of convexity and discuss their relations to each other. Let us begin by recalling that a function $f : A \to \mathbb{R}$ on a convex subset $A \subset E$ of a Banach space E is called **strictly convex** if, for all $x_1, x_2 \in A$ with $x_1 \ne x_2$ and all $\alpha \in (0,1)$, we have

$$f\bigl(\alpha x_1 + (1-\alpha)x_2\bigr) \;<\; \alpha f(x_1) + (1-\alpha) f(x_2)\,.$$

Furthermore, the **modulus of convexity** of f is defined by

$$\delta_f(\varepsilon) := \inf\left\{\frac{f(x_1)+f(x_2)}{2} - f\Bigl(\frac{x_1+x_2}{2}\Bigr) : x_1, x_2 \in A \text{ with } |x_1 - x_2| \ge \varepsilon\right\},$$

$\varepsilon > 0$, and we say that f is **uniformly convex** if $\delta_f(\varepsilon) > 0$ for all $\varepsilon > 0$. Obviously, every strictly convex function is also convex. The following lemma describes some less trivial relations between the different notions of convexity.

Lemma A.6.17. *Given an interval I and a function $f : I \to \mathbb{R}$, we have:*

i) If f is convex and satisfies $f(\alpha_0 x_1 + (1-\alpha_0)x_2) = \alpha_0 f(x_1) + (1-\alpha_0)f(x_2)$ for some $x_1, x_2 \in I$, $\alpha_0 \in [0,1]$, then, for all $\alpha \in [0,1]$, we have

$$f\big(\alpha x_1 + (1-\alpha)x_2\big) \;=\; \alpha f(x_1) + (1-\alpha)f(x_2)\,. \tag{A.58}$$

ii) If f is continuous, then f is convex if and only if, for all $x_1, x_2 \in I$, we have

$$f\Big(\frac{x_1+x_2}{2}\Big) \le \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2)\,. \tag{A.59}$$

iii) If f is continuous, then f is strictly convex if and only if, for all $x_1, x_2 \in I$ with $x_1 \neq x_2$, we have

$$f\Big(\frac{x_1+x_2}{2}\Big) < \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2)\,. \tag{A.60}$$

iv) If f is uniformly convex and continuous, then it is strictly convex. Conversely, if I is compact and f is strictly convex and continuous, then it is actually uniformly convex.

Proof. i). This assertion can be shown using elementary calculations.

ii). This follows from Theorems 8 and 10 of Behringer (1992).

iii). If (A.60) holds, then we have already seen that f is convex. Consequently, if f was not strictly convex, we would have (A.58). However, by *i)*, we could then assume $\alpha_0 = \frac{1}{2}$, which would give a contradiction.

iv) The first assertion follows from *iii)*, and the second one is trivial. □

Our next aim is to investigate the modulus of convexity. Although this concept, in an equivalent formulation, has already been investigated by Polyak (1966) and Levitin and Polyak (1966), almost nothing that is useful for our purposes seems to be known (however, see Butnariu and Iusem, 2000; Zălinescu, 2002; and the references therein for some general information on the modulus). Therefore, we present the following two lemmas, which provide some ways to simplify the computation of $\delta_f(\varepsilon)$.

Lemma A.6.18. *Let $I \subset \mathbb{R}$ be a non-empty interval, $f : I \to \mathbb{R}$ be strictly convex, and $\varepsilon > 0$. Then we have*

$$\delta_f(2\varepsilon) = \inf\left\{\frac{f(x-\varepsilon)+f(x+\varepsilon)}{2} - f(x) : x \text{ satisfies } x-\varepsilon \in I \text{ and } x+\varepsilon \in I\right\}.$$

Proof. For fixed $x_1 \in I$, we define $h_{x_1} : I \to [0,\infty)$ by $h_{x_1}(x_2) := \frac{f(x_1)+f(x_2)}{2} - f(\frac{x_1+x_2}{2})$, $x_2 \in I$. Theorem A.6.6 then shows that the derivative $h'_{x_1}(x_2)$ exists for almost all x_2, and an easy calculation shows $h'_{x_1}(x_2) = \frac{f'(x_2)}{2} - \frac{1}{2}f'(\frac{x_1+x_2}{2})$ for such x_2. Furthermore, f is strictly convex and thus h_{x_1} has a unique minimum at x_1. This yields $h'_{x_1}(x_2) < 0$ if $x_2 < x_1$, and $h'_{x_1}(x_2) > 0$ if $x_2 > x_1$.

Theorem A.6.6 then shows that h_{x_1} is strictly decreasing on $(-\infty, x_1) \cap I$ and strictly increasing on $(x_1, \infty) \cap I$, and thus we have

$$\delta_f(2\varepsilon) = \inf_{\substack{x_1 \in I \\ x_1 \pm 2\varepsilon \in I}} h_{x_1}(x_1 \pm 2\varepsilon) = \inf_{\substack{x_1+\varepsilon \in I \\ x_1-\varepsilon \in I}} h_{x_1-\varepsilon}(x_1+\varepsilon)\,,$$

where in the last step we used $h_{x_1-\varepsilon}(x_1+\varepsilon) = h_{x_1+\varepsilon}(x_1-\varepsilon)$. □

With the help of the following lemma, we can often estimate the modulus of convexity.

Lemma A.6.19. *Let $I \subset \mathbb{R}$ be a symmetric interval, i.e., $x \in I$ implies $-x \in I$. Then, for all strictly convex, symmetric $f : I \to [0, \infty)$ and all $\varepsilon > 0$, we have*

$$\delta_f(2\varepsilon) = \inf_{\substack{x \geq 0 \\ x+\varepsilon \in I}} \frac{f(x-\varepsilon) + f(x+\varepsilon)}{2} - f(x) = \frac{1}{2} \inf_{\substack{x \geq 0 \\ x+\varepsilon \in I}} \int_x^{x+\varepsilon} \big(f'(t) - f'(t-\varepsilon)\big)dt\,.$$

Furthermore, if $I = \mathbb{R}$, then, for all $x \geq 12\varepsilon$, we have

$$f(x) \geq \frac{\delta_f(2\varepsilon)x^2}{8\varepsilon^2}\,.$$

Proof. The first equation follows from Lemma A.6.18 and the symmetry assumptions. Furthermore, by Theorem A.6.6, we obtain

$$f(x+\varepsilon) + f(x-\varepsilon) = 2f(x) + \int_x^{x+\varepsilon} \big(f'(t) - f'(t-\varepsilon)\big)dt\,, \tag{A.61}$$

and hence the second equation follows. Finally, in order to show the last assertion, we first observe that f has a minimum at 0, and hence we have $f'(t) \geq 0$ for all $t \geq 0$ for which the derivative exists. We write $b := 2\delta_f(2\varepsilon)$, and $x_n := 2\varepsilon n$ for $n \geq 1$. These definitions together with (A.61) yield real numbers $t_n \in [x_n, x_n + \varepsilon]$, $n \geq 1$, that satisfy

$$b \leq \int_{x_n}^{x_n+\varepsilon} \big(f'(t) - f'(t-\varepsilon)\big)dt \;\leq\; \varepsilon\big(f'(t_n) - f'(t_n-\varepsilon)\big)\,.$$

Therefore we obtain $f'(t_n) \geq f'(t_n - \varepsilon) + \frac{b}{\varepsilon}$ for all $n \geq 1$. Furthermore, we have $t_n - \varepsilon \geq x_n - \varepsilon = x_{n-1} + \varepsilon \geq t_{n-1}$ and hence $f'(t_n - \varepsilon) \geq f'(t_{n-1})$, $n \geq 2$. By induction, we thus find $f'(t_{n+1}) \geq f'(t_1) + \frac{bn}{\varepsilon} \geq \frac{bn}{\varepsilon}$ for all $n \geq 1$. Now let $t \geq 6\varepsilon$ be a real number at which $f'(t)$ exists. Then there is an $n \geq 3$ with $2\varepsilon n \leq t < 2\varepsilon(n+1)$, and hence we get

$$f'(t) \geq f'(t_{n-1}) \geq \frac{b(n-2)}{\varepsilon} > \frac{b(t-6\varepsilon)}{2\varepsilon^2}\,.$$

Consequently, for $x \geq 12\varepsilon$, Theorem A.6.6 gives

$$f(x) = f(6\varepsilon) + \int_{6\varepsilon}^{x} f'(t)dt \geq \frac{b}{2\varepsilon^2} \int_{6\varepsilon}^{x} (t - 6\varepsilon)dt = \frac{b(x-6\varepsilon)^2}{4\varepsilon^2} \geq \frac{bx^2}{16\varepsilon^2}\,. \qquad \square$$

A.6.4 The Fenchel-Legendre Bi-conjugate

In this subsection, we establish some important properties of the Fenchel-Legendre bi-conjugate operation ** defined in Definition 3.20. To this end, we first note that the Fenchel-Legendre bi-conjugate $g^{**} : I \to [0,\infty)$ of a function $g : I \to [0,\infty)$ is determined by (see, e.g., p. 474 of Rockafellar and Wets, 1998)

$$\operatorname{Epi} g^{**} = \overline{\operatorname{co} \operatorname{Epi} g}\,,$$

where $\operatorname{Epi} g := \{(t,y) \in I \times [0,\infty) : g(t) \le y\}$ denotes the **epigraph** of g and $\operatorname{co} A$ denotes the convex hull of a set A. Now our first result reads as follows.

Lemma A.6.20. *Let $B > 0$ and $g : [0,B] \to [0,\infty)$ be an increasing function with $g(0) = 0$ and $g(t) > 0$ for all $t \in (0,B]$. Then the Fenchel-Legendre bi-conjugate $g^{**} : [0,B] \to [0,\infty)$ of g satisfies $g^{**}(t) > 0$ for all $t \in [0,B]$.*

Proof. Let us assume that there exists an $0 < t \le B$ with $g^{**}(t) = 0$. Then we have $(t,0) \in \operatorname{Epi} g^{**} = \overline{\operatorname{co} \operatorname{Epi} g}$, and hence there exists a sequence $(t_n, y_n) \in \operatorname{co} \operatorname{Epi} g$ with $t_n \to t$ and $y_n \to 0$. Furthermore, we have $\operatorname{co} \operatorname{Epi} g \subset \mathbb{R}^2$, and hence Carathéodory's theorem (see, e.g., Rockafellar and Wets, 1998, p. 55) guarantees that for all $n \ge 1$ there exist $t_{n,1}, t_{n,2}, t_{n,3} \in [0,B]$, $y_{n,1}, y_{n,2}, y_{n,3} \in [0,\infty)$, and $\alpha_{n,1}, \alpha_{n,2}, \alpha_{n,3} \in [0,1]$ with

$$\begin{aligned} t_n &= \alpha_{n,1} t_{n,1} + \alpha_{n,2} t_{n,2} + \alpha_{n,3} t_{n,3}\,, \\ y_n &= \alpha_{n,1} y_{n,1} + \alpha_{n,2} y_{n,2} + \alpha_{n,3} y_{n,3}\,, \\ 1 &= \alpha_{n,1} + \alpha_{n,2} + \alpha_{n,3}\,, \\ y_{n,i} &\ge g(t_{n,i})\,, \qquad\qquad i = 1,\ldots,3. \end{aligned}$$

In addition, we may assume $t_{n,1} \le t_{n,2} \le t_{n,3}$ without loss of generality. Since this yields $t_n = \alpha_{n,1} t_{n,1} + \alpha_{n,2} t_{n,2} + \alpha_{n,3} t_{n,3} \le t_{n,3}$ we find $y_{n,3} \ge g(t_{n,3}) \ge g(t_n) \ge g(\frac{t}{2}) > 0$ for large n. Recalling $y_n \to 0$, we thus obtain $\alpha_{n,3} \to 0$, which implies both $\alpha_{n,1} + \alpha_{n,2} \to 1$ and $\alpha_{n,1} t_{n,1} + \alpha_{n,2} t_{n,2} \to t$. However, the latter convergence gives $\frac{t}{2} \le \alpha_{n,1} t_{n,1} + \alpha_{n,2} t_{n,2} \le (\alpha_{n,1} + \alpha_{n,2}) t_{n,2}$ for large n, and hence we have $t_{n,2} \ge \frac{t}{4}$ for large n. Again this shows $y_{n,2} \ge g(t_{n,2}) \ge g(\frac{t}{4}) > 0$ for large n, and thus we find $\alpha_{n,2} \to 0$. Obviously, this yields both $\alpha_{n,1} \to 1$ and $\alpha_{n,1} t_{n,1} \to t$, and hence we obtain $t_{n,1} \ge \frac{t}{4}$ for large n. Finally, this gives $y_{n,1} \ge g(t_{n,1}) \ge g(\frac{t}{4}) > 0$ for large n, and therefore we find $\alpha_{n,1} \to 0$, which contradicts the convergence $\alpha_{n,1} \to 1$ already found. □

Lemma A.6.21. *Let $B > 0$ and $g : [0,B] \to [0,\infty)$ be a continuous function with $g(0) = 0$. We define $\tilde{g} : [0,B] \to [0,\infty)$ by $\tilde{g}(t) := \inf_{t' \ge t} g(t)$, $t \in [0,B]$. Then $\tilde{g}$ is increasing, and, for all $t \in [0,B]$, we have*

$$g^{**}(t) = \tilde{g}^{**}(t)\,.$$

In addition, if $g(t) > 0$ for all $t \in (0,B]$, then $\tilde{g}(t) > 0$ for all $t \in (0,B]$.

Proof. The first assertion is trivial and the third assertion directly follows from the continuity of g. Therefore, it remains to show

$$\operatorname{co}\operatorname{Epi} g = \operatorname{co}\operatorname{Epi} \tilde{g} \tag{A.62}$$

since this equation immediately yields $g^{**} = \tilde{g}^{**}$. To establish (A.62), we first observe that $\tilde{g}(t) \leq g(t)$ for all $t \in [0, B]$ and hence we have $\operatorname{co}\operatorname{Epi} g \subset \operatorname{co}\operatorname{Epi} \tilde{g}$. To prove the converse inclusion, observe that it suffices to show $(t, \tilde{g}(t)) \in \operatorname{co}\operatorname{Epi} g$ for all $t \in [0, B]$. Furthermore, we have $\tilde{g}(0) = 0 = g(0)$, and $\tilde{g}(B) = g(B)$, and hence we can restrict our considerations to pairs $(t, g(t))$ for $t \in (0, B)$. Therefore let us fix an $t \in (0, B)$. By the definition of $\tilde{g}$, we then find an $t_+ \in [t, B]$ with $g(t_+) = \tilde{g}(t)$. Furthermore, we have $g(0) \leq \tilde{g}(t) \leq g(t)$, and hence the intermediate value theorem applied to the continuous function g gives us an $t_- \in [0, t]$ with $g(t_-) = \tilde{g}(t)$. Now, there exists an $\alpha \in [0, 1]$ with $t = \alpha t_+ + (1 - \alpha)t_-$, and since our previous considerations showed

$$\tilde{g}(t) = \alpha\tilde{g}(t) + (1 - \alpha)\tilde{g}(t) = \alpha g(t_+) + (1 - \alpha)g(t_-)\,,$$

we obtain $(t, \tilde{g}(t)) \in \operatorname{co}\operatorname{Epi} g$. □

A.6.5 Convex Programs and Lagrange Multipliers

The following results on convex programs and Lagrange multipliers are needed in Chapter 11, where we investigate methods to compute the decision functions $f_{\mathrm{D},\lambda}$ of general support vector machines. This subsection is based on Rockafellar (1970, Chapter 28).

Let $S \subset \mathbb{R}^n$ and $g : S \to \bar{\mathbb{R}}$ be a function. Then g is convex if the epigraph $\operatorname{Epi} g$ is convex as a subset of $\mathbb{R}^{n+1}$. The **effective domain** of a convex function g on S is the projection on $\mathbb{R}^n$ of the epigraph of g; i.e.,

$$\operatorname{dom} g = \left\{z \in S : \exists\, \mu \in \mathbb{R}, (z, \mu) \in \operatorname{Epi} g\right\} = \left\{z \in S : g(x) < +\infty\right\}.$$

A convex function $g : \mathbb{R}^n \to \mathbb{R}$ is called **proper** if its epigraph is non-empty and contains no vertical lines; i.e., if $g(z) < +\infty$ for at least one $z \in \mathbb{R}^n$ and $g(z) > -\infty$ for all $z \in \mathbb{R}^n$. The **relative interior** of a convex set $C \subset \mathbb{R}^n$ is defined as the interior that results when C is regarded as a subset of its affine hull $\operatorname{aff} C$; i.e.,

$$\operatorname{ri} C := \left\{z \in \operatorname{aff} C : \exists\, \varepsilon > 0, (z + \varepsilon B) \cap (\operatorname{aff} C) \subset C\right\}.$$

A set $A \subset \mathbb{R}^n$ is called **affine**, if $\alpha x + (1 - \alpha)y \in A$ for all $x, y \in A$ and all $\alpha \in \mathbb{R}$. The **affine hull** $\operatorname{aff} C$ of $C \subset \mathbb{R}^n$ is the smallest affine set that includes C. Let $A \subset \mathbb{R}^n$. A function $g : A \to \mathbb{R}$ is called **affine** if it is finite, convex, and concave; i.e., if there exists a vector $a \in \mathbb{R}^n$ and a constant $b \in \mathbb{R}$ such that $g(x) = a^{\mathsf{T}}x + b$ for all $x \in A$.

Definition A.6.22. *Let $n \in \mathbb{N}$ and $m, r \in \mathbb{N}_0$, $r \le m$. A **convex program** (P) is an $(m+3)$ tuple $(C, g_0, g_1, \ldots, g_m, r)$, where $C \subset \mathbb{R}^n$ is a non-empty convex set, $g_0, g_1, \ldots, g_r : C \to \mathbb{R}$ are finite convex functions, $g_{r+1}, \ldots, g_m : C \to \mathbb{R}$ are affine functions, and the optimization problem is given by*

$$\min_{z \in C} g_0(z)$$

subject to the constraints

$$g_1(z) \le 0, \ldots, g_r(z) \le 0, \quad g_{r+1}(z) = 0, \ldots, g_m(z) = 0. \tag{A.63}$$

We will assume that each function g_i, $i \in \{0, \ldots, m\}$, is defined on $\mathbb{R}^n$ in such a way that *(i)* g_0 is a proper convex function with $\operatorname{dom} g_0 = C$, *(ii)* $g_1, \ldots, g_r$ are proper convex functions with $\operatorname{ri}(\operatorname{dom} g_i) \supset \operatorname{ri} C$ and $\operatorname{dom} g_i \supset C$, and *(iii)* $g_1, \ldots, g_m$ are affine functions throughout $\mathbb{R}^n$ such that $g_1, \ldots, g_m$ are affine on C.

A vector z is called a **feasible solution** of the convex program (P) if $z \in C$ and z satisfies the constraints (A.63). The set of feasible solutions of (P) is denoted by C_0. Note that C_0 is a (possibly empty) convex set. The convex function $g : \mathbb{R}^n \to \mathbb{R}$ defined by

$$g(z) = g_0(z)\mathbf{1}_{C_0}(z) + \infty\, \mathbf{1}_{C_0^c}(z)$$

will be called the **objective function** for (P). Thus, minimizing g over $\mathbb{R}^n$ is equivalent to minimizing $g_0(x)$ over all feasible solutions $z \in C_0$. The infimum of g will be called the **optimal value** in (P). The points where the infimum of g is attained will be called the **optimal solutions** to (P) if $C_0 \neq \emptyset$. The set of all optimal solutions of (P) is thus a possibly empty convex subset of the set of all feasible solutions.

We call $\xi = (\xi_1, \ldots, \xi_m) \in \mathbb{R}^m$ a **Karush-Kuhn-Tucker (KKT) vector** for (P) if $\xi_i \ge 0$ for $i = 1, \ldots, r$ and the infimum of the proper convex function

$$g := g_0 + \sum_{i=1}^{m} \xi_i g_i$$

is finite and equal to the optimal value in (P).

Theorem A.6.23. *Let (P) be a convex program and let $\xi \in \mathbb{R}^m$ be a KKT vector for (P). Define $g := g_0 + \sum_{i=1}^m \xi_i g_i$ and $B = \arg\inf_{z \in \mathbb{R}^n} g(z)$. Let $I := \{i : 1 \le i \le r \text{ and } \xi_i = 0\}$ and $J := \{1, \ldots, m\} \setminus I$. Then*

$$B_0 := \left\{ \bar{z} \in B : g_i(\bar{z}) = 0 \text{ for } i \in J,\ g_i(\bar{z}) \le 0 \text{ for } i \in I \right\}$$

is the set of all optimal solutions of (P).

Proof. See Rockafellar (1970, Theorem 28.1). □

A function $g : \mathbb{R}^n \to \mathbb{R}^m$ is called **closed** if for any closed set $A \subset \mathbb{R}^n$ the image $g(A) \subset \mathbb{R}^m$ is closed.

Corollary A.6.24. *Let* (P) *be a convex program, and let* $\xi \in \mathbb{R}^m$ *be a KKT vector for* (P). *Assume that the functions* g_i *are all closed. If the infimum of* $g := g_0 + \sum_{i=1}^m \xi_i g_i$ *is attained at a unique point* $\bar{z}$, *this* $\bar{z}$ *is the unique optimal solution to* (P).

Proof. See Rockafellar (1970, Corollary 28.1.1). □

The next result shows that a KKT vector exists under mild conditions.

Theorem A.6.25. *Let* (P) *be a convex program, and let* I *be the set of indices* $i \neq 0$ *such that* g_i *is not affine. Assume that the optimal value in* (P) *is not* $-\infty$ *and that* (P) *has at least one feasible solution in* $\operatorname{ri} C$ *that satisfies with strict inequality all the inequality constraints for* $i \in I$. *Then a KKT vector (not necessarily unique) exists for* (P).

Proof. See Rockafellar (1970, Theorem 28.2). □

KKT vectors and optimal solutions in a convex program (P) can be characterized in terms of the saddle point extrema of the **Lagrangian**[3] $L^* : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ of (P), where

$$L^*(\alpha, z) = \begin{cases} g_0(z) + \sum_{i=1}^m \alpha_i g_i(z) & \text{if } \alpha \in E_r\,, z \in C \\ -\infty & \text{if } \alpha \notin E_r\,, z \in C \\ +\infty & \text{if } z \notin C \end{cases} \tag{A.64}$$

and $E_r := \{\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m : \alpha_i \geq 0, i = 1, \ldots, r\}$. The coefficient α_i is called the **Lagrange multiplier** associated with the i-th constraint in (P). The function L^* is concave in α for each z and convex in z for each α. The Lagrangian obviously contains all the structure of (P) because the $(m+3)$-tuple of (P) can be recovered completely from L^*.

A vector $(\bar{\alpha}, \bar{z})$ is said to be a **saddle point** of L^* with respect to maximizing in α and minimizing in z if for all $\alpha \in \mathbb{R}^m$ and all $z \in C$

$$L^*(\alpha, \bar{z}) \leq L^*(\bar{\alpha}, \bar{z}) \leq L^*(\bar{\alpha}, z).$$

The next result shows how the optimal solutions to (P) and KKT vectors for (P) can be characterized in terms of the Lagrangian L^*.

Theorem A.6.26. *Let* (P) *be a convex program,* $\bar{\alpha} \in \mathbb{R}^m$, *and* $\bar{z} \in \mathbb{R}^n$. *In order that* $\bar{\alpha}$ *be a KKT vector for* (P) *and* $\bar{z}$ *be an optimal solution to* (P), *it is necessary and sufficient that* $(\bar{\alpha}, \bar{z})$ *be a saddle point of the Lagrangian* L^* *of* (P). *Furthermore, this condition holds if and only if* $\bar{z}$ *and the components* α_i *of* $\bar{\alpha}$ *satisfy*

[3] We denote the Lagrangian by L^* instead of by the usual symbol L, because we denote loss functions by L.

i) $\alpha_i \geq 0$, $g_i(\bar{z}) \leq 0$, *and* $\alpha_i g_i(\bar{z}) = 0$, $i = 1, \ldots, r$,
ii) $g_i(\bar{z}) = 0$ *for* $i \in \{r+1, \ldots, m\}$,
iii) $0 \in \left[\partial g_0(\bar{z}) + \sum_{i=1}^{m} \alpha_i \partial g_i(\bar{z})\right]$. *(Omit terms with* $\alpha_i = 0$.*)*

Proof. See Rockafellar (1970, Theorem 28.3). □

It follows that, provided the KKT vectors are known to exist, solving the constrained minimization problem in (P) is equivalent to finding a saddle point of L^*. The so-called **Karush-Kuhn-Tucker (KKT) conditions** state that, at the point of the solution of the convex program, the product between the dual variables and the constraints has to vanish.

Corollary A.6.27 (Karush-Kuhn-Tucker theorem). *Let* (P) *be a convex program satisfying the assumptions of Theorem A.6.25. In order that a given vector* $\bar{z}$ *be an optimal solution to* (P)*, it is necessary and sufficient that there exists a vector* $\bar{\alpha}$ *such that* $(\bar{\alpha}, \bar{z})$ *is a saddle point of the Lagrangian* L^* *of* (P)*. Equivalently,* $\bar{z}$ *is an optimal solution if and only if there exist Lagrange multiplier values* α_i *that, together with* $\bar{z}$*, satisfy the KKT conditions for* (P)*.*

Proof. See Rockafellar (1970, Corollary 28.3.1). □

We mention that the KKT conditions of a convex program can also be derived from the theory of subdifferentiation. The next theorem shows how the optimal value in (P) can be characterized in terms of the Lagrangian L^*.

Theorem A.6.28. *Let* (P) *be a convex program with Lagrangian* L^**. If* $\bar{\alpha}$ *is a KKT vector for* (P) *and* $\bar{z}$ *is an optimal solution, the saddle value* $L^*(\bar{\alpha}, \bar{z})$ *is the optimal value in* (P)*. More generally,* $\bar{\alpha}$ *is a KKT vector for* (P) *if and only if*

$$-\infty < \inf_{z \in \mathbb{R}^n} L^*(\bar{\alpha}, z) = \sup_{\alpha \in \mathbb{R}^m} \inf_{z \in \mathbb{R}^n} L^*(\alpha, z) = \inf_{z \in \mathbb{R}^n} \sup_{\alpha \in \mathbb{R}^m} L^*(\alpha, z),$$

in which case the common extremum value in the latter equation is the optimal value in (P)*.*

Proof. See Rockafellar (1970, Theorem 28.4). □

The following result shows that the problem of determining a KKT vector for a convex problem (P) can be reduced to the numerical problem of maximizing a certain concave function g on $\mathbb{R}^m$.

Corollary A.6.29. *Let* (P) *be a convex program with Lagrangian* L^* *having at least one KKT vector. Let* $h : \mathbb{R}^m \to \mathbb{R}$ *be the concave function defined by*

$$h(\alpha) = \inf_{z \in \mathbb{R}^m} L^*(\alpha, z),$$

where L^* *is the Lagrangian of* (P)*. The KKT vectors for* (P) *are then the points* $\bar{\alpha}$ *where* h *attains its supremum over* $\mathbb{R}^m$*.*

Proof. See Rockafellar (1970, Corollary 28.4.1). □

A.7 Complex Analysis

In this section, we briefly recall some facts from complex analysis. We begin with the basic definition.

Definition A.7.1. *Let $D \subset \mathbb{C}^d$ be an open subset and $f : D \to \mathbb{C}$ be a function. We say that f is **holomorphic** at the point $z_0 \in D$ if*

$$f'(z_0) := \lim_{z \to z_0} \frac{f(z_0) - f(z)}{z_0 - z}$$

*exists. Moreover, f is called **holomorphic** if it is holomorphic at every $z_0 \in D$. Finally, f is called an **entire function** if f is holomorphic and $D = \mathbb{C}^d$.*

The following result shows that the set of holomorphic functions over D is closed under compact convergence.

Theorem A.7.2. *Let $D \subset \mathbb{C}^d$ be an open subset and $f_n : D \to \mathbb{C}$, $n \geq 1$, be holomorphic functions. Furthermore, let $f : D \to \mathbb{C}$ be a function such that*

$$\lim_{n \to 0} \sup_{z \in K} |f_n(z) - f(z)| = 0$$

for all compact subsets $K \subset D$. Then f is holomorphic.

Proof. See p. 10 of the book by Range (1986). □

Finally, we recall a simple version of Hardy's convexity theorem, which can be found on p. 9 of the book by Duren (1970).

Theorem A.7.3 (Hardy's convexity theorem). *Let $\mathring{B}_{\mathbb{C}^d}$ be the open unit ball of $\mathbb{C}^d$ and $f : \mathring{B}_{\mathbb{C}^d} \to \mathbb{C}$ be a holomorphic function. Then*

$$r \mapsto \frac{1}{2\pi} \int_0^{2\pi} \left| f(re^{i\theta}) \right|^2 d\theta$$

is non-decreasing on $[0, 1)$.

A.8 Inequalities Involving Rademacher Sequences

In this section, we present some results on Rademacher sequences and empirical Rademacher averages.

In the following, Z always denotes a non-empty set equipped with some σ-algebra. Moreover, whenever we consider a distribution on Z, it is supposed to live on this σ-algebra. Furthermore, given a metric space (T, d), we call $(h_t)_{t \in T} \subset \mathcal{L}_0(Z)$ a **Carathéodory family** if $t \mapsto h_t(z)$ is continuous for all $z \in Z$. Moreover, if T is separable or complete, we say that $(h_t)_{t \in T}$ is a **separable** or **complete Carathéodory family**, respectively. In addition,

we call a subset $\mathcal{H} \subset \mathcal{L}_0(Z)$ a (separable or complete) **Carathéodory set** if there exists a (separable or complete) metric space (T, d) and a Carathéodory family $(h_t)_{t\in T} \subset \mathcal{L}_0(Z)$ such that $\mathcal{H} = \{h_t : t \in T\}$.

Furthermore, a sequence $\varepsilon_1, \ldots, \varepsilon_n$ of independent random variables defined on some probability space $(\Theta, \mathcal{C}, \nu)$ is called a **Rademacher sequence** with respect to ν if $\nu(\varepsilon_i = 1) = \nu(\varepsilon_i = -1) = 1/2$ for all $i = 1, \ldots, n$.

The first result, whose proof (besides different measurability notions) is essentially a copy of the proof of Lemma 2.3.1 of van der Vaart and Wellner (1996), recalls the so-called symmetrization procedure.

Theorem A.8.1 (Symmetrization). *Let $\Psi : [0, \infty) \to [0, \infty)$ be a convex and non-decreasing function. Furthermore, let E be a separable Banach space, $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, and $\xi_1, \ldots, \xi_n : \Omega \to E$ be i.i.d. P-integrable random variables. Finally, let $\varepsilon_1, \ldots, \varepsilon_n$ be a Rademacher sequence with respect to some distribution ν. Then we have*

$$\mathbb{E}_{\mathrm{P}}\, \Psi\biggl(\Bigl\| \frac{1}{n} \sum_{i=1}^{n} \bigl(\xi_i - \mathbb{E}_{\mathrm{P}}\xi_i\bigr) \Bigr\| \biggr) \;\le\; \mathbb{E}_{\mathrm{P}} \mathbb{E}_{\nu} \Psi\biggl(2 \Bigl\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \xi_i \Bigr\| \biggr) .$$

Proof. For fixed $\omega \in \Omega$, we have

$$\Bigl\| \frac{1}{n} \sum_{i=1}^{n} \bigl(\xi_i(\omega) - \mathbb{E}_{\mathrm{P}}\xi_i\bigr) \Bigr\| \le \int_{\Omega} \Bigl\| \frac{1}{n} \sum_{i=1}^{n} \bigl(\xi_i(\omega) - \xi_i(\omega')\bigr) \Bigr\| \, d\mathrm{P}(\omega') .$$

The monotonicity and convexity of Ψ together with Jensen's inequality hence imply

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}}\, \Psi\biggl(\Bigl\| \frac{1}{n} \sum_{i=1}^{n} \bigl(\xi_i(\omega) - \mathbb{E}_{\mathrm{P}}\xi_i\bigr) \Bigr\| \biggr) \\
&\le \int_{\Omega} \Psi\biggl(\int_{\Omega} \Bigl\| \frac{1}{n} \sum_{i=1}^{n} \bigl(\xi_i(\omega) - \xi_i(\omega')\bigr) \Bigr\| \, d\mathrm{P}(\omega') \biggr) d\mathrm{P}(\omega) \\
&\le \mathbb{E}_{\mathrm{P}\otimes\mathrm{P}} \Psi\left(\Bigl\| \frac{1}{n} \sum_{i=1}^{n} \bigl(\xi_i \circ \pi_1 - \xi_i \circ \pi_2\bigr) \Bigr\| \right),
\end{aligned}$$

where $\pi_j : \Omega \times \Omega \to \Omega$, $j = 1, 2$, denotes the j-th-coordinate projection. Moreover, the independence of $\xi_1 \circ \pi_1, \ldots, \xi_n \circ \pi_1, \xi_1 \circ \pi_2, \ldots, \xi_n \circ \pi_2$ together with $\varepsilon_1, \ldots, \varepsilon_n \in \{-1, 1\}$ yields

$$\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}} \Psi\biggl(\Bigl\| \frac{1}{n} \sum_{i=1}^{n} \bigl(\xi_i \circ \pi_1 - \xi_i \circ \pi_2\bigr) \Bigr\| \biggr) = \mathbb{E}_{\mathrm{P}\otimes\mathrm{P}} \Psi\biggl(\Bigl\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \bigl(\xi_i \circ \pi_1 - \xi_i \circ \pi_2\bigr) \Bigr\| \biggr) .$$

Using the monotonicity and convexity of Ψ, we hence obtain

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Psi\biggl(\Bigl\|\frac{1}{n}\sum_{i=1}^n\bigl(\xi_i\circ\pi_1-\xi_i\circ\pi_2\bigr)\Bigr\|\biggr)\\
&\leq \mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Psi\biggl(\Bigl\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\cdot(\xi_i\circ\pi_1)\Bigr\|+\Bigl\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\cdot(\xi_i\circ\pi_2)\Bigr\|\biggr)\\
&\leq \frac{1}{2}\,\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Psi\biggl(2\Bigl\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\cdot(\xi_i\circ\pi_1)\Bigr\|\biggr)+\frac{1}{2}\,\mathbb{E}_{\mathrm{P}\otimes\mathrm{P}}\Psi\biggl(2\Bigl\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\cdot(\xi_i\circ\pi_2)\Bigr\|\biggr)\\
&= \mathbb{E}_{\mathrm{P}}\Psi\biggl(2\Bigl\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\xi_i\Bigr\|\biggr)\,.
\end{aligned}
$$

By averaging over $\varepsilon_1,\ldots,\varepsilon_n$ and changing the integration order, we then obtain the assertion. □

Corollary A.8.2. *Let $(\Omega,\mathcal{A},\mathrm{P})$ be a probability space, Z be a measurable space, and $\xi_1,\ldots,\xi_n:\Omega\to Z$ be i.i.d. random variables. Moreover, let $\mathcal{H}\subset\mathcal{L}_\infty(Z)$ be a separable Carathéodory set with $\sup_{h\in\mathcal{H}}\|h\|_\infty<\infty$. Finally, let $\varepsilon_1,\ldots,\varepsilon_n$ be a Rademacher sequence with respect to some distribution ν. Then we have*

$$
\mathbb{E}_{\mathrm{P}}\sup_{h\in\mathcal{H}}\biggl|\frac{1}{n}\sum_{j=1}^n\bigl(h(\xi_i)-\mathbb{E}_{\mathrm{P}}h(\xi_i)\bigr)\biggr| \;\leq\; 2\mathbb{E}_{\mathrm{P}}\mathbb{E}_\nu\sup_{h\in\mathcal{H}}\biggl|\frac{1}{n}\sum_{i=1}^n\varepsilon_i h(\xi_i)\biggr|\,. \tag{A.65}
$$

Proof. By a simple limit argument, we can obviously assume without loss of generality that $\mathcal{H}$ is finite. We write $E:=\ell_\infty(\mathcal{H})$, i.e., E is the vector space of functions $g:\mathcal{H}\to\mathbb{R}$ equipped with the supremum norm

$$
\|g\|_\infty:=\sup_{h\in\mathcal{H}}|g(h)|\,,\qquad\qquad g\in E\,.
$$

Then E is a finite-dimensional Banach space and hence separable. For $\omega\in\Omega$ and $i=1,\ldots,n$, we further define $\eta_i(\omega)\in E$ by $\eta_i(\omega)(h):=h(\xi_i(\omega))$, $h\in\mathcal{H}$. Then $\eta_1,\ldots,\eta_n:\Omega\to E$ are i.i.d. P-integrable random variables. Moreover, we obviously have

$$
\biggl\|\frac{1}{n}\sum_{i=1}^n(\eta_i-\mathbb{E}_{\mathrm{P}}\eta_i)\biggr\|_E \;=\; \sup_{h\in\mathcal{H}}\biggl|\frac{1}{n}\sum_{j=1}^n\bigl(h(\xi_i)-\mathbb{E}_{\mathrm{P}}h(\xi_i)\bigr)\biggr|\,,
$$

and hence we obtain the assertion by Theorem A.8.1. □

The following inequality, whose proof can be found in Chapter 11 of Diestel *et al.* (1995), shows that we can replace the $L_1(\nu)$-norm on the right-hand side of the symmetrization inequality by any other $L_p(\nu)$-norm.

Theorem A.8.3 (Kahane's inequality). *Let $\varepsilon_1,\ldots,\varepsilon_n$ be a Rademacher sequence with respect to some distribution ν. Then, for all $p,q\in(0,\infty)$, there*

exists a constant $K_{p,q} > 0$ independent of n such that, for all Banach spaces E and all $x_1, \ldots, x_n \in E$, we have

$$\left(\mathbb{E}_\nu \Big\| \sum_{i=1}^n \varepsilon_i x_i \Big\|^p\right)^{1/p} \leq K_{p,q} \left(\mathbb{E}_\nu \Big\| \sum_{i=1}^n \varepsilon_i x_i \Big\|^q\right)^{1/q}.$$

The following result compares the Rademacher average of a composed function class with the Rademacher average of the original function class.

Theorem A.8.4 (Contraction principle). *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous function with $|\varphi|_1 \leq 1$ and $\varphi(0) = 0$, and let $\Psi : [0,\infty) \to [0,\infty)$ be a convex and non-decreasing function. Furthermore, let $\mathcal{H} \subset \mathcal{L}_0(Z)$ be a separable Carathéodory set satisfying $\sup_{h\in\mathcal{H}} |h(z)| < \infty$ for all $z \in Z$. Moreover, let $\varepsilon_1, \ldots, \varepsilon_n$ be a Rademacher sequence with respect to some distribution ν, and $D := (z_1, \ldots, z_n) \in Z^n$. Then we have*

$$\mathbb{E}_\nu \Psi\left(\frac{1}{2} \sup_{h\in\mathcal{H}} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi\big(h(z_i)\big)\right|\right) \leq \mathbb{E}_\nu \Psi\left(\sup_{h\in\mathcal{H}} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i h(z_i)\right|\right).$$

In particular, for $\Psi = \mathrm{id}_{[0,\infty)}$, we obtain $\mathrm{Rad}_D(\varphi \circ \mathcal{H}, n) \leq 2\mathrm{Rad}_D(\mathcal{H}, n)$.

Proof. Apply Theorem 4.12 of Ledoux and Talagrand (1991) to the set $T := \{(h(z_1), \ldots, h(z_n)) : h \in \mathcal{H}\}$. □

With the contraction principle, we can show the following corollary, which will be useful when bounding Rademacher averages from above.

Corollary A.8.5. *Let $\mathcal{H} \subset \mathcal{L}_0(Z)$ be a separable Carathéodory set and P be a distribution on Z. Suppose that there exist constants $B \geq 0$ and $\sigma \geq 0$ such that $\|h\|_\infty \leq B$ and $\mathbb{E}_\mathrm{P} h^2 \leq \sigma^2$ for all $h \in \mathcal{H}$. Then we have*

$$\mathbb{E}_{(z_1,\ldots,z_n)\sim \mathrm{P}^n} \sup_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^n h^2(z_i) \leq \sigma^2 + 8\, B\, \mathbb{E}_{D\sim\mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}, n).$$

Proof. For $h_0 \in \mathcal{H}$, we have

$$\frac{1}{n} \sum_{i=1}^n h_0^2(z_i) \leq \left|\frac{1}{n} \sum_{i=1}^n h_0^2(z_i) - \mathbb{E}_\mathrm{P} h_0^2\right| + \mathbb{E}_\mathrm{P} h_0^2 \leq \sup_{h\in\mathcal{H}} \left|\frac{1}{n} \sum_{i=1}^n h^2(z_i) - \mathbb{E}_\mathrm{P} h^2\right| + \sigma^2.$$

Taking the supremum over all $h_0 \in \mathcal{H}$ on the left-hand side of this inequality and applying Corollary A.8.2, we hence find

$$\begin{aligned}\mathbb{E}_{(z_1,\ldots,z_n)\sim \mathrm{P}^n} \sup_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^n h^2(z_i) &\leq \mathbb{E}_{(z_1,\ldots,z_n)\sim \mathrm{P}^n} \sup_{h\in\mathcal{H}} \left|\frac{1}{n} \sum_{i=1}^n h^2(z_i) - \mathbb{E}_\mathrm{P} h^2\right| + \sigma^2 \\ &\leq 2\mathbb{E}_{(z_1,\ldots,z_n)\sim \mathrm{P}^n} \mathbb{E}_\nu \sup_{h\in\mathcal{H}} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i h^2(z_i)\right| + \sigma^2.\end{aligned}$$

Let us define $\varphi : \mathbb{R} \to \mathbb{R}$ by

$$\varphi(t) := \begin{cases} -t - \frac{B}{2} & \text{if } t < -B \\ \frac{t^2}{2B} & \text{if } t \in [-B, B] \\ t - \frac{B}{2} & \text{if } t > B\,. \end{cases}$$

Obviously, φ is a convex function and hence Lemma A.6.8 shows that φ is Lipschitz continuous with $|\varphi|_1 = 1$. Applying Theorem A.8.4 now yields

$$\begin{aligned} \mathbb{E}_{(z_1,\ldots,z_n)\sim \mathrm{P}^n} \mathbb{E}_\nu \sup_{h\in\mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i h^2(z_i) \right| &= 2B \mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_D(\varphi \circ \mathcal{H}, n) \\ &\leq 4B \mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}, n), \end{aligned}$$

and combining this estimate with the previous one, we find the assertion. □

A.9 Talagrand's Inequality

In this section, we prove the following concentration inequality, which is due to Talagrand.

Theorem A.9.1 (Talagrand's inequality). *Let $B \geq 0$, $\sigma \geq 0$, and $n \geq 1$. Moreover, let $(\Omega, \mathcal{A}, \mu)$ be a probability space and $\mathcal{F} \subset \mathcal{L}_0(\Omega)$ be a countable subset such that $\mathbb{E}_\mu f = 0$, $\mathbb{E}_\mu f^2 \leq \sigma^2$, and $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. We write $Z := \Omega^n$ and $\mathrm{P} := \mu^n$. Furthermore, we define $g : Z \to \mathbb{R}$ by*

$$g(\omega_1, \ldots, \omega_n) := \sup_{f\in\mathcal{F}} \left| \sum_{j=1}^n f(\omega_j) \right|, \qquad z = (\omega_1, \ldots, \omega_n) \in Z.$$

Then, for all $\tau > 0$, we have

$$\mathrm{P}\left(\left\{ z \in Z : g(z) \geq \mathbb{E}_\mathrm{P} g + \sqrt{2\tau(n\sigma^2 + 2B\mathbb{E}_\mathrm{P} g)} + \frac{2\tau B}{3} \right\} \right) \;\leq\; e^{-\tau}\,.$$

The proof of this inequality requires quite a few preparations, which mainly deal with estimating $\mathbb{E}_\mathrm{P} e^{\lambda g}$. Let us begin by introducing some basic concepts.

Definition A.9.2. *A function $\Psi : [0, \infty) \to \mathbb{R}$ is called an **entropy function** if it is strictly convex, continuous, and bounded from below. Moreover, for $k \geq 1$, it is said to be a **k-times continuously differentiable entropy function** if it is an entropy function and $\Psi_{|(0,\infty)}$ is k-times continuously differentiable.*

For our purposes, the most important entropy function is the function $\Psi : [0, \infty) \to \mathbb{R}$ defined by $\Psi(0) := 0$ and $\Psi(t) := t \ln t$ for $t > 0$. Recall that this function is used to define the **Shannon entropy**, which is a central concept

in information theory (see also the definition in (A.66), which is related to this concept). In addition to this classical entropy function, we will also need the entropy function $\Psi : [0,\infty) \to \mathbb{R}$ defined by $\Psi(t) := t^2$, $t \geq 0$.

With the help of an entropy function, we can define a functional on the set of all non-negative, P-integrable functions.

Definition A.9.3. *For a given probability space $(\Omega, \mathcal{A}, \mathrm{P})$, we write $L_1^+(\mathrm{P}) := \{f : \Omega \to [0,\infty) \,|\, f \in L_1(\mathrm{P})\}$. Then, for an entropy function Ψ, we define the* ***Ψ-entropy functional*** *by*

$$H_{\Psi,\mathrm{P}}(f) := \mathbb{E}_\mathrm{P}\, \Psi\circ f - \Psi(\mathbb{E}_\mathrm{P} f)\,.$$

Note that the (not necessarily finite) expectation $\mathbb{E}_\mathrm{P}\, \Psi\circ f$ in the definition of $H_{\Psi,\mathrm{P}}$ is actually well-defined since Ψ is assumed to be bounded from below. Moreover, $\Psi(\mathbb{E}_\mathrm{P} f)$ is always finite, and hence the difference in the definition of the entropy functional is well-defined. Finally, the convexity of Ψ together with Jensen's inequality gives $H_{\Psi,\mathrm{P}}(f) \geq 0$ for all $f \in L_1^+(\mathrm{P})$.

Obviously, the function $\Psi : [0,\infty) \to \mathbb{R}$ defined by $\Psi(t) := t \ln t$ for $t > 0$ and $\Psi(0) := 0$ gives the entropy functional

$$H_{\Psi,\mathrm{P}}(f) = \mathbb{E}_\mathrm{P}(f \ln f) - (\mathbb{E}_\mathrm{P} f) \cdot \ln(\mathbb{E}_\mathrm{P} f)\,, \qquad f \in L_1^+(\mathrm{P})\,, \tag{A.66}$$

which equals Shannon's entropy up to the term $(\mathbb{E}_\mathrm{P} f) \cdot \ln(\mathbb{E}_\mathrm{P} f)$. Moreover, the function $\Psi : [0,\infty) \to \mathbb{R}$ defined by $\Psi(t) := t^2$, $t \geq 0$, gives the variance

$$H_{\Psi,\mathrm{P}}(f) = \mathbb{E}_\mathrm{P} f^2 - (\mathbb{E}_\mathrm{P} f)^2\,, \qquad f \in L_1^+(\mathrm{P}).$$

The following theorem presents a sufficient condition for ensuring the convexity of certain entropy functionals. It shows in particular that both (A.66) and the variance are convex functionals.

Lemma A.9.4 (Convexity of the entropy functional). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and $\Psi : [0,\infty) \to \mathbb{R}$ be a twice continuously differentiable entropy function such that $1/\Psi'' : (0,\infty) \to (0,\infty)$ is concave. Then the Ψ-entropy functional $H_{\Psi,\mathrm{P}} : L_1^+(\mathrm{P}) \to [0,\infty]$ is convex.*

Proof. We first show the assertion for *bounded* functions. To this end, let us fix two bounded functions $f, g \in L_1^+(\mathrm{P})$. For $t \in (0,1)$, we define $\alpha_t := tf + (1-t)g$ and

$$h(t) := H_{\Psi,\mathrm{P}}\big(\alpha_t\big) = \mathbb{E}_\mathrm{P}\Psi\circ\alpha_t - \Psi(\mathbb{E}_\mathrm{P}\alpha_t)\,.$$

For later use, note that $f \geq 0$ and $g \geq 0$ imply $\alpha_t \geq 0$ and $\mathbb{E}_\mathrm{P}\alpha_t \geq 0$ for all $t \in (0,1)$. Moreover, using the concavity of $1/\Psi'' : (0,\infty) \to (0,\infty)$, it is easy to check that $\Psi''(0) := \lim_{s\to 0} \Psi''(s)$ exists and satisfies $\Psi''(0) \neq 0$. By a simple limit argument, it is then trivial to prove that the extended function $1/\Psi'' : [0,\infty) \to [0,\infty)$ is concave.

Since f and g were taken arbitrarily, it is now straightforward to see that it suffices to show the convexity of $h : (0,1) \to [0,\infty)$. To this end, let us fix

an $\omega \in \Omega$ and write $\beta(t) := \Psi(tf(\omega)+(1-t)g(\omega))$, $t \in (0,1)$. Then elementary calculus shows

$$\beta'(t) = \Psi'\big(tf(\omega) + (1-t)g(\omega)\big)\big(f(\omega) - g(\omega)\big) = \Psi' \circ \alpha_t(\omega) \cdot \big(f(\omega) - g(\omega)\big)$$

and

$$\beta''(t) = \Psi''\big(tf(\omega) + (1-t)g(\omega)\big)\big(f(\omega) - g(\omega)\big)^2 = \Psi' \circ \alpha_t(\omega) \cdot \big(f(\omega) - g(\omega)\big)^2$$

for all $t \in (0,1)$. Since f and g were assumed to be bounded, we hence find $h'(t) = \mathbb{E}_{\mathrm{P}}(\Psi' \circ \alpha_t \cdot (f-g)) - \Psi'(\mathbb{E}_{\mathrm{P}}\alpha_t) \cdot \mathbb{E}(f-g)$ and

$$h''(t) = \mathbb{E}_{\mathrm{P}}\big(\Psi'' \circ \alpha_t \cdot (f-g)^2\big) - \Psi''\big(\mathbb{E}_{\mathrm{P}}\alpha_t\big) \cdot \big(\mathbb{E}(f-g)\big)^2 \tag{A.67}$$

for all $t \in (0,1)$. Moreover, using the concavity of the extended function $1/\Psi'' : [0,\infty) \to [0,\infty)$ and Jensen's inequality, we get

$$\mathbb{E}_{\mathrm{P}}\Big(\frac{1}{\Psi'' \circ \alpha_t}\Big) \;\le\; \frac{1}{\Psi''(\mathbb{E}_{\mathrm{P}}\alpha_t)}\,, \qquad\qquad t \in (0,1),$$

and the latter is equivalent to $\Psi''(\mathbb{E}_{\mathrm{P}}\alpha_t) \le (\mathbb{E}_{\mathrm{P}}(\Psi'' \circ \alpha_t)^{-1})^{-1}$. Consequently, we obtain

$$\Psi''\big(\mathbb{E}_{\mathrm{P}}\alpha_t\big) \cdot \big(\mathbb{E}(f-g)\big)^2 \le \frac{(\mathbb{E}(f-g))^2}{\mathbb{E}_{\mathrm{P}}(\Psi'' \circ \alpha_t)^{-1}} \le \mathbb{E}_{\mathrm{P}}\big(\Psi'' \circ \alpha_t \cdot (f-g)^2\big)\,,$$

where we used $\mathbb{E}_{\mathrm{P}}(f-g) \le (\mathbb{E}(\Psi'' \circ \alpha_t)^{-1})^{1/2} \cdot (\mathbb{E}_{\mathrm{P}}(f-g)^2 \cdot \Psi'' \circ \alpha_t)^{1/2}$ and $\Psi'' \ge 0$. By (A.67), we then find $h''(t) \ge 0$ for all $t \in (0,1)$, and since this implies the convexity of $h : (0,1) \to [0,\infty)$, we have shown the assertion for bounded functions, i.e., we have

$$\begin{aligned} &\mathbb{E}_{\mathrm{P}}\Psi\big(tf + (1-t)g\big) - \Psi\big(\mathbb{E}(tf + (1-t)g)\big) \\ &\le t\mathbb{E}_{\mathrm{P}}(\Psi \circ f) - t\,\Psi(\mathbb{E}_{\mathrm{P}}f) + (1-t) \cdot \mathbb{E}_{\mathrm{P}}(\Psi \circ g) - (1-t) \cdot \Psi(\mathbb{E}_{\mathrm{P}}g) \end{aligned} \tag{A.68}$$

for all bounded $f, g \in L_1^+(\mathrm{P})$ and all $t \in (0,1)$.

In order to prove the general case, we fix arbitrary $f, g \in L_1^+(\mathrm{P})$. If $\Psi \circ f \notin L_1(\mathrm{P})$ or $\Psi \circ g \notin L_1(\mathrm{P})$, the inequality (A.68) is trivially satisfied. Therefore we can additionally assume $\Psi \circ f, \Psi \circ g \in L_1(\mathrm{P})$. Moreover, since Ψ is bounded from below, there exists a real number $a \ge 0$ such that $\Psi + a \ge 0$, and because of $H_{\Psi+a,\mathrm{P}} = H_{\Psi,\mathrm{P}}$, we may assume without loss of generality that $\Psi \ge 0$. Together with the convexity of Ψ, the latter assumption implies $\Psi(s) \le \Psi(0) + \Psi(r)$ for all $0 \le s \le r$. Let us now fix two sequences of bounded functions $(f_n), (g_n) \subset L_1^+(\mathrm{P})$ such that $f_n \le f$, $g_n \le g$ for all $n \ge 1$, and $\lim_{n\to\infty} f_n(\omega) = f(\omega)$, $\lim_{n\to\infty} g_n(\omega) = g(\omega)$ for P-almost all $\omega \in \Omega$. Since f_n and g_n are bounded, we have already established (A.68) for these functions. Moreover, we have $\Psi \circ f_n \le \Psi(0) + \Psi \circ f$ and $\Psi \circ g_n \le \Psi(0) + \Psi \circ g$, and since we assumed $\Psi \circ f, \Psi \circ g \in L_1(\mathrm{P})$, a simple application of Lebesgue's theorem now shows that (A.68) holds for f and g. □

The following lemma provides an alternative and often useful way to compute the entropy functional.

Lemma A.9.5 (Variational formula for the entropy functional). *Let $\Psi : [0,\infty) \to \mathbb{R}$ be a continuously differentiable entropy function and $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space. Then, for all $f \in L_1^+(\mathrm{P})$, we have*

$$H_{\Psi,\mathrm{P}}(f) \;=\; \inf_{a>0} \mathbb{E}_\mathrm{P}\big(\Psi \circ f - \Psi(a) - \Psi'(a)\cdot(f-a)\big)\,. \tag{A.69}$$

Before we prove this lemma, we note that we only consider $a > 0$ in (A.69) since formally we have not defined $\Psi'(0)$.

Proof. For $\Psi \circ f \notin L_1(\mathrm{P})$, the assertion is trivially satisfied, and hence we may assume without loss of generality that $\Psi \circ f \in L_1(\mathrm{P})$. Now recall (see Definition A.6.10 and Proposition A.6.12) that $\Psi'(a)$ satisfies the subdifferential inequality $\Psi'(a)(t-a) \le \Psi(t) - \Psi(a)$, $a > 0$, $t \ge 0$, and hence we have $\Psi'(a)(\mathbb{E}_\mathrm{P} f - a) \le \Psi(\mathbb{E}_\mathrm{P} f) - \Psi(a)$. This yields

$$H_{\Psi,\mathrm{P}}(f) \;=\; \mathbb{E}_\mathrm{P}\Psi\circ f - \Psi(\mathbb{E}_\mathrm{P} f) \;\le\; \mathbb{E}_\mathrm{P}\Psi\circ f - \Psi'(a)(\mathbb{E}_\mathrm{P} f - a) - \Psi(a)\,,$$

i.e., we have shown that the infimum in (A.69) is not smaller than $H_{\Psi,\mathrm{P}}(f)$. The converse inequality follows from considering $a_n := \mathbb{E}_\mathrm{P} f + 1/n$, $n \ge 1$. □

Our next goal is to investigate entropy functionals that are defined on product spaces. In order to make these considerations as transparent as possible, we need some additional notations. To this end, let $\Omega_1, \ldots, \Omega_n$ be non-empty sets. We write $Z := \Omega_1 \times \cdots \times \Omega_n$ and

$$Z_i' := \Omega_1 \times \cdots \times \Omega_{i-1} \times \Omega_{i+1} \times \cdots \times \Omega_n\,, \qquad i = 1, \ldots, n,$$

i.e., Z_i' is the space we obtain by omitting the i-th coordinate from Z. Moreover, $\pi_i' : Z \to Z_i'$ denotes the projection of Z onto Z_i'. Finally, for fixed $i \in \{1, \ldots, n\}$ and $z := (\omega_1, \ldots, \omega_n) \in Z$, the **replacement operator** $I_{i,z} : \Omega_i \to Z$ is defined by

$$I_{i,z}(\omega) := (\omega_1, \ldots, \omega_{i-1}, \omega, \omega_{i+1}, \ldots, \omega_n)\,, \qquad \omega \in \Omega_i\,.$$

With the help of these notations, we can now formulate the following lemma, which bounds the entropy functional of a function $g : Z \to \mathbb{R}$ by entropy functionals on the single coordinates.

Lemma A.9.6. *Let $\Psi : [0,\infty) \to \mathbb{R}$ be a twice continuously differentiable entropy function such that $1/\Psi'' : (0,\infty) \to (0,\infty)$ is concave. Moreover, let $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1, \ldots, n$, be probability spaces and $\mathrm{P} := \mu_1 \otimes \cdots \otimes \mu_n$ be the product measure on $Z := \Omega_1 \times \cdots \times \Omega_n$. Then we have*

$$H_{\Psi,\mathrm{P}}(g) \;\le\; \sum_{i=1}^n \int_Z H_{\Psi,\mu_i}(g \circ I_{i,z})\, d\mathrm{P}(z)\,, \qquad g \in L_1^+(\mathrm{P}).$$

Proof. The proof is based on induction over n. For $n = 1$, there is obviously nothing to prove. For the induction step from n to $n+1$, we write $Z_n := \Omega_1 \times \cdots \times \Omega_n$ and $Z_{n+1} := \Omega_1 \times \cdots \times \Omega_{n+1}$, as well as $\mathrm{P}_n := \mu_1 \otimes \cdots \otimes \mu_n$ and $\mathrm{P}_{n+1} := \mu_1 \otimes \cdots \otimes \mu_{n+1}$. Note that, unlike in the rest of this book, in this proof P_n and P_{n+1} do *not* denote empirical distributions. Furthermore, elements of Z_n are denoted by $z_n = (\omega_1, \ldots, \omega_n)$, and for elements of Z_{n+1} we write $z_{n+1} = (\omega_1, \ldots, \omega_{n+1})$. Note that these conventions give the identification $z_{n+1} = (z_n, \omega_{n+1})$, which we will heavily employ. Moreover, for a function $g \in L_1^+(\mathrm{P}_{n+1})$ and fixed $i = 1, \ldots, n$ and $z_n := (\omega_1, \ldots, \omega_n) \in Z_n$, we define $g_{i,z_n} : \Omega_i \times \Omega_{n+1} \to [0, \infty)$ by

$$g_{i,z_n}(\omega, \omega_{n+1}) := g(\omega_1, \ldots, \omega_{i-1}, \omega, \omega_{i+1}, \ldots, \omega_{n+1})$$

for all $\omega \in \Omega_i$ and $\omega_{n+1} \in \Omega_{n+1}$. Note that this definition gives the identities $g \circ I_{i,z_{n+1}} = g_{i,z_n}(\,\cdot\,, \omega_{n+1})$ for $i = 1, \ldots, n$ and $g \circ I_{n+1,z_{n+1}} = g(z_n, \,\cdot\,)$. Now observe that $g \in L_1^+(\mathrm{P}_{n+1})$ implies that $g(\,\cdot\,, \omega_{n+1}) \in L_1^+(\mathrm{P}_n)$ for μ_{n+1}-almost all $\omega_{n+1} \in \Omega_{n+1}$, and consequently the induction hypothesis gives

$$H_{\Psi,\mathrm{P}_n}\bigl(g(\,\cdot\,, \omega_{n+1})\bigr) \;\le\; \sum_{i=1}^{n} \int_{Z_n} H_{\Psi,\mu_i}\bigl(g_{i,z_n}(\,\cdot\,, \omega_{n+1})\bigr)\, d\mathrm{P}_n(z_n)$$

for μ_{n+1}-almost all $\omega_{n+1} \in \Omega_{n+1}$. In other words, we have

$$\begin{aligned}\int_{Z_n} \Psi\bigl(g(z_n, \omega_{n+1})\bigr)\, d\mathrm{P}_n(z_n) \le \sum_{i=1}^{n} \int_{Z_n} H_{\Psi,\mu_i}\bigl(g_{i,z_n}(\,\cdot\,, \omega_{n+1})\bigr)\, d\mathrm{P}_n(z_n) \\ + \Psi\biggl(\int_{Z_n} g(z_n, \omega_{n+1})\, d\mathrm{P}_n(z_n)\biggr)\end{aligned}$$

for μ_{n+1}-almost all $\omega_{n+1} \in \Omega_{n+1}$, and hence we obtain

$$\begin{aligned}\int_{Z_{n+1}} \Psi \circ g\, d\mathrm{P}_{n+1} &= \int_{\Omega_{n+1}} \int_{Z_n} \Psi\bigl(g(z_n, \omega_{n+1})\bigr)\, d\mathrm{P}_n(z_n)\, d\mu_{n+1}(\omega_{n+1}) \\ &\le \sum_{i=1}^{n} \int_{Z_{n+1}} H_{\Psi,\mu_i}\bigl(g \circ I_{i,z_{n+1}}\bigr)\, d\mathrm{P}_{n+1}(z_{n+1}) \\ &\quad + \int_{\Omega_{n+1}} \Psi\biggl(\int_{Z_n} g(z_n, \omega_{n+1})\, d\mathrm{P}_n(z_n)\biggr) d\mu_{n+1}(\omega_{n+1}). \qquad \text{(A.70)}\end{aligned}$$

Moreover, Lemma A.9.4 shows that $H_{\Psi,\mu_{n+1}} : L_1^+(\mu_{n+1}) \to [0, \infty]$ is a convex functional, and hence we have

$$\int_{Z_{n+1}} H_{\Psi,\mu_{n+1}}\bigl(g \circ I_{n+1,z_{n+1}}\bigr)\, d\mathrm{P}_{n+1}(z_{n+1}) \ge H_{\Psi,\mu_{n+1}}\biggl(\int_{Z_n} g(z_n, \,\cdot\,)\, d\mathrm{P}_n(z_n)\biggr).$$

Now the definition of $H_{\Psi,\mu_{n+1}}$ gives

$$H_{\Psi,\mu_{n+1}}\biggl(\int\limits_{Z_n} g(z_n,\,\cdot\,)\,d\mathrm{P}_n(z_n)\biggr) = \int\limits_{\Omega_{n+1}} \Psi\biggl(\int\limits_{Z_n} g(z_n,\omega_{n+1})\,d\mathrm{P}_n(z_n)\biggr) d\mu_{n+1}(\omega_{n+1}) - \Psi\biggl(\int_{Z_{n+1}} g(z_{n+1})\,d\mathrm{P}_{n+1}(z_{n+1})\biggr),$$

and consequently, combining the last two equations, we obtain

$$\begin{aligned}&\int_{\Omega_{n+1}} \Psi\biggl(\int_{Z_n} g(z_n,\omega_{n+1})\,d\mathrm{P}_n(z_n)\biggr)\,d\mu_{n+1}(\omega_{n+1})\\ &\le \int_{Z_{n+1}} H_{\Psi,\mu_{n+1}}\bigl(g\circ I_{n+1,z_{n+1}}\bigr)\,d\mathrm{P}_{n+1}(z_{n+1}) + \Psi\biggl(\int_{Z_{n+1}} g(z_{n+1})\,d\mathrm{P}_{n+1}(z_{n+1})\biggr).\end{aligned}$$

Combining this estimate with (A.70), then yields the assertion. □

With these preparations, we can now prove the main result on the entropy functional.

Theorem A.9.7 (Tensorization of the entropy functional). *For $n \in \mathbb{N}$, let $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1,\dots,n$, be probability spaces and $\Psi : [0,\infty) \to \mathbb{R}$ be a twice continuously differentiable entropy function such that $1/\Psi'' : (0,\infty) \to (0,\infty)$ is concave. Then, for $\mathrm{P} := \mu_1 \otimes \cdots \otimes \mu_n$ on $Z := \Omega_1 \times \cdots \times \Omega_n$, we have*

$$H_{\Psi,\mathrm{P}}(g) \;\le\; \sum_{i=1}^{n} \int_Z \Psi\circ g - \Psi\bigl(g_i\circ\pi_i'\bigr) - \Psi'\bigl(g_i\circ\pi_i'\bigr)\cdot\bigl(g - g_i\circ\pi_i'\bigr)\,d\mathrm{P}$$

for all $g \in L_1^+(\mathrm{P})$ and all measurable $g_i : Z_i' \to (0,\infty)$ with $\Psi\bigl(g_i\circ\pi_i'\bigr) \in L_1(\mathrm{P})$ and $\Psi'\bigl(g_i\circ\pi_i'\bigr)\cdot\bigl(g - g_i\circ\pi_i'\bigr) \in L_1(\mathrm{P})$.

Proof. By Lemma A.9.6, we have

$$H_{\Psi,\mathrm{P}}(g) \;\le\; \sum_{i=1}^{n} \int_Z H_{\Psi,\mu_i}(g\circ I_{i,z})\,d\mathrm{P}(z)\,.$$

Moreover, for $a := g_i\circ\pi_i'(z)$, the variational formula in Lemma A.9.5 yields

$$\begin{aligned}H_{\Psi,\mu_i}(g\circ I_{i,z}) \le \int_{\Omega_i} &\Psi\bigl(g\circ I_{i,z}(\omega)\bigr) - \Psi'\bigl(g_i\circ\pi_i'(z)\bigr)\cdot\bigl(g\circ I_{i,z}(\omega)\bigr)\,d\mu_i(\omega)\\ &-\Psi\bigl(g_i\circ\pi_i'(z)\bigr) + \Psi'\bigl(g_i\circ\pi_i'(z)\bigr)\cdot\bigl(g_i\circ\pi_i'(z)\bigr)\end{aligned}$$

for all $i = 1,\dots,n$ and all $z \in Z$. By combining both estimates, we then obtain the assertion. □

Applying Theorem A.9.7 to the two entropy functions considered at the beginning of this section, we obtain the following two corollaries.

Corollary A.9.8. *Let $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1, \ldots, n$, be probability spaces and $\mathrm{P} := \mu_1 \otimes \cdots \otimes \mu_n$ be the product measure on $Z := \Omega_1 \times \cdots \times \Omega_n$. Then we have*

$$\mathbb{E}_{\mathrm{P}}(ge^g) - \mathbb{E}_{\mathrm{P}}(e^g) \cdot \ln(\mathbb{E}_{\mathrm{P}} e^g) \;\leq\; \sum_{i=1}^{n} \int_Z ge^g - e^g + e^{g_i \circ \pi_i'} - e^g \cdot (g_i \circ \pi_i') \, d\mathrm{P}$$

for all bounded measurable functions $g : Z \to \mathbb{R}$ and $g_i : Z_i' \to \mathbb{R}$, $i = 1, \ldots, n$.

Proof. Let us define $\Psi : [0, \infty) \to \mathbb{R}$ by $\Psi(0) := 0$ and $\Psi(t) := t \ln t$ for $t > 0$. Then Ψ is a twice continuously differentiable entropy function with $\Psi''(t) = 1/t$ for $t > 0$. Consequently, Ψ satisfies the assumptions of Theorem A.9.7. Moreover, we have $H_{\Psi,\mathrm{P}}(e^g) = \mathbb{E}_{\mathrm{P}}(ge^g) - \mathbb{E}_{\mathrm{P}}(e^g) \cdot \ln(\mathbb{E}_{\mathrm{P}} e^g)$ and

$$\Psi(e^g) - \Psi\bigl(e^{g_i \circ \pi_i'}\bigr) - \Psi'\bigl(e^{g_i \circ \pi_i'}\bigr) \cdot \bigl(e^g - e^{g_i \circ \pi_i'}\bigr) = ge^g - e^g + e^{g_i \circ \pi_i'} - e^g \cdot (g_i \circ \pi_i') \,.$$

Applying Theorem A.9.7, then yields the assertion. □

Corollary A.9.9 (Efron-Stein inequality). *Let $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1, \ldots, n$, be probability spaces and $\mathrm{P} := \mu_1 \otimes \cdots \otimes \mu_n$ be the product measure on $Z := \Omega_1 \times \cdots \times \Omega_n$. Then we have*

$$\mathbb{E}_{\mathrm{P}} g^2 - (\mathbb{E}_{\mathrm{P}} g)^2 \;\leq\; \sum_{i=1}^{n} \int_Z \bigl(g - g_i \circ \pi_i'\bigr)^2 \, d\mathrm{P}$$

for all 2-integrable functions $g : Z \to \mathbb{R}$ and $g_i : Z_i' \to \mathbb{R}$, $i = 1, \ldots, n$.

For the last corollary of Theorem A.9.7, we need the following simple lemma.

Lemma A.9.10. *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and $V, g : \Omega \to \mathbb{R}$ be two bounded measurable functions. Then we have*

$$\mathbb{E}_{\mathrm{P}}\bigl(Ve^g\bigr) - \mathbb{E}_{\mathrm{P}} e^g \ln\bigl(\mathbb{E}_{\mathrm{P}} e^V\bigr) \;\leq\; \mathbb{E}_{\mathrm{P}}\bigl(ge^g\bigr) - \mathbb{E}_{\mathrm{P}} e^g \ln\bigl(\mathbb{E}_{\mathrm{P}} e^g\bigr) \,.$$

Proof. Since $\mathbb{E}_{\mathrm{P}} e^g > 0$ exists, we immediately see that $\mathrm{Q} = \frac{e^g}{\mathbb{E}_{\mathrm{P}} e^g} P$ is a probability measure on Ω. Moreover, the definition of Q, the concavity of the logarithm, and Jensen's inequality yield

$$\frac{\mathbb{E}_{\mathrm{P}} Ve^g}{\mathbb{E}_{\mathrm{P}} e^g} - \frac{\mathbb{E}_{\mathrm{P}} ge^g}{\mathbb{E}_{\mathrm{P}} e^g} = \mathbb{E}_{\mathrm{Q}}(V - g) \leq \ln\bigl(\mathbb{E}_{\mathrm{Q}} e^{V-g}\bigr) = \ln\bigl(\mathbb{E}_{\mathrm{P}} e^V\bigr) - \ln\bigl(\mathbb{E}_{\mathrm{P}} e^g\bigr) \,. \qquad \square$$

Now we can establish the last corollary of Theorem A.9.7, which establishes an estimate similar to that of Corollary A.9.8.

Corollary A.9.11. *Let $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1, \ldots, n$, be probability spaces and $\mathrm{P} := \mu_1 \otimes \cdots \otimes \mu_n$ be the product measure on $Z := \Omega_1 \times \cdots \times \Omega_n$. Then we have*

$$\sum_{i=1}^{n} \mathbb{E}_{\mathrm{P}}\big(e^g - e^{g_i\circ\pi_i'}\big) \;\le\; \mathbb{E}_{\mathrm{P}}(e^g)\cdot\ln(\mathbb{E}_{\mathrm{P}}e^g)$$

for all $g : Z \to \mathbb{R}$ *with* $e^g \in L_1(\mathrm{P})$ *and all* $g_i : Z_i' \to (0,\infty)$ *satisfying*

$$\sum_{i=1}^{n}\big(g - g_i\circ\pi_i'\big) \;\le\; g\,. \tag{A.71}$$

Proof. Let us first assume that g and g_i, $i = 1,\dots,n$, are bounded. Writing $V := \sum_{i=1}^n g - g_i\circ\pi_i'$ and $f_i := g_i\circ\pi_i' + \frac{1}{n}\ln\mathbb{E}_{\mathrm{P}}e^V$, $i = 1,\dots,n$, we get

$$ge^g - e^g + e^{f_i} - e^g f_i = ge^g - e^g + e^{g_i\circ\pi_i'}\cdot(\mathbb{E}_{\mathrm{P}}e^V)^{\frac{1}{n}} - e^g(g_i\circ\pi_i') - e^g\ln(\mathbb{E}_{\mathrm{P}}e^V)^{\frac{1}{n}}.$$

By applying Corollary A.9.8 to g and $g_i + \frac{1}{n}\ln\mathbb{E}_{\mathrm{P}}e^V$, we then obtain

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}}ge^g - \mathbb{E}_{\mathrm{P}}e^g\ln(\mathbb{E}_{\mathrm{P}}e^g)\\
&\le \sum_{i=1}^{n}\int_Z ge^g - e^g + e^{g_i\circ\pi_i'}\cdot(\mathbb{E}_{\mathrm{P}}e^V)^{1/n} - e^g\cdot(g_i\circ\pi_i') - e^g(\ln(\mathbb{E}_{\mathrm{P}}e^V)^{1/n})\,d\mathrm{P}\\
&= \mathbb{E}_{\mathrm{P}}Ve^g - n\mathbb{E}_{\mathrm{P}}e^g - \mathbb{E}_{\mathrm{P}}e^g\ln(\mathbb{E}_{\mathrm{P}}e^V) + \sum_{i=1}^{n}\mathbb{E}_{\mathrm{P}}e^{g_i\circ\pi_i'}\cdot(\mathbb{E}_{\mathrm{P}}e^V)^{1/n}\\
&\le \mathbb{E}_{\mathrm{P}}ge^g - \mathbb{E}_{\mathrm{P}}e^g\ln(\mathbb{E}_{\mathrm{P}}e^g) - n\mathbb{E}_{\mathrm{P}}e^g + \sum_{i=1}^{n}\mathbb{E}_{\mathrm{P}}e^{g_i\circ\pi_i'}\cdot(\mathbb{E}_{\mathrm{P}}e^V)^{1/n},
\end{aligned}$$

where in the last step we used Lemma A.9.10. Using $\sum_{i=1}^n \mathbb{E}_{\mathrm{P}}e^g = n\mathbb{E}_{\mathrm{P}}e^g$, we then find

$$\sum_{i=1}^{n}\mathbb{E}_{\mathrm{P}}\big(e^g - e^{g_i\circ\pi_i'}\big) \le n\mathbb{E}_{\mathrm{P}}e^g\big(1-(\mathbb{E}_{\mathrm{P}}e^V)^{-\frac{1}{n}}\big) \le \mathbb{E}_{\mathrm{P}}e^g\ln(\mathbb{E}_{\mathrm{P}}e^V) \le \mathbb{E}_{\mathrm{P}}e^g\ln(\mathbb{E}_{\mathrm{P}}e^g),$$

where in the second step we used $1-t \le -\ln t$ for $t := (\mathbb{E}_{\mathrm{P}}e^V)^{-1/n}$, and in the last step we used (A.71) and the definition of V. Consequently, we have shown the assertion for bounded g and g_i. Finally, observe that $e^{g_i\circ\pi_i'}$, $i = 1,\dots,n$, are positive functions, and hence applying a simple limit argument for g and then for g_i, $i = 1,\dots,n$, shows that the assertion is still valid if we only assume $e^g \in L_1(\mathrm{P})$. □

Besides the results on entropy functionals, we also need some technical yet elementary results for the proof of Talagrand's inequality.

Lemma A.9.12. *Let* $\psi(x) := e^{-x} - 1 + x$ *and* $\varphi(x) := 1 - (1+x)e^{-x}$ *for* $x \in \mathbb{R}$. *Moreover, let* $\alpha \ge 0$, $f(0) := 0$, *and* $f(t) := \frac{\varphi(-t)}{\psi(-t)+\alpha t}$, $t \in \mathbb{R}\setminus\{0\}$. *Then, for all* $x \in (-\infty, 1]$ *and* $t \ge 0$, *we have*

$$\psi(tx) \le f(t)\big(\varphi(tx) + \alpha t x^2 e^{-tx}\big)\,.$$

Proof. Since $\psi(0) = \varphi(0) = 0$, the assertion is obviously satisfied for $t = 0$. Consequently, we fix a $t > 0$ and define $h_t : \mathbb{R} \to \mathbb{R}$ by

$$h_t(x) := \psi(tx) - f(t)\big(\varphi(tx) + \alpha t x^2 e^{-tx}\big)\,, \qquad x \in \mathbb{R}.$$

Clearly, our goal is to show $h_t(x) \leq 0$ for all $x \in (-\infty, 1]$. To this end, we first observe that $\psi(0) = \varphi(0) = 0$ implies $h_t(0) = 0$. Moreover, a simple calculation shows $h_t(1) = 0$. In addition, we have

$$\begin{aligned} h_t'(x) &= \big(\alpha t^2 x^2 - t^2 x - 2\alpha t x\big) f(t) e^{-tx} - t e^{-tx} + t\,, \\ h_t''(x) &= -(\alpha t^3 x^2 - t^3 x - 4\alpha t^2 x + t^2 + 2\alpha t) f(t) e^{-tx} + t^2 e^{-tx}\,. \end{aligned}$$

From this it is easy to check that $\lim_{x\to-\infty} h_t'(x) = \infty$ and $\lim_{x\to\infty} h_t'(x) = t$. Moreover, the second derivative h_t'' is of the form $h_t''(x) = p_t(x)e^{-tx}$, where $p_t(\,\cdot\,)$ is a second-degree polynomial whose leading term has a non-positive coefficient since $f(t) > 0$. Consequently, there exist at most two solutions of $h_t''(x) = 0$. If there was no such solution, then we would have $p_t(x) < 0$ for all $x \in \mathbb{R}$, and hence h_t' would be decreasing. However, since $h_t'(0) = 0$, this contradicts the behavior of h_t' for $x \to \infty$. Let us now denote the solutions of $h_t''(x) = 0$ by x_1 and x_2, where we additionally assume $x_1 \leq x_2$. By the shape of $p_t(\,\cdot\,)$, we then see that h_t' is decreasing on $(-\infty, x_1] \cup [x_2, \infty)$ and increasing on (x_1, x_2). Recalling $\lim_{x\to\pm\infty} h_t'(x) > 0$, there can therefore be at most two x with $h_t'(x) = 0$. We have already seen that $x = 0$ is such a solution. Let x^* denote the other solution. Then we have $x^* > 0$ since otherwise $h_t'(0) = 0$ and $\lim_{x\to\infty} h_t'(x) = t$ would imply $h_t'(x) > 0$ for all $x > 0$. Obviously, the latter contradicts $h_t(0) = h_t(1) = 0$. Now $x^* > 0$ together with $\lim_{x\to\pm\infty} h_t'(x) > 0$ shows that h_t is increasing on $(-\infty, 0] \cup [x^*, \infty)$. Moreover, we have already seen that $h_t'(x) > 0$ for all $x > 0$ contradicts $h_t(0) = h_t(1) = 0$, and hence h_t is decreasing on $(0, x^*)$. This shows $h_t(x) \leq 0$ for all $x \in [0, 1]$. □

We also need the following lemma, which shows that solutions of certain differential inequalities are non-positive functions.

Lemma A.9.13. *Let $G : \mathbb{R} \to \mathbb{R}$ be a twice continuously differentiable function that satisfies $G(0) = G'(0) = 0$ and $G''(0) < 0$. Moreover, let $h : \mathbb{R} \to \mathbb{R}$ be a continuous function such that*

$$xG'(x) - h(x)G(x) \leq 0\,, \qquad x \geq 0. \tag{A.72}$$

Then we have $G(x) \leq 0$ for all $x \geq 0$.

Proof. Let us assume that there exists an $x_0 > 0$ with $G(x_0) > 0$. We define

$$x^* := \sup\big\{x \in [0, x_0) : G(x) = 0\big\}\,.$$

Note that x^* actually exists since $G(0) = 0$. By the continuity of G, we immediately see that $G(x^*) = 0$, and hence we have $x^* < x_0$. Moreover, the continuity of G together with the definition of x^* also yields

$$G(x) > 0\,, \qquad x \in (x^*, x_0]. \tag{A.73}$$

Let us first show that $x^* > 0$. To this end, observe that there exists a $\delta > 0$ with $G''(x) < 0$ for all $x \in [0, \delta]$ by the continuity of G'' and $G''(0) < 0$. Using $G'(0) = 0$ and the fundamental theorem of calculus, we then see $G'(x) < 0$ for all $x \in (0, \delta]$. Another application of the fundamental theorem together with $G(0) = 0$ then shows $G(x) < 0$ for all $x \in (0, \delta]$. By the continuity of G and the definition of x^*, we then find $x^* > 0$. Now, combining (A.72) with (A.73), we find

$$\frac{G'(x)}{G(x)} \;\le\; \frac{h(x)}{x}\,, \qquad x \in (x^*, x_0].$$

Integrating both sides, then yields

$$G(x_0) \;\le\; G(x) \cdot \exp\biggl(\int_x^{x_0} \frac{h(t)}{t}\, dt \biggr)\,, \qquad x \in (x^*, x_0]. \tag{A.74}$$

Now observe that $x^* > 0$ ensures that $t \mapsto \frac{h(t)}{t}$ is integrable on $[x^*, x_0]$ and consequently we obtain $G(x_0) \le 0$ by letting $x \to x^*$ in (A.74). □

With the help of these preparations, we can finally establish upper bounds for certain expectations of the form $\mathbb{E}_{\mathrm{P}} e^g$.

Theorem A.9.14. *Let $n \ge 1$ and $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1, \ldots, n$, be probability spaces. We adopt the notations introduced earlier and assume that we have bounded measurable functions $g : Z \to \mathbb{R}$, $g_i : Z_i' \to \mathbb{R}$, and $u_i : Z \to \mathbb{R}$ such that*

$$u_i(z) \;\le\; g(z) - g_i \circ \pi_i'(z) \;\le\; 1\,, \tag{A.75}$$

$$\sum_{i=1}^n \bigl(g(z) - g_i \circ \pi_i'(z) \bigr) \;\le\; g(z)\,, \tag{A.76}$$

$$\int_{\Omega_i} u_i \circ I_{i,z}(\omega)\, d\mu_i(\omega) \;\ge\; 0\,, \tag{A.77}$$

$$\int_{\Omega_i} \bigl| u_i \circ I_{i,z}(\omega) \bigr|^2\, d\mu_i(\omega) \;\le\; \sigma^2\,, \tag{A.78}$$

for some constant $\sigma > 0$ and all $i = 1, \ldots, n$, $z \in Z$. Then we have

$$\mathbb{E}_{\mathrm{P}} g^2 - (\mathbb{E}_{\mathrm{P}} g)^2 \le n\sigma^2 + 2\mathbb{E}_{\mathrm{P}} g \tag{A.79}$$

and

$$\ln\bigl(\mathbb{E}_{\mathrm{P}} e^{\lambda(g - \mathbb{E}_{\mathrm{P}} g)} \bigr) \le (n\sigma^2 + 2\mathbb{E}_{\mathrm{P}} g)\bigl(e^\lambda - 1 - \lambda \bigr)\,, \qquad \lambda \ge 0. \tag{A.80}$$

Proof. We first show (A.79). To this end, observe that $t \mapsto t^2 - 2t$ is decreasing on $[0, 1]$, and consequently (A.75), (A.77), and (A.78) yield

$$\int_Z (g - g_i \circ \pi_i')^2 - 2(g - g_i \circ \pi_i')\, d\mathrm{P} \;\leq\; \int_{Z_i'} \biggl(\int_{\Omega_i} u_i^2 - 2u_i \, d\mu_i \biggr) d\mathrm{P}_i' \;\leq\; \sigma^2$$

for $i = 1, \ldots, n$. By Corollary A.9.9, we thus find

$$\begin{aligned} \mathbb{E}_\mathrm{P} g^2 - (\mathbb{E}_\mathrm{P} g)^2 \;\leq\; \sum_{i=1}^n \int_Z \bigl(g - g_i \circ \pi_i'\bigr)^2 \, d\mathrm{P} &\leq n\sigma^2 + 2 \sum_{i=1}^n \int_Z (g - g_i \circ \pi_i')\, d\mathrm{P} \\ &\leq n\sigma^2 + 2\mathbb{E}_\mathrm{P} g \,, \end{aligned}$$

where in the last step we used (A.76).

In order to show (A.80), we write $\varphi(t) := 1 - (1+t)e^{-t}$ and $\psi(t) := e^{-t} - 1 + t$ for $t \in \mathbb{R}$. Moreover, we define $f(0) := 0$ and

$$f(t) \;:=\; \frac{\varphi(-t)}{\psi(-t) + t/2} \;=\; \frac{te^t - e^t + 1}{e^t - 1 - t/2} \,, \qquad\qquad t \in \mathbb{R} \backslash \{0\} \,.$$

For $t := \lambda \geq 0$, $\alpha := 1/2$, and $x := g - g_i \circ \pi_i'$, Lemma A.9.12 then yields

$$\begin{aligned} &\psi\bigl(\lambda(g - g_i \circ \pi_i')\bigr) e^{\lambda g} \\ &\leq f(\lambda) \Bigl(\varphi\bigl(\lambda(g - g_i \circ \pi_i')\bigr) + \alpha\lambda (g - g_i \circ \pi_i')^2 e^{-\lambda(g - g_i \circ \pi_i')} \Bigr) e^{\lambda g} \\ &= f(\lambda) \bigl(e^{\lambda g} - e^{\lambda g_i \circ \pi_i'} \bigr) + \lambda f(\lambda) e^{\lambda g_i \circ \pi_i'} \bigl(\alpha (g - g_i \circ \pi_i')^2 - (g - g_i \circ \pi_i') \bigr) \\ &\leq f(\lambda) \bigl(e^{\lambda g} - e^{\lambda g_i \circ \pi_i'} \bigr) + \lambda f(\lambda) e^{\lambda g_i \circ \pi_i'} \bigl(\alpha u_i^2 - u_i \bigr) \,, \end{aligned} \tag{A.81}$$

where in the last step we used (A.75) together with the fact that $t \mapsto \alpha t^2 - t$ is decreasing on $[0, 1]$. Now observe that combining (A.75) and (A.77) yields

$$g_i \circ \pi_i' \;=\; \int_{\Omega_i} g_i \circ \pi_i' \, d\mu_i \;\leq\; \int_{\Omega_i} g - u_i \, d\mu_i \;\leq\; \int_{\Omega_i} g \, d\mu_i \,,$$

from which we conclude $\mathbb{E}_\mathrm{P} e^{\lambda g_i \circ \pi_i'} \leq \mathbb{E}_\mathrm{P} e^{\lambda g}$, $\lambda \geq 0$, by Jensen's inequality. Using this estimate together with (A.77), (A.78), and $\alpha = 1/2$, we then obtain

$$\begin{aligned} \int_Z e^{\lambda g_i \circ \pi_i'} \bigl(\alpha u_i^2 - u_i \bigr) \, d\mathrm{P} &= \int_{Z_i'} e^{\lambda g_i} \biggl(\int_{\Omega_i} \alpha u_i^2 - u_i \, d\mu_i \biggr) d\mathrm{P}_i' \\ &\leq \frac{\sigma^2}{2} \cdot \int_Z e^{\lambda g} \, d\mathrm{P} \,. \end{aligned}$$

Combining this inequality with (A.81), we now find

$$\int_Z \lambda g e^{\lambda g} - e^{\lambda g} + e^{\lambda g_i \circ \pi_i'} - \lambda e^{\lambda g} g_i \circ \pi_i' \, d\mathrm{P} \leq \frac{f(\lambda)}{2} \int_Z 2\bigl(e^{\lambda g} - e^{\lambda g_i}\bigr) + \lambda \sigma^2 e^{\lambda g} \, d\mathrm{P}.$$

Summing over $i = 1, \ldots, n$ and applying both Corollary A.9.8 and Corollary A.9.11, we thus obtain

$$\begin{aligned}\lambda\mathbb{E}_{\mathrm{P}} g e^{\lambda g} - \mathbb{E}_{\mathrm{P}} e^{\lambda g} \ln(\mathbb{E}_{\mathrm{P}} e^{\lambda g}) &\leq \sum_{i=1}^n \int_Z \lambda g e^{\lambda g} - e^{\lambda g} + e^{\lambda g_i \circ \pi_i'} - \lambda e^{\lambda g} g_i \circ \pi_i' \, d\mathrm{P} \\ &\leq \frac{f(\lambda)}{2} \biggl(\sum_{i=1}^n \int_Z 2\bigl(e^{\lambda g} - e^{\lambda g_i}\bigr) d\mathrm{P} + n\lambda\sigma^2 \mathbb{E}_{\mathrm{P}} e^{\lambda g} \biggr) \\ &= \frac{f(\lambda)\, \mathbb{E}_{\mathrm{P}} e^{\lambda g}}{2} \Bigl(2\ln(\mathbb{E}_{\mathrm{P}} e^{\lambda g}) + n\lambda\sigma^2 \Bigr) . \qquad \text{(A.82)}\end{aligned}$$

Let us now define $F(\lambda) := \ln\bigl(\mathbb{E}_{\mathrm{P}} e^{\lambda(g - \mathbb{E}_{\mathrm{P}} g)}\bigr)$, $\lambda \in \mathbb{R}$. Then we obviously have $F(\lambda) = \ln\bigl(\mathbb{E}_{\mathrm{P}} e^{\lambda g}\bigr) - \lambda \mathbb{E}_{\mathrm{P}} g$, and hence we obtain $F'(\lambda) = \frac{\mathbb{E}_{\mathrm{P}} g e^{\lambda g}}{\mathbb{E}_{\mathrm{P}} e^{\lambda g}} - \mathbb{E}_{\mathrm{P}} g$ for all $\lambda \in \mathbb{R}$. Consequently, (A.82) translates into

$$\lambda F'(\lambda) - F(\lambda) = \frac{\lambda \mathbb{E}_{\mathrm{P}} g e^{\lambda g} - \mathbb{E}_{\mathrm{P}} e^{\lambda g} \ln(\mathbb{E}_{\mathrm{P}} e^{\lambda g})}{\mathbb{E}_{\mathrm{P}} e^{\lambda g}} \leq \frac{f(\lambda)}{2} \Bigl(2\ln(\mathbb{E}_{\mathrm{P}} e^{\lambda g}) + n\lambda\sigma^2 \Bigr)$$

i.e., we have derived

$$\lambda F'(\lambda) - \bigl(1 + f(\lambda)\bigr) F(\lambda) \;\leq\; \frac{\lambda f(\lambda)}{2} \bigl(2\mathbb{E}_{\mathrm{P}} g + n\sigma^2 \bigr) , \qquad \lambda \geq 0. \qquad \text{(A.83)}$$

Let us now define $F_0(\lambda) := (2\mathbb{E}_{\mathrm{P}} g + n\sigma^2)(e^\lambda - 1 - \lambda)$, $\lambda \in \mathbb{R}$. Then, using $\lambda f(\lambda)/2 + f(\lambda)(e^\lambda - 1 - \lambda) = \lambda e^\lambda - e^\lambda + 1$, we obtain

$$\lambda F_0'(\lambda) - \bigl(1 + f(\lambda)\bigr) F_0(\lambda) = \frac{\lambda f(\lambda)}{2} \bigl(2\mathbb{E}_{\mathrm{P}} g + n\sigma^2 \bigr) , \qquad \lambda \in \mathbb{R}.$$

For $G(\lambda) := F(\lambda) - F_0(\lambda)$, $\lambda \in \mathbb{R}$, the last equation together with (A.83) then yields $\lambda G'(\lambda) - \bigl(1 + f(\lambda)\bigr) G(\lambda) \leq 0$ for $\lambda \geq 0$. Moreover, we have $G(0) = G'(0) = 0$. Let us now assume that (A.78) is actually a strict inequality. Then it is trivial to see that (A.79) becomes a strict inequality, and hence we have

$$G''(0) = \mathbb{E}_{\mathrm{P}} g^2 - (\mathbb{E}_{\mathrm{P}} g)^2 - (2\mathbb{E}_{\mathrm{P}} g + n\sigma^2) < 0 .$$

Applying Lemma A.9.13, we then see that $F(\lambda) \leq F_0(\lambda)$ for all $\lambda \geq 0$, i.e., we have shown (A.80) in the case where (A.78) is a strict inequality. The general case follows from a simple limit argument. □

Proof (of Talagrand's inequality). By a simple limit argument, we may assume without loss of generality that $\mathcal{F}$ is finite. Moreover, we first consider the case $B = 1$. For $i = 1, \ldots, n$, we then define Z_i', π_i', and $I_{i,z} : \Omega_i \to Z$ as we did before Lemma A.9.6. In addition, we write $\mu_i := \mu$ and define $g_i : Z_i' \to \mathbb{R}$ by

$$g_i(z_i') := \max_{f \in \mathcal{F}} \left| \sum_{\substack{j=1 \\ j \neq i}}^n f(\omega_j) \right| \qquad \text{(A.84)}$$

for $z_i' := (\omega_1, \dots, \omega_{i-1}, \omega_{i+1}, \dots, \omega_n) \in Z_i'$. Moreover, it is elementary to find mutually disjoint, measurable sets $A_f \subset Z_i'$, $f \in \mathcal{F}$, such that $\bigcup_{f\in\mathcal{F}} A_f = Z_i'$ and

$$\left| \sum_{\substack{j=1 \\ j\neq i}}^{n} f(\omega_j) \right| = g_i(z_i') \tag{A.85}$$

for all $z_i' := (\omega_1, \dots, \omega_{i-1}, \omega_{i+1}, \dots, \omega_n) \in A_f$ and $f \in \mathcal{F}$. In other words, the function f realizes the maximum in (A.84) for all $z_i' \in A_f$. Let us define functions $m_i : Z_i' \times \Omega \to \mathbb{R}$ by

$$m_i(z_i', \omega) := \sum_{f\in\mathcal{F}} \mathbf{1}_{A_f}(z_i') \cdot f(\omega)\,, \qquad\qquad z_i' \in Z_i',\, \omega \in \Omega$$

for all $i = 1, \dots, n$. Note that by (A.85) this construction immediately shows

$$\left| \sum_{\substack{j=1 \\ j\neq i}}^{n} m_i(z_i', \omega_j) \right| = g_i(z_i') \tag{A.86}$$

for all $z_i' := (\omega_1, \dots, \omega_{i-1}, \omega_{i+1}, \dots, \omega_n) \in Z_i'$ and $i = 1, \dots, n$, i.e., the function $m_i(z_i', \cdot\,)$ realizes the maximum in (A.84). With the help of these functions, we finally define the functions $u_i : Z \to \mathbb{R}$ by

$$u_i(z) := \left| \sum_{j=1}^{n} m_i\bigl(\pi_i'(z), \omega_j\bigr) \right| - g_i \circ \pi_i'(z) \tag{A.87}$$

for all $i = 1, \dots, n$ and all $z = (\omega_1, \dots, \omega_n) \in Z$. Our first goal is to show that the functions g, g_i, and u_i satisfy the conditions (A.75)–(A.78).

In order to check (A.75), we first observe that $m_i(\pi_i'(z), \cdot\,) \in \mathcal{F}$ implies

$$u_i(z) \leq \max_{f\in\mathcal{F}} \left| \sum_{j=1}^{n} f(\omega_j) \right| - g_i \circ \pi_i'(z) = g(z) - g_i \circ \pi_i'(z)$$

for all $z = (\omega_1, \dots, \omega_n) \in Z$. Moreover, the inverse triangle inequality and $\|f\|_\infty \leq 1$ yields

$$g(z) - g_i \circ \pi_i'(z) = \max_{f\in\mathcal{F}} \left| \sum_{j=1}^{n} f(\omega_j) \right| - \max_{f\in\mathcal{F}} \left| \sum_{\substack{j=1 \\ j\neq i}}^{n} f(\omega_j) \right| \leq \max_{f\in\mathcal{F}} \bigl|f(\omega_i)\bigr| \leq 1,$$

and hence we also have the right-hand side of (A.75).

Let us now show (A.76). Analogously to the construction of the functions m_i, we first construct a measurable function $m : Z \times \Omega \to \mathbb{R}$ such that $m(z, \cdot\,) \in \mathcal{F}$ and

$$\left|\sum_{i=1}^{n} m(z,\omega_i)\right| = g(z)\,, \qquad\qquad z = (\omega_1,\ldots,\omega_n) \in Z.$$

With the help of this function and $m(z,\,\cdot\,) \in \mathcal{F}$, we then obtain

$$(n-1)g(z) = \left|\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} m(z,\omega_j)\right| \leq \sum_{i=1}^{n} \max_{f\in\mathcal{F}} \left|\sum_{\substack{j=1\\ j\neq i}}^{n} f(\omega_j)\right| = \sum_{i=1}^{n} g_i \circ \pi_i'(z)$$

for all $z = (\omega_1,\ldots,\omega_n) \in Z$. From this we easily deduce (A.76).

In order to check (A.77), we first observe $\pi_i' \circ I_{i,z}(\omega) = \pi_i'(z)$, and hence we obtain

$$\begin{aligned}
\int_{\Omega_i} u_i \circ I_{i,z}(\omega)\, d\mu_i(\omega) &\geq \left| \int_{\Omega_i} \sum_{\substack{j=1\\ j\neq i}}^{n} m_i\big(\pi_i'(z),\omega_j\big) + m_i\big(\pi_i'(z),\omega\big)\, d\mu_i(\omega) \right| - g_i \circ \pi_i'(z) \\
&= \left| \sum_{\substack{j=1\\ j\neq i}}^{n} m_i\big(\pi_i'(z),\omega_j\big) + \int_{\Omega_i} m_i\big(\pi_i'(z),\omega\big)\, d\mu_i(\omega) \right| - g_i \circ \pi_i'(z) \\
&= 0\,,
\end{aligned}$$

where in the last step we used $\mathbb{E}_{\mu_i} f = 0$, $f \in \mathcal{F}$, together with the fact that $m_i(\pi_i'(z),\,\cdot\,) \in \mathcal{F}$ is a fixed element.

Finally, inequality (A.78) follows from

$$\int_{\Omega_i} \big|u_i \circ I_{i,z}(\omega)\big|^2\, d\mu_i(\omega) \leq \int_{\Omega_i} \Big|m_i\big(\pi_i'(z),\omega\big)\Big|^2 d\mu_i(\omega) \leq \sigma\,,$$

where in the first step we used (A.86), (A.87), and $|a+b| - |b| \leq |a|$, and in the last step we used $\mathbb{E}_{\mu_i} f^2 \leq \sigma$, $f \in \mathcal{F}$, together with the fact that $m_i(\pi_i'(z),\,\cdot\,) \in \mathcal{F}$ is a fixed element.

Consequently, we have checked (A.75)–(A.78), and thus Theorem A.9.14 yields

$$\ln\big(\mathbb{E}_{\mathrm{P}} e^{\lambda(g-\mathbb{E}_{\mathrm{P}} g)}\big) \;\leq\; (n\sigma^2 + 2\mathbb{E}_{\mathrm{P}} g)\big(e^{\lambda} - 1 - \lambda\big)\,, \qquad\qquad \lambda \geq 0.$$

Applying Markov's inequality, we hence obtain

$$\mathrm{P}\big(\{z \in Z : g(z) - \mathbb{E}_{\mathrm{P}} g \geq \varepsilon\}\big) \leq \exp\Big((n\sigma^2 + 2\mathbb{E}_{\mathrm{P}} g)\big(e^{\lambda} - 1 - \lambda\big) - \lambda\varepsilon\Big)$$

for all $\lambda \geq 0$ and $\varepsilon > 0$. Let us write $a := n\sigma^2 + 2\mathbb{E}_{\mathrm{P}} g$. Simple calculus then shows that the right-hand side of the estimate above is minimized for $\lambda^* := \ln(1 + \frac{\varepsilon}{a})$, and hence we conclude

$$\mathrm{P}\big(\{z \in Z : g(z) - \mathbb{E}_{\mathrm{P}} g \geq \varepsilon\}\big) \leq \exp\left(-a\left(\Big(1+\frac{\varepsilon}{a}\Big)\ln\Big(1+\frac{\varepsilon}{a}\Big) - \frac{\varepsilon}{a}\right)\right) \leq e^{-\frac{3}{2}\cdot\frac{\varepsilon^2}{3a+\varepsilon}}\,,$$

where in the last step we used Lemma 6.11. Now we fix a $\tau > 0$ and define $\varepsilon := \sqrt{2a\tau + \tau^2/9} + \tau/3$. Elementary calculations then show $\tau = \frac{3}{2} \cdot \frac{\varepsilon^2}{3a+\varepsilon}$, and since $\sqrt{2a\tau + \tau^2/9} + \tau/3 \leq \sqrt{2a\tau} + 2\tau/3$ we hence obtain

$$\mathrm{P}\left(\left\{z \in Z : g(z) \geq \mathbb{E}_{\mathrm{P}} g + \sqrt{2a\tau} + \frac{2\tau}{3}\right\}\right) \leq e^{-\tau}\,.$$

Consequently, we have shown the assertion for $B = 1$. The general case finally follows from this specific case by a simple rescaling argument. □

References

Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*. Academic Press, New York, 2nd edition.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, New York.

Aizerman, M., Braverman, E., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Autom. Remote Control*, **25**, 821–837.

Akerkar, R. (1999). *Nonlinear Functional Analysis*. Narosa Publishing House, New Dehli.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**, 1–10.

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, **44**, 615–631.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rodgers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, NJ.

Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, **68**, 337–404.

Ash, R. B. and Doléans-Dade, C. A. (2000). *Probability & Measure Theory*. Academic Press, San Diego, 2nd edition.

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.*, **35**, 608–633.

Averbukh, V. I. and Smolyanov, O. G. (1967). The theory of differentiation in linear topological spaces. *Russian Math. Surveys*, **22**, 201–258.

Averbukh, V. I. and Smolyanov, O. G. (1968). The various definitions of the derivative in linear topological spaces. *Russian Math. Surveys*, **23**, 67–113.

Bakır, G. H., Bottou, L., and Weston, J. (2005). Breaking svm complexity with cross-training. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances*

in Neural Information Processing Systems 17, pages 81–88. MIT Press, Cambridge, MA.

Bargmann, V. (1961). On a Hilbert space of analytic functions and an associated integral transform, part 1. *Comm. Pure Appl. Math.*, **14**, 187–214.

Barnett, V. D. and Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley & Sons, New York, 3rd edition.

Bartlett, P. L. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Trans. Inf. Theory*, **44**, 525–536.

Bartlett, P. L. (2008). Fast rates for estimation error and oracle inequalities for model selection. *Econometric Theory*, **24**, 545–552.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, **3**, 463–482.

Bartlett, P. L. and Tewari, A. (2004). Sparseness versus estimating conditional probabilities: Some asymptotic results. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory*, pages 564–578. Springer, New York.

Bartlett, P. L. and Tewari, A. (2007). Sparseness vs estimating conditional probabilities: some asymptotic results. *J. Mach. Learn. Res.*, **8**, 775–790.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2002). Localized Rademacher complexities. In J. Kivinen and R. H. Sloan, editors, *Proceedings of the 15th Annual Conference on Learning Theory*, pages 44–58. Springer, New York.

Bartlett, P. L., Mendelson, S., and Philips, P. (2004). Local complexities for empirical risk minimization. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory*, pages 270–284. Springer, New York.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *Ann. Statist.*, **33**, 1497–1537.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, **101**, 138–156.

Bauer, H. (2001). *Measure and Integration Theory*. De Gruyter, Berlin.

Becker, C. and Gather, U. (1999). The Masking Breakdown Point of Multivariate Outlier Identification Rules. *J. Amer. Statist. Assoc.*, **94**, 947–955.

Beckmann, R. J. and Cook, R. D. (1983). Outlier.....s. *Technometrics*, **25**, 119–163.

Bednarski, T., Clarke, B. R., and Kolkiewicz, W. (1991). Statistical exansions and locally uniform Fréchet differentiability. *J. Aust. Math. Soc. (Series A)*, **50**, 88–97.

Behringer, F. A. (1992). Convexity is equivalent to midpoint convexity combined with strict quasiconvexity. *Optimization*, **24**, 219–228.

Ben-David, S. and Lindenbaum, M. (1997). Learning distributions by their density levels: a paradigm for learning without a teacher. *J. Comput. System Sci.*, **55**, 171–182.

Bennett, C. and Sharpley, R. (1988). *Interpolation of Operators*. Academic Press, Boston.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.*, **57**, 33–45.

Berg, C., Christensen, J. P. R., and Ressel, P. (1984). *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer, New York.

Bergh, J. and Löfström, J. (1976). *Interpolation Spaces, An Introduction*. Springer-Verlag, New York.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, Boston.

Bernstein, S. N. (1946). *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow.

Berry, M. J. A. and Linoff, G. (1997). *Data Mining Techniques for Marketing, Sales, and Customer Support*. John Wiley & Sons, New York.

Bianco, A. M. and Yohai, V. J. (1996). Robust estimation in the logistic regression model. In H. Rieder, editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter J. Huber's 60th Birthday*, pages 17–34. Springer, New York.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Birman, M. S. and Solomyak, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes W_p^α (Russian). *Mat. Sb.*, **73**, 331–355.

Bishop, C. M. (1996). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Blanchard, G., Bousquet, O., and Massart, P. (2008). Statistical performance of support vector machines. *Ann. Statist.*, **36**, 489–531.

Bochner, S. (1932). Vorlesungen über Fouriersche Integrale. In *Akademische Verlagsgesellschaft*, Leipzig.

Bochner, S. (1959). *Lectures on Fourier integrals. With an author's supplement on Monotonic Functions, Stieltjes Integrals, and Harmonic Analysis*. Princeton University Press, Princeton, NJ.

Boente, G., Fraiman, R., and Yohai, V. J. (1987). Qualitative robustness for stochastic processes. *Ann. Statist.*, **15**, 1293–1312.

Borwein, J. M. and Lewis, A. S. (2000). *Convex Analysis and Nonlinear Optimization*. Springer-Verlag, New York.

Boser, B. E., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM, Madison, WI.

Bottou, L. and Lin, C.-J. (2006). Support vector machine solvers. Technical report, National Taiwan University, Taipei, Department of Computer Science.

Bottou, L., Chapelle, O., DeCoste, D., and Weston, J., editors (2007). *Large-Scale Kernel Machines*. MIT Press, Cambridge, MA.

Boucheron, S., Lugosi, G., and Massart, P. (2003). Concentration inequalities using the entropy method. *Ann. Probab.*, **31**, 1583–1614.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, **9**, 323–375.

Bousquet, O. (2002a). A Bennet concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, **334**, 495–500.

Bousquet, O. (2002b). Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms. Ph.D. thesis, Ecole Polytechnique.

Bousquet, O. (2003a). Concentration inequalities for sub-additive functions using the entropy method. In E. Giné, C. Houdré, and D. Nualart, editors, *Stochastic Inequalities and Applications*, pages 213–247. Birkhäuser, Boston.

Bousquet, O. (2003b). New approaches to statistical learning theory. *Ann. Inst. Statist. Math.*, **55**, 371–389.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *J. Mach. Learn. Res.*, **2**, 499–526.

Breiman, L. (1998). Arcing classifiers. *Ann. Statist.*, **26**, 801–824.

Breiman, L. (1999a). Pasting bites together for prediction in large data sets. *Mach. Learn.*, **36**, 85–103.

Breiman, L. (1999b). Predicting games and arcing algorithms. *Neural Comput.*, **11**, 1493–1517.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International, Belmont, CA.

Brown, A. and Pearcy, C. (1977). *Introduction to Operator Theory I*. Springer, New York.

Bunea, F. and Nobel, A. (2005). Sequential procedures for aggregating arbitrary estimators of a conditional mean. Technical Report M984, Department of Statistics, Florida State University.

Bunea, F., Wegkamp, M., and Tsybakov, A. B. (2007). Aggregation for Gaussian regression. *Ann. Statist.*, **35**, 1674–1697.

Butnariu, D. and Iusem, A. N. (2000). *Totally Convex Functions for Fixed Points Computation and Infinite Dimensional Optimization*. Kluwer, Dordrecht.

Cai, D. M., Gokhale, M., and Theiler, J. (2007). Comparison of feature selection and classification algorithms in identifying malicious executables. *Comput. Statist. Data Anal.*, **51**, 3156–3172.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear model. *J. Amer. Statist. Assoc.*, **96**, 1022–1030.

Caponnetto, A. (2005). A note on the role of squared loss in regression. Technical Report CBCL Paper, MIT, Cambridge, MA.

Caponnetto, A. and De Vito, E. (2005). Fast rates for regularized least-squares algorithm. Technical Report CBCL Paper #248, AI Memo #2005-013, MIT, Cambridge, MA.

Caponnetto, A. and Rakhlin, A. (2006). Stability properties of empirical risk minimization over Donsker classes. *J. Mach. Learn. Res.*, **7**, 2565–2583.

Carl, B. and Stephani, I. (1990). *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge.

Carroll, R. J. and Pederson, S. (1993). On robust estimation in the logistic regression model. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **55**, 693–706.

Castaing, C. and Valadier, M. (1977). *Convex Analysis and Measurable Multifunctions*. Springer, Berlin.

Chafaï, D. (2004). Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities. *J. Math. Kyoto Univ.*, **44**, 326–363.

Chang, C.-C. and Lin, C.-J. (2004). LIBSVM: a library for support vector machines. `www.csie.ntu.edu.tw/~cjlin/papers/libsvm.ps.gz`.

Chapelle, O. (2007). Training a support vector machine in the primal. *Neural Comput.*, **19**, 1155–1178.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0. Step-by-step data mining guide. `www.crisp-dm.org/CRISPWP-0800.pdf`.

Chawla, N. V., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. (2004). Learning ensembles for bites: a scalable and accurate approach. *J. Mach. Learn. Res.*, **5**, 421–451.

Chen, P.-H., Lin, C.-J., and Schölkopf, B. (2005). A tutorial on ν-support vector machines. *Appl. Stoch. Models Bus. Ind.*, **21**, 111–136.

Chen, P.-H., Fan, R.-E., and Lin, C.-J. (2006). A study on SMO-type decomposition methods for support vector machines. *IEEE Trans. Neural Networks*, **17**, 893–908.

Cherkassky, V. and Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, **17**, 113–126.

Christmann, A. (1992). Ausreißeridentifikation und robuste Schätzer im logistischen Regressionsmodell. Dissertation, Department of Statistics, University of Dortmund.

Christmann, A. (1994). Least median of weighted squares in logistic regression with large strata. *Biometrika*, **81**, 413–417.

Christmann, A. (1998). On positive breakdown point estimators in regression models with discrete response variables. Unpublished habilitation thesis, Department of Statistics, University of Dortmund.

Christmann, A. (2002). Classification based on the support vector machine and on regression depth. In Y. Dodge, editor, *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, Statistics for Industry and Technology, pages 341–352. Birkhäuser, Basel.

Christmann, A. (2004). On Propertiechristmann2004bs of Support Vector Machines for Pattern Recognition in Finite Samples. In M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, editors, *Theory and Applications*

of Recent Robust Methods, Statistics for Industry and Technology, pages 49–58. Birkhäuser, Basel.

Christmann, A. (2005). On a Strategy to Develop Robust and Simple Tariffs from Motor Vehicle Insurance Data. *Acta Math. Appl. Sinica (English Ser.)*, **21**, 193–208.

Christmann, A. and Rousseeuw, P. (2001). Measuring overlap in logistic regression. *Comput. Statist. Data Anal.*, **37**, 65–75.

Christmann, A. and Steinwart, I. (2004). On robust properties of convex risk minimization methods for pattern recognition. *J. Mach. Learn. Res.*, **5**, 1007–1034.

Christmann, A. and Steinwart, I. (2007). Consistency and robustness of kernel based regression. *Bernoulli*, **13**, 799–819.

Christmann, A. and Steinwart, I. (2008). Consistency of kernel based quantile regression. *Appl. Stoch. Models Bus. Ind.* DOI:10.1002/asmb.700.

Christmann, A. and Van Messem, A. (2008). Bouligand derivatives and robustness of support vector machines. *J. Mach. Learn. Res.* (tentatively accepted).

Christmann, A., Fischer, P., and Joachims, T. (2002). Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Comput. Statist.*, **17**, 273–287.

Christmann, A., Lübke, K., Marin-Galiano, M., and Rüping, S. (2005). Determination of hyper-parameters for kernel based classification and regression. Technical Report TR-38, University of Dortmund, SFB-475.

Christmann, A., Steinwart, I., and Hubert, M. (2007). Robust learning from bites for data mining. *Comput. Statist. Data Anal.*, **52**, 347–361.

Clarke, B. R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Ann. Statist.*, **11**, 1196–1205.

Clarke, B. R. (1986). Nonsmooth analysis and Fréchet differentiability of M-functionals. *Probab. Theory Related Fields*, **73**, 197–209.

Coakley, C. W. and Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *J. Amer. Statist. Assoc.*, **88**, 872–880.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York, 3rd edition.

Conway, J. B. (1990). *A Course in Functional Analysis*. Springer, New York, 2nd edition.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Mach. Learn.*, **20**, 273–297.

Courant, R. and Hilbert, D. (1953). *Methods of Mathematical Physics*. Interscience Publishers, New York, 1st English edition.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. Chapman and Hall, London, 2nd edition.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.

Croux, C., Rousseeuw, P. J., and Hössjer, O. (1994). Generalized S-estimators. *J. Amer. Statist. Assoc.*, **89**, 1271–1281.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, **39**, 1–49.

Cucker, F. and Zhou, D. X. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge.

Cuevas, A. (1988). Qualitative robustness in abstract inference. *J. Statist. Plann. Inference*, **18**, 277–289.

Dahmen, W. and Micchelli, C. A. (1987). Some remarks on ridge functions. *Approx. Theory Appl.*, **3**, 139–143.

Dalalyan, A. S. and Tsybakov, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In N. Bshouty and C. Gentile, editors, *Proceedings of the 20th Conference on Learning Theory*, pages 97–111. Springer, New York.

Daubechies, I. (1991). *Ten Lectures in Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia.

Davies, P. L. (1990). The asymptotics of S-estimators in the linear regression model. *Ann. Statist.*, **18**, 1651–1675.

Davies, P. L. (1993). Aspects of robust linear regression. *Ann. Statist.*, **21**, 1843–1899.

Davies, P. L. (1994). Desirable properties, breakdown and efficiency in the linear regression model. *Statist. Probab. Lett.*, **19**, 361–370.

Davies, P. L. and Gather, U. (1993). The Identification of Multiple Outliers (with discussion and rejoinder). *J. Amer. Statist. Assoc.*, **88**, 782–801.

Davies, P. L. and Gather, U. (2004). Robust statistics. In J. E. Gentle, W. Härdle, and Y. Mori, editors, *Handbook of Computational Statistics, III*, pages 656–695. Springer, Berlin.

Davies, P. L. and Gather, U. (2005). Breakdown and Groups (with discussion and rejoinder). *Ann. Statist.*, **33**, 977–1035.

Davies, P. L. and Gather, U. (2006). Addendum to the discussion of 'Breakdown and Groups'. *Ann. Statist.*, **34**, 1577–1579.

Davies, P. L., Gather, U., and Weinert, H. (2008). Nonparametric Regression as an Example of Model Choice. *Comm. Statist. Simulation Comput.*, **37**, 274–289.

De Vito, E., Rosasco, L., Caponnetto, A., Piana, M., and Verri, A. (2004). Some properties of regularized kernel methods. *J. Mach. Learn. Res.*, **5**, 1363–1390.

Debruyne, M. (2007). Robustness of censored depth quantiles, PCA and kernel based regression, with new tools for model selection. Unpublished Ph.D. thesis, Faculteit Wetenschappen, Katholieke Universiteit Leuven.

Debruyne, M., Christmann, A., Hubert, M., and Suykens, J. (2007). Robustness of Reweighted Kernel Based Regression. Technical report, Department of Statistics, Katholieke Universiteit Leuven.

Desu, M. M. and Raghavarao, D. (1990). *Sample Size Methodology*. Academic Press, Boston.

Dette, H., Neumeyer, N., and Pilz, K. F. (2006). A simple non-parametric estimator of a monotone regression function. *Bernoulli*, **12**, 469–490.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Devroye, L. P. (1982). Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Trans. Pattern Anal. Mach. Intell.*, **4**, 154–157.

Diestel, J. and Uhl, J. J. (1977). *Vector Measures*. American Mathematical Society, Providence.

Diestel, J., Jarchow, H., and Tonge, A. (1995). *Absolutely Summing Operators*. Cambridge University Press, Cambridge.

Dinculeanu, N. (2000). *Vector Integration and Stochastic Integration in Banach Spaces*. John Wiley & Sons, New York.

Dinuzzo, F., Neve, M., and G. De Nicolao, U. P. G. (2007). On the representer theorem and equivalent degrees of freedom of SVR. *J. Mach. Learn. Res.*, **8**, 2467–2495.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In P. J. Bickel, K. A. Doksum, and J. L. Hodges, Jr., editors, *A Festschrift for Erich L. Lehmann*, pages 157–184. Wadsworth, Belmont, CA.

Donoho, D. L. and Johnstone, I. (1989). Projection-based approximation and a duality with kernel methods. *Ann. Statist.*, **17**, 58–106.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, New York, 2nd edition.

Dudley, R. M. (1967). The sizes of compact subsets of Hilbert spaces and continuity of Gaussian processes. *J. Funct. Anal.*, **1**, 290–330.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, Cambridge.

Dunford, N. and Schwartz, J. T. (1988). *Linear Operators, Part I: General Theory*. John Wiley & Sons, New York, Wiley Classics Library edition.

Duren, P. L. (1970). *Theory of H^p Spaces*. Academic Press, New York.

Duren, P. L. and Schuster, A. (2004). *Bergman Spaces*. American Mathematical Society, Providence.

Edmunds, D. E. and Triebel, H. (1996). *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge.

Einmahl, U. and Li, D. (2008). Characterization of LIL behavior in Banach space. *Trans. Amer. Math. Soc. (to appear)*.

Ekeland, I. and Turnbull, T. (1983). *Infinite-Dimensional Optimization and Convexity*. University of Chicago Press, Chicago.

Elisseeff, A., Evgeniou, T., and Pontil, M. (2005). Stability of randomized learning algorithms. *J. Mach. Learn. Res.*, **6**, 55–79.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 973–978.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear model. *Ann. Statist.*, **13**, 342–368.

Fahrmeir, L. and Kaufmann, H. (1986). Correction: Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear model. *Ann. Statist.*, **14**, 1643.

Fan, R.-E., Chen, P.-H., and Lin, C.-J. (2005). Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.*, **6**, 1889–1918.

Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Lecture Notes in Statistics, 19. Springer, New York.

Fernholz, L. T. and Morgenthaler, S. (2005). Conditionally least-informative distributions: the case of the location parameter. *J. Stat. Plann. Inference*, **135**, 357–370.

Fix, E. and Hodges, J. (1951). Discriminatory analysis. nonparametric discrimination: Consistency properties. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randoph Field, TX.

Fix, E. and Hodges, J. (1952). Discriminatory analysis: small sample performance. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randoph Field, TX.

Floret, K. (1981). *Maß- und Integrationstheorie*. Teubner, Stuttgart.

Folland, G. B. (1989). *Harmonic Analysis in Phase Space*. Princeton University Press, Princeton, NJ.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Inform. and Comput.*, **121**, 256–285.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In L. Saitta, editor, *Machine Learning: Proceedings of the 13th International Conference*, pages 148–156. Morgan Kauffman, San Francisco.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.

Gather, U. (1984). Tests und Schätzungen in Ausreißermodellen. Unpublished habilitation thesis, Department of Mathematics, RWTH Aachen.

Gather, U. (1990). Modelling the Occurrence of Multiple Outliers. *Allg. Statist. Archiv*, **74**, 413–428.

Gather, U. and Fried, R. (2004). Methods and Algorithms for Robust Filtering. In A. Jaromir, editor, *Proceedings of COMPSTAT 2004, 16th Symposium of the International Association for Statistical Computing*, pages 159–170. Physica, Heidelberg.

Gather, U. and Hilker, T. (1997). A Note on Tyler's Modification of the MAD for the Stahel-Donoho Estimator. *Ann. Statist.*, **25**, 2024–2026.

Gather, U. and Pawlitschko, J. (2004). Outlier Detection. In B. Sundt and J. Teugels, editors, *The Encyclopedia of Actuarial Science*, volume 3, pages 1230–1237. John Wiley & Sons, New York.

Gather, U. and Schettlinger, K. (2007). Robust and adaptive methods in analysing online monitoring data. In *Proceedings of the 56th Session of the ISI*, Lisbon.

Gather, U., Bauer, M., and Fried, R. (2002). The Identification of Multiple Outliers in Online Monitoring data. *Estadistica*, **54**, 289–338.

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Comput.*, **7**, 219–269.

Graves, L. M. (1956). *The Theory of Functions of Real Variables*. McGraw-Hill, New York.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52.

Hall, P. and Huang, L.-J. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.*, **29**, 624–647.

Hallin, M. and Paindaveine, D. (2004). Rank-based optimal tests of the adequacy of an elliptic VARMA model. *Ann. Statist.*, **32**, 2642–2678.

Hallin, M. and Paindaveine, D. (2005). Affine-invariant aligned rank tests for the multivariate general linear model with ARMA errors. *J. Multivariate Anal.*, **93**, 122–163.

Hammer, B. and Gersmann, K. (2003). A note on the universal approximation capability of support vector machines. *Neural Process. Lett.*, **17**, 43–53.

Hampel, F. R. (1968). Contributions to the theory of robust estimation. Unpublished Ph.D. thesis, Department of Statistics, University of California, Berkeley.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.*, **42**, 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–393.

Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods (with discussion). In G. S. Maddala and C. R. Rao, editors, *Proceedings of the 40th Session of the ISI*, volume 46, pages 375–391.

Hampel, F. R. (2005). Discussion: Breakdown and groups. *Ann. Statist.*, **33**, 977–1035.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. MIT Press, Cambridge, MA.

Harter, H. L. (1983). Least squares. In S. Kotz and C. B. R. N. L. Johnson, editors, *Encyclopedia of Statistical Sciences*, page 595. John Wiley & Sons, New York.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, New York.

Hartigan, J. A. (1987). Estimation of a convex density contour in 2 dimensions. *J. Amer. Statist. Assoc.*, **82**, 267–270.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

He, X. (1997). Quantile curves without crossing. *Amer. Statist.*, **51**, 186–192.

He, X. and Liang, H. (2000). Quantile regression estimates for a class of linear and partially linear errors-in-variables models. *Statist. Sinica*, **10**, 129–140.

He, X. and Ng, P. (1999). COBS: Qualitatively constrained smoothing via linear programming. *Comput. Statist.*, **14**, 315–337.

He, X. and Simpson, D. G. (1993). Lower bounds for contamination bias: globally minimax versus locally linear estimation. *Ann. Statist.*, **21**, 313–337.

He, X. and Wang, G. (1997). Convergence of depth contours for multivariate datasets. *Ann. Statist.*, **52**, 495–504.

Hedenmalm, H., Korenblum, B., and Zhu, K. (2000). *Theory of Bergman Spaces*. Springer, New York.

Hein, M. and Bousquet, O. (2004). Kernels, associated structures and generalizations. Technical Report 127, Max-Planck-Institute for Biological Cybernetics.

Hettmansperger, T. P., McKean, J. W., and Sheather, S. J. (1997). Rank-based analyses of linear models. In G. S. Maddala and C. R. Rao, editors, *Handbook of Statistics, 15*, pages 145–173. Elsevier Science B.V., Amsterdam.

Hille, E. (1972). Introduction to general theory of reproducing kernels. *Rocky Mountain J. Math.*, **2**, 321–368.

Hille, E. and Phillips, R. S. (1957). *Functional Analysis and Semi-groups*. American Mathematical Society Colloquium Publications Vol. XXXI, Providence, revised edition.

Hipp, J., Güntzer, U., and Grimmer, U. (2001). Data quality mining–making a virtue of necessity. Workshop on Research Issues in Data Mining and Knowledge Discovery, Santa Barbara, CA, `www.cs.cornell.edu/johannes/papers/dmkd2001-papers/p5_hipp.pdf`.

Hochreiter, S. and Obermayer, K. (2004). Gene selection for microarray data. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, pages 319–355. MIT Press, Cambridge, MA.

Hoeffding, W. (1963). Probability-inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13–30.

Höffgen, K. U., Simon, H.-U., and van Horn, K. S. (1995). Robust trainability of single neurons. *J. Comput. Syst. Sci.*, **50**, 114–125.

Hoffmann-Jørgensen, J. (1974). Sums of independent Banach space valued random variables. *Studia Math.*, **52**, 159–186.

Hoffmann-Jørgensen, J. (1977). *Probability in Banach space*. Lecture Notes in Math. 598. Springer, New York.

Huang, T., Kecman, V., and Kopriva, I. (2006). *Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-supervised, and Unsupervised Learning*. Springer, Berlin.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. 5th Berkeley Symp.*, **1**, 221–233.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, New York.

Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.*, **13**, 435–525.

Huber, P. J. (1993). Projections pursuit and robustness. In S. Morgenthaler, E. Ronchetti, and W. A. Stahel, editors, *New Directions in Statistical Data Analysis and Robustness*, pages 139–146. Birkhäuser, Basel.

Hush, D. and Scovel, C. (2003). Polynomial-time decomposition algorithms for support vector machines. *Mach. Learn.*, **51**, 51–71.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods–Support Vector Learning*, pages 41–56, MIT Press, Cambridge, MA.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers, Boston.

Joachims, T. (2006). Training linear SVMs in linear time. In *ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD)*, pages 217–226.

Jurečková, J. and Sen, P. K. (1996). *Robust Statistical Procedures. Asymptotics and Interrelations*. John Wiley & Sons, New York.

Kahane, J.-P. (1968). *Some Random Series of Functions*. D.C. Heath and Company, Lexington, MA.

Kardaun, O. J. W. F., Salomé, D., Schaafsma, W., Steerneman, A. G. M., Willems, J. C., and Cox, D. R. (2003). Reflections on fourteen cryptic issues concerning the nature of statistical inference. *Int. Statist. Rev.*, **71**, 277–318.

Kaufman, L. and Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction of Cluster Analysis*. John Wiley & Sons, Hoboken, NJ.

Keerthi, S., Chapelle, O., and DeCoste, D. (2006). Building support vector machines with reduced classifier complexity. *J. Mach. Learn. Res.*, **7**, 1493–1515.

Keerthi, S. S. and DeCoste, D. (2005). A modified finite Newton method for fast solution of large scale linear SVMs. *J. Mach. Learn. Res.*, **6**, 341–361.

Keerthi, S. S., Shevade, S. K., Battacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput.*, **13**, 637–649.

Keerthi, S. S., Duan, K., Shevade, S. K., and Poo, A. N. (2005). A Fast Dual Algorithm for Kernel Logistic Regression. *Mach. Learn.*, **61**, 151–165.

Keerthi, S. S., Sindhwani, V., and Chapelle, O. (2007). An efficient method for gradient-based adaptation of hyperparameters in SVM models. In *Advances in Neural Information Processing Systems 19*, pages 673–680. MIT Press, Cambridge, MA.

Kelley, J. L. (1955). *General Topology*. D. Van Nostrand, Toronto.

Kent, J. T. and Tyler, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *Ann. Statist.*, **19**, 2102–2119.

Kent, J. T. and Tyler, D. E. (1996). Constrained M-estimates of multivariate location and scatter. *Ann. Statist.*, **24**, 1346–1370.

Kent, J. T. and Tyler, D. E. (2001). Regularity and uniqueness for constrained M-estimates and redescending M-estimates. *Ann. Statist.*, **29**, 252–265.

Kestelman, H. (1960). *Modern Theories of Integration*. Dover, New York.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, **33**, 82–95.

Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, **33**, 1060–1077.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, Cambridge.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, **81**, 673–680.

Koenker, R. W. (1986). Strong consistency of regression quantiles and related empirical processes. *Econometric Theory*, **2**, 191–201.

Koenker, R. W. and Bassett, G. W. (1978). Regression quantiles. *Econometrica*, **46**, 33–50.

Koenker, R. W. and Xiao, Z. (2006). Quantile autoregression (with discussion and rejoinder). *J. Amer. Statist. Assoc.*, **475**, 980–1006.

Kolmogorov, A. N. (1956). Asymptotic characteristics of some completely bounded metric spaces. *Dokl. Akad. Nauk. SSSR*, **108**, 585–589.

Kolmogorov, A. N. and Tikhomirov, V. M. (1961). ε-entropy and ε-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.*, **17**, 277–364.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, **47**, 1902–1914.

Koltchinskii, V. and Beznosova, O. (2005). Exponential convergence rates in classification. In P. Auer and R. Meir, editors, *Proceedings of the 18th Annual Conference on Learning Theory*, pages 295–307. Springer, New York.

Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In D. M. M. E. Giné and J. A. Wellner, editors, *High Dimensional Probability II*, pages 443–459. Birkhäuser, Boston.

Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, **30**, 1–50.

Krishnapuram, B., Carin, L., and Hartemink, A. (2004). Gene expression analysis: Joint feature selection and classifier design. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, pages 299–317. MIT Press, Cambridge, MA.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. In *Proceedings of the 2nd Berkeley Symposium of Mathematical Statistics and Probability*, pages 481–492, University of California Press, Berkeley.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Amer. Statist. Assoc.*, **84**, 460–466.

Lax, P. D. (2002). *Functional Analysis*. John Wiley & Sons, New York.

LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

Lecué, G. (2007a). Optimal rates of aggregation in classification. *Bernoulli*, **13**, 1000–1022.

Lecué, G. (2007b). Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.*, **35**, 1698–1721.

Ledoux, M. (1996). On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, **1**, 63–87.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer, Berlin.

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1998). The importance of convexity in learning with squared loss. *IEEE Trans. Inform. Theory*, **44**, 1974–1980.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *J. Amer. Statist. Assoc.*, **99**, 67–81.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York, 2nd edition.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, 3rd edition.

Lehto, O. (1952). Some remarks on the kernel functions in Hilbert spaces. *Ann. Acad. Sci. Fenn., Ser. A I*, **109**, 6.

Lemeshow, S., Hosmer, D. W., Klar, J., and Lwanga, S. K. (1990). *Adequacy of Sample Size in Health Studies*. WHO, John Wiley & Sons, Chichester.

Levitin, E. S. and Polyak, B. T. (1966). Convergence of minimizing sequences in conditional extremum problems. *Sov. Math. Dokl.*, **7**, 764–767.

Levy, P. S. and Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*. John Wiley & Sons, New York.

Lin, C. J. (2001). On the convergence of the decomposition method for support vector machines. *IEEE Trans. Neural Networks*, **12**, 1288–1298.

Lin, C. J. (2002a). Asymptotic convergence of an SMO algorithm without any assumptions. *IEEE Trans. Neural Networks*, **13**, 248–250.

Lin, Y. (2002b). Support vector machines and the Bayes rule in classification. *Data Min. Knowl. Discov.*, **6**, 259–275.

Lin, Y. (2004). A note on margin-based loss functions in classification. *Statist. Probab. Lett.*, **68**, 73–82.

Lin, Y., Lee, Y., and Wahba, G. (2002). Support vector machines for classification in nonstandard situations. *Mach. Learn.*, **46**, 191–202.

List, N. and Simon, H.-U. (2004). A general convergence theorem for the decomposition method. In *Proceedings of the 17th Annual Conference on Learning Theory*, pages 363–377.

List, N. and Simon, H. U. (2007). General polynomial time decomposition algorithms. *J. Mach. Learn. Res.*, **8**, 303–321.

List, N., Hush, D., Scovel, C., and Steinwart, I. (2007). Gaps in support vector optimization. In N. Bshouty and C. Gentile, editors, *Proceedings of the 20th Conference on Learning Theory*, pages 336–348. Springer, New York.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Liu, R. (1990). On a notion of data depth based on random simplices. *Ann. Statist.*, **18**, 405–414.

Liu, R. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Ann. Statist.*, **27**, 783–858.

Lozano, F. (2000). Model selection using Rademacher penalization. In *Proceedings of the Second ICSC Symposium on Neural Computation (NC2000)*. ICSC Academic Press, New York.

Lugosi, G. and Wegkamp, M. (2004). Complexity regularization via localized random penalties. *Ann. Statist.*, **32**, 1679–1697.

Mammen, E. and Tsybakov, A. (1999). Smooth discrimination analysis. *Ann. Statist.*, **27**, 1808–1829.

Mangasarian, O. L. (2002). A finite Newton method for classification. *Optim. Methods Softw.*, **17**, 913–929.

Maronna, R. A. and Yohai, V. J. (1981). Asymptotic behaviour of general M-estimates for regression and scale with random carriers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **58**, 7–20.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics. Theory and Methods*. John Wiley & Sons, New York.

Martin, R. D., Yohai, V. J., and Zamar, R. H. (1989). Min-max bias robust regression. *Ann. Statist.*, **17**, 1608–1630.

Mason, L., Baxter, J., and Bartlett, P. L. (2000). Improved generalization through explicit optimization of margins. *Mach. Learn.*, **38**, 243–255.

Massart, P. (2000a). About the constants in Talagrand's concentration inequality for empirical processes. *Ann. Probab.*, **28**, 863–884.

Massart, P. (2000b). Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse, VI. Sr., Math.*, **9**, 245–303.

Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *Ann. Statist.*, **34**, 2326–2366.

Mattera, D. and Haykin, S. (1999). Support Vector Machines for Dynamic Reconstruction of a Chaotic System. In B. Schölkopf, J. Burger, and A. Smola, editors, *Advances in Kernel Methods: Support Vector Machine*, pages 211–241. MIT Press, Cambridge, MA.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2nd edition.

McCullogh, W. and Pitts, W. (1943). A logical calculus of the ideas immenent in nervous activity. *Bull. Math. Biophys.*, **5**, 115–133.

Mendelson, S. (2001a). Geometric methods in the analysis of Glivenko-Cantelli classes. In D. Helmbold and B. Williamson, editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory*, pages 256–272. Springer, New York.

Mendelson, S. (2001b). Learning relatively small classes. In D. Helmbold and B. Williamson, editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory*, pages 273–288. Springer, New York.

Mendelson, S. (2002). Improving the sample complexity using global data. *IEEE Trans. Inform. Theory*, **48**, 1977–1991.

Mendelson, S. (2003a). A few notes on statistical learning theory. In S. Mendelson and A. Smola, editors, *Advanced Lectures on Machine Learning: Machine Learning Summer School 2002, Canberra, Australia*, pages 1–40. Springer, Berlin.

Mendelson, S. (2003b). On the performance of kernel classes. *J. Mach. Learn. Res.*, **4**, 759–771.

Mendes, B. and Tyler, D. E. (1996). Constrained M-estimation for regression. In H. Rieder, editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter J. Huber's 60th Birthday*, pages 299–320, Lecture Notes in Statistics, Springer, New York.

Mercer, J. (1909). Functions of positive and negative type and their connection with theory of integral equations. *Philos. Trans. R. Soc. London, Ser. A*, **209**, 415–446.

Meschkowski, H. (1962). *Hilbertsche Räume mit Kernfunktion*. Springer, Berlin.

Micchelli, C. A., Xu, Y., and Zhang, H. (2006). Universal kernels. *J. Mach. Learn. Res.*, **7**, 2651–2667.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Mizera, I. (2002). On depth and deep points: a calculus. *Ann. Statist.*, **30**, 1681–1736.

Momma, M. and Bennett, K. P. (2002). A Pattern Search Method for Model Selection of Support Vector Regression. In R. Grossman, J. Han, V. Kumar, H. Mannila, and R. Motwani, editors, *Proceedings of SIAM Conference on Data Mining*. SIAM, Philadelphia.

Moore, E. H. (1935). *General Analysis, Part I*. Memoirs of the American Philosophical Society, Philadelphia.

Moore, E. H. (1939). *General Analysis, Part II*. Memoirs of the American Philosophical Society, Philadelphia.

Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear model. *Biometrika*, **79**, 747–754.

Morgenthaler, S. and Tukey, J. W. (1991). *Configural Polysampling: A Route to Practical Robustness*. John Wiley & Sons, New York.

Mosler, K. (2002). *Multivariate Dispersion, Central Regions, and Depth: The Lift Zonoid Approach*. Springer, New York.

Mukherjee, S., Niyogi, P., Poggio, T., and Rifkin, R. (2006). Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.*, **25**, 161–193.

Müller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.*, **86**, 738–746.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A*, **135**, 370–384.

Neumann, J., Schnörr, C., and Steidl, G. (2005). Combined SVM-based feature selection and classification. *Mach. Learn.*, **61**, 129–150.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.*, **1**, 327–332.

Ollila, E., Hettmansperger, T. P., and Oja, H. (2002). Estimates of regression coefficients based on sign covariance matrix. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**, 447–466.

Ollila, E., Oja, H., and Koivunen, V. (2003). Estimates of regression coefficients based on rank covariance matrix. *J. Am. Statist. Assoc.*, **98**, 90–98.

Osuna, E. and Girosi, F. (1999). Reducing the run-time complexity in support vector regression. In B. Schölkopf, C. J. C. Burges, and A. Smola, editors, *Advances in Kernel Methods–Support Vector Learning*, pages 271–284. MIT Press, Cambridge, MA.

Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines: An application to face detection. In *Proceedings of CVPR'97*, pages 130–136. IEEE Computer Society, Washington, DC.

Pedersen, G. K. (1988). *Analysis Now*. Springer, New York.

Pelckmans, K., De Brabanter, J., Suykens, J. A. K., and De Moor, B. (2005). Maximal Variation and Missing Values for Componentwise Support Vector Machines. In *Proceedings of the International Joint Conference on Neural Networks, Montreal, Canada*.

Phelps, R. R. (1993). *Convex Functions, Monotone Operators and Differentiability*. Lecture Notes in Math. 1364. Springer, Berlin.

Pietsch, A. (1987). *Eigenvalues and s-Numbers*. Geest & Portig K.-G., Leipzig.

Pinkus, A. (1985). *n-widths in Approximation Theory*. Springer, Berlin.

Pinkus, A. (2004). Strictly positive definite functions on a real inner product space. *Adv. Comput. Math.*, **20**, 263–271.

Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods–Support Vector Learning*, pages 185–208. MIT Press, Cambridge, MA.

Poggio, T. (1975). On optimal nonlinear associative recall. *Biol. Cybernet.*, **19**, 201–209.

Poggio, T. and Girosi, F. (1990). A theory of networks for approximation and learning. *Proc. IEEE*, **78**, 1481–1497.

Poggio, T., Rifkin, R., Mukherjee, S., and Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, **428**, 419–422.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass aproach. *Ann. Statist.*, **23**, 855–881.

Polyak, B. T. (1966). Existence theorems and convergence of minimizing sequences for extremal problems with constraints. *Sov. Math. Dokl.*, **7**, 72–75.

Portnoy, S. (2003). Censored regression quantiles. *J. Amer. Statist. Assoc.*, **98**, 1001–1012.

Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-Error versus Absolute-Error Estimators. *Statist. Sci.*, **12**(4), 279–300.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, **38**, 485–498.

Prohorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.*, **1**, 157–214.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, `www.R-project.org`.

Range, R. M. (1986). *Holomorphic Functions and Integral Representations in Several Complex Variables*. Springer, New York.

Rieder, H. (1994). *Robust Asymptotic Statistics*. Springer, New York.

Rieder, H., Kohl, M., and Ruckdeschel, P. (2008). The cost of not knowing the radius. *Statistical Methods and Applications*, **17**, 13–40.

Riesz, F. and Nagy, B. S. (1990). *Functional Analysis*. Dover Publications, New York, 2nd edition.

Rio, E. (2002). Une inégalité de Bennett pour les maxima de processus empiriques. *Ann. Inst. Henri Poincaré Probab. Statist.*, **38**, 1053–1057.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Ritter, K. (2000). *Average-Case Analysis of Numerical Problems*. Lecture Notes in Math. 1733. Springer, Berlin.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ.

Rockafellar, R. T. (1976). Integral functionals, normal integrands and measurable selections. In *Nonlinear Operators and the Calculus of Variations*, Lecture Notes in Math. 543, pages 157–207. Springer, Berlin.

Rockafellar, R. T. and Wets, R. J.-B. (1998). *Variational Analysis*. Springer, Berlin.

Rosenblatt, F. (1956). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408.

Rosenblatt, R. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan, Washington, DC.

Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, **79**, 871–880.

Rousseeuw, P. J. (1994). Unconventional features of positive-breakdown estimators. *Statist. Probab. Lett.*, **19**, 417–431.

Rousseeuw, P. J. (1997a). Introduction to positive-breakdown methods. In G. S. Maddala and C. R. Rao, editors, *Handbook of Statistics*, volume 15, pages 101–121. North-Holland.

Rousseeuw, P. J. (1997b). Least median of squares regression. In S. Kotz and N. L. Johnson, editors, *Breakthroughs in Statistics*, volume 3, pages 440–462, Springer, New York.

Rousseeuw, P. J. and Bassett, G. W. (1990). The remedian: a robust averaging method for large data sets. *J. Amer. Statist. Assoc.*, **85**, 97–104.

Rousseeuw, P. J. and Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Comput. Statist. Data Anal.*, **43**, 315–332.

Rousseeuw, P. J. and Hubert, M. (1999). Regression depth. *J. Amer. Statist. Assoc.*, **94**, 388–402.

Rousseeuw, P. J. and Leroy, A. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.

Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

Rousseeuw, P. J. and Van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps. In W. Gaul, O. Opitz, and M. Schader, editors, *Data Analysis: Scientific Modeling and Practical Application*, pages 335–346. Springer, New York.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.*, **85**, 633–651.

Rousseeuw, P. J. and van Zomeren, B. C. (1991). Robust distances: simulations and cutoff values. In W. Stahel and S. Weisberg, editors, *Directions in robust statistics and diagnostics, Part II*, pages 195–204. Springer, New York.

Rousseeuw, P. J. and Yohai, V. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle, and D. Martin, editors, *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics, 26, pages 256–272. Springer, New York.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

Ruckdeschel, P. and Rieder, H. (2004). Optimal influence curves for general loss functions. *Statist. Decisions*, **22**, 201–223.

Rudin, W. (1973). *Functional Analysis*. McGraw-Hill, New York.

Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill International Company, Singapore.

Rüping, S. (2000). *mySVM-Manual*. Department of Computer Science, University of Dortmund. `www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM`.

Rüping, S. (2003). *myKLR*. Department of Computer Science, University of Dortmund. `www-ai.cs.uni-dortmund.de/SOFTWARE/MYKLR`.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Saitoh, S. (1988). *Theory of Reproducing Kernels and Applications*. Longman Scientific & Technical, Harlow.

Saitoh, S. (1997). *Integral Transforms, Reproducing Kernels and Their Applications*. Longman Scientific & Technical, Harlow.

Salibian-Barrera, M., Van Aelst, S., and Willems, G. (2008). Fast and Robust Bootstrap. *Statistical Methods and Applications*, **17**, 41–71.

Santner, T. J. and Duffy, D. E. (1986). A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **73**, 755–758.

Sawitzki, G. (1996). The excess mass approach and the analysis of multimodality. In *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis and Knowledge Organization*, Proceedings of the 18th Annual Conference of the GfKl, pages 203–211. Springer, New York.

Schapire, R. (1990). The strength of weak learnability. *Mach. Learn.*, **5**, 197–227.

Schoenberg, I. J. (1938). Metric spaces and completely monotone functions. *Ann. Math. (2)*, **39**, 811–841.

Schölkopf, B. (1997). *Support Vector Learning*. Oldenbourg, München.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

Schölkopf, B., Smola, A. J., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.

Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). New support vector algorithms. *Neural Comput.*, **12**, 1207–1245.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., and Smola, A. J. (2001a). Estimating the support of a high-dimensional distribution. *Neural Comput.*, **13**, 1443–1471.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001b). A generalized representer theorem. In D. Helmbold and B. Williamson, editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory*, pages 416–426. Springer, New York.

Schölkopf, B., Tsuda, K., and Vert, J. P. (2004). *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA.

Scott, C. D. and Nowak, R. D. (2006). Learning minimum volume sets. *J. Mach. Learn. Res.*, **7**, 665–740.

Scovel, C., Hush, D., and Steinwart, I. (2005). Learning rates for density level detection. *Anal. Appl.*, **3**, 356–371.

Seeger, M. W. (2007). Cross-validation optimization for large scale hierarchical classification kernel methods. In *Advances in Neural Information Processing Systems 19*, pages 1233–1240. MIT Press, Cambridge, MA.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.

Shawe-Taylor, J., Williams, C. K. I., Cristianini, N., and Kandola, J. (2002). On the eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In N. Cesa-Bianchi, M. Numao, and R. Reischuk, editors, *Algorithmic Learning Theory, 13th International Conference*, pages 23–40. Springer, New York.

Shawe-Taylor, J., Williams, C. K. I., Cristianini, N., and Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Trans. Inform. Theory*, **51**, 2510–2522.

Simon, H. A. (1983). Why should machines learn? In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning: An Artificial Intelligence Approach*, pages 25–38. Tioga Press, Palo Alto, CA.

Simpson, D. G., Carroll, R. J., and Ruppert, D. (1987). M-estimation for discrete data: asymptotic distribution theory and implications. *Ann. Statist.*, **15**, 657–669.

Smale, S. and Zhou, D.-X. (2003). Estimating the approximation error in learning theory. *Anal. Appl.*, **1**, 17–41.

Smola, A. J. and Schölkopf, B. (1998). On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, **22**, 221–231.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Stat. Comput.*, **14**, 199–222.

Smola, A. J., Murata, N., Schölkopf, B., and Müller, K.-R. (1998). Asymptotically optimal choice of ε-loss for support vector machines. In L. Niklasson, M. Boden, and T. Ziemke, editors, *Perspectives in Neural Computing*, Proceedings of the International Conference on Artificial Neural Networks, pages 105–110. Springer, Berlin.

Song, L., Smola, A., Gretton, A., Borgwardt, K., and Bedo, J. (2007). Supervised feature selection via dependence estimation. In C. Sammut and Z. Ghahramani, editors, *Proceedings of the 24th International Conference on Machine Learning*, pages 823–830. ACM Press, New York.

Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986). Optimally bounded score functions for generalized linear model with applications to logistic regression. *Biometrika*, **73**, 413–424.

Stein, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, Princeton, NJ.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, **2**, 67–93.

Steinwart, I. (2002). Support vector machines are universally consistent. *J. Complexity*, **18**, 768–791.

Steinwart, I. (2003). Sparseness of support vector machines. *J. Mach. Learn. Res.*, **4**, 1071–1105.

Steinwart, I. (2004). Sparseness of support vector machines—some asymptotically sharp bounds. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 1069–1076. MIT Press, Cambridge, MA.

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel machines. *IEEE Trans. Inform. Theory*, **51**, 128–142.

Steinwart, I. (2007). How to compare different loss functions. *Constr. Approx.*, **26**, 225–287.

Steinwart, I. and Anghel, M. (2008). An SVM approach for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. *Ann. Statist.* to appear.

Steinwart, I. and Christmann, A. (2008). How SVMs can estimate quantiles and the median. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA.

Steinwart, I. and Scovel, C. (2005a). Fast rates for support vector machines. In P. Auer and R. Meir, editors, *Proceedings of the 18th Annual Conference on Learning Theory*, pages 279–294. Springer, New York.

Steinwart, I. and Scovel, C. (2005b). Fast rates to Bayes for kernel machines. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1345–1352. MIT Press, Cambridge, MA.

Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, **35**, 575–607.

Steinwart, I., Hush, D., and Scovel, C. (2005). A classification framework for anomaly detection. *J. Mach. Learn. Res.*, **6**, 211–232.

Steinwart, I., Hush, D., and Scovel, C. (2006a). An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory*, **52**, 4635–4643.

Steinwart, I., Hush, D., and Scovel, C. (2006b). Function classes that approximate the Bayes risk. In G. Lugosi and H. U. Simon, editors, *Proceedings of the 19th Annual Conference on Learning Theory*, pages 79–93. Springer, New York.

Steinwart, I., Hush, D., and Scovel, C. (2006c). A new concentration result for regularized risk minimizers. In E. Giné, V. Koltchinskii, W. Li, and J. Zinn, editors, *High Dimensional Probability IV*, pages 260–275. Institute of Mathematical Statistics, Beachwood, OH.

Steinwart, I., Hush, D., and Scovel, C. (2007). An oracle inequality for clipped regularized risk minimizers. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1321–1328. MIT Press, Cambridge, MA.

Steinwart, I., Hush, D., and Scovel, C. (2008). Learning from dependent observations. *J. Multivariate Anal.* to appear.

Stigler, S. M. (1981). Gauss and the invention of least squares. *Ann. Statist.*, **9**, 465–474.

Stone, C. (1977). Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–645.

Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.*, **36**, 424–439.

Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Lett.*, **9**, 293–300.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., and Vandewalle, J. (2002). *Least Squares Support Vector Machines*. World Scientific, Singapore.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. *J. Mach. Learn. Res.*, **7**, 1231–1264.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, **22**, 28–76.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.*, **126**, 505–563.

Tarigan, B. and van de Geer, S. A. (2006). Classifiers of support vector machine type with l_1 complexity regularization. *Bernoulli*, **12**, 1045–1076.

Tewari, A. and Bartlett, P. L. (2005). On the consistency of multiclass classification methods. In P. Auer and R. Meir, editors, *Proceedings of the 18th Annual Conference on Learning Theory*, pages 143–157. Springer, New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **58**, 267–288.

Triebel, H. (1978). *Interpolation Theory, Function Spaces, Differential Operators*. North Holland, Amsterdam.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.*, **25**, 948–969.

Tsybakov, A. B. (2003). Optimal rates of aggregation. In M. Warmuth and B. Schölkopf, editors, *Proceedings of the 16th Annual Conference on Learning Theory*, pages 303–313. Springer, New York.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, **32**, 135–166.

Tukey, J. W. (1975). Mathematics and the picturing of data. In R. D. James, editor, *Proceedings of the 1974 International Congress of Mathematics*, pages 523–531, Vancouver.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Vapnik, V. N. (1982). *Estimation of Dependicies Based on Empirical Data*. Springer, New York.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons, New York.

Vapnik, V. N. and Chervonenkis, A. (1974). *Theory of Pattern Recognition*. Nauka, Moscow. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.

Vapnik, V. N. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Autom. Remote Control*, **24**, 774–780.

Vert, R. and Vert, J.-P. (2006). Consistency and convergence rates of one-class SVMs and related algorithms. *J. Mach. Learn. Res.*, **7**, 817–854.

Vidyasagar, M. (2002). *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*. Springer, London, 2nd edition.

von Mises, R. (1937). Sur les fonctions statistiques. In *Conférence de la Réunion Internationale des Mathématiciens*. Gauthier-Villars, Paris.

Wahba, G. (1990). *Spline Models for Observational Data*. Series in Applied Mathematics 59, SIAM, Philadelphia.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Schölkopf, C. J. C. Burges, and A. Smola, editors, *Advances in Kernel Methods–Support Vector Learning*, pages 69–88. MIT Press, Cambridge, MA.

Werner, D. (1995). *Funktionalanalysis*. Springer, Berlin.

Willard, S. (1970). *General Topology*. Addison-Wesley, Reading, MA.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Witting, H. (1985). *Mathematische Statistik I*. Teubner, Stuttgart.

Wu, M., Schölkopf, B., and Bakır, G. H. (2005). Building sparse large margin classifiers. In L. D. Raedt and S. Wrobel, editors, *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*, pages 996–1003. ACM Press, New York.

Wu, Q., Ying, Y., and Zhou, D.-X. (2007). Multi-kernel regularized classifiers. *J. Complexity*, **23**, 108–134.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.*, **15**, 642–656.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *J. Amer. Statist. Assoc.*, **83**, 406–413.

Yohai, V. J., Stahel, W. A., and Zamar, R. H. (1991). A procedure for robust estimation and inference in linear regression. In W. Stahel and S. Weisberg, editors, *Directions in Robust Statistics and Diagnostics, Part II*, pages 365–374. Springer, New York.

Yurinsky, V. (1995). *Sums and Gaussian Vectors*. Lecture Notes in Math. 1617. Springer, Berlin.

Zălinescu, C. (2002). *Convex Analysis in General Vector Spaces*. World Scientific, Singapore.

Zeidler, E. (1986). *Nonlinear Functional Analysis and Its Applications I: Fixed-Point Theorems*. Springer, New York.

Zhang, T. (2001). Convergence of large margin separable linear classification. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 357–363. MIT Press, Cambridge, MA.

Zhang, T. (2004a). Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.*, **5**, 1225–1251.

Zhang, T. (2004b). Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, **32**, 56–134.

Ziemer, W. P. (1989). *Weakly Differentiable Functions*. Springer, New York.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Ann. Statist.*, **28**, 461–482.

Zwald, L., Bousquet, O., and Blanchard, G. (2004). Statistical properties of kernel principal component analysis. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory*, pages 594–608. Springer, New York.

Notation and Symbols

Miscellaneous

$a := b$, $b =: a$	a is defined by b
c, c_1, c_2	unspecified constants
d	dimension of the vector of explanatory variables X
γ	width of the Gaussian RBF kernel
λ	regularizing constant
n	sample size
(a_i)	shortform for the sequence $(a_i)_{i\geq 1}$

Sets

$\emptyset$	empty set
$\mathbb{N}$, $\mathbb{N}_0$	set of positive integers, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$
$\mathbb{Z}$	set of positive or negative integers including 0
$\mathbb{Q}$	set of rational numbers
$\mathbb{R}$, $[0, \infty)$	set of real numbers, set of non-negative real numbers
$\bar{\mathbb{R}}$	$\mathbb{R} \cup \{-\infty, +\infty\}$
$\mathbb{R}^{n\times d}$	set of $n \times p$ matrices with coefficients in $\mathbb{R}^d$
$\partial\mathbb{R}^d$	$\bar{\mathbb{R}}^d \backslash \mathbb{R}^d$
$\mathbb{C}$	set of complex numbers
$\mathbb{K}$	template for either $\mathbb{R}$ or $\mathbb{C}$
(a, b), $[a, b]$	open or closed intervals in $\mathbb{R}$ or $\mathbb{R}^d$
A^c, $\mathring{A}$, $\overline{A}$	complement, interior, and closure of a set A
∂A	boundary of A, i.e., $\partial A = \bar{A} \backslash \mathring{A}$
$A_1 \uplus A_2$	disjoint union of sets A_1 and A_2
$\lvert A \rvert$	number of elements of a set A

Functions

$\mathbf{1}_A(x)$	indicator function, $\mathbf{1}_A(x) = 1$, if $x \in A$, $\mathbf{1}_A(x) = 0$, else
$f_{\|A}$	restriction of the function f to the set A
$f \le g$	the functions f and g satisfy $f(x) \le g(x)$ for all x
$f^{-1}(A)$	pre-image, i.e., $f^{-1}(A) := \{x \in X : f(x) \in A\}$
$\lfloor \cdot \rfloor$	$\lfloor x \rfloor = \max\{y \in \mathbb{Z} : y \le x\}$, $x \in \mathbb{R}$
$\mathcal{T}$	clipping operation $\mathcal{T} := \min\{-M, \max\{M, t\}\}$
A_2	approximation error function
id	identity map $x \mapsto x$
IF	influence function
SC_n	sensitivity curve

Spaces

X	space of input values (sometimes a closed subset of $\mathbb{R}^d$)
Y	space of output values (a closed subset of $\mathbb{R}$)
H	RKHS or generic Hilbert space
H_γ, $H_\gamma(X)$	RKHS of Gaussian kernel with width γ
$\mathsf{HS}(H)$	space of Hilbert-Schmidt operators on H
$H_1 \oplus H_2$	sum of Hilbert spaces H_1 and H_2
E, F	usually Banach spaces
$C(X)$	space of continuous functions $f : X \to \mathbb{R}$
$C_b(X)$, $C_c(X)$	subspaces of $C(X)$, see Section A.5.5
$C^m(X)$, $C_b^m(X)$	spaces of differentiable functions, see Section A.5.5
$c_0(X)$	space of functions $f : X \to \mathbb{R}$ vanishing at infinity
$\mathcal{L}(E, F)$, $\mathcal{L}(E)$	spaces of bounded linear $S : E \to F$ or $S : E \to E$
$\mathcal{K}(E, F)$, $\mathcal{K}(E)$	spaces of compact operators $S : E \to F$ or $S : E \to E$
$\ell_p(X)$, ℓ_p	space of p-summable functions or sequences
ℓ_p^d	$\mathbb{R}^d$ equipped with the p-norm
$\mathcal{L}_p(\mu)$	set of p-integrable functions (w.r.t. μ)
$L_p(\mu)$	space of equivalence classes of p-integrable functions
$\mathcal{L}_0(X)$	set of measurable functions on X
$\mathcal{L}_0(\mu)$	space of measurable functions equipped with the convergence in measure μ
$L_0(\mu)$	space of equivalence classes of measurable functions equipped with the convergence in measure μ

Norms

$\\| \cdot \\|_2$	Euclidean norm
$\\| \cdot \\|_p$	p-norm
$\\| \cdot \\|_{L_p}$	L_p-norm
$\\| \cdot \\|_\infty$	supremum norm
$\\| \cdot \\|_E$	norm of (generic) Banach space E
$\\| \cdot \\|_H$	norm of RKHS H
$\\| \cdot \\|_{\mathsf{HS}}$	Hilbert-Schmidt norm
$\\| \cdot \\|_{\mathcal{M}}$	norm of total variation
$\\| \cdot \\|_{\mathsf{nuc}}$	nuclear norm

Other symbols related to Banach and metric spaces

$\langle \cdot, \cdot \rangle$, $\langle \cdot, \cdot \rangle_H$	inner product (in Hilbert space H)
B_E	closed unit ball in (Banach) space E
$\dim E$	dimension of the space E
$e_i(A)$	entropy number of A
$\ker S$	kernel of operator S, i.e., $\ker S := \{x \in E : Sx = 0\}$
$\mathcal{N}(A, d, \varepsilon)$	covering number
$\operatorname{rank} S$	rank of operator S
$\operatorname{span} A$	linear span of A
S^*	adjoint of operator S

Measures, probability distributions, and distribution functions

$(\Omega, \mathcal{A})$	generic measurable space with σ-algebra $\mathcal{A}$
$(\Omega, \mathcal{A}, \mathrm{P})$	probability space with distribution P
$(\Omega, \mathcal{A}, \mathcal{P})$	statistical space with family of distributions $\mathcal{P}$
$\sigma(X)$	generic σ-algebra on non-empty set X
$\mathcal{B}$, $\mathcal{B}^d$, $\mathcal{B}(\tau)$	Borel σ-algebra on $\mathbb{R}$, $\mathbb{R}^d$, or w.r.t. topology τ
μ	unspecified measure, sometimes signed measure
$\mu \otimes \nu$	product measure of the measures μ and ν
$\operatorname{supp} \mu$	support of the measure μ (also defined for functions)
λ, λ^d, vol_d	Lebesgue measure on $(\mathbb{R}, \mathcal{B})$, $(\mathbb{R}^d, \mathcal{B}^d)$
$\#$	counting measure
P, $\tilde{\mathrm{P}}$, Q, $\tilde{\mathrm{Q}}$	probability distributions
P_X	marginal distribution
$\lvert \mathrm{P} \rvert_p$	(average) p-moment of a distribution
$\mathrm{P}(\cdot \lvert x)$, $\mathrm{P}(\cdot \lvert X = x)$	regular conditional distribution
D, D_n	empirical distribution associated to the data set D
δ_x, $\delta_{\{x\}}$	Dirac measure at the point x
$\mathcal{M}_1$, $\mathcal{M}_1(Z)$	set of all probability distributions on a measurable space
$\mathcal{P}$, $\mathcal{Q}$	set of probability distributions
$N_\varepsilon(\mathrm{P})$	contamination neighborhood
Q_ε	mixture distribution in $N_\varepsilon(\mathrm{P})$, i.e., $\mathrm{Q}_\varepsilon = (1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}$
$\mathrm{Bi}(m, \pi)$	binomial distribution, $m \in \mathbb{N}$, $\pi \in (0, 1)$
$\mathrm{N}_{\mu,\sigma^2}$	Gaussian distribution, $\mu \in \mathbb{R}$, $\sigma \in (0, \infty)$
$\mathrm{Poi}(\beta)$	Poisson distribution, $\beta \in (0, \infty)$
Λ	$\Lambda(z) = 1/(1 + e^{-z})$, $z \in \mathbb{R}$, c.d.f. of logistic distribution

Random variables, random vectors, and related quantities

X, X_i	inputs, explanatory variables
Y, Y_i	outputs, response variables
Z, Z_i	$Z_i = (X_i, Y_i)$
$\mathbb{E}(X)$, $\mathbb{E}_\mathrm{P}(X)$	expectation of X w.r.t. P
$\mathbb{E}_{x \sim \mathrm{P}} f(x)$, $\mathbb{E}_\mathrm{P} f$	expectation of f w.r.t. P
$\mathbb{E}_\mathrm{D}(X)$	empirical average of X w.r.t. data set D
$\mathrm{Var}(X)$	variance of X
$\mathrm{Cov}(X)$	covariance matrix of X

Estimators

$f_{\mathrm{P},\lambda}$	general SVM decision function w.r.t. P
$f_{\mathrm{P},\lambda,\gamma}$	general SVM decision function in Gaussian RKHS
$f_{\mathrm{D},\lambda}$, $f_{\mathrm{D},\lambda}$	empirical SVM decision function w.r.t. data set D
$f_{\mathrm{D},\lambda,\gamma}$, $f_{\mathrm{D},\lambda,\gamma}$	empirical SVM decision function in Gaussian RKHS
$\hat{f}_{\mathrm{P},\lambda}$	estimator for $f_{\mathrm{P},\lambda}$
$b_{\mathrm{P},\lambda}$	bias or offset of solution of regularized empirical risk
$b_{\mathrm{D},\lambda}$	bias or offset of solution of regularized empirical risk
$S(\mathrm{P})$	value of statistic S at distribution P, often $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$

Kernels and related functions

k	kernel
k_{linear}	linear kernel, $k_{\mathrm{linear}}(x, x') := \langle x, x' \rangle$
k_{RBF}, k_γ	Gaussian RBF kernel
Φ	(canonical) feature map of RKHS H
S_k, T_k	integral operators associated with k

Loss functions

L	generic loss function
$L \circ f$	loss composed with f, i.e., $(x, y) \mapsto L(x, y, f(x))$
$\breve{L}$	loss function used for self-calibration
L_{class}, $L_{\alpha\text{-class}}$	(weighted) classification loss function
L_{hinge}	hinge loss function
$L_{\text{c-logist}}$	logistic loss function for classification
L_{LS}	least squares loss function
$L_{\text{trunc-ls}}$	truncated least squares for classification
L_{DLD}	density level loss function
$L_{\epsilon\text{-insens}}$	ϵ-insensitive loss function for regression
$L_{\text{r-logist}}$	logistic loss function for regression
$L_{\alpha\text{-Huber}}$	Huber's loss function for regression, $\alpha > 0$
$L_{\tau\text{-pin}}$	pinball loss function for quantile regression, $\tau \in (0, 1)$

Risks

$\mathcal{R}_{L,\mathrm{P}}(\cdot)$	L-risk w.r.t. P
$\mathcal{R}^*_{L,\mathrm{P}}$	L-Bayes risk w.r.t. distribution P
$\mathcal{R}_{L,\mathrm{D}}(\cdot)$	empirical L-risk w.r.t. data set D
$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(\cdot)$	regularized L-risk w.r.t. P
$\mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(\cdot)$	regularized empirical L-risk w.r.t. D

Abbreviations

Abbreviation	Explanation
ϵ-CR-ERM	ϵ-approximate CR-ERM
CR-ERM	clipped regularized empirical risk minimization
DLD	density level detection
ERM	empirical risk minimization
GLIM	generalized linear model
IF	influence function
i.i.d.	independent and identically distributed
l.s.c.	lower semi-continuous
KBQR	kernel based quantile regression
KLR	kernel logistic regression
ONB	orthonormal basis
ONS	orthonormal system
RBF	radial basis function
RKHS	reproducing kernel Hilbert space
RLB	robust learning from bites
SMO	sequential minimal optimization
s.t.	subject to
SVM	support vector machine
TV-SVM	training validation support vector machine
w.r.t.	with respect to

Author Index

Subject Index