Vrije Universiteit Brussel

Faculty of Sciences and
Bio-engineering Sciences

Department of Mathematics

Robustness and Consistency Results for Support Vector Machines

Dissertation submitted to the Vrije Universiteit Brussel
in fulfillment of the requirements for the degree of Doctor
in Science

Arnout Van Messem

Supervisor: Prof. Dr. Uwe Einmahl
Co-Supervisor: Prof. Dr. Andreas Christmann

Academic Year: 2010 – 2011

VRIJE UNIVERSITEIT BRUSSEL • SCIENTIA VINCERE TENEBRAS •

Print: Silhouet, Maldegem

2011 Uitgeverij VUBPRESS Brussels University Press
VUBPRESS is an imprint of ASP nv (Academic and Scientific Publishers nv)
Ravensteingalerij 28
B-1000 Brussels
Tel. +32 (0)2 289 26 50
Fax +32 (0)2 289 26 59
E-mail: info@vubpress.be
www.vubpress.be

ISBN 978 90 5487 910 7
NUR 916/919/984
Legal deposit D/2011/11.161/073

Acknowledgments

First of all I would like to express my fondest gratitude to my supervisor Andreas Christmann who introduced me to the world of support vector machines and who motivated me during my research. His support and guidance were an invaluable help in accomplishing this work. Even after his return to Germany, he never failed to make time for me and to answer my questions. I also would like to thank Andreas for his hospitality during my visits in Bayreuth, for introducing me to his colleagues and for giving me the opportunity to present my work at several conferences.

Secondly, I would like to thank Uwe Einmahl, who took over the role of supervisor after Andreas got a position in Bayreuth, and who guided me during the final part of my thesis.

A special word of thanks goes to Christophe Croux without who I probably never would have written this thesis. It was Christophe whose courses increased my interest in statistics and who, as supervisor of my licentiate thesis, introduced me to robust statistics.

I also would like to thank the members of the jury – Christophe Croux , Julia Dony, Tetyana Kadankova, Davy Paindaveine, Mark Sioen, and Stefan Van Aelst – for their insightful questions, comments and remarks, which helped me to improve this work.

I can also not forget to mention my colleagues here at the VUB for creating a nice and enjoyable working atmosphere. In particular I think of my two office mates Ann and Julia without whom these years would have been a lot duller.

I also think back fondly at the many conferences I attended and the people I met there. I want to thank Anneleen, Christel, Dina, Ellen, Géraldine, Koen, Leen, Martin, Michiel, Monique, Sabine, Sarah, Šárka, Thomas, Tim, and Yvik (and all others I might have forgotten to mention) for the many pleasant hours spent both during the lectures as well as afterwards.

A word of thanks is in order for everyone who made the hours outside my working environment fun and who were curious about my progress and encouraged me when things slacked or didn't seem to advance. I particularly want to mention Dimi, Fre, Koen, and Tara as important members in this category. Also training at the Hybrid Dojo and Tenzan Dojo helped me to relax and stay *zen* in more troubled times.

And last, but certainly not least, I would like to thank my family and my girlfriend. First of all I cannot thank my parents Rudi and Immi enough. Without them I would never have stood here. They gave me the opportunity to do the things I wanted to do and have always encouraged me in everything I tried. My brothers as well as my grandparents were always interested in my work and supported me along the way. A big, warm, heartfelt thanks is also in order to my girlfriend Debbie, who stood beside me on every step of the road in the completion of this work, who never failed to motivate me and who was always there for me. I love you.

Arnout Van Messem
May 2011

Summary

Classification and regression problems have been known and studied for a long time. Recent technological developments made it possible to implement algorithms that will do the work for us. This method is called *statistical machine learning. Support vector machines* (SVMs) are one of the techniques of machine learning, which also have the advantage of being a non-parametric method, and therefore do not assume any, or very little, prior knowledge on the underlying distribution of the data. For prediction purposes, such as classification or regression, we will always assume to be in the possession of a set of data points, the training data set $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, with $\mathcal{X}$ the set of input variables and $\mathcal{Y}$ the set of possible output variables. The statistical method will then use these (observed) data in order to obtain a prediction function, that will assign an output to new unobserved inputs. In theory, statistical procedures work well given a data set, however, there always exists the possibility that errors entered in the data, which can make the prediction obtained by our statistical method unreliable. Even small errors might cause huge differences in the prediction. These errors might be extremes or outliers in the data, measurement errors, or just simply typing errors when entering the data in the system, but the result is the same: if the method is not robust, the predicted function might be very different from the true underlying function. Also data obtained from distributions with heavy tails might pose a problem for a statistical prediction method.

In the first chapter, we will give a short introduction to the field of statistical learning as well as an overview on the different stages of the development of SVMs. We will first recreate the original support vector machines as introduced by Vapnik and Lerner (1963), Boser *et al.* (1992) and Cortes and Vapnik (1995) which were based on the underlying geometrical idea and we will then move on to study the *empirical risk minimization* method that led to the analytic description of support vector machines.

Analytically, support vector machines are kernel methods that are inspired by convex risk minimization in infinite dimensional Hilbert spaces. This technique is quite popular for three reasons. Firstly, SVMs are very flexible methods. They were introduced as a linear classification method, but by using an appropriate kernel, their use can easily be extended to the non-linear case. Secondly, due to their sparseness, they are computationally efficient and thirdly, they can easily deal with large data sets with unknown, complex and high-dimensional dependency structures, which can occur for example in bioinformatics or genetics.

Empirical risk minimization (ERM) will look for a function $f : \mathcal{X} \to \mathbb{R}$ that minimizes the L-risk

$$\mathcal{R}_{L,\mathrm{P}}(f) := \mathbb{E}_{\mathrm{P}} L(X, Y, f(X)) \, ,$$

where $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called a loss function and gives an indication of how well the prediction $f(x)$ approximates the true observed output y. The smallest possible risk $\mathcal{R}^*_{L,\mathrm{P}}$ is called the Bayes risk, and the function f^* that achieves the Bayes risk is called the Bayes function. It is however impossible to optimize the risk over all functions, and so a reasonably large subclass of functions will be considered, namely the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. Furthermore, optimizing the risk without further ado might produce a very wiggly or irregular function that will indeed give excellent approximations $f(x)$ of y for observed data points x, but will perform very bad for previously unseen inputs. To avoid this overfitting of the data, a regularization term will be added to the risk, which will then be called the regularized risk

$$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) := \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \left\| f \right\|^2_{\mathcal{H}} \, .$$

This additional term will restrict the function, since very wiggly functions will have large norms and are thus less likely to be selected as optimal function. The support vector machine is then defined as

$$f_{L,\mathrm{P},\lambda} := \arg \inf_{f \in \mathcal{H}} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \left\| f \right\|^2_{\mathcal{H}} \, .$$

Next, we will take a closer look at the loss function and its associated risk and explain why convex and Lipschitz continuous losses are of special importance. Also kernels and the RKHS will be treated more in detail.

An important part of Chapter 1 is Section 1.7 in which the simple concept of the *shifted loss* function $L^\star$ is introduced. From a non-parametric point of view, it is difficult to know in supervised machine learning whether the moment condition $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ is fulfilled. However, some recent results on consistency and statistical robustness properties of SVMs, which are based on a Lipschitz continuous loss function and a bounded kernel, for unbounded output spaces were derived under the assumption that this absolute moment is finite, see e.g., Christmann and Steinwart (2007, 2008), and Steinwart and Christmann (2008b). Unfortunately, this condition excludes distributions with heavy tails, such as the Cauchy distribution, and extreme value distributions that can occur in financial or actuarial problems. In order to enlarge the applicability of SVMs to situations where the output space $\mathcal{Y}$ is unbounded, e.g., $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = [0, \infty)$, without the need for the above mentioned moment condition, we will introduce the $L^\star$-trick. This trick, previously used in robust statistics by, e.g., Huber (1967), consists in shifting the loss function L downwards to obtain the new, shifted loss $L^\star : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ defined by

$$L^\star(x, y, t) := L(x, y, t) - L(x, y, 0)\,.$$

Some properties of $L^\star$ and its associated risk $\mathcal{R}_{L^\star,\mathrm{P}}(f)$ and regularized risk $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f)$ are examined. We also show that, under rather weak conditions that do *not* depend on the data and are thus easily verifiable, the SVM $f_{L^\star,\mathrm{P},\lambda}$ based on the shifted loss function does exist and is unique. This explains the importance of the shifted loss, because it allows us to investigate statistical properties of support vector machines for *all* probability measures P, since no moment assumptions on the conditional distribution of Y given x are needed to guarantee the existence of such SVMs. In this section we also give a representation of the support vector machine based on $L^\star$ in function of the subdifferential. Assumption 1.7.1 contains the general assumptions for this thesis.

In Chapter 2 we will take a look at the *consistency* of support vector machines. Consistency is an important concept for learning methods, since it will guarantee that the method will effectively learn. For an L-risk consistent method, the empirical risk $\mathcal{R}_{L,\mathrm{P}}(f_D)$ of a decision function f_D that is based on the training data set D, will converge to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$ and the decision function produced by the method will thus be near optimal. A method is called consistent if the function f_D converges to the Bayes function f^*. Christmann and Steinwart (2007) and Steinwart and Christmann (2008b) already showed that SVMs are L-risk consistent, and that the SVM for quantile regression based on the pinball loss is consistent. However, they needed the moment condition $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ to obtain these results. Their results therefore exclude heavy-tailed distributions.

In Section 2.2 we will use the $L^\star$-trick to enlarge the applicability of these results to also include SVMs based on data coming from heavy-tailed distributions that thus violate the moment condition $\mathbb{E}_\mathrm{P}|Y| < \infty$. Given some weak conditions we are able to show that support vector machines based on shifted loss functions are L-risk consistent and we also obtain a consistency result for the specific case of the pinball loss. The necessary conditions are satisfied, if the researcher uses a convex Lipschitz continuous loss function, a bounded continuous kernel, and a null-sequence of positive regularization parameters λ_n such that $\lambda_n^2 n$ converges to infinity.

In Chapter 3 we will investigate the *robustness* properties of support vector machines. As mentioned before, outliers or extreme data points might have a bad influence on the prediction function. To avoid this effect, robust methods have been developed. These methods will look for the model that fits the majority of the data and will hence be more robust against possible outliers in the data. The method will thus still perform (reasonably) well, even in the presence of outliers or bad data points. One method to verify the robustness of a statistic $T : \mathcal{M}_1 \to \mathcal{H}$[1] is to look at its *influence function* (IF) at a point $z \in Z$ for a distribution P, which was defined by Hampel (1968) as

$$\mathrm{IF}(z; T, \mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z) - T(\mathrm{P})}{\varepsilon} \, ,$$

if this limit exists. From the definition it is clear that the IF is related to Gâteaux-derivatives. Furthermore, the IF has become a cornerstone of robust statistics, see Hampel (1974), Hampel *et al.* (1986), and Maronna *et al.* (2006). Within the approach of influence functions, a method is called robust if its IF is bounded, thus if the contaminating distribution only has a limited influence on the statistical method. Recent results on the robustness of SVMs by using the influence function can be found, e.g., in Christmann and Steinwart (2004, 2007) and Steinwart and Christmann (2008a). One of the conditions however is that the loss function has to be twice Fréchet-differentiable, and of the most often used losses, only the logistic loss fulfills this property. Therefore we introduce a generalization of the IF that is based on *Bouligand-derivatives* (Robinson, 1991). Bouligand-derivatives and strong Bouligand-derivatives have successfully been used in approximation theory for non-smooth functions, see for example Clarke (1983), Robinson (1987, 1991), Ip and Kyparisis (1992), and

[1] Here $\mathcal{M}_1$ is the set of all probability distributions on some measurable space $(Z, \mathcal{B}(Z))$.

the references cited therein. To the best of our knowledge however, these directional derivatives were not yet used in robust statistics. We first calculate the Bouligand-derivative of some loss functions and then, using these Bouligand-derivatives, we introduce the *Bouligand influence function* (BIF) as a modification of Hampel's influence function. The Bouligand influence function of the function $T : \mathcal{M}_1 \to \mathcal{H}$ for a distribution P in the direction of a distribution $\mathrm{Q} \neq \mathrm{P}$ is the special Bouligand-derivative (if it exists)

$$\lim_{\varepsilon \downarrow 0} \frac{\big\| T\big((1-\varepsilon)\mathrm{P} + \varepsilon \mathrm{Q}\big) - T(\mathrm{P}) - \mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) \big\|_{\mathcal{H}}}{\varepsilon} = 0\,.$$

The BIF will allow us to study the robustness of a broad class of support vector machines with non-smooth loss functions. Examples of these are the support vector machine based on the ϵ-insensitive loss function, and kernel based quantile regression based on the pinball loss function. Another advantage of the BIF, besides being able to work with non-smooth functions, is that it is positive homogeneous which is in general not true for Hampel's influence function. A third advantage of the BIF is that the contaminating distribution is not restricted to a point distribution as is the case for the IF, but that it can be any distribution $\mathrm{Q} \neq \mathrm{P}$. Furthermore, if the BIF exists, then the IF exists and both are equal.

In Sections 3.3 and 3.4 we will show that many support vector machines for regression have a bounded Bouligand influence function and are thus robust. Section 3.3 contains our first robustness result: under some weak assumptions, the BIF for support vector machines based on a bounded kernel and a twice Bouligand-differentiable, Lipschitz continuous and convex loss function exists, it is bounded and an explicit formula for the BIF of $T(\mathrm{P}) := f_{L,\mathrm{P},\lambda}$ in the direction of $\mathrm{Q} \neq \mathrm{P}$ is given by:

$$\begin{aligned} \mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) \quad &= \quad S^{-1}\big(\mathbb{E}_{\mathrm{P}} \nabla_3^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \Phi(X)\big) \\ &\qquad -S^{-1}\big(\mathbb{E}_{\mathrm{Q}} \nabla_3^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \Phi(X)\big)\,, \end{aligned}$$

where $S : \mathcal{H} \to \mathcal{H}$ is defined as

$$S(\,\cdot\,) := 2\lambda\, \mathrm{id}_{\mathcal{H}}(\,\cdot\,) + \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \langle \Phi(X), \,\cdot\, \rangle_{\mathcal{H}} \Phi(X)\,.$$

We then go on to prove that this result covers the important special cases of SVMs based on the loss functions L_ϵ, $L_{\tau-pin}$, and $L_{c-Huber}$, which were not covered by earlier results on the influence function, as well as L_{r-log}. The IF of support vector machines based on the logistic loss was recently derived by Christmann and Steinwart (2007) and Steinwart and Christmann (2008b).

We also show that the asymptotic bias $f_{L,(1-\alpha\varepsilon)\mathrm{P}+\alpha\,\varepsilon\mathrm{Q},\,\lambda} - f_{L,\mathrm{P},\lambda}$ will be of the form $\alpha\,\mathrm{BIF}(\mathrm{Q};T,\mathrm{P}) + o(\alpha h)$, with $h = \varepsilon(\mathrm{Q}-\mathrm{P})$. Even though we are able to avoid that the loss function needs to be Fréchet-differentiable, we still have the condition $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ as an assumption. This means that heavy-tailed distributions are not covered by this theorem, which leads to the second part of our robustness results.

In Section 3.4 we will use the $L^\star$-trick to extend some existing robustness results to heavy-tailed distributions. First we adapt the robustness result on the influence function derived in Christmann and Steinwart (2007) and Steinwart and Christmann (2008b). We show that, using a shifted loss function, the IF of a support vector machine $f_{L^\star,\mathrm{P},\lambda}$ based on a bounded, continuous kernel and a twice Fréchet-differentiable, Lipschitz continuous and convex loss function exists and is bounded, even for heavy-tailed distributions. We obtain a formula for the influence function similar to the one obtained by Christmann and Steinwart (2007) and Steinwart and Christmann (2008b). A bound for the bias $\left\| f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{L^\star,\mathrm{P},\lambda} \right\|_{\mathcal{H}}$ is also derived. This difference will increase at most linearly. In the next step we recalculate the Bouligand influence function for the SVM, but this time using the $L^\star$-trick. We show that also here the BIF of $f_{L^\star,\mathrm{P},\lambda}$ exists, that it is bounded and given by the above formulation if we use a bounded, continuous kernel and a twice Bouligand-differentiable, Lipschitz continuous and convex loss. Hence SVMs are robust in the sense of influence functions, even when the underlying distribution has heavy tails or is an extreme value distribution. We conclude this chapter with some numerical examples in which we demonstrate the usefulness of SVMs even for heavy-tailed distributions by applying support vector machines to a simulated data set with Cauchy errors and to a data set of large fire insurance claims of Copenhagen Re.

Appendix A contains all mathematical prerequisites that might be needed to fully understand this thesis. I tried to present this work as self-containing, so the topics here range from topology over functional analysis and probability theory to convex optimization. Appendix B gives an overview of the notations and abbreviations used in this work.

To summarize, in this thesis we investigate the robustness and consistency of support vector machines. We have introduced a new notion on robustness, called the Bouligand influence function, which is a modification of the classical influence function. Using this method, we show that SVMs that use non-smooth loss functions are robust in the sense of influence functions. Furthermore, by using the $L^\star$-trick, we are also able to proof that SVMs based on the shifted loss are both $L^\star$-risk consistent and robust. Using this trick allows us to treat even SVMs for which the underlying distribution has heavy tails or is an extreme value distribution, something that was excluded from previous results. In the special case of the pinball loss, we also show consistency of the decision function.

Other recent work on SVMs includes Christmann and Hable (2011) in which the authors investigate the robustness – by using the BIF – and consistency of SVMs for additive models, Hable and Christmann (2011) where they look at the qualitative robustness of support vector machines, Steinwart and Christmann (2009b) on the sparsity of SVMs that use the ϵ-insensitive loss function, and Xu *et al.* (2009) who investigate the equivalence between the regularization of the SVM and robust optimization.

Let us conclude with some considerations that might be made after reading this work. First, we decided to consider only non-negative loss functions L (although the shifted loss function $L^\star$ can have negative values), because almost all loss functions used in practice are non-negative and no results on SVMs seem available for loss functions with negative values. Secondly, it might be possible to derive results similar to ours for convex, but locally Lipschitzian loss functions, including the least squares loss, although Lipschitz continuous loss functions can offer better robustness properties, see Christmann and Steinwart (2004, 2007) and Steinwart and Christmann (2008b). And thirdly, from a robustness point of view, *bounded* and *non-convex* loss functions might also be of interest. We did not consider these loss functions for two reasons. The first reason is that existence, uniqueness, consistency, and availability of efficient numerical algorithms are widely accepted as necessary properties which SVMs should have to avoid numerically intractable problems for large and high-dimensional data sets, say for $n > 10^5$ and $d > 100$, see e.g. Vapnik (1998) or Schölkopf and Smola (2002). These properties will be achieved if the risk is convex which is the case for convex loss functions. Secondly, there are currently – as far as we know – no general results on support vector machines available which guarantee that the risk will be convex although the loss function is non-convex and bounded. However, the convexity of the risk plays a key role in the proofs of the existence and uniqueness of SVMs. From our point of view, such results should therefore first be examined before investigating bounded and non-convex loss functions.

Finally, we would like to point out that, even though the Bouligand-derivative has not been used in robust statistics before, we think that it is a promising concept for the following reason. Many robust estimators that are proposed in the literature are implicitly defined as solutions of some minimization problem where the objective function or loss function is continuous or Lipschitz continuous, but not necessarily twice Fréchet-differentiable. Examples are not only support vector machines, but also M-estimators of Huber-type and certain maximum likelihood estimators under non-standard conditions. Bouligand-differentiation nicely fills the gap between Fréchet-differentiation, which is too strong for many robust estimators, and Gâteaux-differentiation which is the basis for the robustness approach based on influence functions. Bouligand-derivatives fulfill a chain rule as well as an implicit function theorem which is in general not true for Gâteaux-derivatives.

This thesis has left some possibilities for further research. A first option might be to investigate the consistency of SVMs that are based on other loss functions than the pinball loss, such as, e.g., the support vector machine based on the ϵ-insensitive loss. Another interesting point could be to look more in detail at robustness properties of SVMs for classification. We did obtain results on the Bouligand influence function that, given our assumptions, are valid for all support vector machines, but we were so far only able to check the assumption concerning the strong Bouligand-derivative for loss functions for regression. A final road to investigate might be to consider either locally Lipschitz loss functions instead of Lipschitz continuous losses or to look at bounded and non-convex losses instead of convex losses.

The work presented in this thesis is based on the following publications:

- ⋆ Christmann, A. and Van Messem, A. (2008). Bouligand Derivatives and Robustness of Support Vector Machines for Regression. *Journal of Machine Learning Research*, **9**, 915–936.
- ⋆ Christmann, A., Van Messem, A., and Steinwart, I. (2009). On Consistency and Robustness Properties of Support Vector Machines for Heavy-Tailed Distributions. *Statistics and Its Interface*, **2**, 311–327.
- ⋆ Van Messem, A. and Christmann, A. (2010). A review on consistency and robustness properties of support vector machines for heavy-tailed distributions. *Advances in Data Analysis and Classification*, **4**(2-3), 199–220.

Samenvatting

Classificatie- en regressieproblemen worden al lange tijd bestudeerd, maar recente technologische ontwikkelingen hebben het mogelijk gemaakt om algoritmes te ontwikkelen en te implementeren die deze problemen voor ons kunnen oplossen. Dit wordt *statistical machine learning* genoemd. Een voorbeeld van zulk een techniek zijn *support vector machines* (SVMs). Deze methode is niet-parametrisch en heeft bijgevolg geen, of erg weinig, voorafgaande kennis over de onderliggende verdeling van de data nodig. Om voorspellingen te kunnen maken, zoals bij classificatie of regressie, zullen we steeds veronderstellen dat we een verzameling van datapunten hebben, de verzameling van trainingsdata $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, waarbij $\mathcal{X}$ de verzameling van de input variabelen is en $\mathcal{Y}$ de verzameling van de mogelijke output variabelen. De statistische methode zal vervolgens deze (geobserveerde) data gebruiken om een predictiefunctie op te stellen. Deze predictor zal een outputwaarde toekennen aan elke nieuwe, nog niet geobserveerde input. In theorie werken statistische methodes goed met een gegeven data set, maar in de praktijk bestaat er altijd de mogelijkheid dat de data fouten bevatten waardoor de voorspelling door onze statistische methode onbetrouwbaar kan zijn: zelfs kleine foutjes kunnen voor een groot verschil zorgen in de voorspelde waarden. Er kunnen zich op verschillende manieren problemen in de data voordoen: extreme waarden of uitschieters kunnen in de data set voorkomen, of er kunnen zich simpelweg typfouten hebben voorgedaan bij het ingeven van de gegevens in het systeem. Het resultaat is echter telkens hetzelfde: als de gebruikte methode niet robuust is, kan de gevonden functie enorm verschillen van de echte, onderliggende functie. Ook data komende van verdelingen met zware staarten kunnen een probleem vormen voor een statistische voorspellingsmethode.

Het eerste hoofdstuk bevat een korte inleiding over statistische leermethodes alsook een overzicht van de verschillende stages die SVMs doorlopen hebben tijdens hun ontwikkeling. We beginnen met het opstellen van de

originele support vector machines zoals ingevoerd door Vapnik en Lerner (1963), Boser *et al.* (1992) en Cortes en Vapnik (1995) en die gebaseerd waren op een geometrisch idee. Daarna gaan we over naar de methode via *empirische risicominimalisatie* die leidde tot de analytische beschrijving van support vector machines. Analytisch gezien zijn support vector machines kern-methodes die geïnspireerd zijn op convexe risicominimalisatie in oneindigdimensionale Hilbert ruimten. Er zijn drie redenen waarom deze techniek zo populair is. Ten eerste zijn SVMs erg flexibel. Oorspronkelijk konden ze enkel lineaire classificatie uitvoeren, maar door gebruik te maken van een geschikte kern werd dit al snel uitgebreid naar het niet-lineaire geval. Een tweede voordeel is dat de methode rekenkundig erg efficiënt is doordat het maar een kleine fractie van het totale aantal datapunten nodig heeft om mee te werken. En ten derde kunnen SVMs makkelijk overweg met grote data sets waarin zich ongekende, complexe en hoogdimensionale afhankelijkheden bevinden, zoals bij voorbeeld in de genetica of bij bio-informatica.

Empirische risicominimisatie (ERM) zal een functie $f : \mathcal{X} \to \mathbb{R}$ zoeken die het L-risico

$$\mathcal{R}_{L,\mathrm{P}}(f) := \mathbb{E}_{\mathrm{P}} L(X, Y, f(X))$$

minimaliseert. Hierbij is $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ een verliesfunctie die een indicatie geeft van hoe goed de voorspelling $f(x)$ de geobserveerde output y benadert. Het kleinst mogelijke risico $\mathcal{R}^*_{L,\mathrm{P}}$ wordt het Bayes risico genoemd, en de functie f^* waarvoor het Bayes risico bereikt wordt, is de Bayes functie. Het is echter onmogelijk om het risico te minimaliseren over alle functies, en dus zal er een voldoend grote deelverzameling van functies, de reproducerende kern Hilbert ruimte (RKHS) $\mathcal{H}$, beschouwd worden. Bovendien kan het klakkeloos optimaliseren van het risico leiden tot zeer onregelmatige of oscillerende functies. Deze functies zullen zeer goede benaderingen $f(x)$ van y produceren voor geobserveerde datapunten x, maar zullen het extreem slecht doen voor nog niet eerder waargenomen inputs. Om dit overfitten te vermijden, wordt een regularisatieterm toegevoegd aan het risico. Het geregulariseerde risico is dan

$$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) := \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \, \|f\|^2_{\mathcal{H}} \, .$$

Deze extra term beperkt de doelfunctie f aangezien zeer onregelmatige functies een grote norm zullen hebben en dus minder in aanmerking komen als optimale functie. De support vector machine zelf is gedefinieerd als

$$f_{L,\mathrm{P},\lambda} := \arg \inf_{f \in \mathcal{H}} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \, \|f\|^2_{\mathcal{H}} \, .$$

Vervolgens zullen we de verliesfunctie en het daarbij horende risico meer in detail bestuderen. We leggen tevens uit waarom convexe en Lipschitz continue functies zo belangrijk zijn en kijken ten slotte nog naar kernen en de RKHS.

Een belangrijk deel van Hoofdstuk 1 is Sectie 1.7 waarin het concept van de *verschoven verliesfunctie* $L^\star$ wordt ingevoerd. Vanuit een niet-parametrisch standpunt is het een moeilijke opgave om te weten of de momentvoorwaarde $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ al dan niet voldaan is bij machine learning. Recente resultaten in oneindigdimensionale output-ruimten over de consistentie en statistische robuustheid van SVMs, gebaseerd op een Lipschitz continue verliesfunctie en een begrensde kern, maken echter gebruik van het feit dat dit absolute moment eindig is, zie b.v. Christmann en Steinwart (2007, 2008) en Christmann en Steinwart (2008b). Jammer genoeg sluit deze voorwaarde verdelingen met zware staarten, zoals de Cauchy verdeling, alsook extreme waarde verdelingen die bij financiële of actuariële problemen kunnen voorkomen, uit. Om de bruikbaarheid van SVMs dus uit te breiden naar situaties waar de output-ruimte $\mathcal{Y}$ onbegrensd is, b.v. $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = [0, \infty)$, zonder hierbij de hierboven genoemde momentvoorwaarde te vereisen, zullen we gebuik maken van de $L^\star$-truc. Deze truc, die al eerder gebruikt is in robuuste statistiek door b.v. Huber (1967), bestaat erin om de verliesfunctie L neerwaarts te verschuiven en zo de nieuwe, verschoven verliesfunctie

$$L^\star(x, y, t) := L(x, y, t) - L(x, y, 0)$$

te bekomen. We bestuderen de eigenschappen van $L^\star$ en het daaraan geassocieerde risico $\mathcal{R}_{L^\star,\mathrm{P}}(f)$ en geregulariseerde risico $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f)$. We tonen tevens dat, gegeven enkele zwakke voorwaarden die *niet* afhangen van de data en die dus makkelijk na te gaan zijn, de SVM $f_{L^\star,\mathrm{P},\lambda}$ gebaseerd op de verschoven verliesfunctie effectief bestaat en zelfs uniek is. Het verschuiven van de verliesfunctie laat dus toe om de statistische eigenschappen van support vector machines te bestuderen voor *alle* kansmaten P aangezien er geen momentvoorwaarde op de voorwaardelijke verdeling van Y gegeven x meer nodig is om het bestaan van zulke SVMs te garanderen. In deze sectie geven we ook een representatie van de support vector machine gebaseerd op $L^\star$ in functie van de subdifferentiaal. Aanname 1.7.1 bevat de algemene assumpties voor deze thesis.

In Hoofdstuk 2 bestuderen we de *consistentie* van support vector machines. Consistentie is een belangrijk concept voor leermethodes daar het garandeert dat de methode effectief zal leren. Voor een L-risico consistente methode zal het empirisch risico $\mathcal{R}_{L,\mathrm{P}}(f_D)$ voor een beslissingsfunctie f_D gebaseerd op de trainingsdata D, convergeren naar het Bayes risico $\mathcal{R}^*_{L,\mathrm{P}}$. De door de methode bekomen beslissingsfunctie zal dus bijna optimaal zijn. Een methode wordt consistent genoemd als de functie f_D convergeert naar de Bayes functie f^*. Christmann en Steinwart (2007) en Steinwart en Christmann (2008b) toonden reeds aan dat SVMs L-risico consistent zijn, en dat de SVM voor kwantielregressie gebaseerd op de pinball verliesfunctie consistent is. Hiervoor maakten ze echter wel gebruik van de momentvoorwaarde $\mathbb{E}_{\mathrm{P}}|Y| < \infty$. Hun resultaten sluiten bijgevolg verdelingen met zware staarten uit.

In Sectie 2.2 zullen we de $L^\star$-truc gebruiken om de toepasbaarheid van deze resultaten uit te breiden naar SVMs gebaseerd op data komende van verdelingen met zware staarten die dus de momentvoorwaarde $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ violeren. Onder enkele zwakke voorwaarden kunnen we aantonen dat support vector machines gebaseerd op verschoven verliesfuncties L-risico consistent zijn en bekomen we eveneens een consistentieresultaat voor het specifieke geval van de pinball verliesfunctie. De benodigde voorwaarden zijn voldaan als de onderzoeker een convexe en Lipschitz continue verliesfunctie, een begrensde continue kern, en een nulrij van positieve regularisatieparameters λ_n zodat $\lambda_n^2 n$ naar oneindig convergeert gebruikt.

In Hoofdstuk 3 zullen we *robuustheidseigenschappen* van support vector machines onderzoeken. Zoals reeds eerder gezegd, kunnen uitschieters of extreme waarden een slechte invloed uitoefenen op de voorspellingsfunctie. Om dit te vermijden, zijn er robuuste methodes ontwikkeld. Deze methodes zullen een model zoeken dat goed aansluit bij het merendeel van de data, maar dat beter bestand is tegen mogelijke uitschieters in de data. De methode zal dus nog steeds (redelijk) goed presteren, zelfs in de aanwezigheid van uitschieters of slechte datapunten. Een manier of de robuustheid van een statistiek $T : \mathcal{M}_1 \to \mathcal{H}$[2] te testen, is door te kijken naar haar *invloedsfunctie* (IF) in een punt $z \in Z$ en voor een verdeling P. De IF is door Hampel (1968) gedefinieerd als

$$\mathrm{IF}(z; T, \mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z) - T(\mathrm{P})}{\varepsilon},$$

als deze limiet bestaat. Uit de definitie is het duidelijk dat de IF direct gerelateerd is aan Gâteaux-afgeleiden. Bovendien is de IF een hoeksteen geworden van de robuuste statistiek, zie hiervoor Hampel (1974), Hampel *et al.*

[2] Hierbij is $\mathcal{M}_1$ de verzameling van alle kansverdelingen op een meetbare ruimte $(Z, \mathcal{B}(Z))$.

(1986), en Maronna *et al.* (2006). Binnen de benadering door invloedsfuncties wordt een methode robuust genoemd als haar IF begrensd is, dus als de contaminerende verdeling slechts een beperkte invloed uitoefent op de statistische methode. Recente resultaten aangaande de robuustheid van SVMs, gebruik makend van de invloedsfunctie, worden gegeven in b.v. Christmann en Steinwart (2004, 2007) en Steinwart en Christmann (2008a). Een van de voorwaarden hier is echter dat de verliesfunctie twee maal Fréchet-afleidbaar moet zijn, een eigenschap waaraan bij de veelgebruikte verliesfuncties alleen door de logistische verliesfunctie voldaan wordt. Om die reden introduceren we een veralgemening van de IF die gebaseerd is op *Bouligand-afgeleiden* (Robinson, 1991). Bouligand-afgeleiden en sterke Bouligand-afgeleiden zijn reeds succesvol gebruikt in approximatietheorie voor niet-gladde functies, zie b.v. Clarke (1983), Robinson (1987, 1991), Ip en Kyparisis (1992) en de referenties daarin geciteerd, maar voor zover wij weten zijn deze richtingsafgeleiden nog niet gebruikt in robuuste statistiek. Om te beginnen berekenen we de Bouligand-afgeleide van enkele verliesfuncties waarna we, gebruik makend van deze afgeleiden, de *Bouligand invloedsfunctie* (BIF) invoeren als een aanpassing van Hampel's invloedsfunctie. De Bouligand invloedsfunctie van de functie $T : \mathcal{M}_1 \to \mathcal{H}$ voor een verdeling P in de richting van een verdeling $\mathrm{Q} \neq \mathrm{P}$ is de speciale Bouligand-afgeleide (als deze bestaat)

$$\lim_{\varepsilon \downarrow 0} \frac{\big\| T\big((1-\varepsilon)\mathrm{P} + \varepsilon \mathrm{Q}\big) - T(\mathrm{P}) - \mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) \big\|_{\mathcal{H}}}{\varepsilon} = 0\,.$$

De BIF zal ons toelaten om de robuustheid van een brede klasse van support vector machines met niet-gladde verliesfuncties te bestuderen. Voorbeelden hiervan zijn de support vector machine gebaseerd op de ϵ-insensitive verliesfunctie, of kern-gebaseerde kwantielregressie die gebruik maakt van de pinball verliesfunctie. Een ander voordeel van de BIF, naast de mogelijkheid om met niet-gladde functies te kunnen werken, is dat ze positief homogeen is, een eigenschap die meestal niet geldt voor Hampel's invloedsfunctie. Een derde voordeel van de BIF is dat de contaminerende verdeling niet beperkt wordt tot een puntverdeling, zoals het geval is voor de IF, maar dat deze om het even welke verdeling $\mathrm{Q} \neq \mathrm{P}$ kan zijn. Bovendien hebben we aangetoond dat als de BIF bestaat ook de IF bestaat en dat beide gelijk zijn.

In Secties 3.3 and 3.4 bewijzen we dat veel support vector machines voor regressie een begrensde Bouligand invloedsfunctie bezitten en dus robuust zijn. Sectie 3.3 bevat ons eerste robuustheidsresultaat: onder enkele zwakke voorwaarden zal de BIF voor support vector machines gebaseerd op een begrensde kern en een twee maal Bouligand-afleidbare, Lipschitz continue en convexe verliesfunctie bestaan, is ze begrensd en wordt de BIF van $T(\mathrm{P}) := f_{L,\mathrm{P},\lambda}$ in de richting van $\mathrm{Q} \neq \mathrm{P}$ expliciet gegeven door:

$$\begin{aligned}\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) \quad &= \quad S^{-1}\big(\mathbb{E}_{\mathrm{P}} \nabla_3^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \Phi(X)\big) \\ &\quad\; -S^{-1}\big(\mathbb{E}_{\mathrm{Q}} \nabla_3^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \Phi(X)\big)\,,\end{aligned}$$

waarbij $S : \mathcal{H} \to \mathcal{H}$ gedefinieerd is als

$$S(\,\cdot\,) := 2\lambda\, \mathrm{id}_{\mathcal{H}}(\,\cdot\,) + \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \langle \Phi(X), \,\cdot\, \rangle_{\mathcal{H}} \Phi(X)\,.$$

We tonen vervolgens dat dit resultaat de belangrijke speciale gevallen van SVMs gebaseerd op de verliesfuncties L_ϵ, $L_{\tau-pin}$, and $L_{c-Huber}$ – dewelke uitgesloten waren bij eerdere resultaten omtrent de invloedsfunctie – alsook L_{r-log} omvat. De IF van support vector machines gebaseerd op de logistische verliesfunctie werd recent bekomen door Christmann en Steinwart (2007) en Steinwart en Christmann (2008b). We bewijzen tevens dat de asymptotische bias $f_{L,(1-\alpha\varepsilon)\mathrm{P}+\alpha\,\varepsilon\mathrm{Q},\,\lambda} - f_{L,\mathrm{P},\lambda}$ van de vorm $\alpha\,\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) + o(\alpha h)$ zal zijn waarbij $h = \varepsilon(\mathrm{Q} - \mathrm{P})$. Ook al kunnen we in onze resultaten de voorwaarde dat de verliesfunctie Fréchet-afleidbaar moet zijn vermijden, we hebben nog steeds de aanname dat $\mathbb{E}_{\mathrm{P}}|Y| < \infty$. Dit betekent dat verdelingen met zware staarten niet gedekt worden door deze stelling, wat leidt naar het tweede deel van onze robuustheidsresulaten.

In Sectie 3.4 zullen we de $L^\star$-truc gebruiken om enkele bestaande robuustheidsresultaten uit te breiden naar verdelingen met zware staarten. De resultaten over de invloedsfunctie bekomen in Christmann en Steinwart (2007) en Steinwart en Christmann (2008b) worden aangepast. We tonen dat, door gebruik te maken van de verschoven verliesfunctie, de IF van een support vector machine $f_{L^\star,\mathrm{P},\lambda}$ gebaseerd op een begrensde, continue kern en een twee maal Fréchet-afleidbare, Lipschitz continue en convexe verliesfunctie bestaat en is begrensd, zelfs indien de data komen van een verdeling met zware staarten. We verkrijgen tevens een formule voor de invloedsfunctie naar analogie met Christmann en Steinwart (2007) en Steinwart en Christmann (2008b). Daarnaast bekomen we ook een begrenzing voor de bias $\left\| f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{L^\star,\mathrm{P},\lambda} \right\|_{\mathcal{H}}$, waaruit blijkt dat dit verschil ten hoogste lineair zal stijgen. Daarna herberekenen we de Bouligand invloedsfunctie voor de SVM, maar deze keer gebruik makend van de $L^\star$-truc. We tonen dat ook hier de BIF van $f_{L^\star,\mathrm{P},\lambda}$ bestaat, dat deze begrensd is en dat ze gegeven wordt door de eerder vermelde formule indien we werken met een begrensde, continue kern en een twee maal Bouligand-afleidbare, Lipschitz continue en convexe verliesfunctie. Bijgevolg mogen we stellen dat SVMs robuust zijn in de zin van invloedsfuncties, zelfs wanneer de onderliggende verdeling zware staarten heeft of een extreme waarden verdeling is. We sluiten dit hoofdstuk af met enkele numerieke voorbeelden waarin we de bruikbaarheid van SVMs aantonen voor dit soort verdelingen. Eerst passen we een SVM toe op een gesimuleerde data set met Cauchy errors (verdeling met zware staarten), daarna op een data set met grote brandschadeclaims (extreme waardenverdeling) komende van Copenhagen Re.

Appendix A bevat alle wiskundige voorkennis die nodig kan zijn om deze thesis volledig te begrijpen. Ik heb geprobeerd ervoor te zorgen dat dit werk gelezen kan worden zonder al te veel opzoekwerk te moeten verrichten, en bijgevolg lopen de onderwerpen hier van topologie over functionaalanalyse en kanstheorie tot convexe optimalisatie. Appendix B geeft een overzicht van alle notaties en afkortingen die in dit werk gebruikt zijn.

Samengevat kunnen we dus stellen dat in deze thesis robuustheid en consistentie van support vector machines werd onderzocht. We voeren een nieuw begrip in voor robuustheid, namelijk de Bouligand invloedsfunctie, dewelke een aanpassing is van de klassieke invloedsfunctie. Door gebruik te maken van deze methode zijn we in staat aan te tonen dat SVMs die een niet-gladde verliesfunctie gebruiken ook robuust zijn in de zin van invloedsfuncties. Bovendien, door gebruik te maken van de $L^{\star}$-truc, kunnen we tevens bewijzen dat SVMs gebaseerd op de verschoven verliesfunctie zowel $L^{\star}$-risico consistent als robuust zijn. Deze truc laat ons toe om ook SVMs waarvoor de onderliggende verdeling zware staarten heeft of een extreme waarden verdeling is, te beschouwen, iets wat in eerdere resultaten niet mogelijk was. In het geval van de pinball verliesfunctie tonen we ook de consistentie van de beslissingsfunctie aan.

Ander recent onderzoek over SVMs omvat Christmann en Hable (2011) waarin de auteurs de robuustheid – aan de hand van de BIF – en de consistentie van SVMs voor additieve modellen nagaan, Hable en Christmann (2011) waarin ze kijken naar kwalitatieve robuustheid van support vector machines, Steinwart en Christmann (2009b) over de ijlheid van SVMs gebaseerd op de ϵ-insensitive verliesfunctie en Xu *et al.* (2009) die de equivalentie tussen het regulariseren van de SVM en robuuste optimalisatie bestudeert.

Laat ons afsluiten met enkele nabeschouwingen die gemaakt kunnen worden na dit werk gelezen te hebben. Ten eerste, we besloten om enkel niet-negatieve verliesfuncties L te beschouwen (alhoewel de verschoven verliesfunctie $L^\star$ wel negatieve waarden kan aannemen) omdat bijna alle verliesfuncties die in de praktijk gebruikt worden niet-negatief zijn en er geen onderzoek over SVMs voorhanden lijkt te zijn voor verliesfuncties met negatieve waarden. Ten tweede, het zou mogelijk kunnen zijn om gelijkaardige resultaten als de onze te bekomen voor convexe, maar lokaal Lipschitze verliesfuncties, zoals de least squares verliesfunctie, maar Lipschitz continue verliesfuncties leveren betere robuustheidseigenschappen op, zie hiervoor Christmann en Steinwart (2004, 2007) en Steinwart en Christmann (2008b). En ten derde, vanuit een robuustheidsstandpunt zouden *begrensde* en *niet-convexe* verliesfuncties ook interessant kunnen zijn. Wij hebben deze echter niet beschouwd om twee redenen. De eerste reden is dat het bestaan, de uniekheid, de consistentie en de beschikbaarheid van efficiënte numerieke algoritmes wijd en zijd aanvaard wordt als nodige voorwaarden waaraan SVMs zouden moeten voldoen om ervoor te zorgen dat numerieke problemen voor grote en hoog-dimensionale data sets, laat ons zeggen voor $n > 10^5$ en $d > 100$ oplosbaar zijn, zie b.v. Vapnik (1998) of Schölkopf en Smola (2002). Deze eigenschappen worden bereikt als het risico convex is, wat het geval is voor convexe verliesfuncties. Ten tweede bestaan er momenteel – voor zover we weten – geen algemene resultaten betreffende support vector machines die ons garanderen dat het risico convex zal zijn als de verliesfunctie niet-convex en begrensd is. Het feit dat het risico convex is speelt echter een belangrijke rol in de bewijzen van het bestaan en de uniekheid van SVMs. Vanuit ons standpunt zouden er dus eerst zulke resultaten bestudeerd moeten worden vooraleer onderzoek te verrichten naar SVMs met begrensde en niet-convexe verliesfuncties.

Ten slotte willen we nogmaals benadrukken dat de Bouligand-afgeleide, ook al was deze hiervoor nog niet gebruikt in robuuste statistiek, een veelbelovend concept is. Veel robuuste schatters die in de literatuur worden voorgesteld worden impliciet gedefinieerd als de oplossing van een zeker minimalisatieprobleem waarbij de objectieffunctie of verliesfunctie wel continu of Lipschitz continu is, maar niet noodzakelijk twee maal Fréchet-afleidbaar. Voorbeelden hier zijn niet enkel support vector machines, maar ook M-schatters van het Huber-type en bepaalde maximum likelihood schatters onder niet-standaard voorwaarden. Bouligand-afleidbaarheid valt mooi tussen Fréchet-afleidbaarheid, wat een te sterk begrip is voor vele robuuste schatters, en Gâteaux-afleidbaarheid, die de basis vormt voor de robuuste benadering gebaseerd op invloedsfuncties. Bovendien voldoen Bouligand-afgeleiden aan een kettingregel en bestaat er een impliciete functie stelling, iets wat in het algemeen niet waar is voor Gâteaux-afgeleiden.

Enkele mogelijke onderwerpen voor verder onderzoek die uit deze thesis voortvloeien zijn de volgende: een eerste optie zou verder onderzoek naar de consistentie van SVMs zijn. De consistentie is bewezen voor het specifieke geval van een SVM gebaseerd op de pinball verliesfunctie, maar nog niet voor SVMs die gebruik maken van andere verliesfuncties zoals b.v. de ϵ-insensitive verliesfunctie. Een ander interessant onderwerp zou kunnen zijn om robuustheid van SVMs voor classificatie meer in detail te bestuderen. De resultaten die bekomen zijn aan de hand van de Bouligand invloedsfunctie zijn, gegeven onze veronderstellingen, geldig voor alle support vector machines, maar tot nog toe hebben we de aanname aangaande de sterkte Bouligand-afgeleide enkel kunnen verifiëren voor verliesfuncties voor regressie. Een laatste pad dat mogelijks bewandeld kan worden zou zijn om ofwel lokaal Lipschitze verliesfuncties in plaats van Lipschitz continue verliesfuncties te beschouwen, ofwel om naar begrensde en niet-convexe verliesfuncties te kijken in plaats van naar convexe verliesfuncties.

Deze thesis is gebaseerd op de volgende publicaties:

- ⋆ Christmann, A. en Van Messem, A. (2008). Bouligand Derivatives and Robustness of Support Vector Machines for Regression. *Journal of Machine Learning Research*, **9**, 915–936.
- ⋆ Christmann, A., Van Messem, A. en Steinwart, I. (2009). On Consistency and Robustness Properties of Support Vector Machines for Heavy-Tailed Distributions. *Statistics and Its Interface*, **2**, 311–327.
- ⋆ Van Messem, A. en Christmann, A. (2010). A review on consistency and robustness properties of support vector machines for heavy-tailed distributions. *Advances in Data Analysis and Classification*, **4**(2-3), 199–220.

Contents

Chapter 1

An Introduction to Support Vector Machines

Let us start with a short overview of the development of support vector machines (SVMs). In 1936, R. A. Fisher (1952) suggested the first algorithm for pattern recognition, and in 1956 Frank Rosenblatt (1958, 1962) invented a linear classifier which he called the perceptron and which can be seen as the simplest kind of a feedforward neural network. Vapnik and Lerner (1963) introduced the Generalized Portrait Algorithm (see Section 1.2.1), which formed the basis for support vector machines. The algorithm implemented by support vector machines is namely a non-linear generalization of the Generalized Portrait Algorithm by using the theory of reproducing kernels as given by Aronszajn (1950). Next, Aizerman *et al.* (1964) introduced the geometrical interpretation of kernels as inner products in a feature space, and in the same year Vapnik and Chervonenkis (1964) further developed the Generalized Portrait Algorithm. Cover (1965) discussed large margin hyperplanes in the input space as well as sparseness, while similar optimization techniques were used in pattern recognition by Mangasarian (1965). The introduction of slack variables to overcome the problem of noise and non-separability was done by Smith (1968) and large margin hyperplanes in the input space were discussed by Duda and Hart (1973). The field of 'statistical learning theory' began – in Russian – with Vapnik and Chervonenkis (1974) with a book on pattern recognition that was translated into German in 1979. SVMs can be said to really have started when Vapnik (1979) – still in Russian – continued to develop statistical learning theory in 'Estimation of Dependences Based on Empirical Data', a book that was translated in English in 1982. Several statistical mechanics papers, such as the one of Anlauf and Biehl (1989) suggested using large margin hyperplanes in the input space, while Poggio and Girosi (1990) and Wahba (1990) discussed the use of kernels. Improvement upon Smith's work on slack variables was done by Bennett and Mangasarian (1992), while in the same year Boser *et al.* (1992) first introduced SVMs close to their current form (see Section 1.2.2) in a paper at the COLT 1992 conference. In 1995 the soft margin classifier (see Section 1.2.3) was introduced by Cortes and Vapnik (1995) and in the same year the algorithm was extended to the case of regression by Vapnik (1995) in his book 'The Nature of Statistical Learning Theory'. Further early work on SVMs comprises of the first rigorous statistical bound on the generalization of hard margin SVMs (Bartlett, 1998, Shawe-Taylor *et al.*, 1998) and statistical bounds on the generalization of soft margin algorithms and for the regression case (Shawe-Taylor and Cristianini, 1999).

In the next section, we will give a short introduction to the field of statistical learning theory, while Section 1.2 will describe the (geometrical) history of SVMs. Section 1.3 will shift the focus from the geometrical interpretation of SVMs to the, nowadays more common, interpretation via empirical risk minimization. Next, we will take a closer look at the loss function and the reproducing kernel Hilbert space in Sections 1.4 and 1.5, we will give conditions for the existence and uniqueness of the SVM in Section 1.6, and we will conclude this chapter with a note on shifting the loss function in Section 1.7.

1.1 An Introduction to Statistical Learning Theory

1.1.1 Statistical Machine Learning

Support vector machines belong to the large class of techniques from modern *statistical machine learning theory.* They are non-parametric methods and can be used both for classification and regression purposes. For more details on statistical learning, we refer the reader to such books as Vapnik (1995, 1998), and Hastie *et al.* (2001).

The aim in non-parametric statistical machine learning is to find a functional relationship between an $\mathcal{X}$-valued *input* random variable X and a $\mathcal{Y}$-valued *output*, or *response*, random variable Y, under the assumption that the joint distribution P of (X, Y) is (almost) completely unknown. Knowing whether this dependence between the input and the output variables exists, and if so what function will describe it, can be of value in real-life applications, such as:

i) predicting whether a client will pay back a loan to a bank based on data of previous clients of the bank;

ii) estimating the total yearly amount of claims of car insurances based on insurance data over the previous years.

Since nowadays the structure of the measurements can be too complex to be reasonably understood by humans (think, e.g., of DNA strings) or the amount of data can be too large to manually find the relationship, people will depend on computers to do the work for them and hence the name *machine learning*. One expects a machine to use observed data to "learn" the unknown dependency via some algorithm as well as automatically assign a response to future input values. So it is not sufficient to find a good description of the relationship of the observed data, it is also needed to find a prediction rule that works well for new, unseen inputs. For both of the above mentioned examples the input space $\mathcal{X}$ could be the set of all personal information gathered on the client. The output set $\mathcal{Y}$ for the first example would be $\{$"no", "yes"$\}$ or $\{-1, +1\}$, while in the other example it would be the set of all possible claim amounts, most likely $\mathbb{R}$.

In order to model the relationship between the input and response variables, one therefore typically assumes that one has a finite training data set $D = ((x_1, y_1), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ consisting of observations from independent and identically distributed (i.i.d.) random variables (X_i, Y_i), $i = 1, \dots, n$, which all have the same, but unknown, distribution P on $\mathcal{X} \times \mathcal{Y}$ equipped with the corresponding Borel σ-algebra. This Borel σ-algebra will exist, since both $\mathcal{X}$ and $\mathcal{Y}$ will be topological spaces, as will become clear later on when we state our assumptions. The method is called *non-parametric* since no, or very little, additional assumptions are made on the distribution P. This means that we do not assume the existence of densities, symmetry, or a parametric model. The goal is then to build a predictor $f : \mathcal{X} \to \mathcal{Y}$, based solely on the observations, which will assign to each input vector, sometimes also called risk vector, x a prediction $f(x)$ which hopefully will be a good approximation of the observed output y.

As can be seen from the two examples mentioned above, the type of output values can be categorized in two classes, either they are quantitative (such as the claims amount) or they are categorical (yes/no). For quantitative measurements, there exists a ranking of the values, and measurements that are close in value will be similar in nature. Categorical, or qualitative or discrete, responses can only take on values in a finite set of possibilities and usually have no ranking among themselves. Usually, these different classes are described by labels rather than by numbers. Most often there are only two possibilities and a $0/1$ or $-1/+1$ coding is sufficient. If there are more than two categories, coding can be done, e.g., by using dummy variables.

The difference in output type defines the different prediction tasks: *classification* when we predict categorical values, and *regression* when the responses are quantitative. Common choices of input and output spaces are $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \{-1, +1\}$ in the binary classification case and $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ in the regression case. Most of the results in this thesis are meant for regression, but classification can be seen as a special case of regression where the responses can only take on a limited number of values.

Two basic examples of non-parametric pattern recognition are the least squares method and the k-nearest neighbors method, which we will briefly describe below. The least squares method is most often used for regression purposes, but is also applicable for classification of data. The k-nearest neighbors method is mainly meant for classification. Of course there are a lot of other prediction methods in existence, some examples of which are discriminant analysis, logistic regression, and neural networks.

1.1.2 Ordinary Least Squares

In this subsection, we will discuss the least squares method for a linear model, also called *ordinary least squares* (OLS). This method is probably the most widely known and used technique for pattern recognition in existence.

Given some input vector $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$, and outcome $Y \in \mathbb{R}$, we will assume that the following linear model

$$Y = \beta_0 + \sum_{j=1}^{d} \beta_j X_j$$

holds. The term β_0 is called the *intercept* or *bias* term. In order to facilitate calculation, a constant 1 is often incorporated in the vector X and $\beta = (\beta_0, \dots, \beta_d)$ is written for the vector of the coefficients. This allows us to rewrite the model as

$$Y = X^T \beta \,,$$

where X^T is the transposition of the column vector X. The function $f(X) = X^T \beta$ is the linear predictor function as described earlier. However, β is unknown, and will be estimated using observed data. In that case, we will write $\hat{\beta}$ and $\hat{y}$ for the predictions of β and y respectively.

Given a set of training data $D = ((x_1, y_1), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, the assumption tells us that

$$y_i = x_i^T \beta + \varepsilon_i \,, \qquad i = 1, \dots, n \,,$$

with ε_i the *error* term. The linear model can be fitted to the data by many different methods, but the *least squares* method will look at minimizing the residual sum of squares

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \,. \tag{1.1}$$

$\mathrm{RSS}(\beta)$ is clearly a quadratic function, and hence will have a minimum, though a priori this minimum is not necessarily unique. If we write $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$, $x = (x_1, \dots, x_n)^T \in \mathbb{R}^{n \times (d+1)}$, and $\beta = (\beta_0, \dots, \beta_d)^T \in \mathbb{R}^{d+1}$, we can rewrite (1.1) in matrix notation as

$$\mathrm{RSS}(\beta) = (y - x\beta)^T (y - x\beta) \,. \tag{1.2}$$

Differentiating (1.2) with respect to β and imposing stationarity (meaning we will set the value of this derivative to zero), gives the *normal equations*

$$x^T (y - x\beta) = 0 \,.$$

For $x^T x$ non-singular, the (unique) solution is given by

$$\hat{\beta} = (x^T x)^{-1} x^T y \,,$$

and the fitted value for x_i is $\hat{y}_i = \hat{y}(x_i) = x_i^T \hat{\beta} = x_i^T (x^T x)^{-1} x^T y$. Similarly, the prediction for some unknown input x_0 is given by $\hat{y}(x_0) = x_0^T \hat{\beta}$.

The assumption of a linear model is a quite important model assumption, but helps the method yield rather stable, but possibly quite inaccurate, results. Extension to a non-linear model is naturally possible, but we will not discuss this here. Furthermore, we only looked at the case where the response variable Y is a scalar, but of course Y could also be a vector of outputs. In that case β will be a matrix instead of a vector, but the method remains unchanged.

1.1.3 k-Nearest Neighbors

The second method for pattern recognition that we will shortly describe, is the k-nearest neighbors (k-NN) method. This method is one of the simplest of all learning algorithms: an object will be classified by a majority vote among its neighbors.

The idea of k-NN methods for classification is to construct a decision function locally for each x by first determining the $k > 0$ points of the training data set D that are closest to x and then making the prediction for y based on the (possibly weighted) average of the k corresponding y-values. This concept of closeness implies the use of a metric, which can influence the outcome. Depending on the used metric, the k closest vectors can vary, and thus so can the prediction. Also the choice of k itself is of importance, and will depend on the data provided. The smaller k is chosen, the less errors the method will make, but the more irregular the decision boundary will become. Larger values of k will reduce the effect of noise, but decrease the distinction between the classes.

The k-NN method makes no real structural assumption such as the least squares fit does, which improves the accuracy of the method, but reduces its stability, since it depends heavily on the local structure of the data.

1.2 History of Support Vector Machines

The original idea for support vector machines is a geometrical one: the aim of support vector classification is to find a computationally efficient way of learning 'good'[1] separating hyperplanes in a high-dimensional feature space. Support vector machines produce sparse dual representations

[1] 'Good' can be defined in different ways. We will use the notion of the maximal margin separating hyperplane, but another possibility would be to find the separating hyperplane that minimizes the number of support vectors.

of the hypothesis, which results in extremely efficient algorithms due to the Karush-Kuhn-Tucker conditions which hold for the solution and play a very important role in the practical implementation and analysis of SVMs. A second important feature is that, due to Mercer's conditions on the kernels, the corresponding optimization problem is convex, and hence there are no local extremes. This fact, in combination with the strongly reduced number of non-zero parameters, is exactly why support vector machines distinguish themselves from other learning algorithms such as neural networks.

The following description of the geometrical evolution of support vector machines is based upon Vapnik (1998, 2000, Chapter 10 resp. Chapter 5.5), Burges (1998, Chapters 3 and 4), Cristianini and Shawe-Taylor (2000, Chapter 6), and Steinwart and Christmann (2008b, Chapter 1.3).

1.2.1 The Generalized Portrait Algorithm

As remarked in the introduction to this chapter, the foundation for SVMs was laid by the *Generalized Portrait Algorithm* (GPA) as introduced by Vapnik and Lerner (1963). The GPA, which is a binary classification algorithm, consideres the simplest case possible: a linear machine trained on separable[2] data. This means that the set of data points will be completely separable by a linear hyperplane. To distinguish between the two classes, the data points will be labeled +1 and −1.

The set-up is the following: since we consider the case of binary classification, $\mathcal{Y} = \{-1, +1\}$, and let $\mathcal{X} \subset \mathbb{R}^d$. We also posses a training data set $D = ((x_1, y_1), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ that is perfectly separable. This means that there exists a hyperplane

$$H_0 \, : \, \langle w, x \rangle + b = 0 \tag{1.3}$$

characterized by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that the training data satisfy

$$\langle w, x_i \rangle + b \quad \geq \quad +1 \qquad \forall\, i \text{ with } y_i = +1\,, \tag{1.4}$$

$$\langle w, x_i \rangle + b \quad \leq \quad -1 \qquad \forall\, i \text{ with } y_i = -1\,, \tag{1.5}$$

which can be rewritten as

$$y_i(\langle w, x_i \rangle + b) \geq 1 \qquad \forall\, i = 1, \dots, n\,. \tag{1.6}$$

[2] In this case 'separable' means that both classes of the data can be separated and has nothing to do with the mathematical definition of a separable space as on p. 110.

Since the hyperplane (1.3) separates the positive from the negative samples without error, it is called the *separating hyperplane*. The parameter w is the normal to the hyperplane and $b/\|w\|_2$ is the perpendicular distance from the hyperplane to the origin. The *geometrical margin* γ_g of the separating hyperplane will be defined as the distance from the closest vector to the hyperplane (Vapnik, 2000, p. 131). Let us call d_+ the shortest distance from the separating hyperplane to a positive point, and d_- the distance to the closest negative point.

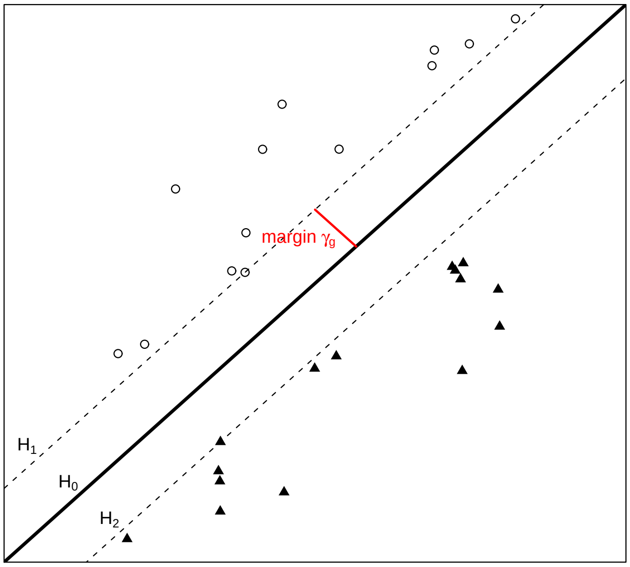

Figure 1.1: A linear hyperplane H_0 in $\mathbb{R}^2$, separating the two classes. The decision boundaries are H_1 and H_2, the geometrical margin is shown in red and is equal on both sides.

The GPA will then search for the perfectly separating hyperplane (w_D, b_D) that has *maximal geometrical margin*. This hyperplane will be called the *optimal hyperplane* or *maximal margin hyperplane*. Once this hyperplane is found, the resulting *decision function* is defined by

$$f_D(x) := \operatorname{sign}(\langle w_D, x\rangle + b_D) \qquad \forall x \in \mathbb{R}^d \,,$$

which means that f_D will assign positive labels to one affine half-space and negative labels to the other.

This decision function will be found by solving a quadratic problem. Let us therefore take a look at the margin γ_g, which needs to be maximized. The points for which the equality in (1.4) holds, will lie on the hyperplane

$$H_1 \,:\, \langle w, x_i \rangle + b = +1\,.$$

The points satisfying equation (1.5), lie on

$$H_2 \,:\, \langle w, x_i \rangle + b = -1\,.$$

The perpendicular distance of H_1 to the origin is $|1 - b|/\|w\|_2$, that of H_2 is $|1 + b|/\|w\|_2$. Therefore, using that H_0, H_1, and H_2 are parallel since their normals are the same, we see that $d_+ = d_- = 1/\|w\|_2$ and thus the geometrical margin $\gamma_g = 1/\|w\|_2$. This implies that maximizing this margin will be equivalent to minimizing the norm $\|w\|_2^2$ subject to (1.6).

Mathematically, this gives the following optimization problem:

$$\begin{array}{ll} \text{minimize } \dfrac{1}{2}\langle w, w \rangle & \text{over } w \in \mathbb{R}^d, b \in \mathbb{R} \\ \text{subject to } y_i(\langle w, x_i \rangle + b) \geq 1\,, & i = 1, \ldots, n\,, \end{array} \tag{1.7}$$

which forces the hyperplane to make no classification errors on D.

The optimization problem (1.7) will be solved using the Lagrangian approach, as described in Subsection A.4.3. There are two reasons for doing so. First of all, the constraints in (1.7) will be replaced by constraints on the Lagrange multipliers, and it is much easier to work with these. Secondly, the training data will only appear in the form of inner products between vectors, which will allow us to later on generalize this algorithm to the non-linear case via kernels.

By applying Lagrangian theory to our problem, we obtain the following *primal Lagrangian*:

$$\mathfrak{L}_P(w, b, \alpha) = \frac{1}{2}\,\langle w, w \rangle - \sum_{i=1}^{n} \alpha_i \big(y_i(\langle w, x_i \rangle + b) - 1\big) \tag{1.8}$$

with $\alpha = (\alpha_1, \ldots, \alpha_n)$ the vector of the positive Lagrange multipliers.

Since the SVM problem is a *convex* problem with linear constraints, the Karush-Kuhn-Tucker (KKT) conditions are *necessary* and *sufficient* for some w_D, b_D and α^* to be an optimal solution to the problem, see Theorem A.4.9. Thus solving the (primal) problem is equivalent to finding a solution to the KKT conditions:

$$\begin{aligned}
\frac{\partial \mathfrak{L}_P(w,b,\alpha)}{\partial w} = w - \sum_{i=1}^{n} y_i \alpha_i x_i &= 0\,, && \qquad (1.9)\\
\frac{\partial \mathfrak{L}_P(w,b,\alpha)}{\partial b} = -\sum_{i=1}^{n} y_i \alpha_i &= 0\,, && \qquad (1.10)\\
y_i(\langle w, x_i\rangle + b) - 1 &\geq 0\,, && \forall\, i = 1,\ldots,n\,,\\
\alpha_i &\geq 0\,, && \forall\, i = 1,\ldots,n\,,\\
\alpha_i\big(y_i(\langle w, x_i\rangle + b) - 1\big) &= 0\,, && \forall\, i = 1,\ldots,n\,. \quad (1.11)
\end{aligned}$$

One of the possible approaches to implement these conditions, is the approach using the dual Lagrangian, with addition of the *complementarity condition*(1.11).

Substituting (1.9) and (1.10) in the primal (1.8) produces the *dual* formulation

$$\begin{aligned}
\mathfrak{L}_D(w,b,\alpha) &= \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle - \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle + \sum_{i=1}^{n} \alpha_i\\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle\,,
\end{aligned}$$

and the corresponding optimization problem is then given by

$$\begin{aligned}
&\text{maximize } \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j\rangle \quad \text{over } \alpha_i \in \mathbb{R}\\
&\text{subject to } \sum_{i=1}^{n} y_i \alpha_i = 0\,, \qquad (1.12)\\
&\qquad\qquad \alpha_i \geq 0\,, \qquad\qquad i = 1,\ldots,n\,.
\end{aligned}$$

Therefore, training the SVM amounts to maximizing $\mathfrak{L}_D$ with respect to the Lagrange parameters α_i. Each training point x_i corresponds to one of the α_i. The x_i for which $\alpha_i \neq 0$ are called the *support vectors* since only these points will add a contribution in the expression (1.9) of the vector w, hence they "support" the hyperplane. Let us define the set of the indices of the support vectors as sv. The complementarity condition makes it clear that the support vectors lie either on H_1 or H_2. Using (1.11), it is easy to see that only for inputs x_i that have $y_i(\langle w, x_i\rangle + b) = 1$, the corresponding α_i can be non-zero. All other α_i will be equal to zero and thus the corresponding data points will not be support vectors. The support vectors are the critical elements of the data set D, since only these vectors determine the decision boundaries. Removing all other points from D and repeating the training process, would give the same separating hyperplane.

Given that the Lagrange parameters $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)$ solve the optimization problem (1.12), the vector

$$w_D = \sum_{i=1}^{n} y_i \alpha_i^* x_i = \sum_{i \in \mathrm{sv}} y_i \alpha_i^* x_i$$

will be the normal to the maximal margin hyperplane. However, since b_D does not appear in the dual problem, its value will need to be computed using the primal constraints:

$$b_D = -\frac{\max_{y_i=-1}\langle w_D, x_i\rangle + \min_{y_i=1}\langle w_D, x_i\rangle}{2}\,.$$

Using this, the maximal margin hyperplane will be given by

$$0 = \langle w_D, x\rangle + b_D = \sum_{i \in \mathrm{sv}} y_i \alpha_i^* \langle x_i, x\rangle + b_D\,. \tag{1.13}$$

The complementarity condition can facilitate the computation of b_D: take an index i for which $\alpha_i^* \neq 0$, and compute b_D using (1.11). Of course it is numerically safer to do this for all i with $\alpha_i^* \neq 0$ and then take the mean of all found b_D.

Also using the complementarity condition, we see that for $i \in \mathrm{sv}$ holds that

$$1 = y_i\big(\langle w_D, x_i\rangle + b_D\big) = y_i\big(\sum_{j \in \mathrm{sv}} y_j \alpha_j^* \langle x_j, x_i\rangle + b_D\big)\,,$$

which, together with (1.10), can be used to calculate the optimal geometrical margin.

$$\begin{aligned} \|w_D\|_2^2 &= \langle w_D, w_D\rangle = \sum_{i,j=1}^{n} \langle y_i \alpha_i^* x_i, y_j \alpha_j^* x_j\rangle \\ &= \sum_{i \in \mathrm{sv}} y_i \alpha_i^* \sum_{j \in \mathrm{sv}} y_j \alpha_j^* \langle x_i, x_j\rangle \\ &= \sum_{i \in \mathrm{sv}} \alpha_i^* (1 - y_i b_D) = \sum_{i \in \mathrm{sv}} \alpha_i^*\,. \end{aligned}$$

Therefore, $\gamma_g = 1/\|w_D\|_2 = \big(\sum_{i \in \mathrm{sv}} \alpha_i^*\big)^{-1/2}$.

There are however two major issues with the setup of the GPA:

i) The first problem is that we use a *linear* hyperplane to separate the data. However, a linear decision function may not always be suitable for the classification we wish to perform. This is, for example, the case if the set D is not linearly separable.

ii) The second issue has to do with *noise*. Due to the effect of noise on our data, we might actually allow that some points will be misclassified in order to avoid overfitting. Especially for the case where the dimension d is greater than the sample size n overfitting can form a serious problem.

1.2.2 The Hard Margin SVM

In order to extend the problem to the case of a non-linear decision function, Boser *et al.* (1992) showed that a rather old trick (Aizerman *et al.*, 1964) can be used to accomplish this feat in a pretty easy way.

Note that the data x_i only appear in the form of inner products in the optimization problem. The idea is to map the input data $(x_1, \ldots, x_n)$ into some (possibly infinite-dimensional) Hilbert space $\mathcal{H}_0$, called the *feature space*, by a typically non-linear map $\Phi : \mathcal{X} \to \mathcal{H}_0$, which is called the *feature map*, such that the mapped data $\Phi(x_i)$ can be separated in the feature space $\mathcal{H}_0$. Next, the GPA will be applied to the mapped data set $((\Phi(x_1), y_1), \ldots, (\Phi(x_n), y_n))$. The training algorithm will then depend on the data through inner products in the feature space $\mathcal{H}_0$, which are of the form $\langle \Phi(x_i), \Phi(x_j) \rangle$, which is exactly the definition of the kernel k, see Definition 1.5.1, and therefore we can work with only the kernel k in the training algorithm, without explicitly knowing the function Φ. For more details on the kernel, the feature space and the feature map, we refer to Section 1.5. Although w will now live in the feature space $\mathcal{H}_0$ and no longer in $\mathbb{R}^d$, and calculating w requires knowledge of the feature map due to its form

$$w = \sum_{i \in \mathrm{sv}} y_i \alpha_i^* \Phi(x_i)\,,$$

the separating hyperplane (and thus also the decision function) can be calculated using only the kernel by adapting the expression (1.13):

$$\begin{aligned} 0 &= \sum_{i \in \mathrm{sv}} y_i \alpha_i^* \langle \Phi(x_i), \Phi(x) \rangle + b_D \\ &= \sum_{i \in \mathrm{sv}} y_i \alpha_i^* k(x_i, x) + b_D\,. \end{aligned}$$

This method was originally called the *maximal margin classifier* and later also the *hard margin SVM*. Given that there are no contradictory data in D, i.e., there are no (x_i, y_i) and (x_j, y_j) with $x_i = x_j$ and $y_i \neq y_j$, and by choosing a suitable feature map Φ, see Steinwart and Christmann (2008b, Sect. 4.6), then for *every* training data set this method will be able to perfectly separate the training data by a hyperplane in the feature space.

There is however a price to pay for this high flexibility: the hyperplane now lies in a high- or even infinite-dimensional space, and hence the problem of overfitting becomes even more prominent.

1.2.3 The Soft Margin SVM

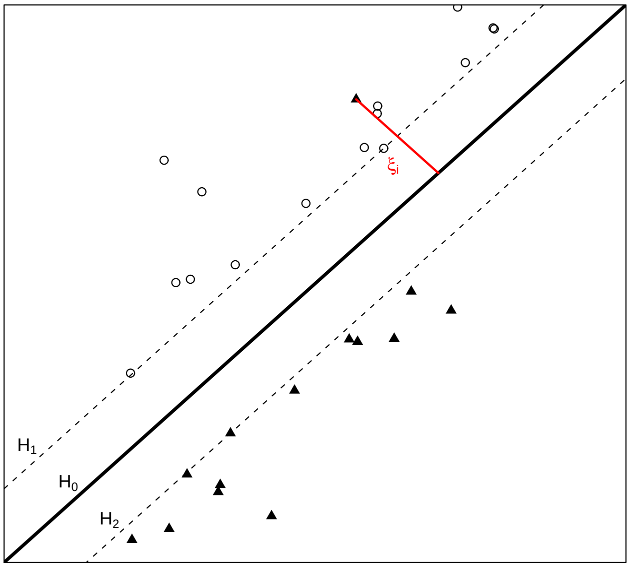

Figure 1.2: Interpretation of the slack variable in the case where $\mathcal{X} = \mathcal{H}_0 = \mathbb{R}^2$ and a linear kernel $k(x, x') = \langle x, x' \rangle$ is used.

The problem of overfitting was solved by the introduction of the *soft margin* SVM by Cortes and Vapnik (1995). The idea of the soft margin SVM is to relax the constraints in (1.7) by introducing positive *slack variables* ξ_i, $i = 1, \ldots, n$. These slack variables will allow for the margin constraints to be violated, but only when necessary, i.e., they will add an extra cost to the primal objective function. Combining this with the idea of the feature map then gives:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\langle w,w\rangle + C\sum_{i=1}^{n}\xi_i && \text{over } w\in\mathcal{H}_0, b\in\mathbb{R}, \xi\in\mathbb{R}^n \\ &\text{subject to } y_i(\langle w,\Phi(x_i)\rangle + b) \geq 1-\xi_i\,, && i=1,\ldots,n \\ &\qquad\qquad \xi_i \geq 0\,, && i=1,\ldots,n\,, \end{aligned} \tag{1.14}$$

where $C > 0$ is a free, but fixed constant that is used to balance the weight accorded to both parts of the objective function and $\xi = (\xi_1, \ldots, \xi_n)$ is the vector of slack variables. The larger the value of C, the more the training errors are penalized. In practice, the value of C will be found by using a grid search or cross-validation technique. For a classification error to be made, the value of ξ_i has to surpass one – since this means crossing the separating hyperplane – and thus $\sum_{i=1}^{n}\xi_i$ can be seen as an upper bound on the number of training errors. Minimizing this upper bound will assure us to make a minimal number of training errors.

This optimization problem is also called the *1-norm* soft margin, since the slack variables are introduced in the objective function using the norm $\|\xi\|_1$. It is of course also possible to use some other k-norm, with $k > 0$. For example, the *2-norm* soft margin algorithm will have an objective function $\frac{1}{2}\langle w,w\rangle + C\sum_{i=1}^{n}\xi_i^2$ and is also easy to use.

Clearly this problem is still a convex programming problem (and for both the 1-norm and 2-norm soft margin SVM it is even quadratic) with linear constraints, and thus the primal-dual approach can be applied to it. As will become clear, the choice of the 1-norm soft margin algorithm has the advantage that neither the ξ_i nor their associated Lagrange multipliers will appear in the dual problem.

Putting the optimization problem (1.14) in the primal Lagrangian form gives us

$$\begin{aligned} \mathfrak{L}_P(w,b,\xi,\alpha,\beta) = &\frac{1}{2}\langle w,w\rangle + C\sum_{i=1}^{n}\xi_i \\ &- \sum_{i=1}^{n}\alpha_i\big(y_i(\langle w,\Phi(x_i)\rangle + b) - 1 + \xi_i\big) - \sum_{i=1}^{n}\beta_i\xi_i \end{aligned} \tag{1.15}$$

with $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\beta = (\beta_1, \ldots, \beta_n)$ the vectors of the positive Lagrange multipliers. The dual formulation is found by differentiating with respect to w, b and ξ and imposing stationarity,

$$\begin{aligned}\frac{\partial \mathfrak{L}_P(w,b,\xi,\alpha,\beta)}{\partial w} &= w-\sum_{i=1}^n y_i\alpha_i\Phi(x_i)=0\,, && (1.16)\\ \frac{\partial \mathfrak{L}_P(w,b,\xi,\alpha,\beta)}{\partial b} &= -\sum_{i=1}^n y_i\alpha_i=0\,, && (1.17)\\ \frac{\partial \mathfrak{L}_P(w,b,\xi,\alpha,\beta)}{\partial \xi_i} &= C-\alpha_i-\beta_i=0\,, \qquad i=1,\ldots,n\,, && (1.18)\end{aligned}$$

and then plugging these expressions into the primal Lagrangian. The KKT conditions are the equations (1.16)-(1.18) completed with

$$\begin{aligned} y_i(\langle w,\Phi(x_i)\rangle+b)-1+\xi_i &\geq 0\\ \xi_i &\geq 0\\ \alpha_i &\geq 0\\ \beta_i &\geq 0\\ \alpha_i\big(y_i(\langle w,\Phi(x_i)\rangle+b)-1+\xi_i\big) &= 0 && (1.19)\\ \beta_i\xi_i &= 0 && (1.20)\end{aligned}$$

Inserting (1.16)-(1.18) into the primal (1.15) gives us as dual

$$\begin{aligned}\mathfrak{L}_D(w,b,\xi,\alpha,\beta) &= \frac{1}{2}\sum_{i,j=1}^n y_iy_j\alpha_i\alpha_j\langle\Phi(x_i),\Phi(x_j)\rangle + C\sum_{i=1}^n\xi_i\\ &\quad -\sum_{i,j=1}^n y_iy_j\alpha_i\alpha_j\langle\Phi(x_i),\Phi(x_j)\rangle+\sum_{i=1}^n\alpha_i\\ &\quad -\sum_{i=1}^n\alpha_i\xi_i-\sum_{i=1}^n\beta_i\xi_i\\ &= \sum_{i=1}^n\alpha_i-\frac{1}{2}\sum_{i,j=1}^n y_iy_j\alpha_i\alpha_j\langle\Phi(x_i),\Phi(x_j)\rangle\\ &\quad +\sum_{i=1}^n(C-\alpha_i-\beta_i)\xi_i\\ &= \sum_{i=1}^n\alpha_i-\frac{1}{2}\sum_{i,j=1}^n y_iy_j\alpha_i\alpha_j\langle\Phi(x_i),\Phi(x_j)\rangle\,,\end{aligned}$$

which is identical to the dual of the maximal margin algorithm (or the GPA with $\Phi(x_i)$ instead of x_i). The only difference occurs in the constraints.

$C - \alpha_i - \beta_i = 0$, together with $\beta_i \geq 0$ enforces that $0 \leq \alpha_i \leq C$, thus it provides an upper bound for the α_i. Keeping the definition of the kernel k in mind, the optimization problem becomes

$$\begin{aligned} &\text{maximize } \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \ \text{ over } \alpha_i \in \mathbb{R} \\ &\text{subject to } \sum_{i=1}^{n} y_i \alpha_i = 0\,, \\ &\qquad\qquad 0 \leq \alpha_i \leq C\,, \qquad\qquad i = 1, \ldots, n\,. \end{aligned} \tag{1.21}$$

Given that $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)$ is a solution of the dual program (1.21), the normal to the separating hyperplane will be

$$w_D = \sum_{i \in \mathrm{sv}} y_i \alpha_i^* \Phi(x_i)\,.$$

The complementarity condition (1.20) combined with (1.18) shows that if $\xi_i \neq 0$ then $\beta_i = C - \alpha_i^* = 0$ and thus $\alpha_i^* = C$. Also, if $\alpha_i^* < C$, then $\xi_i = 0$. This allows us to calculate the value of b_D by taking any training point for which $0 < \alpha_i^* < C$ and $\xi_i = 0$, and compute b_D via (1.19). As before, it will be safer to take the average over all such training points.

The condition (1.19) shows that the training points for which $\alpha_i \neq 0$ will lie between the boundaries determined by the hyperplanes H_1 and H_2. Their distance from the separating hyperplane is thus less than $1/\|w\|$ and these points are called $1/\|w\|$-margin errors. If $0 < \alpha_i < C$, then they lie at exactly the target distance $1/\|w\|$ from the separating hyperplane and thus on either H_1 or H_2.

If we define $f(x) := \sum_{i=1}^{n} y_i \alpha_i^* k(x, x_i) + b_D$, then the decision function will be $f_D(x) = \operatorname{sign}(f(x))$. We also see that the choice of b_D implies, through (1.19), that $yf(x) = 1$ for the training points with $0 < \alpha_i^* < C$, which shows again that these points lie on H_1 or H_2.

Since $\|w_D\|^2 = \sum_{i,j \in \mathrm{sv}} y_i y_j \alpha_i^* \alpha_j^* k(x_i, x_j)$, the geometrical margin will be $\gamma_g = \big(\sum_{i,j \in \mathrm{sv}} y_i y_j \alpha_i^* \alpha_j^* k(x_i, x_j)\big)^{-1/2}$.

1.3 Empirical Risk Minimization and Support Vector Machines

In this section, we will motivate why support vector machines are defined the way they are, and describe the main goals of SVMs. The precise definition of a support vector machine is given in Definition 1.3.1. We refer to

Vapnik (1998), Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002), and Steinwart and Christmann (2008b) for textbooks on SVMs and related topics.

As already stated in Section 1.1, the predictor $f : \mathcal{X} \to \mathcal{Y}$ will assign to each risk vector x a prediction $f(x)$ which, we hope, will be a good approximation of the observed output y. To formalize the aim of estimating this predictor function f, we call a function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ a *loss function* (or just *loss*) if L is measurable. Some examples and properties of loss functions will be given in Section 1.4. The loss function assesses the quality of a prediction $f(x)$ for an observed output value y by $L(x, y, f(x))$, i.e., it measures the "closeness" between y and $f(x)$. We follow the convention that the smaller $L(x, y, f(x))$ is the better the prediction is. We will also assume that $L(x, y, y) = 0$ for all $y \in \mathcal{Y}$, because the loss is zero, if the forecast $f(x)$ equals the observed value y.

Of course, in order to assess the quality of a predictor f it is not sufficient to only 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. There are various ways to do this, but a common choice in statistical learning theory is to consider the *expected loss* of f, also called the *L-risk*, defined by

$$\mathcal{R}_{L,\mathrm{P}}(f) := \mathbb{E}_{\mathrm{P}} L(X, Y, f(X)) \,.$$

The *learning goal* is then to find a decision function f_D that (approximately) achieves the smallest possible risk, i.e., the Bayes risk

$$\mathcal{R}^*_{L,\mathrm{P}} := \inf\{\mathcal{R}_{L,\mathrm{P}}(f) \,;\, f : \mathcal{X} \to \mathbb{R} \text{ measurable}\} \,. \tag{1.22}$$

However, since the distribution P is unknown, the risk $\mathcal{R}_{L,\mathrm{P}}(f)$ is also unknown and consequently we cannot compute f_D. To resolve this problem, we can replace P by the empirical distribution

$$\mathrm{D} = \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$$

corresponding to the data set D. Here $\delta_{(x_i, y_i)}$ denotes the Dirac distribution in (x_i, y_i). This way we obtain the *empirical L-risk*

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

Although $\mathcal{R}_{L,\mathrm{D}}(f)$ can be considered as an approximation of $\mathcal{R}_{L,\mathrm{P}}(f)$ for each *single* f, solving $\inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_{L,\mathrm{D}}(f)$ will in general not result in a good approximate minimizer of $\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$. This is partially due to the effect of *overfitting*. Overfitting means that the learning method will model the data D too closely and thus will have a poor performance on future data points. The algorithm will construct a very wiggly function that will give good approximations for all observed data in D, but will hence perform quite bad for new and previously unseen inputs.

One way to reduce the danger of overfitting is to not consider all measurable functions $f : \mathcal{X} \to \mathbb{R}$ but to choose a smaller, but still reasonably rich, class $\mathcal{F}$ of functions that is assumed to contain a good approximation of the solution of (1.22). Instead of then looking for the infimum[3] of $\mathcal{R}_{L,\mathrm{D}}(\,\cdot\,)$ over all measurable functions, one only searches over $\mathcal{F}$, i.e., one solves

$$\inf_{f\in\mathcal{F}} \mathcal{R}_{L,\mathrm{D}}(f)\,. \tag{1.23}$$

This approach, called *empirical risk minimization* (ERM), often tends to produce approximate solutions of

$$\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} := \inf_{f\in\mathcal{F}} \mathcal{R}_{L,\mathrm{P}}(f)\,. \tag{1.24}$$

There are, however, two serious issues with ERM. The first one is that our knowledge of the distribution P is often not good enough to identify a set $\mathcal{F}$ such that a solution of (1.24) is a reasonably good approximation of a solution of (1.22). Or, the other way around, we usually cannot guarantee that the *model error* $\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} - \mathcal{R}^*_{L,\mathrm{P}}$ is sufficiently small. The second problem is that solving (1.23) might be computationally infeasible.

A first step of SVMs to make the optimization problem computationally feasible is to use a *convex* loss function, because it is easy to show that then the risk functional $f \mapsto \mathcal{R}_{L,\mathrm{P}}(f)$ is convex. If we further assume that the set $\mathcal{F}$ of functions over which we will have to optimize is also convex, the learning method defined by (1.23) will become a convex optimization problem.

A second step of SVMs towards computational feasibility is to consider only a very specific set of functions, namely the *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ of some measurable *kernel* $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. The kernel can be used to describe all functions contained in $\mathcal{H}$. Moreover, the value $k(x, x')$

[3] Which, in practice, will most often be a minimum.

can often be interpreted as a measure of similarity or dissimilarity between the two vectors x and x'. For more details on the kernel and reproducing kernel Hilbert spaces, we refer to Section 1.5.

A third step of SVMs towards computational feasibility – and also to uniqueness of the SVM – is to use a special Hilbert norm regularization term which we will describe now. It is obvious that the sum of a convex function and of a strictly convex function over a convex set is strictly convex. Let us fix an RKHS $\mathcal{H}$ over the input space $\mathcal{X}$ and denote the norm in $\mathcal{H}$ by $\|\cdot\|_{\mathcal{H}}$. The *regularization term* $\lambda\|f\|^2_{\mathcal{H}}$ also serves to reduce the danger of overfitting, see e.g., Vapnik (1998) and Schölkopf and Smola (2002). It will penalize rather complex functions f which model the output values in the training set D too closely, since these functions will have a large RKHS norm. The term

$$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) := \mathcal{R}_{L,\mathrm{P}}(f) + \lambda\|f\|^2_{\mathcal{H}}$$

is called the *regularized L-risk*, the constant λ is the *regularization parameter*. The regularized empirical L-risk is given by

$$\mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f) := \mathcal{R}_{L,\mathrm{D}}(f) + \lambda\|f\|^2_{\mathcal{H}} .$$

Definition 1.3.1. *Let $\mathcal{X}$ be the input space, $\mathcal{Y} \subset \mathbb{R}$ be the output space, $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ be a loss function, $\mathcal{H}$ be a reproducing kernel Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$, and $\lambda > 0$ be a constant. For a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, a **support vector machine** is defined as the minimizer, if it exists,*

$$f_{L,\mathrm{P},\lambda} := \arg\inf_{f\in\mathcal{H}} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda\|f\|^2_{\mathcal{H}} . \tag{1.25}$$

The empirical SVM will be denoted by

$$\begin{aligned} f_{L,\mathrm{D},\lambda} &:= \arg\inf_{f\in\mathcal{H}} \mathcal{R}_{L,\mathrm{D}}(f) + \lambda\|f\|^2_{\mathcal{H}} \\ &= \arg\inf_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, f(x_i)) + \lambda\|f\|^2_{\mathcal{H}} . \end{aligned}$$

Conditions for the existence and uniqueness of the SVM $f_{L,\mathrm{P},\lambda}$ will be stated in Section 1.6.

If we recall that an *M-estimator* is some estimator $T_n := T_n(X_1, \ldots, X_n)$ defined as

$$\arg\min \sum \rho(x_i; T_n) ,$$

with ρ an arbitrary function on $\mathcal{X} \times \mathbb{R}$, see, e.g., Huber (1981), then the above definition of $f_{L,\mathrm{D},\lambda}$ shows us that SVMs can be seen as M-estimators with a Hilbert norm regularization term for functions.

1.4 Loss Functions and the Risk

As seen in the previous section, the support vector machine uses a loss function to measure the similarity between the observed output y_i and the predicted output $f(x_i)$ for a given risk vector x_i. This similarity will be measured by looking at the L-risk associated with the loss function L. In this section we will take a closer look at the loss and the risk, and state some properties of the L-risk.

1.4.1 Definitions

We will start by defining the loss function and the L-risk, and give some examples of commonly used losses. We will assume that, unless otherwise stated, all subsets of $\mathbb{R}^d$ are equipped with their Borel σ-algebra, and that products of measurable[4] spaces are equipped with the corresponding product σ-algebra.

Definition 1.4.1. *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space, and $\mathcal{Y}$ be a closed subset of $\mathbb{R}$. A function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called a **loss function**, or in short a **loss**, if it is measurable.*

The value $L(x, y, f(x))$ gives the *cost* or *loss* incurred for predicting y by $f(x)$. Therefore, the smaller this value is, the better the prediction will be. In the same reasoning, it is logical to assume that $L(x, y, y) = 0$ for all $y \in \mathcal{Y}$, since in this case the forecast $f(x)$ equals the observed value y and hence there is no loss.

Recall from the previous section that it was our goal to obtain a small average loss for future, so far unseen, observations. Therefore we need the following definition.

Definition 1.4.2. *Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function, and* P *be a probability measure on $\mathcal{X} \times \mathcal{Y}$. For a measurable function $f : \mathcal{X} \to \mathbb{R}$, the **$L$-risk** is then defined as*

$$\mathcal{R}_{L,\mathrm{P}}(f) := \mathbb{E}_{\mathrm{P}} L(X, Y, f(X)) = \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) d\mathrm{P}(x, y) \,.$$

*The **Bayes risk** with respect to* P *and L is given by*

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

[4] If not specified, we mean Borel-measurable.

*and the measurable function $f^*_{L,\mathrm{P}} : \mathcal{X} \to \mathbb{R}$ for which $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ is the **Bayes decision function**.*

We call a loss function L convex, continuous, Lipschitz continuous, or differentiable, if L has this property with respect to its third argument. E.g., L is Lipschitz continuous if there exists a constant $|L|_1 \in (0, \infty)$, called the *Lipschitz constant*, such that, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and all $t_1, t_2 \in \mathbb{R}$,

$$|L(x, y, t_1) - L(x, y, t_2)| \leq |L|_1 \, |t_1 - t_2|.$$

If $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ only depends on its last two arguments, i.e., if there exists a measurable function $\breve{L} : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ such that $L(x, y, t) = \breve{L}(y, t)$ for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$, then L is called a *supervised loss.*

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

$$L(y, t) = \varphi(yt)$$

for all $(y, t) \in \mathcal{Y} \times \mathbb{R}$. A loss function $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called *distance-based*, if there exists a representing function $\psi : \mathbb{R} \to [0, \infty)$ with

$$L(y, t) = \psi(y - t)$$

for all $(y, t) \in \mathcal{Y} \times \mathbb{R}$ and $\psi(0) = 0$. It is called *symmetric* if $\psi(r) = \psi(-r)$ for all $r \in \mathbb{R}$. As will become clear, most losses for classification are margin-based, whereas most loss functions often used in regression are distance-based.

A loss function L is called a *Nemitski loss* if there exists a measurable function $b : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ and an increasing function $h : [0, \infty) \to [0, \infty)$ such that

$$L(x, y, t) \leq b(x, y) + h(|t|) \,, \quad (x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \,.$$

If additionally $b \in \mathcal{L}_1(\mathrm{P})$, we say that L is a P*-integrable Nemitski loss.*

Traditionally, research in non-parametric regression is often based on the *least squares loss*

$$L_{LS}(x, y, t) := (y - t)^2 \,.$$

The least squares loss function is convex in t, is useful to estimate the conditional mean function, and is advantageous from a numerical point of view, but L_{LS} is not Lipschitz continuous. From a practical point of view there are situations in which a different loss function is more appropriate.

In some situations one is actually not interested in modeling the conditional mean, but in fitting a conditional quantile function instead. For this purpose the *pinball loss* function

$$L_{\tau-pin}(x,y,t) := \begin{cases} (\tau-1)(y-t)\,, & \text{if } y-t<0 \\ \tau(y-t)\,, & \text{if } y-t\geq 0 \end{cases} \tag{1.26}$$

is used, where $\tau \in (0,1)$ specifies the desired conditional quantile, see Koenker and Bassett (1978) and Koenker (2005) for parametric quantile regression and Takeuchi *et al.* (2006) for non-parametric quantile regression. The pinball loss is, for example, often used in econometrics.

If the goal is to estimate the conditional median function, then the ϵ-*insensitive loss* given by

$$L_\epsilon(x,y,t) := \max\{|y-t|-\epsilon, 0\}\,,$$

$\epsilon \in (0,\infty)$, promises algorithmic advantages in terms of sparseness compared to the $L1$-*loss* function $L_{L1}(y,t) = |y-t|$, see Vapnik (1998) and Schölkopf and Smola (2002).

And finally, if the conditional distribution of Y given $X = x$ is known to be symmetric, basically all distance-based loss functions of the form $L(y,t) = \psi(r)$ with $r = y-t$, 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, see Steinwart (2007). An example is the *logistic loss* for regression defined as

$$\begin{aligned} L_{r-log}(x,y,t) &:= -\ln\frac{4\exp(y-t)}{(1+\exp(y-t))^2} \\ &= -\ln\big(4\Lambda(y-t)(1-\Lambda(y-t))\big)\,, \end{aligned} \tag{1.27}$$

with $\Lambda(y-t) = 1/\big(1+e^{-(y-t)}\big)$. If one fears outliers in y-direction, then a less steep loss function such as *Huber's loss* function given by

$$L_{c-Huber}(x,y,t) := \begin{cases} 0.5(y-t)^2 & \text{if } |y-t| \leq c \\ c|y-t| - c^2/2 & \text{if } |y-t| > c \end{cases}$$

for some $c \in (0,\infty)$, may be more suitable, see, e.g., Huber (1964) and Christmann and Steinwart (2007).

Since the focus of this work is mainly on regression, we have concentrated the above reasoning on loss functions for (quantile) regression. Plots of these

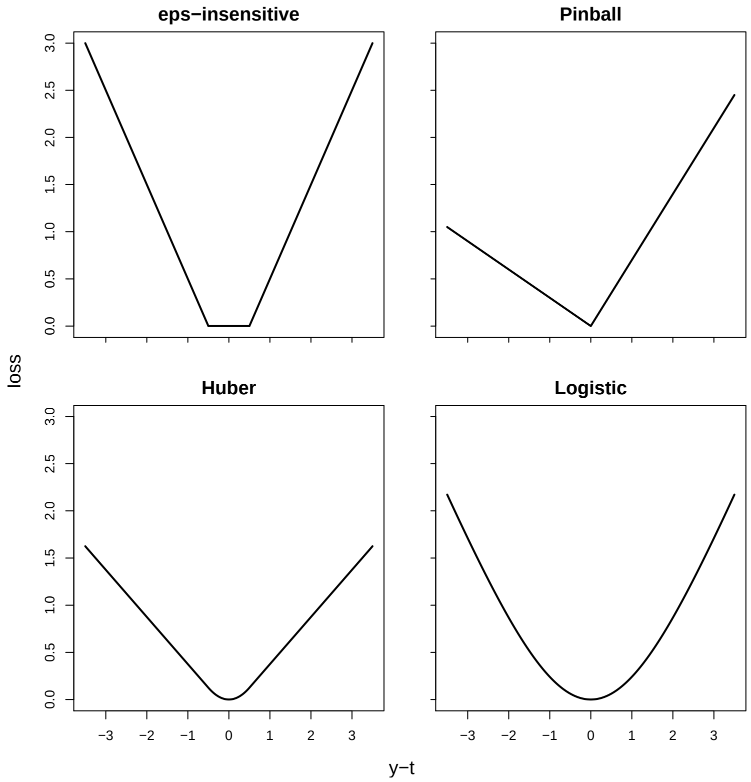

Figure 1.3: The plot shows some commonly used loss functions for regression: ϵ-insensitive loss with $\epsilon = 0.5$, pinball loss with $\tau = 0.7$, Huber loss with $\alpha = 0.5$, and logistic loss.

loss functions can be seen in Figure 1.3. However, for completeness, we will also define two losses that are commonly used in classification problems. These are the the *hinge loss*

$$L_{hinge}(x, y, t) := \max\{0, 1 - yt\}$$

and the *logistic loss* for classification

$$L_{c-log}(x, y, t) := \ln(1 + \exp(-yt)) \tag{1.28}$$

for classification, for which $(x, y, t) \in \mathcal{X} \times \{-1, +1\} \times \mathbb{R}$.

Notice that all six loss functions mentioned above (not counting the least squares loss and the $L1$-loss) are convex and Lipschitz continuous supervised losses, but only the logistic losses are twice continuously Fréchet-differentiable.[5] Note that both the ϵ-insensitive loss and the pinball loss are not even once Fréchet-differentiable. Clearly, the hinge loss and the logistic loss for classification are margin-based losses, whereas the other four are distance-based. All of the mentioned distance-based losses, except for the pinball loss with $\tau \neq 0.5$, are symmetric.

The reason to consider only SVMs based on a *convex* loss function is that then they have, under weak assumptions, at least the following four advantageous properties, see e.g., Vapnik (1998), Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002), and Steinwart and Christmann (2008b) for details.

i) Firstly, the objective function in (1.25) becomes convex in f and a support vector machine $f_{L,\mathrm{D},\lambda}$ exists, is unique, and depends continuously on the data points in basically all situations of interest (see Theorem 1.6.3). The SVM is therefore the solution of a well-posed convex optimization problem in Hadamard's sense. [6] Moreover, this minimizer is of the form

$$f_{L,\mathrm{D},\lambda} = \sum_{i=1}^{n} \alpha_i k(x_i, \,\cdot\,)\,, \tag{1.29}$$

where k is the kernel corresponding to the RKHS $\mathcal{H}$ and the $\alpha_i \in \mathbb{R}$, $i = 1 \ldots n$, are suitable coefficients. We see that the minimizer $f_{L,\mathrm{D},\lambda}$ is thus a weighted sum of (at most) n kernel functions $k(x_i, \,\cdot\,)$, where the weights α_i are data-dependent, cfr. the historical discussion of SVMs in Section 1.2. A consequence of (1.29) is that the SVM $f_{L,\mathrm{D},\lambda}$ is contained in a *finite* dimensional space, even if the space $\mathcal{H}$ itself is considerably larger. This observation makes it possible to consider even *infinite* dimensional spaces $\mathcal{H}$ such as the one corresponding to the popular *Gaussian radial basis function* (RBF) kernel defined in (1.30).

ii) Furthermore, these SVMs are under very weak assumptions L-risk consistent, i.e., for suitable null-sequences (λ_n) with $\lambda_n > 0$ we have

$$\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}\,, \qquad n \to \infty\,,$$

in probability.

[5] See Appendix A.3.3 for the definition of Fréchet-derivatives.

[6] See the definition on p. 128.

iii) Next, support vector machines based on a convex loss function have good statistical robustness properties, if k is continuous and *bounded* in the sense of $\|k\|_\infty := \sup\{\sqrt{k(x,x)} : x \in \mathcal{X}\} < \infty$ *and* if L is *Lipschitz continuous*, see Christmann and Steinwart (2004, 2007). In a nutshell, statistical robustness implies that the SVM $f_{L,\mathrm{P},\lambda}$ only varies in a smooth and bounded manner if P changes slightly in the set $\mathcal{M}_1$ of all probability measures on $\mathcal{X} \times \mathcal{Y}$.

iv) And last but not least, there exist efficient numerical algorithms to determine the vector of the weights[7] $\alpha = (\alpha_1, \ldots, \alpha_n)$ in the empirical representation (1.29) and therefore also the SVM $f_{L,\mathrm{D},\lambda}$ even for large and high-dimensional data sets D. From a numerical point of view, the vector α is usually computed as a solution of the convex dual problem derived from a Lagrange approach, see Section 1.2.

If the loss function is not convex, the SVM may in general be not unique, and the optimization problem might encounter computational difficulties. It needs to be remarked that recently non-convex functions or non-convex optimization problems have also been considered, see, e.g., Wu and Liu (2007), Guillory *et al.* (2009), Masnadi-Shirazi and Vasconcelos (2009) or Ding and Vishwanathan (2011). It should also be noted that, in fact, we do not really need the convexity of the loss function itself, but rather the fact that the risk is convex for all distributions P, because in this case we obtain (by adding the strictly convex regularization term) a strictly convex objective function. However, to the extent of our knowledge, there exist *no* non-convex losses for regression such that the risk is convex for all P.

Often, the Lipschitz continuity of the loss function L is also needed since this will be a condition for most robustness and consistency results. Clearly, the six loss functions defined earlier are all Lipschitz continuous. Another nice feat is that Lipschitz continuous loss functions are trivially Nemitski loss functions for *all* probability measures on $\mathcal{X} \times \mathcal{Y}$, because

$$\begin{aligned} L(x,y,t) &= L(x,y,0) + L(x,y,t) - L(x,y,0) \\ &\leq b(x,y) + |L|_1 \, |t| \, , \end{aligned}$$

where $b(x,y) := L(x,y,0)$ for $(x,y,t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ and $|L|_1 \in (0,\infty)$ denotes the Lipschitz constant of L. Furthermore, Lipschitz continuous L

[7] The vector α is in general not unique. This is easy to see from (1.29) for the case that two data points are identical, say $(x_i, y_i) = (x_j, y_j)$ for some pair of indices $i \neq j$. However, the SVM $f_{L,\mathrm{D},\lambda}$ exists and is unique under the very weak assumptions from Assumption 1.7.1.

are P-integrable if $\mathcal{R}_{L,\mathrm{P}}(0)$ is finite, see Steinwart and Christmann (2008b, p. 31).

As a final remark, we would like to mention that, although the least squares loss is not Lipschitz continuous, there has been extended research on the *least squares support vector machine* (LS-SVM), see Suykens *et al.* (2002a) and the references therein. However, sparseness is lost in the LS-SVM case and this method is not robust. To solve these shortcomings, a weighted LS-SVM has been proposed (Suykens *et al.*, 2002b).

1.4.2 Properties of the Loss and the Risk

In this subsection, we will give some necessary properties of the loss and its associated risk. The first lemma, see, e.g., Steinwart and Christmann (2008b, Lemma 2.13) shows that the use of a convex loss function implies that the risk is also convex.

Lemma 1.4.3 (Convexity of the risks)**.** *Let* $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ *be a (strictly) convex loss, and* P *be a probability measure on* $\mathcal{X} \times \mathcal{Y}$*. Then* $\mathcal{R}_{L,\mathrm{P}} : \mathcal{L}_0(\mathcal{X}) \to [0, \infty]$ *is (strictly) convex.*

It can be shown that, given some minor assumptions (e.g., the loss needs to be Lipschitz continuous), the continuity of the loss function will assure the continuity of the risk. A modified version of this lemma is given in Lemma 1.7.8.

The next lemma (Steinwart and Christmann, 2008b, Lemma 2.19) relates the Lipschitz continuity of L to the Lipschitz continuity of its risk.

Lemma 1.4.4 (Lipschitz continuity of the risks)**.** *Let* $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ *be a Lipschitz continuous loss, and* P *be a probability measure on* $\mathcal{X} \times \mathcal{Y}$*. Then we have, for all* $f, g \in L_\infty(\mathrm{P}_X)$*,*

$$|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)| \leq |L|_1 \cdot \|f - g\|_{L_1(\mathrm{P}_X)} \, .$$

1.5 Kernels and the Reproducing Kernel Hilbert Space

In this section, we examine the kernel and the reproducing kernel Hilbert space more closely. As we have seen in Subsection 1.2.2, the kernel and the associated feature map allowed us to obtain non-linear decision functions by using the linear SVM approach by mapping the original input data in a higher-dimensional feature space $\mathcal{H}$. An example can be seen in Figure 1.4, where the original two-dimensional data are mapped in a three-dimensional space so that the regression can be done by fitting a hyperplane in feature space instead of a (possibly more difficult) non-linear function in the original space.

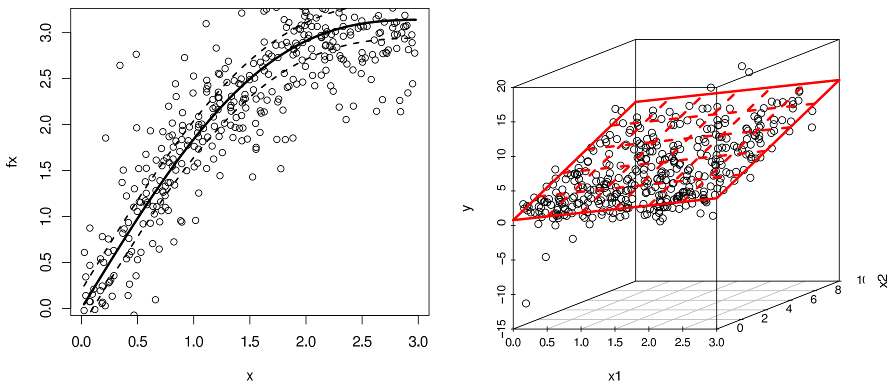

Figure 1.4: Mapping the original data to a higher dimensional space allows for regression by a hyperplane instead of a non-linear function.

We will start by giving a formal definition of the kernel and the feature map, we will then describe the reproducing kernel Hilbert space, and finally we will state some needed properties of kernels and RKHSs.

1.5.1 An Overview on Kernels

We will first introduce the kernel, the feature map and the feature space and give some examples of commonly used kernels. We will only consider real-valued kernels, but for completeness we should mention that most of the theory described below will also be applicable to complex-valued kernels. Since we only work with $\mathbb{R}$-valued kernels, the Hilbert spaces we consider will be $\mathbb{R}$-Hilbert spaces, and we will not explicitly state this further on.

Definition 1.5.1. *Let $\mathcal{X}$ be a non-empty set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We call k a* ***kernel*** *on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}_0$ and a map $\Phi : \mathcal{X} \to \mathcal{H}_0$ such that for all $x, x' \in \mathcal{X}$ we have*

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

In this case, Φ is called the ***feature map*** *and $\mathcal{H}_0$ the* ***feature space*** *of k.*

For a given kernel, neither the feature map, nor the feature space are uniquely defined. To obtain this uniqueness, we will introduce the reproducing kernel Hilbert space in the next subsection.

Before we give some examples of commonly used kernels, we will first discuss how we can create kernels from scratch (Steinwart and Christmann, 2008b, Chapter 4.1).

Lemma 1.5.2 (Construction of kernels)**.** *Let $\mathcal{X}$, $\tilde{\mathcal{X}}$, $\mathcal{X}_1$, $\mathcal{X}_2$ be non-empty sets, $\alpha \geq 0$, k, $k^\dagger$ be kernels on $\mathcal{X}$, k_1 be a kernel on $\mathcal{X}_1$, k_2 be a kernel on $\mathcal{X}_2$, $A : \tilde{\mathcal{X}} \to \mathcal{X}$ be a map, and $f_n : \mathcal{X} \to \mathbb{R}$, $n \in \mathbb{N}$, be functions such that $(f_n(x)) \in \ell_2$ for all $x \in \mathcal{X}$. Then:*

i) The function $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by

$$\tilde{k}(x, x') := \sum_{i=1}^{\infty} f_n(x) f_n(x')\,, \qquad x, x' \in \mathcal{X}\,,$$

defines a kernel on $\mathcal{X}$.

ii) (Restriction of kernels) The function $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $\tilde{k} := k(A(x), A(x'))$, $x, x' \in \mathcal{X}$, is a kernel on $\tilde{\mathcal{X}}$. In particular, if $\tilde{\mathcal{X}} \subset \mathcal{X}$, then the restriction $k_{|\tilde{\mathcal{X}} \times \tilde{\mathcal{X}}}$ is a kernel on $\tilde{\mathcal{X}}$.

iii) Both αk and $k + k^\dagger$ are also kernels on $\mathcal{X}$.

iv) (Product of kernels) $k_1 \cdot k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. In particular, if $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$, then $\tilde{k}(x, x') := k_1(x, x') k_2(x, x')$ defines a kernel on $\mathcal{X}$.

Remark that, in general, differences of kernels are *not* kernels. With these rules,and in some cases with the help of Taylor series, it is possible to construct the following kernels.

Let $m \geq 0$ be an integer and $c \geq 0$ be a real number, and take $x, x' \in \mathcal{X}$. Some kernels that are often used in practice are the *polynomial* kernel defined as

$$k(x, x') := (\langle x, x' \rangle + c)^m\,,$$

with as special case the *linear* kernel ($m = 1$ and $c = 0$), the *exponential* kernel that is given as $k(x, x') := \exp(\langle x, x' \rangle)$, and the *Gaussian radial basis function (RBF)* kernel which is defined as

$$k_{\mathrm{RBF}}(x, x') = \exp(-\gamma^{-2} \|x - x'\|^2)\,, \tag{1.30}$$

where γ is a positive constant, called the width.

Recall that a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *positive definite* if, for all $n \in \mathbb{N}$, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, and all $x_1, \dots, x_n \in \mathcal{X}$, holds that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_j, x_i) \geq 0 \, .$$

If for mutually distinct $x_1, \dots, x_n \in \mathcal{X}$ the equality only holds if $\alpha_1 = \ldots = \alpha_n = 0$, then k is called *strictly positive definite.* k is said to be *symmetric* if $k(x, x') = k(x', x)$ for all $x, x' \in \mathcal{X}$.

The following well-known result shows that symmetry and positive definiteness are sufficient and necessary conditions for kernels, see, for instance, Steinwart and Christmann (2008b, Theorem 4.16).

Theorem 1.5.3 (Symmetric, positive definite functions are kernels)**.** *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.*

1.5.2 The Reproducing Kernel Hilbert Space

We will now introduce reproducing kernel Hilbert spaces and relate them to kernels. It can be shown that the RKHS is in a way the smallest feature space of the kernel and can thus be seen as a canonical feature space.

Definition 1.5.4. *Let $\mathcal{X} \neq \emptyset$ and $\mathcal{H}$ a Hilbert space consisting of functions mapping from $\mathcal{X}$ into $\mathbb{R}$.*

i) a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a ***reproducing kernel*** *of $\mathcal{H}$ if, for all $x \in \mathcal{X}$, $k(\,\cdot\,, x) \in \mathcal{H}$ and the* ***reproducing property***

$$f(x) = \langle f, k(\,\cdot\,, x) \rangle_{\mathcal{H}} \tag{1.31}$$

holds for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$.

ii) The space $\mathcal{H}$ is called a ***reproducing kernel Hilbert space (RKHS)*** *over $\mathcal{X}$ if it possesses a reproducing kernel.*

The following theorem (Yosida, 1974, Theorem 1, p. 96) is based on results from Aronszajn (1950) and Bergman (1950) and tells us a bit more about the existence of reproducing kernels.

Theorem 1.5.5 (Existence of reproducing kernel)**.** *Let $\mathcal{X}$ be a set and $\mathcal{H}_0$ be a Hilbert space of functions over $\mathcal{X}$. $\mathcal{H}_0$ then has a reproducing kernel k if and only if there exists, for all $x \in \mathcal{X}$, a positive constant C_x which depends on x, such that for all $f \in \mathcal{H}_0$*

$$|f(x)| \leq C_x \, \|f\|_{\mathcal{H}_0} \, .$$

Moreover, this reproducing kernel k is unique.

In order to have the existence of a reproducing kernel, all functions in the Hilbert space thus have to be pointwise bounded by a multiple of their norm. Another way to state this, is to say that the Dirac functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined as $\delta_x(f) := f(x)$, $f \in \mathcal{H}$, has to be continuous.

It can be shown, see, e.g., Steinwart and Christmann (2008b, Lemma 4.19), that reproducing kernels are kernels in the sense of Definition 1.5.1.

Lemma 1.5.6. *Let $\mathcal{H}$ be a function Hilbert space over $\mathcal{X}$ that has a reproducing kernel k. Then $\mathcal{H}$ is an RKHS and $\mathcal{H}$ is also a feature space of k. In this case the feature map $\Phi : \mathcal{X} \to \mathcal{H}$ is given by*

$$\Phi(x) := k(\,\cdot\,, x)\,, \qquad x \in \mathcal{X}\,,$$

and is called the ***canonical feature map****.*

From the above lemma follows that every Hilbert function space with a reproducing kernel is an RKHS. Because of (1.31), the kernel can be used to describe all functions contained in $\mathcal{H}$. Moreover, the value $k(x, x')$ can often be interpreted as a measure of similarity or dissimilarity between the two risk vectors x and x'. Due to the above, the reproducing property can be also written as, for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$,

$$f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}} \, .$$

Furthermore, every RKHS is uniquely defined by its kernel k and vice versa. From Theorem 1.5.5 we already know that the reproducing kernel of a given Hilbert space is unique, Steinwart and Christmann (2008b, Theorem 4.21) show us the uniqueness of the RKHS for a given kernel k.

Theorem 1.5.7 (Every kernel has a unique RKHS)**.** *Let $\mathcal{X}$ be a non-empty set and k be a kernel over $\mathcal{X}$ with feature space $\mathcal{H}_0$ and feature map $\Phi_0 : \mathcal{X} \to \mathcal{H}_0$. Then*

$$\mathcal{H} := \{f : \mathcal{X} \to \mathbb{R} \,|\, \exists\, w \in \mathcal{H}_0 \text{ with } f(x) = \langle w, \Phi_0(x) \rangle_{\mathcal{H}_0} \text{ for all } x \in \mathcal{X}\} \tag{1.32}$$

equipped with the norm

$$\|f\|_{\mathcal{H}} := \inf\{\|f\|_{\mathcal{H}_0} \mid \exists\, w \in \mathcal{H}_0 \text{ with } f = \langle w, \Phi_0(\,\cdot\,)\rangle_{\mathcal{H}_0}\}$$

is the only RKHS with k as reproducing kernel. Obviously, both definitions are independent of the choice of $\mathcal{H}_0$ and Φ_0. Moreover, the operator $V : \mathcal{H}_0 \to \mathcal{H}$ defined by

$$Vw := \langle w, \Phi_0(\,\cdot\,)\rangle_{\mathcal{H}_0}\,, \qquad w \in \mathcal{H}_0\,,$$

is a metric surjection, this means that $VB^*_{\mathcal{H}_0} = B^*_{\mathcal{H}}$, where $B^*_{\mathcal{H}_0}$ and $B^*_{\mathcal{H}}$ are the open unit balls of respectively $\mathcal{H}_0$ and $\mathcal{H}$.

Theorems 1.5.5 and 1.5.7 thus tell us that there exists a one-to-one relationship between the kernel and the RKHS. The expression (1.32) tells us that the RKHS associated with a kernel k consists exactly of all possible functions of the given form, which consequently allows us to determine the RKHS of a given kernel. Moreover, it also shows that this set of functions remains unchanged when considering other feature spaces of k. For more details on reproducing kernels and their RKHSs, we refer the interested reader to the book by Berlinet and Thomas-Agnan (2004).

To conclude this part, we will take a look at when the RKHS $\mathcal{H}$ is dense in $L_1(\mu)$ with μ a distribution on $\mathcal{X}$, since this will be a key assumption on $\mathcal{H}$ needed to obtain consistency. There exists a characterization for this fact that is, unfortunately, often hard to verify, see Steinwart and Christmann (2008b, Lemma 4.59). Luckily, Steinwart and Christmann (2008b, Theorem 4.63) shows that the denseness assumption for the special case of the Gaussian RBF kernel with width γ is trivial. Both of these characterizations are based on results from Steinwart *et al.* (2006).

Theorem 1.5.8. *Let $\gamma > 0$, $p \in [1, \infty)$, and μ be a finite measure on $\mathbb{R}^d$. Then the RKHS $\mathcal{H}$ of the Gaussian RBF kernel k_{RBF} is dense in $L_p(\mu)$.*

1.5.3 Properties of the RKHS

In this subsection, we will verify that some properties of the kernel, such as boundedness or measurability, are transferred to the functions of the RKHS.

For k a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$, the Cauchy-Schwarz inequality and (1.31) show us that

$$\begin{aligned} |k(x,x')|^2 = |\langle k(\,\cdot\,,x), k(\,\cdot\,,x')\rangle_{\mathcal{H}}|^2 &\leq \|k(\,\cdot\,,x)\|^2_{\mathcal{H}}\,\|k(\,\cdot\,,x')\|^2_{\mathcal{H}} \\ &= k(x,x)\,\cdot\,k(x',x') \qquad (1.33) \end{aligned}$$

for all $x, x' \in \mathcal{X}$. Therefore $\sup_{x,x' \in \mathcal{X}} |k(x,x')| = \sup_{x \in \mathcal{X}} k(x,x)$, and hence a kernel k is called *bounded*, if

$$\|k\|_\infty := \sup_{x \in \mathcal{X}} \sqrt{k(x,x)} < \infty \,.$$

The equation (1.33) also shows that, for $\Phi : \mathcal{X} \to \mathcal{H}$ a feature map of k, $\|\Phi(x)\|_\mathcal{H} = \sqrt{k(x,x)}$ for all $x \in \mathcal{X}$. Thus Φ is bounded if and only if k is bounded. Using this equality and the reproducing property, we can obtain the well-known inequalities

$$\|f\|_\infty \leq \|k\|_\infty \|f\|_\mathcal{H} \quad \text{and} \quad \|\Phi(x)\|_\infty \leq \|k\|_\infty \|\Phi(x)\|_\mathcal{H} \leq \|k\|_\infty^2 \tag{1.34}$$

for $f \in \mathcal{H}$ and $x \in \mathcal{X}$. As an example of a bounded kernel we mention the Gaussian RBF kernel, a feat that adds to its popularity. Furthermore, this kernel is *universal* in the sense of Steinwart (2001), that is, its RKHS is dense in $C(\mathcal{X})$ for all compact $\mathcal{X} \subset \mathbb{R}^d$. Finally, see Theorem 4.63 of Steinwart and Christmann (2008b), its RKHS is dense in $L_1(\mu)$ for all probability measures μ on $\mathbb{R}^d$. The corresponding RKHS of this kernel has infinite dimension.

We will need the following results, see, e.g., Steinwart and Christmann (2008b, Lemma 4.23 and Lemma 4.24), for our proofs. The first lemma shows that if k is bounded, so are all the functions in its RKHS.

Lemma 1.5.9 (RKHSs of bounded kernels)**.** *Let $\mathcal{X}$ be a set and k be a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$. Then k is bounded if and only if every $f \in \mathcal{H}$ is bounded. Moreover, in this case the inclusion* $\mathrm{id} : \mathcal{H} \to \ell_\infty(\mathcal{X})$ *is continuous and we have* $\|\mathrm{id} : \mathcal{H} \to \ell_\infty(\mathcal{X})\| = \|k\|_\infty$.

The second lemma gives a similar result for k a measurable kernel.

Lemma 1.5.10 (RKHSs of measurable kernels)**.** *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and k be a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$. Then all $f \in \mathcal{H}$ are measurable if and only if $k(\,\cdot\,, x) : \mathcal{X} \to \mathbb{R}$ is measurable for all $x \in \mathcal{X}$.*

1.6 Existence and uniqueness of the SVM

We saw before that the SVM, also called the SVM solution or SVM decision function, was defined as

$$f_{L,\mathrm{P},\lambda} := \arg \inf_{f \in \mathcal{H}} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|_\mathcal{H}^2 \,.$$

Of course this raises the question when exactly the SVM is defined and when is it unique. The following lemmae, see Steinwart and Christmann (2008b, Lemma 5.1 and Lemma 5.2), will provide us with an answer.

Lemma 1.6.1 (Uniqueness of SVM solutions)**.** *Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex loss, $\mathcal{H}$ be the RKHS of a measurable kernel over $\mathcal{X}$, and* P *be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for some $f \in \mathcal{H}$. Then for all $\lambda > 0$ there exists at most one general SVM solution $f_{L,\mathrm{P},\lambda}$.*

Lemma 1.6.2 (Existence of SVM solutions)**.** *Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex loss, $\mathcal{H}$ be the RKHS of a bounded measurable kernel over $\mathcal{X}$, and* P *be a distribution on $\mathcal{X} \times \mathcal{Y}$ such that L is a* P*-integrable Nemitski loss. Then for all $\lambda > 0$ there exists a general SVM solution $f_{L,\mathrm{P},\lambda}$.*

There exist also a number of representation theorems for support vector machines, and we would in particular like to mention the following result for the empirical SVM $f_{L,\mathrm{D},\lambda}$ (Steinwart and Christmann, 2008b, Theorem 5.5).

Theorem 1.6.3. *Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex loss, $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, and $\mathcal{H}$ an RKHS over $\mathcal{X}$. Then, for all $\lambda > 0$, there exists a unique empirical SVM solution $f_{L,\mathrm{D},\lambda} \in \mathcal{H}$ satisfying*

$$\mathcal{R}_{L,\mathrm{D}}(f_{L,\mathrm{D},\lambda}) + \lambda \left\| f_{L,\mathrm{D},\lambda} \right\|_{\mathcal{H}}^2 = \inf_{f \in \mathcal{H}} \mathcal{R}_{L,\mathrm{D}}(f) + \lambda \left\| f \right\|_{\mathcal{H}}^2 .$$

Furthermore, there exist $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$f_{L,\mathrm{D},\lambda}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) , \qquad x \in \mathcal{X} . \tag{1.35}$$

We see that, by using a convex loss function, the empirical SVM solution can be written as a weighted sum of at most n kernel terms, where the weights α_i are data-dependent, cfr. the explanation given after formula line (1.29). This clearly shows the connection with the geometrical interpretation of SVMs derived in Section 1.2: the set of support vectors will consist of the x_i for which α_i is not zero.

1.7 Shifting the Loss Function

Support vector machines are known to be consistent and robust both for classification and regression purposes if they are based on a Lipschitz continuous loss and a bounded kernel with a separable RKHS that is dense in $L_1(\mu)$ for all distributions μ, see e.g., Christmann and Steinwart (2007, 2008), and Steinwart and Christmann (2008b). These facts are even true in the regression context for unbounded output spaces, if the target function f is integrable with respect to the marginal distribution of the input variable X and if the output variable Y has a finite first absolute moment. However, the latter assumption clearly excludes distributions with heavy tails, such as several stable distributions, including the Cauchy distribution, or some extreme value distributions which occur in financial or insurance projects.

In this section we will show that we can enlarge, by considering shifted loss functions, the applicability of SVMs even to heavy-tailed distributions, which violate the previously mentioned moment condition. We will describe the approach of SVMs based on shifted loss functions and list some properties of such SVMs. We will also give results on existence, uniqueness and representation. The consistency and statistical robustness of such SVMs will be proven in the following chapters, thus showing that SVMs can even successfully deal with heavy-tailed conditional distributions of the response variable Y given x. The results for SVMs based on shifted loss functions are mainly meant for regression purposes. Of course, the case of classification where the output space $\mathcal{Y}$ is just a set of a finite number of real numbers is a special case and thus classification is covered by the results we will show. The problem of heavy tails is however *not* present in classification because the conditional distribution of the response variable Y given x has then obviously a bounded support.

The goal of this section is twofold. First we clarify why SVMs for response variables with heavy tails need careful consideration, and secondly we show that SVMs based on shifted loss functions $L^\star$ are defined for *all* distributions and that they are identical to SVMs based on unshifted loss functions for all data sets. Hence no new algorithms need to be established to compute the SVM based on the shifted loss function.

1.7.1 The $L^\star$-Trick

Let us assume that the probability measure P on $\mathcal{X} \times \mathcal{Y}$ can be split up into the marginal distribution P_X on $\mathcal{X}$ and the conditional probability $\mathrm{P}(y|x)$ on $\mathcal{Y}$. This is, e.g., possible when $\mathcal{Y} \subset \mathbb{R}$ is closed, because in that case $\mathcal{Y}$ is a complete separable metric space, and therefore a Polish space (see the text after Definition A.1.2), and $\mathcal{X}$ is a measurable space and thus Lemma A.2.7 is applicable. Furthermore, take $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ a Lipschitz continuous loss function. If we then take a look at the L-risk, keeping in mind that $L(X,Y,Y)=0$, we obtain the inequality

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f) &= \mathbb{E}_{\mathrm{P}}\big(L(X,Y,f(X))-L(X,Y,Y)\big) \qquad (1.36)\\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} L(x,y,f(x))-L(x,y,y)\,d\mathrm{P}(y|x)\,d\mathrm{P}_X(x)\\ &\leq |L|_1\int_{\mathcal{X}}\int_{\mathcal{Y}} |f(x)-y|\,d\mathrm{P}(y|x)\,d\mathrm{P}_X(x)\\ &\leq |L|_1\int_{\mathcal{X}} |f(x)|\,d\mathrm{P}_X(x)+|L|_1\int_{\mathcal{X}}\int_{\mathcal{Y}}|y|d\mathrm{P}(y|x)\,d\mathrm{P}_X(x)\,,\end{aligned}$$

which is finite, if $f\in L_1(\mathrm{P}_X)$ *and* the first absolute moment

$$\mathbb{E}_{\mathrm{P}}|Y| = \int_{\mathcal{X}}\int_{\mathcal{Y}} |y|\,d\mathrm{P}(y|x)\,d\mathrm{P}_X(x)<\infty\,. \qquad (1.37)$$

It is exactly this latter condition that excludes heavy-tailed distributions such as many stable distributions, including the Cauchy distribution, and many extreme value distributions which occur in financial or actuarial problems. The moment condition (1.37) is one of the assumptions made by Christmann and Steinwart (2007) and Steinwart and Christmann (2008b) for their consistency and robustness proofs of SVMs for an *unbounded* output set $\mathcal{Y}$.

As said in the introduction to this section, we would like to enlarge the applicability of SVMs even to heavy-tailed distributions, which violate the moment condition $\mathbb{E}_{\mathrm{P}}|Y|<\infty$. This will be done by using a trick well-known in the literature on robust statistics, see e.g., Huber (1967) for an early use of this trick on M-estimators (without regularization term). The trick consist of shifting the loss $L(x,y,t)$ downwards by the amount of $L(x,y,0)\in[0,\infty)$. We will call the function $L^\star:\mathcal{X}\times\mathcal{Y}\times\mathbb{R}\to\mathbb{R}$ defined by

$$L^\star(x,y,t) := L(x,y,t)-L(x,y,0)$$

the *shifted loss function* or the *shifted version of* L. Using this definition, we obtain, for all $f\in L_1(\mathrm{P}_X)$,

$$\begin{aligned}\mathcal{R}_{L^\star,\mathrm{P}}(f) &= \mathbb{E}_{\mathrm{P}}L^\star(X,Y,f(X)) \qquad (1.38)\\ &= \mathbb{E}_{\mathrm{P}}\big(L(X,Y,f(X))-L(X,Y,0)\big)\\ &\leq \int_{\mathcal{X}\times\mathcal{Y}} |L(x,y,f(x))-L(x,y,0)|\,d\mathrm{P}(x,y)\\ &\leq |L|_1\int_{\mathcal{X}}|f(x)|\,d\mathrm{P}_X(x)<\infty\,,\end{aligned}$$

no matter whether the moment condition (1.37) is fulfilled or violated. We will use this "$L^\star$-trick" to show, either in this chapter or in one of the following chapters, that many important results on the SVM $f_{L,\mathrm{P},\lambda}$, such as existence, uniqueness, representation, consistency, and statistical robustness, can also be shown for

$$f_{L^\star,\mathrm{P},\lambda} := \arg \inf_{f\in\mathcal{H}} \mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda \|f\|_{\mathcal{H}}^2 \, ,$$

where

$$\mathcal{R}_{L^\star,\mathrm{P}}(f) := \mathbb{E}_{\mathrm{P}} L^\star(X,Y,f(X))$$

denotes the $L^\star$-risk of f. Moreover, we will show that

$$f_{L^\star,\mathrm{P},\lambda} = f_{L,\mathrm{P},\lambda}$$

if $f_{L,\mathrm{P},\lambda}$ exists. Hence, there is no need for new algorithms to compute $f_{L^\star,\mathrm{D},\lambda}$ because the empirical SVM $f_{L,\mathrm{D},\lambda}$ exists for all data sets $D = ((x_1,y_1),\dots,(x_n,y_n)) \subset (\mathcal{X}\times\mathcal{Y})^n$. The advantage of $f_{L^\star,\mathrm{P},\lambda}$ over $f_{L,\mathrm{P},\lambda}$ is that $f_{L^\star,\mathrm{P},\lambda}$ is still well-defined and useful for heavy-tailed conditional distributions $\mathrm{P}(y|x)$, for which the first absolute moment $\int_{\mathcal{Y}} |y| d\mathrm{P}(y|x)$ is infinite. In particular, our results will show that even in the case of heavy-tailed distributions, the forecasts $f_{L^\star,\mathrm{D},\lambda}(x) = f_{L,\mathrm{D},\lambda}(x)$ are consistent (Chapter 2) and robust (Chapter 3) with respect to the influence function and the maxbias, if the kernel is bounded and a Lipschitz continuous loss function is used. In this respect, the combination of the Gaussian RBF kernel with the ϵ-insensitive loss function or Huber's loss function for regression purposes or with the pinball loss for quantile regression yields SVMs with good consistency and robustness properties.

1.7.2 Properties of Shifted Loss Functions

In this section we will state some general facts on the function $L^\star$ which will be used to obtain our results in the next section. The general assumptions for the rest of this thesis are summarized in

Assumption 1.7.1. *Let $n \in \mathbb{N}$, $\mathcal{X}$ be a complete separable metric space (e.g., a closed $\mathcal{X} \subset \mathbb{R}^d$ or $\mathcal{X} = \mathbb{R}^d$), $\mathcal{Y} \subset \mathbb{R}$ be a non-empty and closed set (e.g., $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = \{-1,+1\}$), and* P *be a probability distribution on $\mathcal{X}\times\mathcal{Y}$ equipped with its Borel σ-algebra. Since $\mathcal{Y}$ is closed,* P *can be split up into the marginal distribution* P_X *on $\mathcal{X}$ and the conditional probability*

$\mathrm{P}(y|x)$ *on* $\mathcal{Y}$. *Let* $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ *be a loss function and define its* ***shifted loss function*** $L^\star : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ *by*

$$L^\star(x, y, t) := L(x, y, t) - L(x, y, 0)\,.$$

We say that L *(or* $L^\star$*) is convex, Lipschitz continuous, continuous or differentiable, if* L *(or* $L^\star$*) has this property with respect to its third argument. If not otherwise mentioned,* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is a measurable kernel with reproducing kernel Hilbert space* $\mathcal{H}$ *of measurable functions* $f : \mathcal{X} \to \mathbb{R}$*, and* $\Phi : \mathcal{X} \to \mathcal{H}$ *denotes the canonical feature map, i.e.,* $\Phi(x) := k(\,\cdot\,, x)$ *for* $x \in \mathcal{X}$.

Please note that these assumptions are independent of the data set, and therefore can really be checked. The reason why the trick uses the value $L(x, y, 0)$ is simply that the zero function $f(x) = 0$, for all $x \in \mathcal{X}$, is *always* an element of the RKHS $\mathcal{H}$, whereas for other (constant) functions this might not be the case.

Since $L(x, y, t) \in [0, \infty)$, it clearly follows from the definition that $-\infty < L^\star(x, y, t) < \infty$, and thus it is no longer a non-negative loss. As shown in Section 1.7.1, we obtain by (1.36) that the L-risk $\mathbb{E}_{\mathrm{P}} L(X, Y, f(X))$ is finite, if $f \in L_1(\mathrm{P}_X)$ and $\mathbb{E}_{\mathrm{P}}|Y| < \infty$. On the other hand, (1.38) shows us that $\mathbb{E}_{\mathrm{P}} L^\star(X, Y, f(X))$ is finite, if $f \in L_1(\mathrm{P}_X)$ no matter whether $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ is finite or infinite. Therefore, by using the $L^\star$-trick, we can enlarge the applicability of SVMs by relaxing the finiteness of the risk. As a remark, we would like to state that there is no intuitive interpretation of the $L^\star$-function, since in practice negative losses do not make sense. The shift we use is only a mathematical trick to enlarge the domain on which SVMs are defined. In practice this means that for all data sets the SVM based on L and the SVM based on $L^\star$ will give identical results, the trick only shifts the objective function, but *not* the value where this function is minimal.

The following obvious result gives a relationship between L and $L^\star$ in terms of convexity and Lipschitz continuity.

Proposition 1.7.2. *Let* L *be a loss function. Then the following statements are valid.*

i) $L^\star$ *is (strictly) convex, if and only if* L *is (strictly) convex.*

ii) $L^\star$ *is Lipschitz continuous, if and only if* L *is Lipschitz continuous. Furthermore, both Lipschitz constants are equal, i.e.,* $|L|_1 = |L^\star|_1$.

It follows from Proposition 1.7.2 and the strict convexity of the mapping $f \mapsto \lambda \|f\|_{\mathcal{H}}^2$, $f \in \mathcal{H}$, that $L^\star(x,y,\,\cdot\,) + \lambda \|\cdot\|_{\mathcal{H}}^2$ is a strictly convex function if L is convex. This also yields for a convex loss L that the mapping

$$f \mapsto \mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda \|f\|_{\mathcal{H}}^2 \,, \qquad f \in \mathcal{H} \,,$$

is a strictly convex function because it is the sum of the convex risk functional $\mathcal{R}_{L^\star,\mathrm{P}}$ and the strictly convex mapping $f \mapsto \lambda \|f\|_{\mathcal{H}}^2$.

Remark, however, that for L a distance-based loss, $L^\star$ will not necessarily share this property as the following example shows. For the least squares loss $L_{LS}(x,y,t) = (y-t)^2$ we obtain $L^\star(x,y,t) = (y-t)^2 - (y-0)^2 = t(t-2y)$ which clearly cannot be written as a function in $y-t$ only.

Proposition 1.7.3. *The following assertions are valid.*

i) $\inf_{t\in\mathbb{R}} L^\star(x,y,t) \le 0$.

ii) If L is a Lipschitz continuous loss, then for all $f \in \mathcal{H}$:

$$-|L|_1 \mathbb{E}_{\mathrm{P}_X}|f| \le \mathcal{R}_{L^\star,\mathrm{P}}(f) \le |L|_1 \mathbb{E}_{\mathrm{P}_X}|f| \,, \tag{1.39}$$

$$-|L|_1 \mathbb{E}_{\mathrm{P}_X}|f| + \lambda \|f\|_{\mathcal{H}}^2 \le \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \le |L|_1 \mathbb{E}_{\mathrm{P}_X}|f| + \lambda \|f\|_{\mathcal{H}}^2 \,. \tag{1.40}$$

iii) $\inf_{f\in\mathcal{H}} \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \le 0$ *and* $\inf_{f\in\mathcal{H}} \mathcal{R}_{L^\star,\mathrm{P}}(f) \le 0$.

iv) Let L be a Lipschitz continuous loss and assume that $f_{L^\star,\mathrm{P},\lambda}$ exists. Then we have

$$\begin{aligned} \lambda\|f_{L^\star,\mathrm{P},\lambda}\|_{\mathcal{H}}^2 &\le -\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda}) \le \mathcal{R}_{L,\mathrm{P}}(0) \,, \\ 0 &\le -\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \le \mathcal{R}_{L,\mathrm{P}}(0) \,, \\ \lambda\|f_{L^\star,\mathrm{P},\lambda}\|_{\mathcal{H}}^2 &\le \min\bigl\{|L|_1 \mathbb{E}_{\mathrm{P}_X}|f_{L^\star,\mathrm{P},\lambda}|, \mathcal{R}_{L,\mathrm{P}}(0)\bigr\} \,. \end{aligned} \tag{1.41}$$

If the kernel k is additionally bounded, then

$$\|f_{L^\star,\mathrm{P},\lambda}\|_\infty \;\le\; \lambda^{-1}|L|_1\, \|k\|_\infty^2 < \infty \,, \tag{1.42}$$

$$|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda})| \;\le\; \lambda^{-1}|L|_1^2\, \|k\|_\infty^2 < \infty \,. \tag{1.43}$$

v) If the partial Fréchet- and Bouligand-derivatives[8] *of L and $L^\star$ exist for $(x,y) \in \mathcal{X} \times \mathcal{Y}$, then*

$$\nabla_3^F L^\star(x,y,t) \;=\; \nabla_3^F L(x,y,t) \,, \quad \forall\, t \in \mathbb{R} \,, \tag{1.44}$$

$$\nabla_3^B L^\star(x,y,t) \;=\; \nabla_3^B L(x,y,t) \,, \quad \forall\, t \in \mathbb{R} \,. \tag{1.45}$$

[8] See Appendix A.3.3 for Fréchet-derivatives and Section 3.2.1 for Bouligand-derivatives

Proof of Proposition 1.7.3. *(i)* Obviously, $\inf_{t\in\mathbb{R}} L^\star(x,y,t) \le L^\star(x,y,0) = 0$.
(ii) We have for all $f \in \mathcal{H}$ that

$$\begin{aligned}|\mathcal{R}_{L^\star,\mathrm{P}}(f)| &= |\mathbb{E}_{\mathrm{P}} L^\star(X,Y,f(X))| \\ &= |\mathbb{E}_{\mathrm{P}} L(X,Y,f(X)) - L(X,Y,0)| \\ &\le \mathbb{E}_{\mathrm{P}} |L(X,Y,f(X)) - L(X,Y,0)| \\ &\le |L|_1 \, \mathbb{E}_{\mathrm{P}_X} |f| \,,\end{aligned}$$

which proves (1.39). Equation (1.40) follows from $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) = \mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda \left\| f \right\|^2_{\mathcal{H}}$.

(iii) As $0 \in \mathcal{H}$, we obtain

$$\inf_{f\in\mathcal{H}} \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) \le \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(0) = 0$$

and the same reasoning holds for $\inf_{f\in\mathcal{H}} \mathcal{R}_{L^\star,\mathrm{P}}(f)$.

(iv) Due to *(iii)* we have $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \le 0$. Because $L \ge 0$ we obtain

$$\begin{aligned}\lambda \left\| f_{L^\star,\mathrm{P},\lambda} \right\|^2_{\mathcal{H}} &\le -\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda}) \\ &= \mathbb{E}_{\mathrm{P}} \big(L(X,Y,0) - L(X,Y,f_{L^\star,\mathrm{P},\lambda}(X)) \big) \\ &\le \mathbb{E}_{\mathrm{P}} L(X,Y,0) = \mathcal{R}_{L,\mathrm{P}}(0) \,.\end{aligned}$$

Using similar arguments as above, we get that

$$\begin{aligned}0 &\le -\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \\ &= \mathbb{E}_{\mathrm{P}} \big(L(X,Y,0) - L(X,Y,f_{L^\star,\mathrm{P},\lambda}(X)) \big) - \lambda \left\| f_{L^\star,\mathrm{P},\lambda} \right\|^2_{\mathcal{H}} \\ &\le \mathbb{E}_{\mathrm{P}} L(X,Y,0) = \mathcal{R}_{L,\mathrm{P}}(0) \,.\end{aligned}$$

Furthermore, by *(ii)*, we obtain

$$\begin{aligned}-|L|_1 \, \mathbb{E}_{\mathrm{P}_X} |f_{L^\star,\mathrm{P},\lambda}| + \lambda \left\| f_{L^\star,\mathrm{P},\lambda} \right\|^2_{\mathcal{H}} &\le \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \\ &\le \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(0) = 0 \,.\end{aligned}$$

This yields (1.41). Using (1.34) and (1.41), we obtain for $f_{L^\star,\mathrm{P},\lambda} \ne 0$ that

$$\begin{aligned}\left\| f_{L^\star,\mathrm{P},\lambda} \right\|_\infty &\le \left\| k \right\|_\infty \left\| f_{L^\star,\mathrm{P},\lambda} \right\|_{\mathcal{H}} \\ &\le \left\| k \right\|_\infty \sqrt{\lambda^{-1} |L|_1 \mathbb{E}_{\mathrm{P}_X} |f_{L^\star,\mathrm{P},\lambda}|} \\ &\le \left\| k \right\|_\infty \sqrt{\lambda^{-1} |L|_1 \left\| f_{L^\star,\mathrm{P},\lambda} \right\|_\infty} < \infty \,.\end{aligned}$$

Hence $\|f_{L^\star,\mathrm{P},\lambda}\|_\infty \le \|k\|_\infty^2\,\lambda^{-1}|L|_1$. The case $f_{L^\star,\mathrm{P},\lambda}=0$ is trivial.

Using *(ii)* we see that

$$\begin{aligned} |\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda})| &\le |L|_1\,\mathbb{E}_{\mathrm{P}_X}|f_{L^\star,\mathrm{P},\lambda}| \\ &\le |L|_1\,\|f_{L^\star,\mathrm{P},\lambda}\|_\infty \\ &\le \lambda^{-1}|L|_1^2\,\|k\|_\infty^2\,. \end{aligned}$$

(v) By definition of $L^\star$ and of the Fréchet-derivative we immediately obtain

$$\begin{aligned} \nabla_3^F L^\star(x,y,t) &= \lim_{h\to 0,\,h\neq 0} \frac{L^\star(x,y,t+h)-L^\star(x,y,t)}{h} \\ &= \lim_{h\to 0,\,h\neq 0} \frac{L(x,y,t+h)-L(x,y,t)}{h} \\ &= \nabla_3^F L(x,y,t)\,. \end{aligned}$$

An analogous calculation is valid for the Bouligand-derivative because the term $L(x,y,0)$ will cancel out in the definition of the Bouligand-derivative and we obtain $\nabla_3^B L^\star(x,y,t)=\nabla_3^B L(x,y,t)$. □

The following proposition ensures that the optimization problem to determine $f_{L^\star,\mathrm{P},\lambda}$ is well-posed.

Proposition 1.7.4. *Let L be a Lipschitz continuous loss and $f\in L_1(\mathrm{P}_X)$. Then $\mathcal{R}_{L^\star,\mathrm{P}}(f)\notin\{-\infty,+\infty\}$. Moreover, we have $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f)>-\infty$ for all $f\in L_1(\mathrm{P}_X)\cap\mathcal{H}$.*

Proof of Proposition 1.7.4. Using (1.39) we have

$$|\mathcal{R}_{L^\star,\mathrm{P}}(f)|\le |L|_1\mathbb{E}_{\mathrm{P}_X}|f|<\infty$$

for $f\in L_1(\mathrm{P}_X)$. Then (1.40) yields

$$\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f)\ge -|L|_1\mathbb{E}_{\mathrm{P}_X}|f|+\lambda\,\|f\|_{\mathcal{H}}^2>-\infty\,.$$

□

1.7.3 SVMs for Heavy-Tailed Distributions

In this subsection we will show the existence and uniqueness of the SVM $f_{L^\star,\mathrm{P},\lambda}$ together with a representation theorem.

Theorem 1.7.5 (Uniqueness of SVM)**.** *Let L be a convex loss function. Assume that (i) $\mathcal{R}_{L^\star,\mathrm{P}}(f) < \infty$ for some $f \in \mathcal{H}$ and $\mathcal{R}_{L^\star,\mathrm{P}}(f) > -\infty$ for all $f \in \mathcal{H}$ or (ii) L is Lipschitz continuous and $f \in L_1(\mathrm{P}_X)$ for all $f \in \mathcal{H}$. Then for all $\lambda > 0$ there exists at most one SVM solution $f_{L^\star,\mathrm{P},\lambda}$.*

Lemma 1.7.6 (Convexity of risks)**.** *Let L be a (strictly) convex loss. Then $\mathcal{R}_{L^\star,\mathrm{P}} : \mathcal{H} \to [-\infty, \infty]$ is (strictly) convex and $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda} : \mathcal{H} \to [-\infty, \infty]$ is strictly convex.*

Proof of Lemma 1.7.6. Proposition 1.7.2 yields that $L^\star$ is (strictly) convex. Trivially $\mathcal{R}_{L^\star,\mathrm{P}}$ is also convex. Further $f \mapsto \lambda \left\| f \right\|^2_{\mathcal{H}}$ is strictly convex, and hence the mapping $f \mapsto \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) = \mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda \left\| f \right\|^2_{\mathcal{H}}$ is strictly convex. □

Proof of Theorem 1.7.5. Let us assume that the mapping $f \mapsto \lambda \left\| f \right\|^2_{\mathcal{H}} + \mathcal{R}_{L^\star,\mathrm{P}}(f)$ has two minimizers f_1 and $f_2 \in \mathcal{H}$ with $f_1 \neq f_2$. *(i)* By application of Lemma A.3.7, we then find

$$\left\|(f_1 + f_2)/2\right\|^2_{\mathcal{H}} < \left\|f_1\right\|^2_{\mathcal{H}}/2 + \left\|f_2\right\|^2_{\mathcal{H}}/2\,.$$

The convexity of $f \mapsto \mathcal{R}_{L^\star,\mathrm{P}}(f)$, see Lemma 1.7.6, and

$$\lambda \left\|f_1\right\|^2_{\mathcal{H}} + \mathcal{R}_{L^\star,\mathrm{P}}(f_1) = \lambda \left\|f_2\right\|^2_{\mathcal{H}} + \mathcal{R}_{L^\star,\mathrm{P}}(f_2)$$

then shows for $f^* := \frac{1}{2}(f_1 + f_2)$ that

$$\lambda \left\|f^*\right\|^2_{\mathcal{H}} + \mathcal{R}_{L^\star,\mathrm{P}}(f^*) < \lambda \left\|f_1\right\|^2_{\mathcal{H}} + \mathcal{R}_{L^\star,\mathrm{P}}(f_1)\,,$$

i.e., f_1 is *not* a minimizer of $f \mapsto \lambda \left\|f\right\|^2_{\mathcal{H}} + \mathcal{R}_{L^\star,\mathrm{P}}(f)$. Consequently, the assumption that there are two minimizers is false. *(ii)* This condition implies that $|\mathcal{R}_{L^\star,\mathrm{P}}(f)| < \infty$, see Proposition 1.7.4, and the assertion follows from *(i)*. □

Theorem 1.7.7 (Existence of SVM)**.** *Let L be a Lipschitz continuous and convex loss function and let $\mathcal{H}$ be the RKHS of a bounded measurable kernel k. Then for all $\lambda > 0$ there exists an SVM solution $f_{L^\star,\mathrm{P},\lambda}$.*

Lemma 2.17 from Steinwart and Christmann (2008b) gives us a result on the continuity of risks, which we will adapt to our needs.

Lemma 1.7.8 (Continuity of risks)**.** *Let L be a Lipschitz continuous loss function. Then the following statements hold:*

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

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

ii) The mapping $\mathcal{R}_{L^\star,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to \mathbb{R}$ *is well-defined and continuous.*

A consequence of this lemma is that the function $f \mapsto \mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f)$ is continuous, since both mappings $f \mapsto \mathcal{R}_{L^\star,\mathrm{P}}(f)$ and $f \mapsto \lambda \|f\|^2_{\mathcal{H}}$ are continuous.

Proof of Lemma 1.7.8. *(i)* Obviously, f is a bounded and measurable function with $\|f\|_\infty \leq B$. Furthermore, the continuity of L shows

$$\begin{aligned} & \lim_{n\to\infty} |L^\star(x,y,f_n(x)) - L^\star(x,y,f(x))| \\ = & \lim_{n\to\infty} |L(x,y,f_n(x)) - L(x,y,f(x))| = 0 \end{aligned}$$

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

$$\begin{aligned} & |L^\star(x,y,f_n(x)) - L^\star(x,y,f(x))| \\ \leq & |L|_1 \, |f_n(x) - f(x)| \\ \leq & |L|_1 (\|f_n\|_\infty + \|f\|_\infty) \\ \leq & 2B|L|_1 < \infty \end{aligned}$$

for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and all $n \geq 1$. Since the constant function $2B|L|_1$ is P-integrable, Lebesgue's theorem of dominated convergence, see Theorem A.2.4, together with

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

gives the assertion.

(ii) We know from Proposition 1.7.4 that $|\mathcal{R}_{L^\star,\mathrm{P}}(f)| < \infty$ for $f \in L_1(\mathrm{P}_X)$ and thus also for all $f \in L_\infty(\mathrm{P}_X)$, i.e., $\mathcal{R}_{L^\star,\mathrm{P}}(f)$ actually maps L_∞ into $\mathbb{R}$. Moreover, the continuity is a direct consequence of *(i)*. □

Proof of Theorem 1.7.7. Since the kernel k of $\mathcal{H}$ is measurable, $\mathcal{H}$ consists of measurable functions by Lemma 1.5.10. Moreover, k is bounded, and thus Lemma 1.5.9 shows that id : $\mathcal{H} \to L_\infty(\mathrm{P}_X)$ is continuous. In addition we have $L(x,y,t) \in [0,\infty)$, and hence $-\infty < L^\star(x,y,t) < \infty$ for all $(x,y,t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$. Thus $L^\star$ is continuous by the convexity of $L^\star$ and Lemma A.4.1. Therefore, Lemma 1.7.8 shows that $\mathcal{R}_{L^\star,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to \mathbb{R}$ is continuous and hence $\mathcal{R}_{L^\star,\mathrm{P}} : \mathcal{H} \to \mathbb{R}$ is continuous since $\mathcal{H} \subset L_\infty(\mathrm{P}_X)$, see Lemma 1.5.9. In addition, Lemma 1.7.6 provides the convexity of this mapping. These lemmas also yield that $f \mapsto \lambda \|f\|_{\mathcal{H}}^2 + \mathcal{R}_{L^\star,\mathrm{P}}(f)$ is strictly convex and continuous. Proposition A.4.2 shows that if $\mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda \|f\|_{\mathcal{H}}^2$ is convex and continuous and additionally $\mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda \|f\|_{\mathcal{H}}^2 \to \infty$ for $\|f\|_{\mathcal{H}} \to \infty$, then $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(\,\cdot\,)$ will have a minimizer. Therefore we need to show that this limit is infinite. By using (1.34) we obtain

$$\begin{aligned}\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) &\geq -|L|_1 \mathbb{E}_{\mathrm{P}_X}|f| + \lambda \|f\|_{\mathcal{H}}^2 \\ &\geq -|L|_1 \|f\|_\infty + \lambda \|f\|_{\mathcal{H}}^2 \\ &\geq -|L|_1 \|k\|_\infty \|f\|_{\mathcal{H}} + \lambda \|f\|_{\mathcal{H}}^2 \;\; \to \;\infty \;\; \text{for} \;\; \|f\|_{\mathcal{H}} \to \infty\,,\end{aligned}$$

as $|L|_1 \|k\|_\infty \in [0,\infty)$ and $\lambda > 0$. □

Knowing now that a solution $f_{L^\star,\mathrm{P},\lambda}$ of the shifted problem exists and is unique, it is interesting to investigate what relationship it has to the solution $f_{L,\mathrm{P},\lambda}$ of the original problem, given that this solution exists.

The application of the $L^\star$-trick is superfluous if $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, because in this case we obtain

$$\begin{aligned}\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) &= \inf_{f\in\mathcal{H}} \mathbb{E}_{\mathrm{P}}\big(L(X,Y,f(X)) - L(X,Y,0)\big) + \lambda \|f\|_{\mathcal{H}}^2 \\ &= \inf_{f\in\mathcal{H}} \big(\mathbb{E}_{\mathrm{P}} L(X,Y,f(X)) + \lambda \|f\|_{\mathcal{H}}^2\big) - \mathbb{E}_{\mathrm{P}} L(X,Y,0) \\ &= \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{L,\mathrm{P},\lambda}) - \mathcal{R}_{L,\mathrm{P}}(0)\end{aligned}$$

and $\mathcal{R}_{L,\mathrm{P}}(0)$ is finite and independent of f. Hence, $f_{L^\star,\mathrm{P},\lambda} = f_{L,\mathrm{P},\lambda}$, and thus both solutions coincide if $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$. Remark that there is a link between the finiteness of the risk and the moment condition (1.37). If we take a look at (1.36) and plug in $f = 0$, we immediately see that the risk can only be finite if the moment $\mathbb{E}_{\mathrm{P}}|Y|$ is finite. The above calculation thus shows that, when the moment is finite and the application of the trick is thus unnecessary, $f_{L^\star,\mathrm{P},\lambda}$ and $f_{L,\mathrm{P},\lambda}$ are always equal.

The following result gives a useful representation of $f_{L^\star,\mathrm{P},\lambda}$ and shows that the mapping $\mathrm{P} \mapsto f_{L^\star,\mathrm{P},\lambda}$ behaves similar to a Lipschitz continuous function. The subdifferential of $L^\star$ is denoted by $\partial L^\star$, and is defined in

Definition A.4.3. With a slight misuse of notation we will write $\mathbb{E}_{\mathrm{P}} h\Phi$ for $\mathbb{E}_{\mathrm{P}} h(X,Y)\Phi(X)$.

Theorem 1.7.9 (Representer theorem)**.** *Let L be a convex and Lipschitz continuous loss function, k be a bounded and measurable kernel with separable RKHS $\mathcal{H}$. Then, for all $\lambda > 0$, there exists an $h \in \mathcal{L}_\infty(\mathrm{P})$ such that*

$$\begin{aligned} h(x,y) &\in \partial L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x)) \qquad \forall\,(x,y)\,, & (1.46)\\ f_{L^\star,\mathrm{P},\lambda} &= -(2\lambda)^{-1}\mathbb{E}_{\mathrm{P}}(h\Phi)\,, & (1.47)\\ \|h\|_\infty &\le |L|_1\,, & (1.48)\\ \left\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\right\|_{\mathcal{H}} &\le \lambda^{-1}\left\|\mathbb{E}_{\mathrm{P}}(h\Phi) - \mathbb{E}_{\bar{\mathrm{P}}}(h\Phi)\right\|_{\mathcal{H}}\,, & (1.49)\end{aligned}$$

for all distributions $\bar{\mathrm{P}}$ on $\mathcal{X}\times\mathcal{Y}$. If L is additionally distance-based, we obtain for (1.46) that

$$h(x,y) \in -\partial\psi(y - f_{L^\star,\mathrm{P},\lambda}(x)) \qquad \forall\,(x,y)\,. \tag{1.50}$$

Proof of Theorem 1.7.9. The existence and uniqueness of $f_{L^\star,\mathrm{P},\lambda}$ follow from the Theorems 1.7.5 and 1.7.7. As k is bounded, Proposition 1.7.3*(iv)* is applicable and (1.42) and (1.43) yield $\|f_{L^\star,\mathrm{P},\lambda}\|_\infty \le \lambda^{-1}|L|_1\,\|k\|_\infty^2 < \infty$ and $|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda})| \le \lambda^{-1}|L|_1^2\,\|k\|_\infty^2 < \infty$. Further, the shifted loss function $L^\star$ is continuous because L and hence $L^\star$ are Lipschitz continuous. Moreover, $R : L_1(\mathrm{P}) \to \mathbb{R}$ defined by

$$R(f) := \int_{\mathcal{X}\times\mathcal{Y}} L^\star\big(x,y,f(x,y)\big)\,d\mathrm{P}(x,y)\,, \quad f \in L_1(\mathrm{P})\,,$$

is well-defined and continuous. The first property follows by the definition of $L^\star$ and its Lipschitz continuity, because

$$|R(f)| \le |L|_1 \int_{\mathcal{X}\times\mathcal{Y}} |f(x,y)|\,d\mathrm{P}(x,y) < \infty\,, \quad f \in L_1(\mathrm{P})\,, \tag{1.51}$$

and hence R is well-defined. The continuity of R can be shown as follows. Fix $\delta > 0$ and let $f_1, f_2 \in L_1(\mathrm{P})$ with $\|f_1 - f_2\|_{L_1(\mathrm{P})} < \delta$. The Lipschitz continuity of $L^\star$ yields

$$\begin{aligned} |R(f_1) - R(f_2)| &\le \int_{\mathcal{X}\times\mathcal{Y}} \big|L^\star(x,y,f_1(x,y)) - L^\star(x,y,f_2(x,y))\big|\,d\mathrm{P}(x,y)\\ &\le |L|_1 \int_{\mathcal{X}\times\mathcal{Y}} |f_1(x,y) - f_2(x,y)|\,d\mathrm{P}(x,y) < \delta\,|L|_1\,,\end{aligned}$$

which shows the continuity of R. We can now apply Proposition A.4.6 with $p = 1$ because (1.51) guarantees that $R(f)$ exists and is finite for all $f \in L_1(\mathrm{P})$. The subdifferential of R can thus be computed by[9]

$$\partial R(f) = \bigl\{h \in L_\infty(\mathrm{P}) : h(x,y) \in \partial L^\star(x,y,f(x,y)) \\ \text{for P-almost all } (x,y)\bigr\}\,.$$

Now, we infer from Lemma 1.5.9 that the inclusion map $I : \mathcal{H} \to L_1(\mathrm{P})$ defined by $(If)(x,y) := f(x)$, $f \in \mathcal{H}$, $(x,y) \in \mathcal{X} \times \mathcal{Y}$, is a bounded linear operator. Moreover, for $h \in L_\infty(\mathrm{P})$ and $f \in \mathcal{H}$, the reproducing property yields

$$\begin{aligned} \langle h, If\rangle_{L_\infty(\mathrm{P}),L_1(\mathrm{P})} &= \mathbb{E}_\mathrm{P} hIf = \mathbb{E}_\mathrm{P} h\langle f,\Phi\rangle_\mathcal{H} \\ &= \langle f, \mathbb{E}_\mathrm{P} h\Phi\rangle_\mathcal{H} = \langle \iota\mathbb{E}_\mathrm{P} h\Phi, f\rangle_{\mathcal{H}',\mathcal{H}}\,, \end{aligned}$$

where $\iota : \mathcal{H} \to \mathcal{H}'$ is the Fréchet-Riesz isomorphism described in Theorem A.3.8. Consequently, the adjoint operator I' of I is given by $I'h = \iota\mathbb{E}_\mathrm{P} h\Phi$, $h \in L_\infty(\mathrm{P})$. Moreover, the $L^\star$-risk functional $\mathcal{R}_{L^\star,\mathrm{P}} : \mathcal{H} \to \mathbb{R}$ restricted to $\mathcal{H}$ satisfies $\mathcal{R}_{L^\star,\mathrm{P}} = R \circ I$, and hence the chain rule for subdifferentials (see Proposition A.4.5) yields $\partial\mathcal{R}_{L^\star,\mathrm{P}}(f) = \partial(R\circ I)(f) = I'\partial R(If)$ for all $f \in \mathcal{H}$. Applying the formula for $\partial R(f)$ thus yields

$$\partial\mathcal{R}_{L^\star,\mathrm{P}}(f) = \bigl\{\iota\mathbb{E}_\mathrm{P} h\Phi : h \in L_\infty(\mathrm{P}) \text{ with} \\ h(x,y) \in \partial L^\star(x,y,f(x)) \text{ P-a.s.}\bigr\}$$

for all $f \in \mathcal{H}$. In addition, $f \mapsto \|f\|_\mathcal{H}^2$ is Fréchet-differentiable and its derivative at f is $2\iota f$ for all $f \in \mathcal{H}$. By picking suitable representations of $h \in L_\infty(\mathrm{P})$, Proposition A.4.5 thus gives

$$\partial\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f) = 2\lambda\iota f + \bigl\{\iota\mathbb{E}_\mathrm{P} h\Phi : h \in \mathcal{L}_\infty(\mathrm{P}) \text{ with} \\ h(x,y) \in \partial L^\star(x,y,f(x))\ \forall\,(x,y)\bigr\}$$

for all $f \in \mathcal{H}$. Now recall that $\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(\,\cdot\,)$ has a minimum at $f_{L^\star,\mathrm{P},\lambda}$, and therefore we have $0 \in \partial\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda})$ by another application of Proposition A.4.5. This together with the injectivity of ι yields the assertions (1.46) and (1.47).

Let us now show that (1.48) holds. Since k is a bounded kernel, we have by (1.42) and (1.43) that

$$\|f_{L^\star,\mathrm{P},\lambda}\|_\infty \le \lambda^{-1}|L|_1\,\|k\|_\infty^2 := B_\lambda < \infty\,.$$

[9] We have $h \in L_\infty(\mathrm{P})$ since there exists an isometric isomorphism between $(L_1(\mathrm{P}))'$ and $L_\infty(\mathrm{P})$, see Theorem A.3.6.

Now (1.46) and Proposition A.4.4 with $\delta := 1$ yield, for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$|h(x,y)| \le \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \bigl|\partial L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x))\bigr| \le |L|_1$$

and hence we have shown $h \in \mathcal{L}_\infty(\mathrm{P})$ and (1.48).

Let us now establish (1.49). To this end, observe that we have by (1.46) and the definition of the subdifferential

$$\begin{aligned} & h(x,y)\bigl(f_{L^\star,\bar{\mathrm{P}},\lambda}(x) - f_{L^\star,\mathrm{P},\lambda}(x)\bigr) \\ \le\ & L^\star\bigl(x,y,f_{L^\star,\bar{\mathrm{P}},\lambda}(x)\bigr) - L^\star\bigl(x,y,f_{L^\star,\mathrm{P},\lambda}(x)\bigr) \end{aligned}$$

for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. By integrating with respect to $\bar{\mathrm{P}}$, we hence obtain

$$\begin{aligned} & \langle f_{L^\star,\bar{\mathrm{P}},\lambda} - f_{L^\star,\mathrm{P},\lambda}\,,\, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi\rangle_{\mathcal{H}} \\ \le\ & \mathcal{R}_{L^\star,\bar{\mathrm{P}}}(f_{L^\star,\bar{\mathrm{P}},\lambda}) - \mathcal{R}_{L^\star,\bar{\mathrm{P}}}(f_{L^\star,\mathrm{P},\lambda})\,. \end{aligned} \tag{1.52}$$

Moreover, an easy calculation shows

$$\begin{aligned} & 2\lambda\,\langle f_{L^\star,\bar{\mathrm{P}},\lambda} - f_{L^\star,\mathrm{P},\lambda}\,,\, f_{L^\star,\mathrm{P},\lambda}\rangle_{\mathcal{H}} \\ & + \lambda\,\bigl\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\bigr\|_{\mathcal{H}}^2 \\ =\ & \lambda\,\bigl\|f_{L^\star,\bar{\mathrm{P}},\lambda}\bigr\|_{\mathcal{H}}^2 - \lambda\,\|f_{L^\star,\mathrm{P},\lambda}\|_{\mathcal{H}}^2\,. \end{aligned} \tag{1.53}$$

By combining (1.52) and (1.53), we thus find

$$\begin{aligned} & \bigl\langle f_{L^\star,\bar{\mathrm{P}},\lambda} - f_{L^\star,\mathrm{P},\lambda}\,,\, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi + 2\lambda f_{L^\star,\mathrm{P},\lambda}\bigr\rangle_{\mathcal{H}} \\ & + \lambda\,\bigl\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\bigr\|_{\mathcal{H}}^2 \\ \le\ & \mathcal{R}^{reg}_{L^\star,\bar{\mathrm{P}},\lambda}(f_{L^\star,\bar{\mathrm{P}},\lambda}) - \mathcal{R}^{reg}_{L^\star,\bar{\mathrm{P}},\lambda}(f_{L^\star,\mathrm{P},\lambda}) \le 0\,, \end{aligned}$$

and consequently the representation $f_{L^\star,\mathrm{P},\lambda} = -\frac{1}{2\lambda}\,\mathbb{E}_{\mathrm{P}} h\Phi$ yields in combination with the Cauchy-Schwarz inequality that

$$\begin{aligned} & \lambda\,\bigl\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\bigr\|_{\mathcal{H}}^2 \\ \le\ & \bigl\langle f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\,,\, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi - \mathbb{E}_{\mathrm{P}} h\Phi\bigr\rangle_{\mathcal{H}} \\ \le\ & \bigl\|f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,\bar{\mathrm{P}},\lambda}\bigr\|_{\mathcal{H}} \cdot \|\mathbb{E}_{\bar{\mathrm{P}}} h\Phi - \mathbb{E}_{\mathrm{P}} h\Phi\|_{\mathcal{H}}\,. \end{aligned}$$

From this we easily obtain (1.49).

It remains to show (1.50) for the special case of a distance-based loss function. By the definition of the subdifferential we obtain for L and $L^\star$

that, for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned}
& \partial L^\star(x,y,t) \\
= \; & \big\{t' \in \mathbb{R}' : \langle t', v-t\rangle \le L^\star(x,y,v) - L^\star(x,y,t) \;\forall\, v \in \mathbb{R}\big\} \\
= \; & \big\{t' \in \mathbb{R}' : \langle t', v-t\rangle \le L(x,y,v) - L(x,y,t) \;\forall\, v \in \mathbb{R}\big\} \\
= \; & \partial L(x,y,t)\,, \qquad\qquad t \in \mathbb{R}\,.
\end{aligned}$$

Hence $\partial L(f) = \partial L^\star(f)$ for all measurable functions $f : \mathcal{X} \to \mathbb{R}$. If we combine this with Proposition A.4.5, it follows, for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, that $\partial L^\star(x,y,t) = \partial L(x,y,t) = -\partial\psi(y-t)$ for all $t \in \mathbb{R}$, and therefore (1.46) implies (1.50). □

If we take L and $\mathcal{H}$ as in Theorem 1.7.9, then the conditions of Theorem 1.6.3 are also fulfilled. If additionally L is Fréchet-differentiable, and if we take $\mathrm{P} = \mathrm{D}$ the empirical distribution of the data set D as in Theorem 1.6.3, we obtain from (1.46) that $h(x,y) = \nabla_3^F L^\star(x,y,f_{L^\star,\mathrm{D},\lambda}(x))$ and from (1.47) that

$$f_{L^\star,\mathrm{D},\lambda}(x) = \sum_{i=1}^n \alpha_i k(x,x_i)\,, \qquad x \in \mathcal{X}\,,$$

where a small calculation yields that the coefficients are equal to

$$\alpha_i = -\frac{1}{2\lambda n}\nabla_3^F L^\star(x_i,y_i,f_{L^\star,\mathrm{D},\lambda}(x_i))\,.$$

At first this might seem strange, but the values $f_{L^\star,\mathrm{D},\lambda}(x_i)$ are already known from the empirical risk minimization step and can therefore be used to determine the value of $f_{L^\star,\mathrm{D},\lambda}$ in all other points $x \in \mathcal{X}$. In this case we thus obtain an explicit formula for the empirical decision function $f_{L^\star,\mathrm{D},\lambda}$ as a linear combination of the kernel functions $k(\,\cdot\,,x_i) = \Phi(x_i)$, for $i = 1,\dots,n$.

Chapter 2

Consistency of Support Vector Machines

In this chapter we will focus our attention on the consistency of the SVM solution $f_{L^\star,\mathrm{P},\lambda}$ of the shifted problem. As explained in Section 1.7 this is in particular useful for SVMs based on heavy-tailed distributions. We will start by giving an introduction on consistency and then state our consistency results for SVMs based on heavy-tailed distributions. Consistency results for standard SVMs can be found in, e.g., Christmann and Steinwart (2007, 2008), and Steinwart and Christmann (2008b). Due to the presence of the moment condition $\mathbb{E}_\mathrm{P}|Y| < \infty$ in these results, they exclude distributions with heavy tails and extreme value distributions. Our aim is to extend the consistency results to also include these distributions.

2.1 Consistency

As mentioned in the introduction, the aim of a support vector machine in particular, or a statistical learning method in general, is to find a decision function f_D, based on the set D of training data, such that the L-risk $\mathcal{R}_{L,\mathrm{P}}(f_D)$ is as close as possible to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$. Since this set D consists of realizations of i.i.d. random variables from an unknown distribution P, also f_D and $\mathcal{R}_{L,\mathrm{P}}(f_D)$ will be random variables. To verify if the learning method $D \mapsto f_D$ is really capable of learning, we should investigate what the probability is that the empirical risk $\mathcal{R}_{L,\mathrm{P}}(f_D)$ will be close to minimal risk $\mathcal{R}^*_{L,\mathrm{P}}$ or whether the difference between both will tend to zero when the size of the data set increases. Possible answers to this question can be given by two notions, namely *consistency* and *learning rates*. Consistency is of an asymptotical nature and can often be verified without any assumptions on the distribution P, whereas learning rates are more related to practical needs, but will almost always require some assumption on the unknown distribution P. In this work, we will only discuss the consistency of SVMs.

Concerning the rate of convergence of support vector machines, we refer to Steinwart and Christmann (2009a) and the references cited therein. In this paper, the authors consider rates of convergence not only for i.i.d. random variables (X_i, Y_i) but also allow for some kind of weak dependence, so-called alpha-mixing. We would also like to remark that, due to the no-free-lunch theorem (Devroye, 1982, Devroye *et al.*, 1996, Theorem 7.2), there exists no uniform rate of convergence for all distributions, a fact is holds in general and not only for SVMs. It is however possible to obtain uniform rates of convergence within special classes of distributions.

Since consistency is a way to describe the "learning ability" of a statistical method, we will first define the concept of a learning method more formally.

Definition 2.1.1. *Let $\mathcal{X}$ be a set and $\mathcal{Y} \subset \mathbb{R}$. A* ***learning method*** *$\mathcal{L}$ on $\mathcal{X} \times \mathcal{Y}$ maps every data set $D \in (\mathcal{X} \times \mathcal{Y})^n$, $n \geq 1$, to a function $f_D : \mathcal{X} \to \mathbb{R}$. If additionally, $\mathcal{X} \neq \emptyset$ is equipped with some σ-algebra and $\mathcal{Y} \neq \emptyset$ is closed and equipped with the Borel-σ-algebra, then the learning method $\mathcal{L}$ is said to be* ***measurable*** *if for all $n \leq 1$ the map*

$$(\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \to \mathbb{R} : (D, x) \mapsto f_D(x)$$

is measurable with respect to the universal completion of the product σ-algebra on $(\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X}$, and where f_D is the decision function obtained from $\mathcal{L}$.

Recall that the P-completion $\mathcal{A}_{\mathrm{P}}$ of a σ-algebra $\mathcal{A}$ is the smallest σ-algebra that contains both $\mathcal{A}$ and all subsets of P-zero sets in $\mathcal{A}$, and that the universal completion is defined as the intersection of all such completions $\mathcal{A}_{\mathrm{P}}$ over all P in the set of probability measures on $\mathcal{A}$. We will from here on assume that $(\mathcal{X} \times \mathcal{Y})^n$ will be equipped with the universal completion of the product σ-algebra on $(\mathcal{X} \times \mathcal{Y})^n$.

It was shown, see, e.g., Steinwart and Christmann (2008b, Lemma 6.17 and Lemma 6.23), that both ERM and SVMs are measurable learning methods under some minimal assumptions. Since for measurable learning methods the maps $x \mapsto f_D(x)$ are measurable, the risks $\mathcal{R}_{L,\mathrm{P}}(f_D)$ will exist for all fixed $D \in (\mathcal{X} \times \mathcal{Y})^n$ and all $n \geq 1$. This implies, Steinwart and Christmann (2008b, Lemma 6.3), that also the maps $(\mathcal{X} \times \mathcal{Y})^n \to [0, \infty] : D \mapsto \mathcal{R}_{L,\mathrm{P}}(f_D)$ are measurable.

Knowing all this, we can now introduce the concept of consistency.

Definition 2.1.2. *Let* $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ *be a loss,* P *be a distribution on* $\mathcal{X} \times \mathcal{Y}$*, and* $\mathcal{L}$ *be a measurable learning method on* $\mathcal{X} \times \mathcal{Y}$*. Then* $\mathcal{L}$ *is said to be* ***L-risk consistent*** *for* P *if, for all* $\varepsilon > 0$*, we have that*

$$\lim_{n\to\infty} \mathrm{P}^n\Big(D \in (\mathcal{X} \times \mathcal{Y})^n : \mathcal{R}_{L,\mathrm{P}}(f_D) \leq \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon\Big) = 1\,.$$

Furthermore, $\mathcal{L}$ *is called* ***universally L-risk consistent*** *if it is L-risk consistent for all distributions* P *on* $\mathcal{X} \times \mathcal{Y}$*.*

This means that, when the data set becomes sufficiently large, an L-risk consistent method will deliver a decision function f_D whose associated risk will be close to the Bayes risk. Or, that with high probability, f_D will be nearly optimal. The method will thus be able to learn. If the method is even universally L-risk consistent, this learning can be done without any specific knowledge of the underlying distribution P, and is thus a prerequisite for non-parametric methods. The only drawback is that we do not know at what speed the convergence takes place, thus we do not know at what rate the method is able to learn. It just tells us that, in the long run, the method will learn the optimal decision function. Furthermore, we will call a learning method *consistent* if the decision function f_D converges to the Bayes function f^*. For a more detailed treatise of the consistency of several learning methods, we refer to Devroye *et al.* (1996), Koenker (1986), Tewari and Bartlett (2005), Zhang (2004), and Stone (1977). The latter being the article wherein it was shown for the first time that a learning method, namely nearest neighbors, was consistent. For more information about the consistency of SVMs in particular, good references are Christmann and Steinwart (2007), Christmann and Steinwart (2008), Steinwart (2001), Steinwart (2002), and Steinwart (2005).

2.2 Consistency of SVMs Based on Shifted Loss Functions

2.2.1 Some Consistency Results

The first result shows that, when the size n of the training data set $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ goes to infinity, the $L^\star$-risk of the empirical SVM $f_{L^\star,\mathrm{D},\lambda_n}$ stochastically converges to the smallest possible risk, i.e., to the Bayes risk. This is somewhat astonishing at first glance because $f_{L^\star,\mathrm{D},\lambda_n}$ is evaluated by minimizing a *regularized empirical risk* over the RKHS $\mathcal{H}$, whereas the Bayes risk is defined as the minimal *non-regularized* risk over the broader set of *all* measurable functions $f : \mathcal{X} \to \mathbb{R}$. To make this universal consistency achievable, we need the denseness assumption on $\mathcal{H}$ in the following theorem.

Theorem 2.2.1 (Risk consistency)**.** *Let L be a convex, Lipschitz continuous loss function, $L^\star$ its shifted version, and $\mathcal{H}$ be a separable RKHS of a bounded measurable kernel k such that $\mathcal{H}$ is dense in $L_1(\mu)$ for all distributions μ on $\mathcal{X}$. Let (λ_n) be a sequence of strictly positive numbers with $\lambda_n \to 0$.*

i) If $\lambda_n^2 n \to \infty$, then, for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$,

$$\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L^\star,\mathrm{P}}\,, \quad n \to \infty\,, \tag{2.1}$$

in probability P^∞ *for all* $|D| = n$.

ii) If $\frac{\lambda_n^2 n}{\ln(n)} \to \infty$, then the convergence in (2.1) holds even P^∞*-almost surely.*

Proof of Theorem 2.2.1. *(i)* To avoid handling too many constants, let us assume $\|k\|_\infty = 1$. This implies $\|f\|_\infty \le \|k\|_\infty \|f\|_\mathcal{H} \le \|f\|_\mathcal{H}$ for all $f \in \mathcal{H}$. Now we use the Lipschitz continuity of L (and thus also of $L^\star$), $|L|_1 < \infty$, and Lemma 1.4.4 to obtain, for all $g \in \mathcal{H}$,

$$\left|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda_n}) - \mathcal{R}_{L^\star,\mathrm{P}}(g)\right| \le |L|_1 \; \|f_{L^\star,\mathrm{P},\lambda_n} - g\|_\mathcal{H}\,. \tag{2.2}$$

For $n \in \mathbb{N}$ and $\lambda_n > 0$, we write $h_n := h_{L^\star,n} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for the function h obtained by the representer theorem 1.7.9. Let $\Phi : \mathcal{X} \to \mathcal{H}$ be the canonical feature map. We have $f_{L^\star,\mathrm{P},\lambda_n} = -(2\lambda_n)^{-1}\mathbb{E}_\mathrm{P} h_n \Phi$, and for all distributions Q on $\mathcal{X} \times \mathcal{Y}$, we have

$$\|f_{L^\star,\mathrm{P},\lambda_n} - f_{L^\star,\mathrm{Q},\lambda_n}\|_\mathcal{H} \le \lambda_n^{-1} \; \|\mathbb{E}_\mathrm{P} h_n \Phi - \mathbb{E}_\mathrm{Q} h_n \Phi\|_\mathcal{H}\,.$$

Note that $\|h_n\|_\infty \le |L|_1$ due to (1.48). Moreover, let $\varepsilon \in (0,1)$ and D be a training set of n data points and corresponding empirical distribution D such that

$$\|\mathbb{E}_\mathrm{P} h_n \Phi - \mathbb{E}_\mathrm{D} h_n \Phi\|_\mathcal{H} \le \lambda_n \varepsilon\,. \tag{2.3}$$

Then Theorem 1.7.9 gives $\|f_{L^\star,\mathrm{P},\lambda_n} - f_{L^\star,\mathrm{D}_n,\lambda_n}\|_{\mathcal{H}} \leq \varepsilon$ and hence (2.2) yields

$$\begin{aligned}&\big|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda_n}) - \mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D},\lambda_n})\big| \\ &\leq |L|_1 \cdot \|f_{L^\star,\mathrm{P},\lambda_n} - f_{L^\star,\mathrm{D},\lambda_n}\|_{\mathcal{H}} \ \leq\ |L|_1\varepsilon\,.\end{aligned} \tag{2.4}$$

Let us now estimate the probability of D satisfying (2.3). To this end, we first observe that $\lambda_n n^{1/2} \to \infty$ implies that $\lambda_n\varepsilon \geq n^{-1/2}$ for all sufficiently large $n \in \mathbb{N}$. Moreover, Theorem 1.7.9 shows $\|h_n\|_\infty \leq |L|_1$, and our assumption $\|k\|_\infty = 1$ thus yields $\|h_n\Phi\|_\infty \leq |L|_1$. Consequently, Hoeffding's inequality in Hilbert spaces (see Theorem A.3.9) yields for $\xi_i = h_n(X_i,Y_i)\Phi(X_i)$, $B = |L|_1$ and

$$\tau_n = \frac{3}{8}\,\frac{|L|_1^{-2}\varepsilon^2\lambda_n^2 n}{|L|_1^{-1}\varepsilon\lambda_n + 3} = \frac{3}{8}\frac{a_n^2 n}{a_n+3}$$

with $a_n := |L|_1^{-1}\lambda_n\varepsilon$, the bound

$$\begin{aligned}&\mathrm{P}^n\Big(D\in(\mathcal{X}\times\mathcal{Y})^n : \|\mathbb{E}_\mathrm{P} h_n\Phi - \mathbb{E}_\mathrm{D} h_n\Phi\|_{\mathcal{H}} \leq \lambda_n\varepsilon\Big) && (2.5)\\ \geq\ &\mathrm{P}^n\Big(D\in(\mathcal{X}\times\mathcal{Y})^n : \|\mathbb{E}_\mathrm{P} h_n\Phi - \mathbb{E}_\mathrm{D} h_n\Phi\|_{\mathcal{H}} \\ &\qquad\qquad \leq \frac{|L|_1(\sqrt{2\tau_n}+1)}{\sqrt{n}} + \frac{4|L|_1\tau_n}{3n}\Big) && (2.6)\\ \geq\ & 1-\exp\Big(-\frac{3}{8}\cdot\frac{\varepsilon^2\lambda_n^2 n/|L|_1^2}{\varepsilon\lambda_n/|L|_1+3}\Big)\\ =\ & 1-\exp\Big(-\frac{3}{8}\cdot\frac{\varepsilon^2\lambda_n^2 n}{(\varepsilon\lambda_n+3|L|_1)|L|_1}\Big)\end{aligned}$$

for all sufficiently large values of n. In order to go from (2.5) to (2.6), we used the following inequality:

$$\begin{aligned}\frac{\sqrt{2\tau_n}+1}{\sqrt{n}} + \frac{4\tau_n}{3n} &= \frac{a_n}{2}\frac{\sqrt{3}}{\sqrt{a_n+3}} + \frac{1}{\sqrt{n}} + \frac{a_n}{2}\frac{a_n}{a_n+3}\\ &< \frac{a_n}{2} + \frac{1}{\sqrt{n}} + \frac{a_n}{2}\frac{1}{3} < a_n = |L|_1^{-1}\lambda_n\varepsilon\,.\end{aligned}$$

Now using $\lambda > 0$, $\lambda_n \to 0$ and $\lambda_n n^{1/2} \to \infty$, we find that the probability of sample sets D satisfying (2.3) converges to 1 if $|D| = n \to \infty$. As we have seen above, this implies that (2.4) holds true with probability tending to 1. Now, since $\lambda_n > 0$ and $\lambda_n \to 0$, $n \to \infty$, we additionally have $|\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda_n}) - \mathcal{R}^*_{L^\star,\mathrm{P}}| \le \varepsilon$ for all sufficiently large n, and hence we obtain the assertion of $L^\star$-risk consistency of $f_{L^\star,\mathrm{P},\lambda}$.

(ii) First observe that $\frac{\lambda_n^2 n}{\ln(n)} \to \infty$ allows us to rewrite λ_n as

$$\lambda_n = \zeta h(n)\sqrt{\frac{\ln(n)}{n}}\,,$$

with $h(n) : (0,\infty) \to (0,\infty)$ a positive function such that $h(n) \to \infty$ and $\lambda_n \to 0$ for $n \to \infty$ and with $\zeta = \sqrt{\frac{16(|L|_1+3|L|_1^2)}{3}}$. This gives $\frac{\lambda_n^2 n}{\ln(n)} = \big(\zeta h(n)\big)^2$. In order to show the second assertion, we define $\varepsilon_n := 1/h(n)$ and

$$\delta_n := \mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda_n}) - \mathcal{R}^*_{L^\star,\mathrm{P}} + \varepsilon_n\,, \quad n \in \mathbb{N}\,.$$

A small calculation yields that $\varepsilon_n\lambda_n = \zeta\sqrt{\frac{\ln(n)}{n}} \downarrow 0$ for $n \to \infty$ and that $\varepsilon_n^2\lambda_n^2 n = \zeta^2\ln(n)$. Therefore

$$\frac{\varepsilon_n^2\lambda_n^2 n}{(\varepsilon_n\lambda_n + 3|L|_1)|L|_1} = \frac{\zeta^2\ln(n)}{(\varepsilon_n\lambda_n + 3|L|_1)|L|_1} \ge \frac{\zeta^2\ln(n)}{|L|_1 + 3|L|_1^2}$$

for n sufficiently large.

Moreover, for an infinite sample

$$D_\infty := ((x_1,y_1),(x_2,y_2),\dots) \in (\mathcal{X}\times\mathcal{Y})^\infty\,,$$

we write $D_n := ((x_1,y_1),\dots,(x_n,y_n))$. With these notations, we define, for $n \in \mathbb{N}$,

$$A_n := \big\{D_\infty \in (\mathcal{X}\times\mathcal{Y})^\infty : \mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D}_n,\lambda_n}) - \mathcal{R}^*_{L^\star,\mathrm{P}} > \delta_n\big\}\,.$$

Now, our estimates above together with $\frac{\lambda_n^2 n}{\ln(n)} \to \infty$ yield, for n large enough,

$$\begin{aligned}\mathrm{P}(A_n) &\le \exp\left(-\frac{3}{8}\cdot\frac{\varepsilon_n^2\lambda_n^2 n}{(\varepsilon_n\lambda_n + 3|L|_1)|L|_1}\right)\\ &\le \exp\left(-\frac{3}{8}\cdot\frac{\zeta^2\ln(n)}{|L|_1+3|L|_1^2}\right)\\ &= \exp\big(-2\ln(n)\big) = n^{-2}\,,\end{aligned}$$

from which the convergence of the series $\sum_{n\in\mathbb{N}}\mathrm{P}^\infty(A_n)$ follows. We obtain by the Borel-Cantelli lemma A.2.8 that

$$\begin{aligned}\mathrm{P}^\infty\big(\,\{\,D_\infty \in (\mathcal{X}\times\mathcal{Y})^\infty \,:\, &\exists\, n_0\ \forall\, n \ge n_0 \text{ with}\\ &\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D}_n,\lambda_n}) - \mathcal{R}^*_{L^\star,\mathrm{P}} \le \delta_n\}\,\big) = 1\,.\end{aligned}$$

The assertion follows because $\lambda_n \to 0$ implies $\delta_n \to 0$. □

We see from the theorem that, in order to obtain consistency, the regularization parameters can be chosen in a *data-independent* way, as long as they constitute an appropriate null sequence. The regularization parameter λ can also be determined in a *data-dependent* manner for SVMs, e.g., by using a *training-validation support vector machine* (TV-SVM), see Steinwart and Christmann (2008b, Chapter 6.5) and p. 100 for a short description. It has been shown that the training-validation method gives measurable SVMs and even oracle inequalities for TV-SVMs are available, so consistency can be obtained for this method.

Most often the goal in practice is to minimize the risk, and thus in those cases, the knowledge of risk consistency will be enough. However, sometimes practitioners are also interested in explicitly knowing the goodness of the function $f_{L,\mathrm{D},\lambda}$ and the predictions $f_{L,\mathrm{D},\lambda}(x)$ made by it for previously unseen input values $x \in \mathcal{X}$.

In general, it is unclear whether the convergence of the risks in (2.1) implies the convergence of $f_{L^\star,\mathrm{D},\lambda_n}$ to a minimizer $f^*_{L^\star,\mathrm{P}}$ of the Bayes risk $\mathcal{R}^*_{L^\star,\mathrm{P}}$. However, Theorem 2.2.2 will show such a convergence for the important special case of non-parametric quantile regression. Estimation of conditional quantiles instead of estimation of conditional means is especially interesting for heavy-tailed distributions that often have no finite moments. It is known that the pinball loss function defined in (1.26) can be used to estimate the conditional τ-quantiles, $\tau \in (0,1)$,

$$\begin{aligned} f^*_{\tau,\mathrm{P}}(x) := \big\{t^* \in \mathbb{R} :\ & \mathrm{P}\big((-\infty,t^*]\,|x\big) \geq \tau \text{ and} \\ & \mathrm{P}\big([t^*,\infty)\,|x\big) \geq 1-\tau\big\}\,, \end{aligned}$$

$x \in \mathcal{X}$, see Koenker (2005) and Takeuchi *et al.* (2006). For some recent results on support vector machines based on this loss function we refer to Christmann and Steinwart (2008) and Steinwart and Christmann (2008a). Remember that the pinball loss function is convex and Lipschitz continuous, but asymmetric for $\tau \neq \frac{1}{2}$. Before we formulate the next result, we define

$$d_0(f,g) := \mathbb{E}_{\mathrm{P}_X} \min\{1, |f-g|\}\,,$$

where $f, g : \mathcal{X} \to \mathbb{R}$ are arbitrary measurable functions. It is known that d_0 is a translation-invariant metric describing the convergence in probability.

Theorem 2.2.2 (Consistency)**.** *For $\tau \in (0,1)$, let L be the τ-pinball loss and $L^\star$ its shifted version. Moreover, let* P *be a distribution on $\mathcal{X} \times \mathbb{R}$ whose* *conditional τ-quantile $f^*_{\tau,\mathrm{P}} : \mathcal{X} \to \mathbb{R}$ is P_X-almost surely unique. Under the assumptions of Theorem 2.2.1, we then have*

$$d_0(f_{L^\star,\mathrm{D},\lambda_n}, f^*_{\tau,\mathrm{P}}) \to 0\,, \qquad n \to \infty\,,$$

where the convergence is either in probability P^∞ or P^∞-almost surely, depending on whether assumption (i) or (ii) on the null-sequence (λ_n) is taken from Theorem 2.2.1.

Before we can prove Theorem 2.2.2, we need some prerequisites on self-calibrated loss functions and related results, cfr., Steinwart and Christmann (2008b, Chapter 3). Let $\tau \in (0,1)$, L be a pinball loss, and $L^\star$ its shifted version. Hence, L is Lipschitz continuous, convex, and $L(x,y,t) = \psi(y-t) \to \infty$ for $|t| \to \infty$. Our goal is to extend the consistency results derived in Christmann and Steinwart (2008) to *all* distributions P on $\mathcal{X} \times \mathbb{R}$. To this end, we adopt the *inner risk* notation from Steinwart and Christmann (2008b) by writing, for $t \in \mathbb{R}$,

$$\mathcal{C}_{L^\star,\mathrm{Q}}(t) := \int_{\mathbb{R}} L^\star(x,y,t)\, d\mathrm{Q}(y) = \int_{\mathbb{R}} \psi(y-t) - \psi(y)\, d\mathrm{Q}(y)\,,$$

where Q is a distribution on $\mathbb{R}$ that will serve us as a template for the conditional distribution $\mathrm{P}(\,\cdot\,|x)$. Similarly, we write $\mathcal{C}^*_{L^\star,\mathrm{Q}} := \inf_{t\in\mathbb{R}} \mathcal{C}_{L^\star,\mathrm{Q}}(t)$ for the minimal inner $L^\star$-risk. Note that, like for the $L^\star$-risk, we have $|\mathcal{C}^*_{L^\star,\mathrm{Q}}| < \infty$. Finally, for $\varepsilon \in [0,\infty]$, we denote the set of *$\varepsilon$-approximate minimizers* by

$$\mathcal{M}_{L^\star,\mathrm{Q}}(\varepsilon) := \bigl\{t \in \mathbb{R} : \mathcal{C}_{L^\star,\mathrm{Q}}(t) - \mathcal{C}^*_{L^\star,\mathrm{Q}} < \varepsilon\bigr\}$$

and the set of *exact minimizers* by

$$\mathcal{M}_{L^\star,\mathrm{Q}}(0^+) := \bigcap_{\varepsilon>0} \mathcal{M}_{L^\star,\mathrm{Q}}(\varepsilon) = \bigl\{t \in \mathbb{R} : \mathcal{C}_{L^\star,\mathrm{Q}}(t) = \mathcal{C}^*_{L^\star,\mathrm{Q}}\bigr\}\,.$$

Since $|\mathcal{C}^*_{L^\star,\mathrm{Q}}| < \infty$ it is easy to verify that these notations coincide with those of Steinwart and Christmann (2008b, Chapter 3) modulo the fact that we now consider the shifted loss function $L^\star$ rather than L. The following proposition, which is an $L^\star$-analogue to Steinwart and Christmann (2008b, Proposition 3.9), computes the $L^\star$-excess risk and the set of exact minimizers.

Proposition 2.2.3. *For $\tau \in (0,1)$, let L be the τ-pinball loss and $L^\star$ its shifted version. Moreover, let Q be a distribution on $\mathbb{R}$ and t^* be a τ-quantile of Q, i.e., we have*

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

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

$$\mathcal{C}_{L^\star,\mathrm{Q}}(t^* + t) - \mathcal{C}^*_{L^\star,\mathrm{Q}} = tq_+ + \int_0^t \mathrm{Q}\big((t^*, t^* + s)\big)\, ds\,, \tag{2.7}$$

$$\mathcal{C}_{L^\star,\mathrm{Q}}(t^* - t) - \mathcal{C}^*_{L^\star,\mathrm{Q}} = tq_- + \int_0^t \mathrm{Q}\big((t^* - s, t^*)\big)\, ds\,, \tag{2.8}$$

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

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

Proof of Proposition 2.2.3. Let us consider the distribution $\mathrm{Q}^{(t^*)}$ defined by $\mathrm{Q}^{(t^*)}(A) := \mathrm{Q}(t^* + A)$ for all measurable sets $A \subset \mathbb{R}$. Then it is not hard to see that 0 is a τ-quantile of $\mathrm{Q}^{(t^*)}$. Moreover, we obviously have $\mathcal{C}_{L^\star,\mathrm{Q}}(t^* + t) = \mathcal{C}_{L^\star,\mathrm{Q}^{(t^*)}}(t)$. Therefore, 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 $q_+ \in \mathbb{R}$ satisfying $0 \leq q_+ \leq \mathrm{Q}(\{0\})$ and

$$\mathrm{Q}((-\infty, 0]) = \tau + q_+\,. \tag{2.9}$$

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

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

Moreover, using a well-known relationship between expectations and tail bounds, see Lemma A.2.5, we get

$$
\begin{aligned}
t\,\mathrm{Q}([0,t)) - \int\limits_{0\le y<t} y\,d\mathrm{Q}(y) &= \int\limits_0^t \mathrm{Q}([0,t))\,ds - \int\limits_0^t \mathrm{Q}([s,t))\,ds \\
&= t\,\mathrm{Q}(\{0\}) + \int_0^t \mathrm{Q}((0,s))\,ds\,,
\end{aligned}
$$

and since (2.9) implies

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

we thus obtain

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

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 $0 \le q_- \le \mathrm{Q}(\{0\})$ such that $\mathrm{Q}([0,\infty)) = 1-\tau+q_-$ and

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

for all $t \ge 0$. Consequently, $t^* = 0$ is a minimizer of $\mathcal{C}_{L^\star,\mathrm{Q}}(\,\cdot\,)$ and we have $\mathcal{C}^*_{L^\star,\mathrm{Q}} = \mathcal{C}_{L^\star,\mathrm{Q}}(0) = 0$. From this we conclude both (2.7) and (2.8). Moreover, combining $\mathrm{Q}([0,\infty)) = 1-\tau+q_-$ with (2.9), we find $q_+ + q_- = \mathrm{Q}(\{0\})$. Finally, the formula for the set of exact minimizers is an obvious consequence of (2.7) and (2.8). □

In order to investigate how well approximate $L^\star$-risk minimizers approximate the exact $L^\star$-risk minimizers, we further have to adopt the *self-calibration approach* of Steinwart and Christmann (2008b, Chapter 3). Fortunately, the fact that we always have $|\mathcal{C}^*_{L^\star,\mathrm{Q}}| < \infty$ makes our considerations a little easier than those in Steinwart and Christmann (2008b, Chapter 3) for general loss functions. To further decrease the notational burden we assume in the following that the considered distribution Q on $\mathbb{R}$ has a *unique* τ-quantile, denoted by $t^*_{\tau,\mathrm{Q}}$ or simply t^* if no confusion can arise. Fortunately, this uniqueness assumption is by no means necessary, and we refer the interested reader to Steinwart and Christmann (2008b, Chapter 3) for a modification to this general situation.

With these preparations, the $L^\star$-generalization of the *self-calibration function* now reads as follows:

$$\delta_{\max}(\varepsilon,\mathrm{Q}) := \inf_{|t-t^*|\ge\varepsilon} \mathcal{C}_{L^\star,\mathrm{Q}}(t) - \mathcal{C}^*_{L^\star,\mathrm{Q}}\,, \qquad \varepsilon > 0\,.$$

Note that, for $t \in \mathbb{R}$ and $\varepsilon := |t - t^*|$, we have

$$\delta_{\max}(|t - t^*|, \mathrm{Q}) = \delta_{\max}(\varepsilon, \mathrm{Q}) \le \mathcal{C}_{L^\star,\mathrm{Q}}(t) - \mathcal{C}^*_{L^\star,\mathrm{Q}}\,,$$

i.e., as for standard loss functions $\delta_{\max}(\varepsilon, \mathrm{Q})$ measures how well approximate $\mathcal{C}_{L^\star,\mathrm{Q}}(\,\cdot\,)$-minimizers approximate the exact minimizer t^*. Moreover, by Proposition 2.2.3 we conclude that, for all $\varepsilon > 0$, we have

$$\begin{aligned}\delta_{\max}(\varepsilon, \mathrm{Q}) = \min\Big\{ &\varepsilon q_+ + \int_0^\varepsilon \mathrm{Q}\big((t^*, t^* + s)\big)\, ds\,, \\ &\varepsilon q_- + \int_0^\varepsilon \mathrm{Q}\big((t^* - s, t^*)\big)\, ds \Big\} > 0\,,\end{aligned}$$

where we used the assumption that t^* is the only τ-quantile, i.e., the only exact $\mathcal{C}_{L^\star,\mathrm{Q}}(\,\cdot\,)$-minimizer. Since the proofs of Theorem 3.61 and its Corollary 3.62 in Steinwart and Christmann (2008b) only consider excess inner risks and not the underlying loss function itself, a literal repetition of these proofs then yields the following result.

Corollary 2.2.4. *For $\tau \in (0,1)$, let L be the τ-pinball loss and $L^\star$ its shifted version. Moreover, let P be a distribution on $\mathcal{X} \times \mathbb{R}$ whose conditional τ-quantile $f^*_{\tau,\mathrm{P}} : \mathcal{X} \to \mathbb{R}$ is P_X-almost surely unique. Then, for all sequences (f_n) of measurable functions $f_n : \mathcal{X} \to \mathbb{R}$, the convergence*

$$\mathcal{R}_{L^\star,\mathrm{P}}(f_n) \to \mathcal{R}^*_{L^\star,\mathrm{P}}$$

implies

$$f_n \to f^*_{\tau,\mathrm{P}} \qquad \textit{in probability } \mathrm{P}_X\,.$$

Proof of Theorem 2.2.2. Due to the assumptions, Theorem 2.2.1 is applicable and hence $f_{L^\star,\mathrm{D},\lambda_n}$ satisfies $\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L^\star,\mathrm{P}}$ in probability (or almost surely) for $n \to \infty$. The existence of a unique minimizer $f^*_{\tau,\mathrm{P}}$ is guaranteed by the assumptions of Theorem 2.2.2. Hence, Corollary 2.2.4 yields the assertion. □

Chapter 3

Robustness of Support Vector Machines

3.1 Robustness

As shown in the previous section, a strong argument in favor of SVMs is that they are L-risk consistent under weak assumptions, that is SVMs are able to "learn". It is, however, also important to investigate the robustness properties for such statistical learning methods. In almost all cases statistical models are only approximations to the true random process which generated a given data set. Hence the natural question arises what impact such deviations may have on 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."*

Let us consider $T(\mathrm{P}) := f_{L,\mathrm{P},\lambda}$, with P a probability measure, as a mapping $T : \mathrm{P} \mapsto f_{L,\mathrm{P},\lambda}$. In robust statistics we are interested in smooth and bounded functions T, because this will give stable regularized risks within small neighborhoods of P. If an appropriate derivative $\nabla T(\mathrm{P})$ of $T(\mathrm{P})$ is bounded, then the function $T(\mathrm{P})$ cannot increase or decrease unlimited in small neighborhoods of P. We thus expect the value of $T(\mathrm{Q})$ to be close

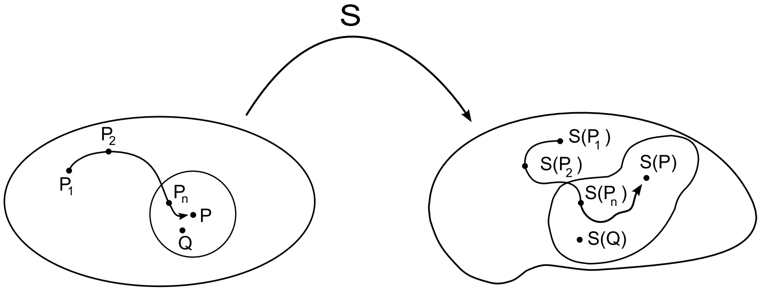

Figure 3.1: Schematic representation of the robustness of a statistic S. On the left, a neighborhood of P in the set $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ is shown. On the right we see that this neighborhood is mapped onto a neighborhood of $S(\mathrm{P})$.

to the value of $T(\mathrm{P})$ for distributions Q in a small neighborhood of P. A schematic representation of the robustness of a statistic $S(\mathrm{P})$ is given in Figure 3.1. Several notions of differentiability have been used for this purpose.[1] From a mathematical point of view, the Fréchet-derivative would probably be the most suitable notion to use for robustness properties, but there exist many interesting statistical methods which don't have a Fréchet-derivative, since Fréchet-differentiability is a rather strict concept. Therefore many approaches use weaker notions on differentiability, such as Hadamard- or Gâteaux-differentiability. It was exactly Gâteaux-differentiability that inspired Hampel to introduce the influence function, which is an even weaker concept since it is defined as a Gâteaux-derivative, but without the assumption of linearity. These other, weaker concepts are especially important if the estimates use non-smooth functions, such as some of the loss functions we saw in Section 1.4.

As said above, one general approach to robustness is the one based on influence functions (Hampel, 1968, 1974) which are closely related to Gâteaux-derivatives. Of course there also exist other notions to verify robustness, such as the breakdown[2] point of an estimator, its maxbias or its sensitivity curve. For more detail on the theory of robustness, we refer to Hampel *et al.* (1986), Huber (1981), and Maronna *et al.* (2006). Robustness results on SVMs can be found in, e.g., Christmann and Steinwart (2007)

[1] For more details on Gâteaux-, Hadamard-, and Fréchet-derivatives, we refer to Appendix A.3.3. Bouligand-derivatives are treated in Subsection 3.2.1.

[2] As far as we know, there exist no published results on the breakdown point of SVMs, but since the breakdown point is a measure of robustness, we conjecture there is a conflict of goals between universal consistency and a positive breakdown point (cfr. p. 97).

and Steinwart and Christmann (2008b).

Definition 3.1.1. *Let $\mathcal{M}_1$ be the set of probability distributions on a measurable space $(Z, \mathcal{B}(Z))$ and let $\mathcal{H}$ be a reproducing kernel Hilbert space. The* ***influence function (IF)*** *of $T : \mathcal{M}_1 \to \mathcal{H}$ at a point $z \in Z$ for a distribution* P *is defined as*

$$\mathrm{IF}(z; T, \mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z) - T(\mathrm{P})}{\varepsilon}\,, \tag{3.1}$$

if the limit exists.

Within this approach robust estimators are those which have a bounded influence function.[3] The influence function is neither supposed to be linear nor continuous. If the influence functions exists for all points $z \in Z$ and if it is continuous and linear, then the IF is a special Gâteaux-derivative.

Christmann and Steinwart (2004, 2007) and Steinwart and Christmann (2008a) showed that SVMs have a bounded influence function in binary classification and in regression problems provided that the kernel is bounded and continuous, L is twice Fréchet-differentiable, and the first and second Fréchet-derivative of L are bounded. Hence Lipschitz continuous loss functions are of special interest from a robustness point of view. An example of a loss function with these properties is the logistic loss for regression given by (1.27). However the important special cases L_ϵ, $L_{\tau-pin}$, and $L_{c-Huber}$ are excluded in these results, because these loss functions are not everywhere (twice) Fréchet-differentiable.

We try to fill this gap by proposing an alternative to the classical influence function in Definition 3.2.3. This alternative is based on Bouligand-derivatives whereas Hampel's influence function was defined having Gâteaux-derivatives in mind. Since Bouligand-derivatives are only supposed to be positive homogeneous instead of linear, our tacit hope what that this notion would allow us to extend the robustness result on SVMs to also those SVMs that are based on non-smooth loss functions. Using then this new notion of robustness, we show in Sections 3.3 and 3.4 that SVMs are robust in this sense even if the loss function has no Fréchet-derivative.

The main objects in this work are the consistency and robustness of SVMs. However, the precision of an estimator $T(\mathrm{D}_n)$ of the statistic $T(\mathrm{P})$, called the *efficiency* of the estimator, is also of importance. This efficiency

[3] In the following we use the term "robust" in this sense, unless otherwise stated.

will often be determined through the *asymptotic variance*. In a parametric context, it is often said that $\sqrt{n}(T(\mathrm{D}_n) - T(\mathrm{P}))$ will be asymptotically normal with mean zero and asymptotic variance

$$V(T, \mathrm{P}) = \int \mathrm{IF}(x; T, \mathrm{P})^2 d\mathrm{P}(x)\,, \tag{3.2}$$

see, e.g., Hampel *et al.* (1986, p. 85) or Huber (1981, p. 39). Davies (1993, p. 1856) remarks however that in general there is no such connection between the influence function and the asymptotic approximation of a statistical method, at least not without further conditions, and he also shows this through a counterexample. As we could also see in Huber (1981, p. 39), the existence of the Fréchet-derivative of T at P is a necessary assumption to obtain asymptotic normality such that the asymptotic variance can be written as (3.2).

For the case of SVMs we see nevertheless two problems with this formulation. First of all, Hable (2011) showed that for a sequence of SVM-estimators

$$(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}\,, \qquad \mathrm{D}_n \mapsto f_{L,\mathrm{D}_n,\lambda_{\mathrm{D}_n}}$$

asymptotic normality holds for the difference between the empirical SVM and the theoretical SVM $f_{L,\mathrm{P},\lambda_0}$ with $\lambda_0 \in (0, \infty)$, that is

$$\sqrt{n}(f_{L,\mathrm{D}_n,\lambda_{\mathrm{D}_n}} - f_{L,\mathrm{P},\lambda_0})$$

converges weakly to a (zero-mean) Gaussian process in the function space $\mathcal{H}$, given that $\sqrt{n}(\lambda_{\mathrm{D}_n} - \lambda_0) \to 0$ in probability. It would of course also be desirable to have asymptotic normality of

$$\sqrt{n}(f_{L,\mathrm{D}_n,\lambda_{\mathrm{D}_n}} - f^*_{L,\mathrm{P}})\,,$$

but in the non-parametric setting where $\mathcal{H}$ is a large infinite-dimensional function space, this is probably not possible because such a result would violate the no-free-lunch theorem, and thus we do not obtain the desired asymptotic normality.

And secondly, even if the asymptotic normality would hold, there are, to our knowledge, no results published so far that the SVM-functional $\mathrm{P} \mapsto f_{L,\mathrm{P},\lambda}$ is Fréchet-differentiable. It is known that, given some assumptions, support vector machines are Bouligand- and Gâteaux-differentiable (see the results on the IF and the BIF), as well as Hadamard-differentiable (Hable, 2011, Theorem 5.8), but the Fréchet-differentiability of the SVM-functional itself remains unknown.

3.2 The Bouligand Influence Function

3.2.1 Bouligand-Derivatives

In this subsection we will recall some facts on Bouligand-derivatives and strong approximation of functions, because these notions will be used to introduce our alternative to the influence function (which we will call the Bouligand influence function, or, in short, BIF) as well as to investigate robustness properties for SVMs for non-smooth loss functions. For the rest of this subsection, let E_1, E_2, W, and Z be normed linear spaces, and let us consider neighborhoods $\mathcal{N}(x_0)$ of x_0 in E_1, $\mathcal{N}(y_0)$ of y_0 in E_2, and $\mathcal{N}(w_0)$ of w_0 in W. Let F and G be functions from $\mathcal{N}(x_0) \times \mathcal{N}(y_0)$ to Z, h_1 and h_2 functions from $\mathcal{N}(w_0)$ to Z, f a function from $\mathcal{N}(x_0)$ to Z and g a function from $\mathcal{N}(y_0)$ to Z. A function f *approximates* F in x at (x_0, y_0), written as $f \; \sim_x \; F$ at (x_0, y_0), if

$$F(x, y_0) - f(x) = o(x - x_0)\,.$$

Similarly, $g \; \sim_y \; F$ at (x_0, y_0) if $F(x_0, y) - g(y) = o(y - y_0)$. This kind of approximation can be seen in the definition of the partial Fréchet-derivative. E.g., F has a partial Fréchet-derivative $\nabla_1^F F$ at (x_0, y_0) is the same as saying that $F(x_0, y_0) + \nabla_1^F F(x_0, y_0)(x - x_0) \; \sim_x \; F$ at (x_0, y_0) (Robinson, 1991).

A function h_1 *strongly approximates* h_2 at w_0, written as $h_1 \; \approx \; h_2$ at w_0, if for each $\varepsilon > 0$ there exists a neighborhood $\mathcal{N}(w_0)$ of w_0 such that whenever w and w' belong to $\mathcal{N}(w_0)$,

$$\left\| \big(h_1(w) - h_2(w)\big) - \big(h_1(w') - h_2(w')\big) \right\| \le \varepsilon \left\| w - w' \right\| \,.$$

Strong approximation amounts to requiring $h_1 - h_2$ to have a strong Fréchet-derivative equal to 0 at w_0, though neither h_1 nor h_2 is assumed to be differentiable in any sense. We define strong approximation for functions of several groups of variables, for example $G \; \approx_{(x,y)} \; F$ at (x_0, y_0), by replacing W by $E_1 \times E_2$ and making the obvious substitutions. A function f *strongly approximates* F in x at (x_0, y_0), written as $f \; \approx_x \; F$ at (x_0, y_0), if for each $\varepsilon > 0$ there exist neighborhoods $\mathcal{N}(x_0)$ of x_0 and $\mathcal{N}(y_0)$ of y_0 such that whenever x and x' belong to $\mathcal{N}(x_0)$ and y belongs to $\mathcal{N}(y_0)$ we have

$$\left\| \big(F(x, y) - f(x)\big) - \big(F(x', y) - f(x')\big) \right\| \le \varepsilon \left\| x - x' \right\| \,.$$

A similar definition is made for strong approximation in y. For example, if $F(x, y)$ is Fréchet-differentiable in x in a neighborhood of (x_0, y_0) and its partial Fréchet-derivative $\nabla_1^F F$ is continuous in both x and y at

(x_0, y_0), then $\nabla_1^F F(x_0, y_0) \approx_x F$ at (x_0, y_0) (Dontchev and Hager, 1994). Note that one has both $f \approx_x F$ and $g \approx_y F$ at (x_0, y_0) exactly if $f(x) + g(y) \approx_{(x,y)} F$ at (x_0, y_0).

Recall that a function $f : E_1 \to Z$ is called *positive homogeneous* if

$$f(\alpha x) = \alpha f(x) \quad \forall\, \alpha \geq 0\,, \quad \forall\, x \in E_1\,.$$

Following Robinson (1987) we can now define the *Bouligand-derivative.*

Definition 3.2.1. *Given a function f from an open subset U of a normed linear space E_1 into another normed linear space Z, we say that f is* ***Bouligand-differentiable*** *at a point $x_0 \in U$, if there exists a positive homogeneous function $\nabla^B f(x_0) : U \to Z$ such that*

$$f(x_0 + h) = f(x_0) + \nabla^B f(x_0)(h) + o(h).$$

which can be rewritten as

$$\lim_{h \to 0} \frac{\big\| f(x_0 + h) - f(x_0) - \nabla^B f(x_0)(h) \big\|_Z}{\|h\|_{E_1}} = 0\,. \tag{3.3}$$

We will sometimes use the abbreviations B-, F-, H-, and G-derivatives for Bouligand-, Fréchet-, Hadamard-, and Gâteaux-derivatives respectively.

Let $F : E_1 \times E_2 \to Z$, and suppose that F has a partial B-derivative[4] $\nabla_1^B F(x_0, y_0)$ with respect to x at (x_0, y_0). We say that $\nabla_1^B F(x_0, y_0)$ is *strong* if

$$F(x_0, y_0) + \nabla_1^B F(x_0, y_0)(x - x_0) \;\approx_x\; F \;\text{ at }\; (x_0, y_0)\,.$$

Robinson (1987) showed that the chain rule holds for Bouligand-derivatives. Let f be a Lipschitzian function from an open set $\Omega \subset \mathbb{R}^m$ to $\mathbb{R}^k$, $x_0 \in \Omega$, and f B-differentiable at x_0. Let g be a Lipschitzian function from an open set $\Gamma \subset \mathbb{R}^k$, with $f(x_0) \in \Gamma$, to $\mathbb{R}^l$ be B-differentiable at $f(x_0)$. Then $g \circ f$ is B-differentiable at x_0 and

$$\nabla^B (g \circ f)(x_0) = \nabla^B g\big(f(x_0)\big) \circ \nabla^B f(x_0)\,.$$

The fact that B-derivatives, just as F- and H-derivatives, fulfill the chain rule is no contradiction to the fact that H-differentiability is the *weakest $\mathcal{S}$-differentiation*[5] which fulfills the chain rule (Averbukh and Smolyanov,

[4] Partial B-derivatives of f are denoted by $\nabla_1^B f$, $\nabla_2^B f$, $\nabla_{2,2}^B f := \nabla_2^B \big(\nabla_2^B f\big)$ etc.

[5] See Appendix A.3.3.

1967, 1968) because the B-derivative is not necessarily a continuous linear function.

In general are Gâteaux- and Bouligand-differentiability not directly comparable, because B-derivatives are by definition positive homogeneous, but not necessarily linear. We will show in Subsection 3.2.3 that the existence of the BIF implies the existence of the IF and that in that case both are equal. Please note however, that this in general still does not imply that the IF is a Gâteaux-derivative.

Since every linear function is trivially also positive homogeneous, it follows directly from the definition that every Fréchet-differentiable function is also Bouligand-differentiable.

The following implicit function theorem for B-derivatives, can be found in Robinson (1991, Corollary 3.4). For a function f from a metric space (X, d_X) to another metric space (Y, d_Y), we define

$$\delta(f, X) = \inf\{d_Y\big(f(x_1), f(x_2)\big) \,/\, d_X(x_1, x_2) \mid x_1 \neq x_2;\, x_1,\, x_2 \in X\}\,.$$

Clearly $\delta(f, X) \neq 0$ only if f is one-to-one on X.

Theorem 3.2.2 (Implicit function theorem)**.** *Let Y be a Banach space and X and Z be normed linear spaces. Let x_0 and y_0 be points of X and Y, respectively, and let $\mathcal{N}(x_0)$ be a neighborhood of x_0 and $\mathcal{N}(y_0)$ be a neighborhood of y_0. Suppose that G is a function from $\mathcal{N}(x_0) \times \mathcal{N}(y_0)$ to Z with $G(x_0, y_0) = 0$. In particular, for some ϕ and each $y \in \mathcal{N}(y_0)$, $G(\,\cdot\,, y)$ is assumed to be Lipschitz continuous on $\mathcal{N}(x_0)$ with modulus ϕ. Assume that G has partial B-derivatives with respect to x and y at (x_0, y_0), and that:*

(i) $\nabla_2^B G(x_0, y_0)(\,\cdot\,)$ *is strong.*

(ii) $\nabla_2^B G(x_0, y_0)(y - y_0)$ *lies in a neighborhood of* $0 \in Z$, $\forall y \in \mathcal{N}(y_0)$.

(iii) $\delta(\nabla_2^B G(x_0, y_0), \mathcal{N}(y_0) - y_0) \;=:\; d_0 > 0$.

Then for each $\xi > d_0^{-1}\phi$ there are neighborhoods U of x_0 and V of y_0, and a function $f^* : U \to V$ satisfying

(a) $f^*(x_0) = y_0$.

(b) f^* *is Lipschitz continuous on* $\mathcal{N}(x_0)$ *with modulus* ξ.

(c) For each $x \in U$, $f^*(x)$ is the unique solution in V of $G(x,y) = 0$.

(d) The function f^* is B-differentiable at x_0 with

$$\nabla^B f^*(x_0)(u) = \left(\nabla_2^B G(x_0, y_0)\right)^{-1} \left(-\nabla_1^B G(x_0, y_0)(u)\right) .$$

3.2.2 The Bouligand-Derivative of Some Loss Functions

In this subsection we will calculate the Bouligand-derivatives of some loss functions.

Least squares loss
Let us start easy with a function we know to be F-differentiable. In this case, the B-derivative should be the same as the F-derivative. We will thus show for the least squares loss

$$L(x,y,t) = L_{LS}(x,y,t) := (y-t)^2$$

that

$$\nabla_3^B L(x,y,t)(h) = -2(y-t)h$$

and $\nabla_{3,3}^F L(x,y,t)(h) = 2h$.

The first B-derivative is

$$\begin{aligned} \nabla_3^B L(x,y,t)(h) + o(h) &= L(x,y,t+h) - L(x,y,t) \\ &= (y-t-h)^2 - (y-t)^2 \\ &= 2th - 2yh + h^2 \\ &= -2h(y-t) + h^2 , \end{aligned}$$

and therefore $\nabla_3^B L(x,y,t) = -2(y-t) = \nabla_3^F L(x,y,t)$.

In the same way, we find the second B-derivative:

$$\begin{aligned} \nabla_{3,3}^B L(x,y,t)(h) + o(h) &= \nabla_3^B L(y,t+h) - \nabla_3^B L(x,y,t) \\ &= -2(y-t-h) - (-2(y-t)) = 2h . \end{aligned}$$

So $\nabla_{3,3}^B L(x,y,t) = 2$, which also corresponds to the second order partial F-derivative.

ϵ-insensitive loss
We shall show for the ϵ-insensitive loss

$$L(x,y,t) = L_\epsilon(x,y,t) := \max\{|y-t| - \epsilon, 0\}$$

that

$$\nabla_3^B L(x,y,t)(h) = \begin{cases} -h & \text{if} \quad \{t < y-\epsilon\} \text{ or } \{y-t=\epsilon, h<0\} \\ 0 & \text{if} \quad \{y-\epsilon < t < y+\epsilon\} \text{ or } \{y-t=\epsilon, h \geq 0\} \\ & \quad\quad \text{or } \{y-t=-\epsilon, h<0\} \\ h & \text{if} \quad \{t > y+\epsilon\} \text{ or } \{y-t=-\epsilon, h \geq 0\} \end{cases}$$

and $\nabla_{3,3}^B L(x,y,t)(h) = 0$.

For the derivation of $\nabla_3^B L(x,y,t)$ we need to consider 5 cases.

i) If $t > y+\epsilon$, we have $t+h > y+\epsilon$ as long as h is small enough. Therefore,

$$\begin{aligned} \nabla_3^B L(x,y,t)(h) + o(h) &= L(x,y,t+h) - L(x,y,t) \\ &= t+h-y-\epsilon-(t-y-\epsilon) = h\,. \end{aligned}$$

ii) If $t < y-\epsilon$, we have $t+h < y+\epsilon$ if h is sufficiently small. Thus

$$\nabla_3^B L(x,y,t)(h) + o(h) = y-t-h-\epsilon-(y-t-\epsilon) = -h\,.$$

iii) If $y-t \in (-\epsilon,\epsilon)$ we have $y-t-h \in (-\epsilon,\epsilon)$ for $h \to 0$. This yields $\nabla_3^B L(x,y,t)(h) + o(h) = 0-0 = 0$.

iv) If $y-t=\epsilon$ we have to consider 2 cases. If $h \geq 0$ and small, then $-\epsilon < y-t-h < \epsilon$ and hence $\nabla_3^B L(x,y,t)(h) + o(h) = 0-0 = 0$. If $h<0$, we have $y-t-h > \epsilon$ and thus

$$\nabla_3^B L(x,y,t)(h) + o(h) = y-t-h-\epsilon-0 = -h\,.$$

v) If $y-t=-\epsilon$ we have again to consider 2 cases. If $h \geq 0$, we have $y-t-h < -\epsilon$. Hence

$$\nabla_3^B L(x,y,t)(h) + o(h) = t+h-y-\epsilon-0 = h\,.$$

If $h<0$, we get $-\epsilon < y-t-h < \epsilon$ which gives $\nabla_3^B L(x,y,t)(h) + o(h) = 0-0 = 0$.

This gives the assertion for the first partial B-derivative. Using the same reasoning we obtain $\nabla_{3,3}^B L(x,y,t)(h) = 0$.

Pinball-loss
It will be shown that for the pinball loss

$$L(x,y,t) = L_{\tau-pin}(x,y,t) := \begin{cases} (\tau-1)(y-t)\,, & \text{if } y-t<0 \\ \tau(y-t)\,, & \text{if } y-t \geq 0 \end{cases}$$

we get

$$\nabla_3^B L(x,y,t)(h) = \begin{cases} (1-\tau)h & \text{if} \quad \{y-t<0\} \text{ or } \{y-t=0, h\geq 0\} \\ -\tau h & \text{if} \quad \{y-t>0\} \text{ or } \{y-t=0, h<0\} \end{cases}$$

and $\nabla_{3,3}^B L(x,y,t)(h) = 0$.

For the calculation of $\nabla_3^B L(x,y,t)$ we consider 3 cases.

i) If $y-t<0$ we have $y-t-h<0$ for sufficiently small values of $|h|$. Hence

$$\begin{aligned} \nabla_3^B L(x,y,t)(h) + o(h) &= L(x,y,t+h) - L(x,y,t) \\ &= (\tau-1)(y-t-h) - (\tau-1)(y-t) \\ &= (1-\tau)h\,. \end{aligned}$$

ii) If $y-t>0$ we have $y-t-h>0$ for sufficiently small values of $|h|$ which yields

$$\nabla_3^B L(x,y,t)(h) + o(h) = \tau(y-t-h) - \tau(y-t) = -\tau h\,.$$

iii) Assume $y-t=0$. If $y-t-h<0$ we have

$$\nabla_3^B L(x,y,t)(h) + o(h) = (1-\tau)h\,.$$

If $y-t-h>0$ it follows

$$\nabla_3^B L(x,y,t)(h) + o(h) = \tau(y-t-h) - \tau(y-t) = -\tau h\,.$$

Together this gives the assertion for $\nabla_3^B L(x,y,t)(h)$. In the same way we get $\nabla_{3,3}^B L(x,y,t)(h) = 0$.

Huber loss
It will be shown that for the Huber loss

$$L(x,y,t) = L_{c-Huber}(x,y,t) := \begin{cases} 0.5(y-t)^2 & \text{if } |y-t| \leq c \\ c|y-t| - c^2/2 & \text{if } |y-t| > c \end{cases}$$

we have

$$\nabla_3^B L(x,y,t)(h) = \begin{cases} -c\,\mathrm{sign}(y-t)h & \text{if} \quad |y-t| > c \\ -(y-t)h & \text{if} \quad |y-t| \leq c \end{cases}$$

and

$$\nabla_{3,3}^B L(x,y,t)(h) \quad = \quad \begin{cases} h & \text{if} \quad \{y-t=c, h \geq 0\} \text{ or } \{y-t=-c, h<0\} \\ & \quad\ \text{ or } \{|y-t|<c\} \\ 0 & \text{if} \quad \text{else}\,. \end{cases}$$

For the derivation of $\nabla_3^B L(x,y,t)$ we consider the following 5 cases.

i) Let $y-t=c$. If $h \geq 0$ or $y-t-h \leq c$ then

$$\begin{aligned} \nabla_3^B L(x,y,t)(h) + o(h) &= L(x,y,t+h) - L(x,y,t) \\ &= \frac{1}{2}(y-t-h)^2 - \frac{1}{2}(y-t)^2 \\ &= -(y-t)h + \frac{h^2}{2}\,. \end{aligned}$$

If $h<0$ or $y-t-h>c>0$ we have

$$\begin{aligned} \nabla_3^B L(x,y,t)(h) + o(h) &= c|y-t-h| - \frac{c^2}{2} - \frac{1}{2}(y-t)^2 \\ &= c(y-t-h) - \frac{c^2}{2} - \frac{c^2}{2} \\ &= c(c-h) - c^2 = -(y-t)h\,. \end{aligned}$$

ii) Now we consider the case $y-t=-c$. If $h \geq 0$ or $y-t-h \leq -c < 0$ we obtain

$$\begin{aligned} \nabla_3^B L(x,y,t)(h) + o(h) &= c|y-t-h| - \frac{c^2}{2} - \frac{1}{2}(y-t)^2 \\ &= c(c+h) - \frac{c^2}{2} - \frac{c^2}{2} = -(y-t)h\,. \end{aligned}$$

If $h<0$ or $y-t-h>-c$ we get

$$\nabla_3^B L(x,y,t)(h) + o(h) = \frac{1}{2}(y-t-h)^2 - \frac{1}{2}(y-t)^2 = -(y-t)h + \frac{h^2}{2}\,.$$

iii) If $y-t>c$, we have $y-t-h>c$ and thus

$$\begin{aligned}\nabla_3^B L(x,y,t)(h)+o(h) &= c|y-t-h|-\frac{c^2}{2}-c|y-t|+\frac{c^2}{2}\\ &= c(y-t-h)-c(y-t)\\ &= -ch=-c\,\mathrm{sign}(y-t)h\,.\end{aligned}$$

iv) If $y-t<-c$, we have $y-t-h<-c$ and obtain analogously to (iii) that

$$\begin{aligned}\nabla_3^B L(x,y,t)(h)+o(h) &= c|y-t-h|-\frac{c^2}{2}-c|y-t|+\frac{c^2}{2}\\ &= c(-y+t+h)-c(-y+t)\\ &= ch=-c\,\mathrm{sign}(y-t)h\,.\end{aligned}$$

v) If $-c<y-t<c$, then $-c<y-t-h<c$ and

$$\nabla_3^B L(x,y,t)(h)+o(h)=\frac{1}{2}(y-t-h)^2-\frac{1}{2}(y-t)^2=-(y-t)h+\frac{h^2}{2}\,.$$

This gives the assertion for $\nabla_3^B L(x,y,t)(h)$. Only the first two cases, where $y-t=\pm c$, were necessary to compute, since in the other 3 parts the function is already F-differentiable, and thus also B-differentiable. For the second partial B-derivative we consider 3 cases.

i) Assume $y-t=c$. If $y-t-h<c$ then

$$\begin{aligned}\nabla_{3,3}^B L(x,y,t)(h)+o(h) &= \nabla_3^B L(x,y,t+h)-\nabla_3^B L(x,y,t)\\ &= -(y-t-h)-(-(y-t))=h\,.\end{aligned}$$

If $y-t-h>c$ then $\nabla_{3,3}^B L(x,y,t)(h)+o(h)=-c-(-(y-t))=0$.

ii) Assume $y-t=-c$. If $y-t-h<-c$ we obtain $\nabla_{3,3}^B L(x,y,t)(h)+o(h)=c-(-(y-t))=0$.
If $y-t-h>-c$ then

$$\nabla_{3,3}^B L(x,y,t)(h)+o(h)=-(y-t-h)-(-(y-t))=h\,.$$

iii) Assume that $|y-t|\neq c$. Then $\nabla_3^B L(x,y,t+h)=\nabla_3^B L(x,y,t)$. The difference, and consequently $\nabla_{3,3}^B L(x,y,t)(h)=0$.

This gives the assertion for Huber's loss function.

Hinge loss
Finally we will show that for the hinge loss

$$L(x,y,t) = L_{hinge}(x,y,t) := \max\{0, 1-yt\}$$

we obtain the following partial B-derivatives:

$$\nabla_3^B L(x,y,t)(h) = \begin{cases} -yh & \text{if} \quad \{yt<1\} \text{ or } \{y=t=1, h<0\} \\ & \qquad \text{or } \{y=t=-1, h\geq 0\} \\ 0 & \text{if} \quad \{yt>1\} \text{ or } \{y=t=1, h\geq 0\} \\ & \qquad \text{or } \{y=t=-1, h<0\} \end{cases}$$

and $\nabla_{3,3}^B L(x,y,t)(h) = 0$.

For the first derivative, we need to differ 3 cases:

i) If $yt < 1$ then for h small enough, we also have $y(t+h) < 1$. Thus

$$\begin{aligned} \nabla_3^B L(x,y,t)(h) + o(h) &= L(x,y,t+h) - L(x,y,t) \\ &= 1 - y(t-h) - (1-yt) = -yh\,. \end{aligned}$$

ii) Likewise if $yt > 1$ then also $y(t+h) > 1$ if h is sufficiently small and so

$$\begin{aligned} \nabla_3^B L(x,y,t)(h) + o(h) &= L(y,t+h) - L(y,t) \\ &= 0 - 0 = 0\,. \end{aligned}$$

iii) The third case is where $yt = 1$. Here we have to consider 2 cases. If $y(t+h) > 1$ we have

$$\nabla_3^B L(x,y,t)(h) + o(h) = 0 - 0 = 0\,.$$

Or else, for $y(t+h) < 1$, we obtain

$$\nabla_3^B L(x,y,t)(h) + o(h) = 1 - y(t+h) - 0 = -yh\,,$$

which gives us the assertion. Following the same reasoning, it is clear that $\nabla_{3,3}^B L(x,y,t)(h) = 0$.

3.2.3 The Bouligand Influence Function

We will now define the Bouligand influence function as a new measure in robust statistics, and we then will use it to show that a broad class of support vector machines based on a Lipschitz continuous, but not necessarily Fréchet-differentiable, loss function are robust in the sense of having a bounded Bouligand influence function. Recall that we denote the set of all probability distributions on some measurable space $(Z, \mathcal{A})$ by $\mathcal{M}_1$ and let $\mathcal{H}$ be a Hilbert space.

Definition 3.2.3. *The* ***Bouligand influence function (BIF)*** *of the function* $T : \mathcal{M}_1 \to \mathcal{H}$ *for a distribution* P *in the direction of a distribution* $\mathrm{Q} \neq \mathrm{P}$ *is the special Bouligand-derivative (if it exists)*

$$\lim_{\varepsilon \downarrow 0} \frac{\left\| T\big((1-\varepsilon)\mathrm{P} + \varepsilon \mathrm{Q}\big) - T(\mathrm{P}) - \mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) \right\|_{\mathcal{H}}}{\varepsilon} = 0\,. \qquad (3.4)$$

The BIF has the interpretation that it measures the impact of an infinitesimal small amount of contamination of the original distribution P in the direction of Q on the quantity of interest $T(\mathrm{P})$. It is thus desirable that the function T has a *bounded* BIF.

Note that (3.4) is indeed a special B-derivative, because we consider the directions $h = \varepsilon(\mathrm{Q} - \mathrm{P})$ and $x_0 = \mathrm{P}$. If Q equals the Dirac distribution δ_z in a point $z \in Z$, that is $\delta_z(\{z\}) = 1$, we write $\mathrm{BIF}(z; T, \mathrm{P})$. The choice of the metric on $\mathcal{M}_1$ is not important for the definition of the BIF, because $\|\varepsilon(\mathrm{Q} - \mathrm{P})\| = \varepsilon \, \|\mathrm{Q} - \mathrm{P}\|$ and $\|\mathrm{Q} - \mathrm{P}\|$ is a positive constant. For the norm of total variation $\|\cdot\|_{\mathcal{M}}$ we obtain for example,

$$\lim_{\varepsilon(\mathrm{Q}-\mathrm{P}) \downarrow 0} \frac{\left\| T\big(\mathrm{P} + \varepsilon(\mathrm{Q} - \mathrm{P})\big) - T(\mathrm{P}) - \mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) \right\|_{\mathcal{H}}}{\|\varepsilon(\mathrm{Q} - \mathrm{P})\|_{\mathcal{M}}} = 0\,,$$

(cf., Equation 3.3). Since $\varepsilon(\mathrm{Q} - \mathrm{P}) \to 0$ if and only if $\varepsilon \to 0$ and by the assumption that $\mathrm{Q} \neq \mathrm{P}$ we obtain (3.4).

The Bouligand influence function is a modification of the influence function given by (3.1). Recall that the Gâteaux-derivative of some mapping f at a point x_0 equals

$$\nabla^G f(x_0)(h) = \lim_{\varepsilon \downarrow 0} \frac{f(x_0 + \varepsilon h) - f(x_0)}{\varepsilon}$$

if it exists for every $h \in X$. Hence the influence function is the special Gâteaux-derivative with $\mathrm{Q} = \delta_z$ and $h = \delta_z - \mathrm{P}$, if the IF is continuous and linear. However, the BIF is always positive homogeneous because it is a Bouligand-derivative, which is in general not true for the influence function. As will be shown in (3.24), this property leads to the result that for $\alpha \geq 0$ and $h := \varepsilon(\mathrm{Q} - \mathrm{P})$ the asymptotic bias $T((1-\alpha\varepsilon)\mathrm{P} + \alpha\varepsilon\mathrm{Q}) - T(\mathrm{P})$ equals $\alpha\,\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) + o(\alpha h)$.

The following simple calculations clarify the connection between the BIF and the IF. In general we have for B-derivatives with $h = \varepsilon\tilde{h}$, where $\varepsilon \in (0, \infty)$ and $\tilde{h} \in X$ with $0 < \|\tilde{h}\| \leq 2$,

$$\begin{aligned} 0 &= \lim_{h \to 0} \frac{\|f(x_0 + h) - f(x_0) - \nabla^B f(x_0)(h)\|}{\|h\|} \\ \Leftrightarrow\ 0 &= \lim_{\varepsilon \downarrow 0} \frac{\|f(x_0 + \varepsilon\tilde{h}) - f(x_0) - \varepsilon\nabla^B f(x_0)(\tilde{h})\|}{\varepsilon\|\tilde{h}\|} \\ \Leftrightarrow\ 0 &= \lim_{\varepsilon \downarrow 0} \left\| \frac{f(x_0 + \varepsilon\tilde{h}) - f(x_0)}{\varepsilon} - \nabla^B f(x_0)(\tilde{h}) \right\| . \end{aligned}$$

Hence

$$\lim_{\varepsilon \downarrow 0} \frac{f(x_0 + \varepsilon\tilde{h}) - f(x_0)}{\varepsilon} = \nabla^B f(x_0)(\tilde{h}) .$$

In particular we obtain for $\mathrm{Q} \neq \mathrm{P}$ and taking $0 < \|\mathrm{Q} - \mathrm{P}\| \leq 2$ into account that, if $\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P})$ exists, then

$$\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}) - T(P)}{\varepsilon} ,$$

which is the definition of the IF, if we choose $\mathrm{Q} = \delta_z$.

Figure 3.2 show the connections between the various types of differentiation and their link to the influence functions.

In the following sections, we will investigate the robustness of a specific learning method, namely the SVM, through the use of influence functions (both IF and BIF) and bounds on the bias.

Using these general concepts of robust statistics, it is possible to compare different methods with respect to their robustness properties. Methods with a smaller norm of the influence function or with a smaller maximal bias are generally considered as more robust. In our case one could consider either the supremum norm or the Hilbert norm.

It would of course also be possible to look at the (Bouligand) influence function of the risk, but this is not done in the scope of this work. This

$$\begin{tikzcd}
 & \text{Hadamard} \arrow[r, Rightarrow] & \text{Gâteaux} \arrow[dr, Rightarrow] & \\
\text{Fréchet} \arrow[ur, Rightarrow] \arrow[dr, Rightarrow] & & & \text{IF} \\
 & \text{Bouligand} \arrow[r, Rightarrow] & \text{BIF} \arrow[ur, Rightarrow] &
\end{tikzcd}$$

Figure 3.2: Overview on the relation between the different types of differentiation and influence functions.

would probably be an easier problem than the one solved in the following sections, since the function $f_{L,\mathrm{P},\lambda} \in \mathcal{H}$, whereas the risk, either $\mathcal{R}_{L,\mathrm{P}}$ or $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}$, is only an element of $\mathbb{R}$.

3.3 Robustness of SVMs Based on Standard Loss Functions

After introducing the Bouligand influence function, we are ready to show that a broad class of support vector machines based on a Lipschitz continuous, but not necessarily Fréchet-differentiable, loss function has a bounded Bouligand influence function.

3.3.1 Robustness of SVMs

We can now give a general result on the BIF of the support vector machine. To this end define

$$T : \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}, \quad T(\mathrm{P}) := f_{L,\mathrm{P},\lambda}\,.$$

We restrict attention to Lipschitz continuous loss functions, because the growth behavior of the loss function L plays an important role to obtain consistency and robustness results as shown by Christmann and Steinwart (2007). For notational convenience we shall often write $\nabla_3^B L(X,Y,f(X))$ instead of $\nabla_3^B L(X,Y,\,\cdot\,)(f(X))$, because $f(X) \in \mathbb{R}$. We will sometimes explicitly write "$\cdot$" for multiplication to avoid misunderstandings.

Theorem 3.3.1 (Bouligand influence function)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be closed,*[6] *$\mathcal{H}$ be an RKHS with a bounded, continuous kernel k, $f_{L,\mathrm{P},\lambda} \in \mathcal{H}$, and $L : \mathcal{X} \times$*

[6] Or more general, $\mathcal{X}$ a complete separable normed linear space.

$\mathcal{Y}\times\mathbb{R}\to[0,\infty)$ *be a convex loss function which is Lipschitz continuous with Lipschitz constant* $|L|_1\in(0,\infty)$. *Further, assume that* L *has measurable partial B-derivatives with*

$$\begin{aligned}\kappa_1 &:= \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\|\nabla_3^B L(x,y,\,\cdot\,)\right\|_\infty \in (0,\infty)\,,\\ \kappa_2 &:= \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\|\nabla_{3,3}^B L(x,y,\,\cdot\,)\right\|_\infty < \infty\,.\end{aligned} \tag{3.5}$$

Let P,Q *be probability measures*[7] *on* $\big(\mathcal{X}\times\mathcal{Y},\mathcal{B}(\mathcal{X}\times\mathcal{Y})\big)$ *with* $\mathbb{E}_\mathrm{P}|Y|<\infty$ *and* $\mathbb{E}_\mathrm{Q}|Y|<\infty$, $\delta_1>0$, $\delta_2>0$,

$$\mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda}) := \{f\in\mathcal{H};\, \|f-f_{L,\mathrm{P},\lambda}\|_\mathcal{H}<\delta_1\}\,,$$

and $\lambda>\frac{1}{2}\kappa_2\,\|k\|_\infty^3$. *Define* $G:(-\delta_2,\delta_2)\times\mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})\to\mathcal{H}$,

$$G(\varepsilon,f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}}\nabla_3^B L(X,Y,f(X))\cdot\Phi(X)\,, \tag{3.6}$$

and assume that $\nabla_2^B G(0,f_{L,\mathrm{P},\lambda})$ *is strong. Then the Bouligand influence function of* $T(\mathrm{P}):=f_{L,\mathrm{P},\lambda}$ *in the direction of* $\mathrm{Q}\neq\mathrm{P}$ *exists,*

$$\begin{aligned}\mathrm{BIF}(\mathrm{Q};T,\mathrm{P}) \;&=\; S^{-1}\big(\mathbb{E}_\mathrm{P}\nabla_3^B L(X,Y,f_{L,\mathrm{P},\lambda}(X))\cdot\Phi(X)\big) &(3.7)\\ &\quad -S^{-1}\big(\mathbb{E}_\mathrm{Q}\nabla_3^B L(X,Y,f_{L,\mathrm{P},\lambda}(X))\cdot\Phi(X)\big)\,, &(3.8)\end{aligned}$$

where $S:\mathcal{H}\to\mathcal{H}$ *with*

$$\begin{aligned}S(\,\cdot\,) \;&:=\; \nabla_2^B G(0,f_{L,\mathrm{P},\lambda})(\,\cdot\,)\\ &=\; 2\lambda\,\mathrm{id}_\mathcal{H}(\,\cdot\,)+\mathbb{E}_\mathrm{P}\nabla_{3,3}^B L(X,Y,f_{L,\mathrm{P},\lambda}(X))\cdot\langle\Phi(X),\,\cdot\,\rangle_\mathcal{H}\Phi(X)\,,\end{aligned}$$

and $\mathrm{BIF}(\mathrm{Q};T,\mathrm{P})$ *is bounded.*

The proof of Theorem 3.3.1 is based upon the implicit function theorem 3.2.2 for Bouligand-derivatives as well as Theorem A.3.2.

Remark 3.3.2. *In order to prove 3.3.1, we additionally show that under the assumptions of Theorem 3.3.1 we have:*

i) For some χ *and each* $f\in\mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$, $G(\,\cdot\,,f)$ *is Lipschitz continuous on* $(-\delta_2,\delta_2)$ *with Lipschitz constant* χ.

[7] Because $\mathcal{X}$ and $\mathcal{Y}$ are assumed to be closed, P can be split up into the marginal distribution P_X and the regular conditional probability $\mathrm{P}(\,\cdot\,|x)$, $x\in\mathcal{X}$, on $\mathcal{Y}$. Same for Q.

ii) *G has partial B-derivatives with respect to ε and f at $(0, f_{L,\mathrm{P},\lambda})$.*

iii) *$\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})(h - f_{L,\mathrm{P},\lambda})$ lies in a neighborhood of $0 \in \mathcal{H}$, $\forall h \in \mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$.*

iv) *The constant d_0 defined as*

$$\inf_{\substack{h_1,h_2 \in \mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda}) - f_{L,\mathrm{P},\lambda} \\ h_1 \neq h_2}} \frac{\left\| \nabla_2^B G(0, f_{L,\mathrm{P},\lambda})(h_1) - \nabla_2^B G(0, f_{L,\mathrm{P},\lambda})(h_2) \right\|_{\mathcal{H}}}{\|h_1 - h_2\|_{\mathcal{H}}}$$

is strictly positive.

v) *For each $\xi > d_0^{-1}\chi$ there exist constants $\delta_3, \delta_4 > 0$, a neighborhood $\mathcal{N}_{\delta_3}(f_{L,\mathrm{P},\lambda}) := \{f \in \mathcal{H}; \|f - f_{L,\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta_3\}$, and a function $f^* : (-\delta_4, \delta_4) \to \mathcal{N}_{\delta_3}(f_{L,\mathrm{P},\lambda})$ satisfying*

v.1) *$f^*(0) = f_{L,\mathrm{P},\lambda}$.*

v.2) *f^* is Lipschitz continuous on $(-\delta_4, \delta_4)$ with Lipschitz constant $|f^*|_1 = \xi$.*

v.3) *For each $\varepsilon \in (-\delta_4, \delta_4)$ is $f^*(\varepsilon)$ the unique solution of $G(\varepsilon, f) = 0$ in $\mathcal{N}_{\delta_3}(f_{L,\mathrm{P},\lambda})$.*

v.4) *$\nabla^B f^*(0)(u) = \left(\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})\right)^{-1} \left(-\nabla_1^B G(0, f_{L,\mathrm{P},\lambda})(u)\right)$, $u \in (-\delta_4, \delta_4)$.*

The function f^* is the same as in the implicit function theorem 3.2.2.

Remark 3.3.3. *It has been shown in Subsection 3.2.2 that $\kappa_2 = 0$ for $L = L_\epsilon$ and $L = L_{\tau-pin}$ and thus in these cases the regularization condition only states that $\lambda > \frac{1}{2}\kappa_2 \|k\|_\infty^3 = 0$.*

Note that S can be interpreted as the (Bouligand-)Hessian of the regularized risk, see (3.9) and (3.12). Further the formula in (3.7) and (3.8) is similar to the one obtained by Christmann and Steinwart (2007) for the IF of $T(\mathrm{P}) = f_{L,\mathrm{P},\lambda}$. The difference is that we used B-derivatives instead of F-derivatives, because we allow for non-smooth loss functions.

Also note that the first summand of the BIF given in (3.7) does *not* depend on the contaminating distribution Q. In contrast to that, the second summand of the BIF given in (3.8) depends on Q and consists of two factors. The first factor depends on the partial B-derivative of the loss function, and is hence bounded due to (3.5). For many loss functions this factor depends only on the residual term $y - f_{L,\mathrm{P},\lambda}(x)$. The second factor is the feature map $\Phi(x)$ which is bounded, because k is bounded. For the Gaussian RBF kernel we expect that the second factor is not only bounded, but that the impact of $\mathrm{Q} \neq \mathrm{P}$ on the BIF is approximately local, because $k(x, x')$ converges exponentially fast to zero if $||x - x'||_2$ is large.

The key ingredient of our proof of Theorem 3.3.1 is of course the map $G : \mathbb{R} \times \mathcal{H} \to \mathcal{H}$ defined by (3.6). If $\varepsilon < 0$ the integration is with respect to a signed measure. The $\mathcal{H}$-valued expectation used in the definition of G is well-defined for all $\varepsilon \in (-\delta_2, \delta_2)$ and all $f \in \mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$, because $\kappa_1 \in (0, \infty)$ by (3.5) and $\|\Phi(x)\|_\infty \leq \|k\|_\infty^2 < \infty$ by (1.34). For F- and B-derivatives holds a chain rule and F-differentiable functions are also B-differentiable. For $\varepsilon \in [0, 1]$ we thus obtain

$$G(\varepsilon, f) = \frac{\partial \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}}{\partial \mathcal{H}}(f) = \nabla_2^B \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}(f)\,. \tag{3.9}$$

Since $f \mapsto \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}(f)$ is convex and continuous for all $\varepsilon \in [0, 1]$ equation (3.9) shows that we have $G(\varepsilon, f) = 0$ if and only if $f = f_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}$ for such ε. Hence

$$G(0, f_{L,\mathrm{P},\lambda}) = 0\,. \tag{3.10}$$

By using the steps in Remark 3.3.2, we shall show that Theorem 3.2.2 is applicable for G and that there exists a B-differentiable function $\varepsilon \mapsto f_\varepsilon$ defined on a small interval $(-\delta_2, \delta_2)$ for some $\delta_2 > 0$ satisfying $G(\varepsilon, f_\varepsilon) = 0$ for all $\varepsilon \in (-\delta_2, \delta_2)$. From the existence of this function we shall obtain $\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) = \nabla^B f_\varepsilon(0)$.

Proof of Theorem 3.3.1. The existence of $f_{L,\mathrm{P},\lambda}$ follows from the convexity of L and the penalizing term, see also Christmann and Steinwart (2007, Proposition 8). The assumption that $G(0, f_{L,\mathrm{P},\lambda}) = 0$ is valid by (3.10). Let us now prove the results of Remark 3.3.2 parts 1 to 5.

Remark 3.3.2 part (i). For $f \in \mathcal{H}$ fixed let $\varepsilon_1, \varepsilon_2 \in (-\delta_2, \delta_2)$. Using $\|k\|_\infty < \infty$ and (3.10) we obtain

$$\begin{aligned}
&\big|\mathbb{E}_{(1-\varepsilon_1)\mathrm{P}+\varepsilon_1\mathrm{Q}} \nabla_3^B L(X, Y, f(X)) \cdot \Phi(X) \\
&\quad - \mathbb{E}_{(1-\varepsilon_2)\mathrm{P}+\varepsilon_2\mathrm{Q}} \nabla_3^B L(X, Y, f(X)) \cdot \Phi(X)\big| \\
= \; &\big|\mathbb{E}_\mathrm{P} \nabla_3^B L(X, Y, f(X))\Phi(X) + \varepsilon_1 \mathbb{E}_{\mathrm{Q}-\mathrm{P}} \nabla_3^B L(X, Y, f(X))\Phi(X) \\
&\quad - \mathbb{E}_\mathrm{P} \nabla_3^B L(X, Y, f(X))\Phi(X) - \varepsilon_2 \mathbb{E}_{\mathrm{Q}-\mathrm{P}} \nabla_3^B L(X, Y, f(X))\Phi(X)\big| \\
= \; &\big|(\varepsilon_1 - \varepsilon_2)\mathbb{E}_{\mathrm{Q}-\mathrm{P}} \nabla_3^B L(X, Y, f(X)) \cdot \Phi(X)\big|
\end{aligned}$$

$$\begin{aligned}
&= \left|(\varepsilon_1-\varepsilon_2)\int \nabla_3^B L(x,y,f(x))\Phi(x)\,d(\mathrm{Q}-\mathrm{P})(x,y)\right| \\
&\le |\varepsilon_1-\varepsilon_2| \int \left|\nabla_3^B L(x,y,f(x))\cdot\Phi(x)\right|\,d|\mathrm{Q}-\mathrm{P}|(x,y) \\
&\le |\varepsilon_1-\varepsilon_2| \int \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |\nabla_3^B L(x,y,f(x))| \sup_{x\in X} |\Phi(x)|\,d|\mathrm{Q}-\mathrm{P}|(x,y) \\
&\le |\varepsilon_1-\varepsilon_2|\, \|\Phi(x)\|_\infty \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\|\nabla_3^B L(x,y,\,\cdot\,)\right\|_\infty \int d|\mathrm{Q}-\mathrm{P}|(x,y) \\
&\le 2\,\|k\|_\infty^2 \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\|\nabla_3^B L(x,y,\,\cdot\,)\right\|_\infty |\varepsilon_1-\varepsilon_2| \\
&= 2\,\|k\|_\infty^2\,\kappa_1\,|\varepsilon_1-\varepsilon_2| < \infty\,.
\end{aligned}$$

Remark 3.3.2 part (ii). We have

$$\begin{aligned}
\nabla_1^B G(\varepsilon,f) &= \nabla_1^B\Big(\mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}}\nabla_3^B L(X,Y,f(X))\cdot\Phi(X)\Big) \\
&= \nabla_1^B\Big((1-\varepsilon)\mathbb{E}_\mathrm{P}\nabla_3^B L(X,Y,f(X))\Phi(X) \\
&\qquad +\varepsilon\mathbb{E}_\mathrm{Q}\nabla_3^B L(X,Y,f(X))\Phi(X)\Big) \\
&= \nabla_1^B\Big(\mathbb{E}_\mathrm{P}\nabla_3^B L(X,Y,f(X))\cdot\Phi(X) \\
&\qquad +\varepsilon\mathbb{E}_{\mathrm{Q}-\mathrm{P}}\nabla_3^B L(X,Y,f(X))\cdot\Phi(X)\Big) \\
&= \mathbb{E}_{\mathrm{Q}-\mathrm{P}}\nabla_3^B L(X,Y,f(X))\cdot\Phi(X) \\
&= \mathbb{E}_\mathrm{Q}\nabla_3^B L(X,Y,f(X))\cdot\Phi(X) \qquad (3.11)\\
&\quad -\mathbb{E}_\mathrm{P}\nabla_3^B L(X,Y,f(X))\cdot\Phi(X)\,.
\end{aligned}$$

These expectations exists due to (1.34) and (3.5). Furthermore, we obtain

$$\begin{aligned}
&\nabla_2^B G(0,f_{L,\mathrm{P},\lambda})(h)+o(h) \\
={}& G(0,f_{L,\mathrm{P},\lambda}+h)-G(0,f_{L,\mathrm{P},\lambda}) \\
={}& 2\lambda h+\mathbb{E}_\mathrm{P}\nabla_3^B L(X,Y,(f_{L,\mathrm{P},\lambda}(X)+h(X)))\cdot\Phi(X) \\
&\quad -\mathbb{E}_\mathrm{P}\nabla_3^B L(X,Y,f_{L,\mathrm{P},\lambda}(X))\cdot\Phi(X) \\
={}& 2\lambda h+\mathbb{E}_\mathrm{P}\Big(\nabla_3^B L\big(X,Y,(f_{L,\mathrm{P},\lambda}(X)+h(X))\big) \\
&\qquad -\nabla_3^B L\big(X,Y,f_{L,\mathrm{P},\lambda}(X)\big)\Big)\cdot\Phi(X)\,.
\end{aligned}$$

This expectation exists, as the term $\nabla_3^B L\big(X,Y,(f_{L,\mathrm{P},\lambda}(X)+h(X))\big) - \nabla_3^B L\big(X,Y,f_{L,\mathrm{P},\lambda}(X)\big)$ is bounded due to (1.34), (3.5), and $\|k\|_\infty < \infty$.

Using $\langle \Phi(x), \cdot \rangle_{\mathcal{H}} \in \mathcal{H}$, for all $x \in \mathcal{X}$, we get

$$\nabla_2^B G(0, f_{\mathrm{P},\lambda})(\,\cdot\,) = 2\lambda \mathrm{id}_{\mathcal{H}}(\,\cdot\,) + \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \langle \Phi(X), \cdot \rangle_{\mathcal{H}} \Phi(X)\,. \tag{3.12}$$

Note that $\mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L\big(X, Y, f(X)\big) = \nabla_3^B \mathbb{E}_{\mathrm{P}} \nabla_3^B L\big(X, Y, f(X)\big)$, because by using the definition of the Bouligand-derivative, we obtain for the partial Bouligand-derivatives

$$\begin{aligned} & \nabla_3^B \mathbb{E}_{\mathrm{P}} \nabla_3^B L\big(X, Y, f(X)\big) - \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L\big(X, Y, f(X)\big) \\ = \;& \mathbb{E}_{\mathrm{P}} \big(\nabla_3^B L(X, Y, (f(X) + h(X))) - \nabla_3^B L(X, Y, f(X)) \big) \\ & - \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X, Y, f(X)) + o(h) \\ = \;& \mathbb{E}_{\mathrm{P}} \big(\nabla_3^B L(X, Y, (f(X) + h(X))) - \nabla_3^B L(X, Y, f(X)) \\ & \quad - \nabla_{3,3}^B L(X, Y, f(X)) \big) + o(h) \\ = \;& o(h)\,, \qquad h \in \mathcal{H}\,. \end{aligned}$$

Remark 3.3.2 part (iii). Let $\mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$ be a δ_1-neighborhood of $f_{L,\mathrm{P},\lambda}$. Because $\mathcal{H}$ is an RKHS and hence a vector space it follows for all $h \in \mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$ that $\|f_{L,\mathrm{P},\lambda} - h - 0\|_{\mathcal{H}} \le \delta_1$ and hence $h - f_{L,\mathrm{P},\lambda} \in \mathcal{N}_{\delta_1}(0) \subset \mathcal{H}$. Note that $\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})(\,\cdot\,)$ computed by (3.12) is a mapping from $\mathcal{H}$ to $\mathcal{H}$. For $\xi := h - f_{L,\mathrm{P},\lambda}$ we have $\|\xi\|_{\mathcal{H}} \le \delta_1$ and the reproducing property yields

$$\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})(\xi) = 2\lambda\xi + \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \xi\Phi(X)\,.$$

Using (1.34) and (3.5) we obtain

$$\begin{aligned} & \big\| 2\lambda\xi + \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \xi\Phi(X) - 0 \big\|_{\mathcal{H}} \\ \le \;& 2\lambda \|\xi\|_{\mathcal{H}} + \big\| \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \xi\Phi(X) \big\|_{\mathcal{H}} \\ \le \;& 2\lambda \|\xi\|_{\mathcal{H}} + \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \big\| \nabla_{3,3}^B L(x, y, \,\cdot\,) \big\|_{\infty} \, \|\xi\|_{\infty} \, \|\Phi(x)\|_{\infty} \\ \le \;& 2\lambda \|\xi\|_{\mathcal{H}} + \kappa_2 \, \|\xi\|_{\mathcal{H}} \, \|k\|_{\infty}^3 \\ \le \;& \big(2\lambda + \kappa_2 \, \|k\|_{\infty}^3 \big)\, \delta_1\,, \end{aligned}$$

which shows that $\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})(h - f_{L,\mathrm{P},\lambda})$ lies in a neighborhood of $0 \in \mathcal{H}$, for all $h \in \mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$.

Remark 3.3.2 part (iv). Due to (3.12) we have to prove that

$$d_0 := \inf_{f_1 \ne f_2} \frac{\big\| 2\lambda(f_1 - f_2) + \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L\big(X, Y, f_{L,\mathrm{P},\lambda}(X)\big) \cdot (f_1 - f_2)\Phi(X) \big\|_{\mathcal{H}}}{\|f_1 - f_2\|_{\mathcal{H}}}$$

is strictly positive. If $f_1 \neq f_2$, then (1.34), (3.5), and $\lambda > \frac{1}{2}\,\kappa_2\,\|k\|_\infty^3$ yield that

$$\begin{aligned}
& \frac{\left\|2\lambda(f_1 - f_2) + \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^{B} L\big(X, Y, f_{L,\mathrm{P},\lambda}(X)\big) \cdot (f_1 - f_2)\Phi(X)\right\|_{\mathcal{H}}}{\|f_1 - f_2\|_{\mathcal{H}}} \\
\geq\; & \frac{\|2\lambda(f_1 - f_2)\|_{\mathcal{H}} - \left\|\mathbb{E}_{\mathrm{P}} \nabla_{3,3}^{B} L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot (f_1 - f_2)\Phi(X)\right\|_{\mathcal{H}}}{\|f_1 - f_2\|_{\mathcal{H}}} \\
\geq\; & 2\lambda - \kappa_2\,\|k\|_\infty^3 > 0
\end{aligned}$$

by our assumption, which gives the assertion.

Remark 3.3.2 part (v). The assumptions of Robinson's implicit function theorem, see Theorem 3.2.2, are valid for G due to the results of Remark 3.3.2 parts (i) to (iv) and the assumption that $\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})$ is strong. This gives part (v).

The result of Theorem 3.3.1 now follows from inserting (3.11) and (3.12) into Remark 3.3.2 part (v.4). Using (3.5) we see that S is bounded. The linearity of S follows from its definition and the inverse of S does exist by Theorem 3.2.2. If necessary we can restrict the range of S to $S(\mathcal{H})$ to obtain a bijective function $S_* : \mathcal{H} \to S(\mathcal{H})$ with $S_*(f) = S(f)$ for all $f \in \mathcal{H}$. Hence S^{-1} is also bounded and linear by Theorem A.3.2. This gives the existence of a bounded BIF specified by (3.7) and (3.8). □

3.3.2 Some Examples

In this subsection we will show that our Theorem 3.3.1 covers some SVMs that are widely used in practice. The following result treat SVMs based on the ϵ-insensitive loss function or Huber's loss function for regression, and SVMs based on the pinball loss function for non-parametric quantile regression. These loss functions have uniformly bounded first and second partial B-derivatives, see Subsection 3.2.2.

Corollary 3.3.4. *Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ be closed, and P, Q be distributions on $\mathcal{X} \times \mathcal{Y}$ with $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ and $\mathbb{E}_{\mathrm{Q}}|Y| < \infty$.*

i) For $L \in \{L_{\tau-pin}, L_\epsilon\}$, assume that for all $\delta > 0$ there exist positive constants ξ_{P}, ξ_{Q}, c_{P}, and c_{Q} such that for all $t \in \mathbb{R}$ with $|t - f_{L,\mathrm{P},\lambda}(x)| \leq \delta\,\|k\|_\infty$ the following inequalities hold for all $a \in [0, 2\delta\,\|k\|_\infty]$ and $x \in \mathcal{X}$:

$$\mathrm{P}\big(Y \in [t, t+a] \,\big|\, x\big) \leq c_{\mathrm{P}} a^{1+\xi_{\mathrm{P}}} \text{ and } \mathrm{Q}\big(Y \in [t, t+a] \,\big|\, x\big) \leq c_{\mathrm{Q}} a^{1+\xi_{\mathrm{Q}}}\,. \tag{3.13}$$

ii) For $L = L_{c-Huber}$, assume for $x \in \mathcal{X}$:

$$\begin{aligned} \mathrm{P}\big(Y \in \big\{f_{L,\mathrm{P},\lambda}(x) - c, f_{L,\mathrm{P},\lambda}(x) + c\big\} \,\big|\, x\big) = 0\,, \\ \mathrm{Q}\big(Y \in \big\{f_{L,\mathrm{P},\lambda}(x) - c, f_{L,\mathrm{P},\lambda}(x) + c\big\} \,\big|\, x\big) = 0\,. \end{aligned} \tag{3.14}$$

Then the assumptions of Theorem 3.3.1 are valid: $\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P})$ *of* $T(\mathrm{P}) := f_{L,\mathrm{P},\lambda}$ *exists, is given by (3.7) to (3.8), and is bounded.*

For the proof of Corollary 3.3.4 we need to know the partial B-derivatives for the three loss functions and also have to check that $\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})$ is strong. We have already computed the partial B-derivatives for these loss functions in advance, see Subsection 3.2.2, so it only remains to show that these partial derivatives are strong.

Proof of Corollary 3.3.4. It has been shown that these loss functions have bounded first and second partial B-derivatives, so now we are ready to check if $\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})$ is strong in these cases. Recall that $\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})$ is strong, if for all $\varepsilon^* > 0$ there exist a neighborhood $\mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$ and an interval $(-\delta_2, \delta_2)$ with $\delta_1, \delta_2 > 0$ such that for all $f_1, f_2 \in \mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$ and for all $\varepsilon \in (-\delta_2, \delta_2)$ we have

$$\big\|\big(G(\varepsilon, f_1) - g(f_1)\big) - \big(G(\varepsilon, f_2) - g(f_2)\big)\big\|_{\mathcal{H}} \leq \varepsilon^* \left\|f_1 - f_2\right\|_{\mathcal{H}}\,, \tag{3.15}$$

where, for $f \in \mathcal{H}$,

$$\begin{aligned} g(f) \;=\; & 2\lambda f_{L,\mathrm{P},\lambda}(X) + \mathbb{E}_{\mathrm{P}} \nabla_3^B L\big(X, Y, f_{L,\mathrm{P},\lambda}(X)\big) \cdot \Phi(X) \\ & + 2\lambda\, \mathrm{id}_{\mathcal{H}}(f(X) - f_{L,\mathrm{P},\lambda}(X)) \\ & + \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L\big(X, Y, f_{L,\mathrm{P},\lambda}(X)\big) \cdot \langle (f(X) - f_{L,\mathrm{P},\lambda}(X)), \Phi(X)\rangle_{\mathcal{H}} \Phi(X)\,. \end{aligned}$$

Fix $\varepsilon^* > 0$. Obviously, (3.15) is valid for $f_1 = f_2$. For the rest of the proof we therefore fix arbitrary functions $f_1, f_2 \in \mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$ with $f_1 \neq f_2$. We obtain for the term on the left hand side of (3.15) that

$$\begin{aligned} \Bigg\| \Big(& 2\lambda f_1(X) + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_3^B L(X, Y, f_1(X)) \cdot \Phi(X) \\ & - 2\lambda f_{L,\mathrm{P},\lambda}(X) - \mathbb{E}_{\mathrm{P}} \nabla_3^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \Phi(X) \\ & - 2\lambda(f_1(X) - f_{L,\mathrm{P},\lambda}(X)) \\ & - \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_{L,\mathrm{P},\lambda}(X))\Phi(X)\Big) \\ - \Big(& 2\lambda f_2(X) + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_3^B L(X, Y, f_2(X)) \cdot \Phi(X) \\ & - 2\lambda f_{L,\mathrm{P},\lambda}(X) - \mathbb{E}_{\mathrm{P}} \nabla_3^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot \Phi(X) \\ & - 2\lambda(f_2(X) - f_{L,\mathrm{P},\lambda}(X)) \\ & - \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot (f_2(X) - f_{L,\mathrm{P},\lambda}(X))\Phi(X)\Big) \Bigg\|_{\mathcal{H}} \end{aligned}$$

$$
\begin{aligned}
= \quad & \Big\| \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \Big(\nabla_3^B L(X,Y,f_1(X)) - \nabla_3^B L(X,Y,f_2(X)) \Big) \cdot \Phi(X) \quad (3.16) \\
& \quad - \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X,Y,f_{L,\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X))\Phi(X) \Big\|_{\mathcal{H}} \\
\leq \quad & |1-\varepsilon| \, \Big\| \mathbb{E}_{\mathrm{P}} \Big(\nabla_3^B L(X,Y,f_1(X)) - \nabla_3^B L(X,Y,f_2(X)) \\
& \qquad - \nabla_{3,3}^B L(X,Y,f_{L,\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X)) \Big) \Phi(X) \Big\|_{\mathcal{H}} \\
& + |\varepsilon| \, \Big\| \mathbb{E}_{\mathrm{Q}} \Big(\nabla_3^B L(X,Y,f_1(X)) - \nabla_3^B L(X,Y,f_2(X)) \Big) \cdot \Phi(X) \Big\|_{\mathcal{H}} \\
& + |\varepsilon| \, \Big\| \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L(X,Y,f_{L,\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X))\Phi(X) \Big\|_{\mathcal{H}} \\
=: \quad & |1-\varepsilon| A + |\varepsilon| B + |\varepsilon| C \, . \qquad (3.17)
\end{aligned}
$$

We shall show that (3.17) is bounded from above by $\varepsilon^* \|f_1 - f_2\|_{\mathcal{H}}$. When we look at the first partial B-derivatives of our loss functions, we see that we can separate them in 2 cases: for L_ϵ and $L_{\tau-pin}$ there are one or more discontinuities in $\nabla_3^B L$, whereas $\nabla_3^B L$ is continuous for $L_{c-Huber}$. Recall that the set $\mathfrak{D}$ of points where Lipschitz continuous functions are *not* Fréchet-differentiable, has Lebesgue measure zero by Rademacher's theorem A.2.3. Define then the function

$$
h\big(y, f_1(x), f_2(x)\big) := \nabla_3^B L\big(x, y, f_1(x)\big) - \nabla_3^B L\big(x, y, f_2(x)\big) \, .
$$

For $L \in \{L_\epsilon, L_{\tau-pin}\}$, denote the set of discontinuity points of $\nabla_3^B L$ by $\mathfrak{D}$. Take f_1, $f_2 \in \mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$. For $\nabla_3^B L(X,Y,f_{L,\mathrm{P},\lambda}(x)) \notin \mathfrak{D}$ we obtain $\nabla_3^B L(X,Y,f_1(x)) = \nabla_3^B L(X,Y,f_2(x))$ for sufficiently small δ_1 and hence $h(y,f_1(x),f_2(x)) = 0$. If, on the other hand, $\nabla_3^B L(X,Y,f_{L,\mathrm{P},\lambda}(x)) \in \mathfrak{D}$ and $f_1(x) < f_{L,\mathrm{P},\lambda}(x) < f_2(x)$ or $f_2(x) < f_{L,\mathrm{P},\lambda}(x) < f_1(x)$, then we have that $\nabla_3^B L(X,Y,f_1(x)) \neq \nabla_3^B L(X,Y,f_2(x))$ and hence $h(y,f_1(x),f_2(x)) \neq 0$. Define $m = 2|\mathfrak{D}|$.

Pinball loss

Using the first part of this proof we see that for the pinball loss $L = L_{\tau-pin}$ we obtain $|h(y,f_1(x),f_2(x))| \leq c_1$, with $c_1 = 1$, $\mathfrak{D} = \{0\}$, $m = 2$, and $\nabla_{3,3}^B L(x,y,t) = 0$, for all $t \in \mathbb{R}$. For all $f \in \mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$ we get

$$
|f(x) - f_{L,\mathrm{P},\lambda}(x)| \leq \|f - f_{L,\mathrm{P},\lambda}\|_\infty \leq \|k\|_\infty \, \|f - f_{L,\mathrm{P},\lambda}\|_{\mathcal{H}} \leq \|k\|_\infty \, \delta_1 \, . \qquad (3.18)
$$

Further

$$
|f_1(x) - f_2(x)| \leq \|f_1 - f_2\|_\infty \leq \|k\|_\infty \, \|f_1 - f_2\|_{\mathcal{H}} \leq 2 \, \|k\|_\infty \, \delta_1 \, . \qquad (3.19)
$$

Using (3.18), (3.19), and (3.13) we obtain

$$\begin{aligned}
A &= \left\|\mathbb{E}_{\mathrm{P}}(\nabla_3^B L(X,Y,f_1(X)) - \nabla_3^B L(X,Y,f_2(X)))\cdot\Phi(X)\right\|_{\mathcal{H}} \\
&\le \mathbb{E}_{\mathrm{P}}|h(Y,f_1(X),f_2(X))|\,|\Phi(X)| \\
&\le \|k\|_\infty^2\,\mathbb{E}_{\mathrm{P}}|h(Y,f_1(X),f_2(X))|\mathbf{1}_{\{h\neq 0\}} \\
&\le \|k\|_\infty^2\,c_1\mathrm{P}\big(\nabla_3^B L(X,Y,f_1(X)) \neq \nabla_3^B L(X,Y,f_2(X))\big) \\
&= \|k\|_\infty^2\,\Big(\mathrm{P}\big(\{Y-f_1(X)<0\}\wedge\{Y-f_2(X)>0\}\big) \\
&\qquad\qquad + \mathrm{P}\big(\{Y-f_2(X)<0\}\wedge\{Y-f_1(X)>0\}\big)\Big) \\
&= \|k\|_\infty^2 \int_{\mathcal{X}} \mathrm{P}\big(Y\in(f_2(x),f_1(x))\,|\,x\big)+\mathrm{P}\big(Y\in(f_1(x),f_2(x))\,|\,x\big)d\mathrm{P}_X(x) \\
&= \|k\|_\infty^2 \int_{\mathcal{X}} \mathrm{P}\big(Y\in(f_2(x),f_2(x)+[f_1(x)-f_2(x)])\,|\,x\big) \\
&\qquad\qquad + \mathrm{P}\big(Y\in(f_1(x),f_1(x)+[f_2(x)-f_1(x)])\,|\,x\big)d\mathrm{P}_X(x) \\
&\le m\,\|k\|_\infty^2 \int_{\mathcal{X}} c_{\mathrm{P}}|f_1(x)-f_2(x)|^{1+\xi_{\mathrm{P}}}d\mathrm{P}_X(x) \\
&\le m\,\|k\|_\infty^2\,c_{\mathrm{P}}\,\|f_1-f_2\|_\infty^{1+\xi_{\mathrm{P}}} \\
&\le m\,c_{\mathrm{P}}\,\|k\|_\infty^{3+\xi_{\mathrm{P}}}\,\|f_1-f_2\|_{\mathcal{H}}^{1+\xi_{\mathrm{P}}}\,,
\end{aligned}$$

where P_X denotes the marginal distribution of X. Similar calculations give that $B \le m\,c_{\mathrm{Q}}\,\|k\|_\infty^{3+\xi_{\mathrm{Q}}}\ \|f_1-f_2\|_{\mathcal{H}}^{1+\xi_{\mathrm{Q}}}$. We obtain $C=0$, because $\nabla_{3,3}^B L(X,Y,f_{L,\mathrm{P},\lambda}(X))=0$. Hence, the term in (3.17) is less than or equal to

$$\begin{aligned}
&|1-\varepsilon|m\,c_{\mathrm{P}}\,\|k\|_\infty^{3+\xi_{\mathrm{P}}}\,\|f_1-f_2\|_{\mathcal{H}}^{1+\xi_{\mathrm{P}}} + |\varepsilon|m\,c_{\mathrm{Q}}\,\|k\|_\infty^{3+\xi_{\mathrm{Q}}}\,\|f_1-f_2\|_{\mathcal{H}}^{1+\xi_{\mathrm{Q}}} \\
=\;& \big(|1-\varepsilon|c_{\mathrm{P}}\,\|k\|_\infty^{\xi_{\mathrm{P}}}\,\|f_1-f_2\|_{\mathcal{H}}^{\xi_{\mathrm{P}}} + |\varepsilon|c_{\mathrm{Q}}\,\|k\|_\infty^{\xi_{\mathrm{Q}}}\,\|f_1-f_2\|_{\mathcal{H}}^{\xi_{\mathrm{Q}}}\,\big) \\
&\cdot m\,\|k\|_\infty^3\,\|f_1-f_2\|_{\mathcal{H}} \\
\le\;& \varepsilon^*\|f_1-f_2\|_{\mathcal{H}}\,,
\end{aligned}$$

where $\varepsilon^* = (|1-\varepsilon|c_{\mathrm{P}}\,\|k\|_\infty^{\xi_{\mathrm{P}}}\,2^{\xi_{\mathrm{P}}}\delta_1^{\xi_{\mathrm{P}}} + |\varepsilon|c_{\mathrm{Q}}\,\|k\|_\infty^{\xi_{\mathrm{Q}}}\,2^{\xi_{\mathrm{Q}}}\delta_1^{\xi_{\mathrm{Q}}})m\,\|k\|_\infty^3\,.$

ϵ-insensitive loss
The proof for the ϵ-insensitive loss $L=L_\epsilon$ is analogous to the proof for $L_{\tau-pin}$, but with $c_1=2$, $\mathfrak{D}=\{-\epsilon,+\epsilon\}$, $m=4$ and thus we must consider 4 cases instead of 2 where $h(y,f_1(x),f_2(x))\neq 0$.

Huber loss

For Huber's loss function $L = L_{c-Huber}$ we have $|\nabla^B_{3,3} L(x,y,t)| \leq 1 := c_2$ and $h(y, f_1(x), f_2(x))$ is bounded by $c_1 = 2c$. Let us define

$$\begin{aligned} h^*(y, f_{L,\mathrm{P},\lambda}(x), f_1(x), f_2(x)) \quad := \quad & \nabla^B_3 L(x,y,f_1(x)) - \nabla^B_3 L(x,y,f_2(x)) \\ & - \nabla^B_{3,3} L(x,y,f_{L,\mathrm{P},\lambda}(x)) \cdot (f_1(x) - f_2(x))\,. \end{aligned}$$

Somewhat tedious calculations show that there exist 8 cases where we obtain that $h^*(y, f_{L,\mathrm{P},\lambda}(x), f_1(x), f_2(x)) \neq 0$ and 6 cases for which we get that $h^*(y, f_{L,\mathrm{P},\lambda}(x), f_1(x), f_2(x)) = 0$. In each of the 8 cases, $y - f_{L,\mathrm{P},\lambda}(x) \in \{-c, c\}$ and $|h^*(y, f_{L,\mathrm{P},\lambda}(x), f_1(x), f_2(x))| \leq |f_1(x) - f_2(x)|$. Due to symmetry of the Huber loss function, the calculations are quite similar, therefore we only consider here some cases.

If $-c < Y - f_{L,\mathrm{P},\lambda}(x) < c$, then $\nabla^B_{3,3} L(X,Y,f_{L,\mathrm{P},\lambda}(x)) \cdot (f_1(x) - f_2(x)) = f_1(x) - f_2(x)$ and for sufficiently small δ_1, $\nabla^B_3 L(X,Y,f_1(x)) = -(Y - f_1(x))$ and $\nabla^B_3 L(X,Y,f_2(x)) = -(Y - f_2(x))$. A small calculation shows that $h^*(Y, f_{L,\mathrm{P},\lambda}(x), f_1(x), f_2(x)) = 0$.

Straightforward calculations give us that $h^*(Y, f_{L,\mathrm{P},\lambda}(x), f_1(x), f_2(x)) = 0$ for the following 5 cases:

i) $Y - f_{L,\mathrm{P},\lambda}(x) < -c$ or $Y - f_{L,\mathrm{P},\lambda}(x) > c$,

ii) $Y - f_{L,\mathrm{P},\lambda}(x) = -c$ and $f_{L,\mathrm{P},\lambda}(x) > f_2(x) > f_1(x)$,

iii) $Y - f_{L,\mathrm{P},\lambda}(x) = -c$ and $f_1(x) > f_2(x) > f_{L,\mathrm{P},\lambda}(x)$,

iv) $Y - f_{L,\mathrm{P},\lambda}(x) = c$ and $f_{L,\mathrm{P},\lambda}(x) > f_2(x) > f_1(x)$,

v) $Y - f_{L,\mathrm{P},\lambda}(x) = c$ and $f_1(x) > f_2(x) > f_{L,\mathrm{P},\lambda}(x)$.

For $Y - f_{L,\mathrm{P},\lambda}(x) = -c$ and for $f_1(x) > f_{L,\mathrm{P},\lambda}(x) > f_2(x)$, we obtain that $\nabla^B_3 L(X,Y,f_1(X)) = c$, $\nabla^B_3 L(X,Y,f_2(x)) = -(Y - f_2(x))$ and $\nabla^B_{3,3} L(X,Y,f_{L,\mathrm{P},\lambda}(x)) \cdot (f_1(x) - f_2(x)) = 0$. Hence,

$$h^*(Y, f_{L,\mathrm{P},\lambda}(x), f_1(x), f_2(x)) = c + Y - f_2(x) = f_{L,\mathrm{P},\lambda}(x) - f_2(x) \neq 0\,,$$

since $f_2(x) < f_{L,\mathrm{P},\lambda}(x)$.

Analogously, some calculations show that $h^*(Y, f_{L,\mathrm{P},\lambda}(x), f_1(x), f_2(x)) \neq 0$ for the following 7 cases:

i) $Y - f_{L,\mathrm{P},\lambda}(x) = -c$ and $f_2(x) > f_{L,\mathrm{P},\lambda}(x) > f_1(x)$,

ii) $Y - f_{L,\mathrm{P},\lambda}(x) = -c$ and $f_{L,\mathrm{P},\lambda}(x) > f_1(x) > f_2(x)$,

iii) $Y - f_{L,\mathrm{P},\lambda}(x) = -c$ and $f_2(x) > f_1(x) > f_{L,\mathrm{P},\lambda}(x)$,

iv) $Y - f_{L,\mathrm{P},\lambda}(x) = c$ and $f_1(x) > f_{L,\mathrm{P},\lambda}(x) > f_2(x)$,

v) $Y - f_{L,\mathrm{P},\lambda}(x) = c$ and $f_2(x) > f_{L,\mathrm{P},\lambda}(x) > f_1(x)$,

vi) $Y - f_{L,\mathrm{P},\lambda}(x) = c$ and $f_{L,\mathrm{P},\lambda}(x) > f_1(x) > f_2(x)$,

vii) $Y - f_{L,\mathrm{P},\lambda}(x) = c$ and $f_2(x) > f_1(x) > f_{L,\mathrm{P},\lambda}(x)$.

Using (3.14) in (3.17) we get for the term A in (3.17) that

$$\begin{aligned} A &= \|\mathbb{E}_{\mathrm{P}} h^*(Y, f_{L,\mathrm{P},\lambda}(X), f_1(X), f_2(X))\Phi(X)\|_{\mathcal{H}} \\ &\leq \|k\|_\infty^2 \int |h^*(y, f_{L,\mathrm{P},\lambda}(x), f_1(x), f_2(x))| \mathbf{1}_{\{h^* \neq 0\}} d\mathrm{P}(x,y) \\ &\leq \|k\|_\infty^2 \int |f_1(x) - f_2(x)| \mathrm{P}\big(Y \in \{-c + f_{L,\mathrm{P},\lambda}(x), c + f_{L,\mathrm{P},\lambda}(x)\} \big| x\big) d\mathrm{P}_X(x) \\ &= 0\,. \end{aligned}$$

Also

$$\begin{aligned} C &= \big\|\mathbb{E}_{\mathrm{P}} \nabla^B_{3,3} L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X))\Phi(X)\big\|_{\mathcal{H}} \\ &\leq \kappa_2 \|k\|_\infty^3 \|f_1 - f_2\|_{\mathcal{H}}\,. \end{aligned}$$

One can compute the analogous terms to A and C, say $A(\mathrm{Q})$ and $C(\mathrm{Q})$, respectively, where the integration is with respect to Q instead of P. Combining these expressions we obtain

$$\begin{aligned} B &= \big\|\mathbb{E}_{\mathrm{Q}} (\nabla^B_3 L(X, Y, f_1(X)) - \nabla^B_3 L(X, Y, f_2(X))) \cdot \Phi(X)\big\|_{\mathcal{H}} \\ &\leq \mathbb{E}_{\mathrm{Q}} \big|\nabla^B_3 L(X, Y, f_1(X)) - \nabla^B_3 L(X, Y, f_2(X)) - \\ &\qquad \nabla^B_{3,3} L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X))\big| \, |\Phi(X)| \\ &\quad + \mathbb{E}_{\mathrm{Q}} \big|\nabla^B_{3,3} L(X, Y, f_{L,\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X))\big| \, |\Phi(X)| \\ &= A(\mathrm{Q}) + C(\mathrm{Q}) \;\leq\; \kappa_2 \|k\|_\infty^3 \|f_1 - f_2\|_{\mathcal{H}}\,. \end{aligned}$$

Hence, the term in (3.17) is less than or equal to $\varepsilon^* \|f_1 - f_2\|_{\mathcal{H}}$ where $\varepsilon^* = 2|\varepsilon|\kappa_2 \|k\|_\infty^3$. This gives the assertion, because $|\varepsilon|$ can be chosen arbitrarily small. □

For the somewhat smoother Huber loss function we only need to exclude by (3.14) that the conditional probabilities of Y given X with respect to P and Q have no point probabilities at the two points $f_{L,\mathrm{P},\lambda}(x) - c$ and

$f_{L,\mathrm{P},\lambda}(x)+c$. Therefore, for this loss function Q can be a Dirac distribution and in this case we have BIF = IF.

For the pinball loss function some calculations give

$$\begin{aligned}\mathrm{BIF}(\mathrm{Q};T,\mathrm{P}) &= \frac{1}{2\lambda}\int_{\mathcal{X}}\big(\mathrm{P}\big(Y \le f_{L,\mathrm{P},\lambda}(x)\,\big|\,x\big)-\tau\big)\Phi(x)\,d\mathrm{P}_X(x)\\ &\quad -\frac{1}{2\lambda}\int_{\mathcal{X}}\big(\mathrm{Q}\big(Y \le f_{L,\mathrm{P},\lambda}(x)\,\big|\,x\big)-\tau\big)\Phi(x)\,d\mathrm{Q}_X(x)\,,\end{aligned}$$

if the BIF exists. We expect the first integral to be small, because $f_{L,\mathrm{P},\lambda}(x)$ approximates the τ-quantile of $\mathrm{P}(\,\cdot\,|x)$ and even rates of convergence are known (Steinwart and Christmann, 2008a,b). As seen in the proof, (3.13) and (3.14) guarantee that the regular conditional probabilities $\mathrm{P}(\,\cdot\,|x)$ and $\mathrm{Q}(\,\cdot\,|x)$ do not have large point masses at those points where the Lipschitz continuous loss function L is *not* F-differentiable or in small neighborhoods around these points. Even for the case of parametric quantile regression, that is for $L = L_{\tau-pin}$, $\lambda = 0$ and the unbounded linear kernel $k(x,x') := \langle x,x'\rangle$, some assumptions on the distribution P seem to be necessary for the existence of the IF, see Koenker (2005, p. 44). He assumes that P has a continuous density which is strictly positive where needed.

Nevertheless, the question arises whether we can proof Theorem 3.3.1 and Corollary 3.3.4 without any assumption on the distributions P and Q. This is—at least with the techniques we used—not possible for non-smooth loss functions as the following counterexample shows.

Let us consider kernel based quantile regression based on the Gaussian RBF kernel, that is $L = L_{\tau-pin}$, $k = k_{RBF}$, and $\lambda > 0$. Hence the set $\mathfrak{D}$ of discontinuity points of $\nabla_3^B L$ is $\mathfrak{D} = \{0\}$. Fix $x \in \mathcal{X}$ and $y, y^* \in \mathcal{Y}$ with $y \neq y^*$. Define $\mathrm{P} = \delta_{(x,y)}$ and $\mathrm{Q} = \delta_{(x,y^*)}$. Consider $f_1, f_2 \in \mathcal{N}_{\delta_1}(f_{L,\mathrm{P},\lambda})$ with $f_1(x) \neq f_2(x)$, $y - f_1(x) > 0$, $y - f_2(x) < 0$, $y^* - f_1(x) > 0$, and $y^* - f_2(x) < 0$. Hence, $\nabla_3^B L(x,y,f_1(x)) = \nabla_3^B L(x,y^*,f_1(x)) = -\tau$ and $\nabla_3^B L(x,y,f_2(x)) = \nabla_3^B L(x,y^*,f_2(x)) = 1-\tau$. Note that $\nabla_{3,3}^B L(x,y,t) = 0$ for all $y,t \in \mathbb{R}$. We thus obtain for the $\mathcal{H}$-norm in (3.16) that

$$\begin{aligned}&\big\|\mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}}\big(\nabla_3^B L(X,Y,f_1(X)) - \nabla_3^B L(X,Y,f_2(X))\big)\cdot\Phi(X)\big\|_{\mathcal{H}}\\ = \;&\|\Phi(x)\|_{\mathcal{H}} > 0\,.\end{aligned}$$

Hence $\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})$ is not strong in this special case, because $\|\Phi(x)\|_{\mathcal{H}}$ is in general greater than $\varepsilon^*\,\|f_1 - f_2\|_{\mathcal{H}}$ for arbitrarily small values of ε^*.

Now we shall show for L_{r-log} that the assumptions (3.13) or (3.14) are not needed to obtain a bounded BIF. It is easy to see that L_{r-log} is strictly convex and Fréchet-differentiable with

$$\begin{aligned}\nabla_3^F L_{r-log}(x,y,t) &= 1-2\Lambda(y-t)\,,\\ \nabla_{3,3}^F L_{r-log}(x,y,t) &= 2\Lambda(y-t)[1-\Lambda(y-t)]\end{aligned}$$

and

$$\nabla_{3,3,3}^F L_{r-log}(x,y,t) = -2\Lambda(y-t)[1-\Lambda(y-t)][1-2\Lambda(y-t)]$$

with $\Lambda(y-t) = 1/\big(1+e^{-(y-t)}\big)$. Obviously, these partial derivatives are bounded for all $y,t \in \mathbb{R}$. Furthermore,

$$\kappa_1 = \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |\nabla_3^F L_{r-log}(x,y,\,\cdot\,)|_1 = 1$$

and

$$\kappa_2 = \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |\nabla_{3,3}^F L_{r-log}(x,y,\,\cdot\,)|_1 \le 1/2\,,$$

because an everywhere F-differentiable function g is Lipschitz continuous with $|g|_1 = ||\nabla^F g||_\infty$ if $\nabla^F g$ is bounded.

Corollary 3.3.5. *Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ be closed, $L = L_{r-log}$, and P,Q be distributions on $\mathcal{X}\times\mathcal{Y}$ with $\mathbb{E}_\mathrm{P}|Y| < \infty$ and $\mathbb{E}_\mathrm{Q}|Y| < \infty$. Then the assumptions of Theorem 3.3.1 are valid, and $\mathrm{BIF}(\mathrm{Q};T,\mathrm{P})$ of $T(\mathrm{P}) := f_{L,\mathrm{P},\lambda}$ exists, is given by (3.7) to (3.8), and $\mathrm{BIF}(\mathrm{Q};T,\mathrm{P})$ is bounded.*

Proof of Corollary 3.3.5. Both partial F-derivatives $\nabla_3^F L_{r-log}(x,y,t) = 1-2\Lambda(y-t)$ and $\nabla_{3,3}^F L_{r-log}(x,y,t) = 2\Lambda(y-t)[1-\Lambda(y-t)]$ are clearly bounded, because $\Lambda(z) \in (0,1)$, $z \in \mathbb{R}$. Thus it only remains to show that $\nabla_2^B G(0, f_{L,\mathrm{P},\lambda})$ is strong for $L = L_{r-log}$, that is that the term in (3.16) is bounded by $\varepsilon^* \|f_1 - f_2\|_\mathcal{H}$ for arbitrary chosen $\varepsilon^* > 0$. A Taylor expansion gives for arbitrary $y, t_1, t_2 \in \mathbb{R}$ that

$$\Lambda(y-t_2) = \Lambda(y-t_1) + (t_1-t_2)\Lambda(y-t_1)\big(1-\Lambda(y-t_1)\big) + O((t_1-t_2)^2)\,. \quad (3.20)$$

Combining (1.34), (3.18), (3.19), and (3.20) we obtain

$$\begin{aligned}&\big|\mathbb{E}_\mathrm{P}\big(\nabla_3^B L(X,Y,f_1(X)) - \nabla_3^B L(X,Y,f_2(X))\\ &\qquad -\nabla_{3,3}^B L(X,Y,f_{L,\mathrm{P},\lambda})\cdot(f_1(X)-f_2(X))\big)\Phi(X)\big|\\ \le\;& 2\,\|k\|_\infty^2\,\mathbb{E}_\mathrm{P}\big|\Lambda(Y-f_2(X)) - \Lambda(Y-f_1(X))\\ &\qquad -\Lambda(Y-f_{L,\mathrm{P},\lambda}(X))(1-\Lambda(Y-f_{L,\mathrm{P},\lambda}(X))\\ &\qquad \cdot\big(f_1(X)-f_2(X)\big)\big|\end{aligned}$$

$$\begin{aligned}
&\leq \; 2\,\|k\|_\infty^2\,\mathbb{E}_{\mathrm{P}}\big|\big(f_1(X)-f_2(X)\big)\big[\Lambda(Y-f_1(X))(1-\Lambda(Y-f_1(X))) \\
&\qquad\qquad -\Lambda(Y-f_{L,\mathrm{P},\lambda}(X))(1-\Lambda(Y-f_{L,\mathrm{P},\lambda}(X)))\big] \\
&\qquad\qquad +O((f_1(X)-f_2(X))^2)\big| \\
&\leq \; 2\,\|k\|_\infty^2\,\mathbb{E}_{\mathrm{P}}\big(\|f_1-f_2\|_\infty\,\big|\Lambda(Y-f_1(X))(1-\Lambda(Y-f_1(X))) \qquad (3.21)\\
&\qquad\qquad -\Lambda(Y-f_{L,\mathrm{P},\lambda}(X))(1-\Lambda(Y-f_{L,\mathrm{P},\lambda}(X)))\big| \\
&\qquad\qquad +c_3\,\|f_1-f_2\|_\infty^2\big)\,.
\end{aligned}$$

A Taylor expansion around $f_{L,\mathrm{P},\lambda}(x)$ shows that $\Lambda(y-f_1(x))(1-\Lambda(y-f_1(x)))$ equals

$$\begin{aligned}
&\Lambda(y-f_{L,\mathrm{P},\lambda}(x))(1-\Lambda(y-f_{L,\mathrm{P},\lambda}(x))) \\
+\;&\big(f_{L,\mathrm{P},\lambda}(x)-f_1(x)\big)\Lambda(y-f_{L,\mathrm{P},\lambda}(x))(1-\Lambda(y-f_{L,\mathrm{P},\lambda}(x))) \\
&\cdot\,(1-2\Lambda(y-f_{L,\mathrm{P},\lambda}(x))) \\
+\;&O((f_1(x)-f_{L,\mathrm{P},\lambda}(x))^2)\,.
\end{aligned}$$

Using this expansion and (1.34), (3.18), and (3.19) it follows that the term in (3.21) is bounded by

$$\begin{aligned}
&2\,\|k\|_\infty^2\,\mathbb{E}_{\mathrm{P}}\big(\|f_1-f_2\|_\infty\,(\frac{\|f_1-f_{L,\mathrm{P},\lambda}\|_\infty}{4}+c_4\delta_1^2\,\|k\|_\infty^2)+c_3\,\|f_1-f_2\|_\infty^2\big) \\
\leq\;&\|k\|_\infty^4\,\big(\delta_1/2+2c_4\delta_1^2\,\|k\|_\infty+4c_3\delta_1\big)\;\|f_1-f_2\|_{\mathcal{H}}\,. \qquad (3.22)
\end{aligned}$$

Using the Lipschitz continuity of $\nabla_3^B L(x,y,\,\cdot\,)$, (1.34), and (3.20) we obtain

$$\begin{aligned}
&|\varepsilon|\,\mathbb{E}_{\mathrm{Q}-\mathrm{P}}\big|\big(\nabla_3^B L(X,Y,f_1(X))-\nabla_3^B L(X,Y,f_2(X))\big)\cdot\Phi(X)\big| \\
\leq\;&|\varepsilon|\,\|k\|_\infty^2\,\mathbb{E}_{|\mathrm{Q}-\mathrm{P}|}\big|\nabla_3^B L(X,Y,f_1(X))-\nabla_3^B L(X,Y,f_2(X))\big| \\
\leq\;&|\varepsilon|\,\|k\|_\infty^3\,\|f_1-f_2\|_{\mathcal{H}}\,. \qquad (3.23)
\end{aligned}$$

Combining (3.22) and (3.23) shows that the term in (3.16) is bounded by $\varepsilon^*\,\|f_1-f_2\|_{\mathcal{H}}$ with the positive constant

$$\varepsilon^* = \|k\|_\infty^3\,\big(\delta_1\,\|k\|_\infty/2+2c_4\delta_1^2\,\|k\|_\infty^2+4c_3\delta_1\,\|k\|_\infty+|\varepsilon|\big)\,,$$

where $\delta_1>0$ and $\varepsilon>0$ can be chosen as small as necessary. □

Corollary 3.3.5 is of course also valid for empirical distributions D_n and Q_m consisting of n and m data points, because no specific assumptions on P and Q are made.

The influence function of $T(\mathrm{P}) = f_{L,\mathrm{P},\lambda}$ based on L_{r-log} and error bounds of the type

$$\left\| T\big((1-\varepsilon)\mathrm{P} + \varepsilon\delta_{(x,y)} - T(\mathrm{P})\big) \right\|_{\mathcal{H}} \leq c^{*}\,\varepsilon$$

where the constant c^{*} is known and depends only on P, $\mathrm{Q} := \delta_{(x,y)}$, and λ, were recently derived by Christmann and Steinwart (2007). We like to mention that Corollary 3.3.5 shows that this influence function is even a Bouligand-derivative, hence it is *positive homogeneous* in $h = \varepsilon(\mathrm{Q}-\mathrm{P})$. Therefore, we immediately obtain from the existence of the BIF that the asymptotic bias of SVMs has the form

$$\begin{aligned} f_{L,(1-\alpha\varepsilon)\mathrm{P}+\alpha\,\varepsilon\mathrm{Q},\,\lambda} - f_{L,\mathrm{P},\lambda} &= T(\mathrm{P}+\alpha h) - T(\mathrm{P}) \\ &= \alpha\,\mathrm{BIF}(\mathrm{Q};T,\mathrm{P}) + o(\alpha h) \qquad (3.24) \\ &= \alpha\big(T(\mathrm{P}+h) - T(\mathrm{P}) + o(h)\big) + o(\alpha h) \\ &= \alpha\big(f_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\,\lambda} - f_{L,\mathrm{P},\lambda}\big) + o(\alpha\varepsilon(\mathrm{Q}-\mathrm{P}))\,, \end{aligned}$$

for $\alpha \geq 0$. This equation nicely describes the behavior of the asymptotic bias term $f_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\,\lambda} - f_{L,\mathrm{P},\lambda}$ if we consider the amount $\alpha\varepsilon$ of contamination instead of ε.

3.4 Robustness of SVMs Based on Shifted Loss Functions

3.4.1 Robustness of SVMs

Let us now consider robustness properties of SVMs based on shifted loss functions. To this end, define the function

$$T : \mathcal{M}_1(\mathcal{X}\times\mathcal{Y}) \to \mathcal{H}, \quad T(\mathrm{P}) := f_{L^{\star},\mathrm{P},\lambda}\,.$$

The same remarks as stated at the beginning of Section 3.3 apply. We will first give a result for the influence function of such SVMs and afterwards show a similar result for the Bouligand influence function.

Theorem 3.4.1 (Influence function)**.** *Let $\mathcal{X}$ be a complete separable metric space and $\mathcal{H}$ be an RKHS of a bounded continuous kernel k. Let L be a convex, Lipschitz continuous loss function with continuous partial Fréchet-derivatives $\nabla_3^F L(x,y,\cdot\,)$ and $\nabla_{3,3}^F L(x,y,\cdot\,)$ which are bounded by*

$$\begin{aligned} \kappa_1 &:= \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\| \nabla_3^F L(x,y,\cdot\,)\right\|_{\infty} \in (0,\infty) \\ \kappa_2 &:= \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\| \nabla_{3,3}^F L(x,y,\cdot\,)\right\|_{\infty} < \infty\,. \end{aligned} \qquad (3.25)$$

Then, for all probability measures P *on* $\mathcal{X}\times\mathcal{Y}$ *and for all* $z := (x,y) \in \mathcal{X}\times\mathcal{Y}$, *the influence function* $\mathrm{IF}(z;T,\mathrm{P})$ *of* $T(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda}$ *exists, is bounded, and equals*

$$\begin{aligned}&\mathbb{E}_{\mathrm{P}}\nabla_3^F L^\star\big(X,Y,f_{L^\star,\mathrm{P},\lambda}(X)\big)S^{-1}\Phi(X)\\ -&\nabla_3^F L^\star\big(x,y,f_{L^\star,\mathrm{P},\lambda}(x)\big)S^{-1}\Phi(x)\,,\end{aligned} \tag{3.26}$$

where $S:\mathcal{H}\to\mathcal{H}$ *is the Hessian of the regularized risk and is given by*

$$S(\,\cdot\,) := 2\lambda\,\mathrm{id}_{\mathcal{H}}(\,\cdot\,) + \mathbb{E}_{\mathrm{P}}\nabla_{3,3}^F L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X))\langle\Phi(X),\,\cdot\,\rangle_{\mathcal{H}}\Phi(X)\,. \tag{3.27}$$

Proof of Theorem 3.4.1. Let $z=(x,y)\in\mathcal{X}\times\mathcal{Y}$. The two key ingredients of our analysis are the function $G:\mathbb{R}\times\mathcal{H}\to\mathcal{H}$ defined by

$$G(\varepsilon,f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z}\nabla_3^F L^\star(X,Y,f(X))\Phi(X)\,, \tag{3.28}$$

and the application of the implicit function theorem for Fréchet-derivatives A.3.11. Let us first check that G is well-defined. Recall that every function $f\in\mathcal{H}$ is bounded because we assumed that $\mathcal{H}$ has a bounded kernel k. By using (1.44) and (3.25) we get $\mathbb{E}_{\mathrm{P}}|\nabla_3^F L^\star(X,Y,f(X))| \le \kappa_1 \in (0,\infty)$ for all $f\in\mathcal{H}$. As $\Phi(x) := k(x,\,\cdot\,)\in\mathcal{H}$ for all $x\in\mathcal{X}$, we obtain that $\Phi:\mathcal{X}\to\mathcal{H}$ is a bounded mapping. Therefore, the $\mathcal{H}$-valued (Bochner) integral used in the definition of G is well-defined for all $\varepsilon\in\mathbb{R}$ and all $f\in\mathcal{H}$. Note that for $\varepsilon\notin[0,1]$ the $\mathcal{H}$-valued integral in (3.28) is with respect to a signed measure. As in Christmann and Steinwart (2007) we obtain for $\varepsilon\in[0,1]$ the equation

$$G(\varepsilon,f) = \frac{\partial\mathcal{R}^{reg}_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}}{\partial\mathcal{H}}(f) = \nabla_2^F\mathcal{R}^{reg}_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}(f)\,. \tag{3.29}$$

Given an $\varepsilon\in[0,1]$, the function $f\mapsto\mathcal{R}^{reg}_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}(f)$ is convex and continuous (see the proof of Theorem 1.7.7) and hence (3.29) shows that $G(\varepsilon,f)=0$ if and only if $f=f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z,\lambda}$. Our aim is to show the existence of a Fréchet-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

$$\mathrm{IF}(z;T,\mathrm{P}) = \nabla^F f_\varepsilon(0)\,.$$

For the existence of $\varepsilon\mapsto f_\varepsilon$ we have to check by Theorem A.3.11 that G is continuously differentiable and that $\nabla_2^F G(0,f_{L^\star,\mathrm{P},\lambda})$ is invertible. Let us start with the first. By the definition of G and by using $\nabla_3^F L^\star(x,y,\,\cdot\,) = \nabla_3^F L(x,y,\,\cdot\,)$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, we get

$$\begin{aligned} & \nabla_1^F G(\varepsilon, f) \qquad (3.30)\\ = \ & -\mathbb{E}_{\mathrm{P}} \nabla_3^F L^\star(X,Y,f(X))\Phi(X) + \nabla_3^F L^\star(x,y,f(x))\Phi(x) \\ = \ & -\mathbb{E}_{\mathrm{P}} \nabla_3^F L(X,Y,f(X))\Phi(X) + \nabla_3^F L(x,y,f(x))\Phi(x)\,. \end{aligned}$$

A similar, but slightly more involved computation using (1.44) and (3.27) yields

$$\begin{aligned} & \nabla_2^F G(\varepsilon, f) \qquad (3.31)\\ = \ & \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\delta_z} \nabla_{3,3}^F L(X,Y,f(X)) \langle \Phi(X),\,\cdot\,\rangle_{\mathcal{H}} \Phi(X) \\ & +2\lambda \mathrm{id}_{\mathcal{H}}\,, \end{aligned}$$

which equals S. To prove that $\nabla_1^F G$ is continuous, we fix $\varepsilon \in \mathbb{R}$ and a sequence $(f_n)_{n\in\mathbb{N}}$ such that $f_n \in \mathcal{H}$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} f_n = f \in \mathcal{H}$. Since k is bounded, the sequence $(f_n)_{n\in\mathbb{N}}$ is uniformly bounded. By (3.25), we have, for all $(x,y,t) \in \mathcal{X}\times\mathcal{Y}\times\mathbb{R}$, that $|\nabla_3^F L(x,y,t)| \leq \kappa_1 + |t|$. Hence $|\nabla_3^F L|$ is a P-integrable Nemitski loss function for *all* probability measures P, because we only have to choose the constant function $b(x,y) \equiv \kappa_1$ in the definition of a P-integrable Nemitski loss defined in Subsection 1.4.1.

We can thus find a bounded, measurable function $g : \mathcal{X} \to \mathbb{R}$ with $|\nabla_3^F L^\star(x,y,f_n(x))| \leq |\nabla_3^F L^\star(x,y,g(x))|$ for all $n \in \mathbb{N}$ and all $(x,y) \in \mathcal{X}\times\mathcal{Y}$. For the function $v : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ with $v(x,y) := L^\star(x,y,g(x))$, we hence obtain by the definition of $L^\star$ and by the Lipschitz continuity of L that

$$\begin{aligned} & \int_{\mathcal{X}\times\mathcal{Y}} |v(X,Y)|\, d\mathrm{P} \\ = \ & \int_{\mathcal{X}\times\mathcal{Y}} |L(X,Y,g(X)) - L(X,Y,0)|\, d\mathrm{P} \leq |L|_1\, \|g\|_\infty \end{aligned}$$

is finite for *all* $\mathrm{P} \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$. Thus, an application of the dominated convergence theorem for Bochner integrals, see Theorem A.3.4, gives the continuity of $\nabla_1^F G$. Because the continuity of G and $\nabla_2^F G$ can be shown analogously, we obtain that G is continuously differentiable, see Theorem A.3.10.

To show that $\nabla_2^F G(0, f_{L^\star,\mathrm{P},\lambda})$ is invertible, it suffices by the Fredholm alternative (see Theorem A.3.1) to show that $\nabla_2^F G(0, f_{L^\star,\mathrm{P},\lambda})$ is injective and that

$$Ag := \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^F L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X)) g(X) \Phi(X)\,, \quad g \in \mathcal{H}\,,$$

defines a compact operator on $\mathcal{H}$. To show the compactness of the operator A, recall that $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{X} \times \mathcal{Y}$ are Polish spaces because $\mathcal{X}$ is a complete separable metric space and $\mathcal{Y} \subset \mathbb{R}$ is closed, see Dudley (2002). Furthermore, Borel probability measures on Polish spaces are regular by Ulam's theorem (Theorem A.2.6), that is, they can be approximated from inside by compact sets. Hence, there exists a sequence of measurable compact subsets $\mathcal{X}_n \times \mathcal{Y}_n \subset \mathcal{X} \times \mathcal{Y}$ with $\mathrm{P}(\mathcal{X}_n \times \mathcal{Y}_n) \geq 1 - \frac{1}{n}$, $n \in \mathbb{N}$. Let us also define a sequence of operators $A_n : \mathcal{H} \to \mathcal{H}$, where $A_n g$ equals

$$\int_{\mathcal{X}_n} \int_{\mathcal{Y}_n} \nabla^F_{3,3} L^\star(x, y, f_{L^\star,\mathrm{P},\lambda}(x))\, \mathrm{P}(dy|x)\, g(x)\Phi(x)\, d\mathrm{P}_X(x)$$

for all $g \in \mathcal{H}$. Note that if $\mathcal{X} \times \mathcal{Y}$ is compact, we can choose $\mathcal{X}_n \times \mathcal{Y}_n := \mathcal{X} \times \mathcal{Y}$, which implies $A = A_n$. Let us now show that A_n, $n \geq 1$, is a compact operator. To this end we assume without loss of generality that $\|k\|_\infty \leq 1$. Denote the closed unit ball in $\mathcal{H}$ by $B_{\mathcal{H}}$. For $g \in B_{\mathcal{H}}$ and $x \in \mathcal{X}$, we have due to the assumption (3.25) that

$$\begin{aligned} h_g(x) &:= \int_{\mathcal{Y}_n} \nabla^F_{3,3} L^\star(x, y, f_{L^\star,\mathrm{P},\lambda}(x))\, |g(x)|\ \mathrm{P}(dy|x) \\ &\leq \kappa_2\, \|g\|_\infty =: h(x)\,. \end{aligned}$$

Therefore, we have $h \in L_1(\mathrm{P})$, which implies $h_g \in L_1(\mathrm{P})$ with $\|h_g\|_1 \leq \|h\|_1 < \infty$ for all $g \in B_{\mathcal{H}}$. Consequently, $\mu_g := h_g \mathrm{P}_X$ and $\mu := h\mathrm{P}_X$ are finite measures. By Theorem A.3.5 we hence obtain

$$\begin{aligned} A_n g &:= \int_{\mathcal{X}_n} \operatorname{sign} g(x)\Phi(x) h_g(x)\, d\mathrm{P}_X(x) \\ &= \int_{\mathcal{X}_n} \operatorname{sign} g(x)\Phi(x)\, d\mu_g(x) \\ &\in \mu_g(\mathcal{X}_n)\, \overline{\operatorname{aco}\Phi(X_n)} \subset \mu(\mathcal{X}_n)\, \overline{\operatorname{aco}\Phi(X_n)}\,, \quad g \in \mathcal{H}\,, \end{aligned}$$

where $\operatorname{aco}\Phi(X_n)$ denotes the absolute convex hull of $\Phi(X_n)$, and the closure is with respect to $\|\cdot\|_{\mathcal{H}}$. The continuity of k yields the continuity of the canonical feature map Φ. Thus, $\Phi(X_n)$ is compact and hence so is the closure of $\operatorname{aco}\Phi(X_n)$. This shows that A_n is a compact operator.

To see that A is compact, it therefore suffices to show $\|A_n - A\| \to 0$ with respect to the operator norm for $n \to \infty$. Recalling that the convexity of $L^\star$ and the existence of its second derivative implies $\nabla^F_{3,3} L^\star(x, y, \cdot\,) \geq 0$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, it follows from (3.25) that

$$0 \leq \int \nabla^F_{3,3} L^\star(x, y, f_{L^\star,\mathrm{P},\lambda}(x))\, d\mathrm{P}(x, y) \leq \kappa_2\,,$$

which shows due to (1.44) that

$$\nabla^F_{3,3} L^\star(\,\cdot\,,\,\cdot\,, f_{L^\star,\mathrm{P},\lambda}(\,\cdot\,)) = \nabla^F_{3,3} L(\,\cdot\,,\,\cdot\,, f_{L^\star,\mathrm{P},\lambda}(\,\cdot\,)) \in L_\infty(\mathrm{P})$$

for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$. Now define $B := (\mathcal{X} \times \mathcal{Y}) \backslash (\mathcal{X}_n \times \mathcal{Y}_n)$. Then the desired convergence follows from (1.34), $\mathrm{P}(\mathcal{X}_n \times \mathcal{Y}_n) \geq 1 - \frac{1}{n}$, and

$$\begin{aligned} & \|A_n g - Ag\|_{\mathcal{H}} \\ \leq\ & \int_B \nabla^F_{3,3} L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x))\, |g(x)|\ \|\Phi(x)\|_{\mathcal{H}}\ d\mathrm{P}(x,y) \\ \leq\ & \|g\|_\infty \, \|\Phi(x)\|_{\mathcal{H}} \int_B \nabla^F_{3,3} L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x))\, d\mathrm{P}(x,y) \\ \leq\ & \frac{\kappa_2\ \|g\|_{\mathcal{H}}\, \|k\|^3_\infty}{n}\,. \end{aligned}$$

Let us now show that $\nabla^F_2 G(0, f_{L^\star,\mathrm{P},\lambda}) = 2\lambda \mathrm{id}_{\mathcal{H}} + A$ is injective. To this end, let us choose $g \in \mathcal{H}\backslash\{0\}$. Then we find

$$\begin{aligned} & \langle (2\lambda \mathrm{id}_{\mathcal{H}} + A)g, (2\lambda \mathrm{id}_{\mathcal{H}} + A)g \rangle_{\mathcal{H}} \\ >\ & 4\lambda \, \langle g, Ag \rangle_{\mathcal{H}} \\ =\ & 4\lambda \, \mathbb{E}_{\mathrm{P}} \nabla^F_{3,3} L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X)) g^2(X) \geq 0\,, \end{aligned}$$

which shows the injectivity. The implicit function Theorem A.3.11 for Fréchet-derivatives guarantees that $\varepsilon \mapsto f_\varepsilon$ is differentiable on $(-\delta, \delta)$ if $\delta > 0$ is small enough. Furthermore, (3.30) and (3.31) yield, for all $z = (x,y) \in \mathcal{X} \times \mathcal{Y}$, that

$$\begin{aligned} & \mathrm{IF}(z;T,\mathrm{P}) \ = \ \nabla^F f_\varepsilon(0) \\ =\ & -S^{-1} \circ \nabla^F_1 G(0, f_{L^\star,\mathrm{P},\lambda}) \\ =\ & S^{-1}\big(\mathbb{E}_{\mathrm{P}}\big(\nabla^F_3 L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X))\Phi(X)\big)\big) \\ & -\nabla^F_3 L^\star(x,y,f_{L^\star,\mathrm{P},\lambda}(x)) S^{-1}\Phi(x)\,, \end{aligned}$$

which yields the existence of the influence function and (3.26). The boundedness follows from (3.25) and (3.26). □

The Lipschitz continuity of L already guarantees $\kappa_1 < \infty$. Some calculations for the logistic loss functions defined in (1.28) and (1.27) give $(\kappa_1, \kappa_2) = (1, \frac{1}{4})$ for classification and $(\kappa_1, \kappa_2) = (1, \frac{1}{2})$ for regression.

Remark 3.4.2. *(i) Note that only the second term of* IF$(z;T,\mathrm{P})$ *in (3.26) depends on* z*, where the contamination of* P *occurs. (ii) All assumptions of*

Theorem 3.4.1 can be verified without *knowledge of* P*, which is not true for Steinwart and Christmann (2008b, Theorem 10.18). It is easy to check that the assumptions of Theorem 3.4.1 on L are fulfilled, e.g., for the logistic loss functions for classification and for regression defined in (1.28) and (1.27). The Gaussian RBF kernel defined in (1.30) is bounded and continuous.*

Unfortunately, the previously mentioned conditions on the existence of partial Fréchet-derivatives of the loss function are *not* fulfilled for some losses that are often used in practice, such as the ϵ-insensitive loss or the pinball loss.

The next result shows that the $\mathcal{H}$-norm of the difference $f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{L^\star,\mathrm{P},\lambda}$, the bias of the SVM, increases at most linearly with the radius $\varepsilon \in [0,1]$ of a mixture contamination neighborhood around P. We denote the norm of total variation of a signed measure μ by $\|\mu\|_{\mathcal{M}}$.

Theorem 3.4.3 (Bounds for bias)**.** *Let L be a convex and Lipschitz continuous loss function and let $\mathcal{H}$ be a separable RKHS of a bounded and measurable kernel k. Then, for all $\lambda > 0$, all $\varepsilon \in [0,1]$, and all probability measures* P *and* Q *on $\mathcal{X} \times \mathcal{Y}$, we have*

$$\left\| f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{L^\star,\mathrm{P},\lambda} \right\|_{\mathcal{H}} \leq c_{\mathrm{P},\mathrm{Q}}\, \varepsilon \,,$$

where

$$c_{\mathrm{P},\mathrm{Q}} = \lambda^{-1} \, \|k\|_\infty \, |L|_1 \, \|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}} \,.$$

Let $\mathrm{Q} = \delta_z$ *be the Dirac measure in* $z = (x,y) \in \mathcal{X} \times \mathcal{Y}$*. If the influence function of* $T(\mathrm{P}) = f_{L^\star,\mathrm{P},\lambda}$ *exists, then*

$$\|\mathrm{IF}(z;T,\mathrm{P})\|_{\mathcal{H}} \leq c_{\mathrm{P},\delta_z} \,.$$

Proof of Theorem 3.4.3. Theorem 1.7.9 guarantees the existence of a bounded measurable function $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $\|h\|_\infty \leq |L|_1$ and

$$\left\| f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} \right\|_{\mathcal{H}} \leq \frac{\varepsilon}{\lambda} \left\| \mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\mathrm{Q}} h\Phi \right\|_{\mathcal{H}} \,.$$

From (1.49) we get

$$\begin{aligned} & \left\| f_{L^\star,\mathrm{P},\lambda} - f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} \right\|_{\mathcal{H}} \\ \leq \;& \frac{\varepsilon}{\lambda} \left\| \mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\mathrm{Q}} h\Phi \right\|_{\mathcal{H}} \leq \frac{1}{\lambda} \|h\|_\infty \|k\|_\infty \|\mathrm{P}-\mathrm{Q}\|_{\mathcal{M}}\, \varepsilon \,, \end{aligned}$$

which gives the assertion. □

Remark that the upper bounds for the bias and the influence function are proportional to λ^{-1}, and thus the bounds will go to infinity for $\lambda \to 0$, which is unfortunate. However, please note that (i) these are only bounds that might help to estimate the bias. (ii) The goodness of these bounds is unknown, and will depend on the distribution P. (iii) Due to the no-free-lunch theorem (Devroye, 1982, Devroye *et al.*, 1996, Theorem 7.2), there will always exist an arbitrary P such that the average risk converges arbitrarily slow to the Bayes risk. Therefore there is no learning method that learns with a uniform rate and confidence for all distributions. (iv) There seems to be a conflict in goals between the universal consistency and qualitative robustness (Hable and Christmann, 2011). The authors showed that SVMs are qualitatively robust for *fixed* regularization parameters $\lambda \in (0, \infty)$. However, if the fixed λ is replaced by a null sequence of parameters $\lambda_n \in (0, \infty)$ – as is the case for universal consistency results – then support vector machines are no longer qualitatively robust under mild conditions. (v) Large values of λ force the support vector machine $f_{L^\star,\mathrm{P},\lambda}$ to be smoother, thereby limiting the influence of perturbations Q. This fact, however, is not linked with the intrinsic robustness of the method, but is more related to the regularization itself. On the other hand, small λ will allow for an interpolation of the data, which leaves room for even a single point to have a large influence on the estimated curve, and thus can lead to large biases.

Recall that the Bouligand influence function as defined in Subsection 3.2.3 is in particular useful to study robustness properties of statistical functionals which are defined as minimizers of *non-Fréchet-differentiable* objective functions, such as, e.g., the ϵ-insensitive loss or the pinball loss.

Theorem 3.4.4 (Bouligand influence function). *Let $\mathcal{X}$ be a complete separable normed linear space*[8] *and $\mathcal{H}$ be an RKHS of a bounded, continuous kernel k. Let L be a convex, Lipschitz continuous loss function with Lipschitz constant $|L|_1 \in (0, \infty)$. Let the partial Bouligand-derivatives $\nabla_3^B L(x, y, \cdot)$ and $\nabla_{3,3}^B L(x, y, \cdot)$ be measurable and bounded by*

$$\begin{aligned} \kappa_1 &:= \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\| \nabla_3^B L(x, y, \cdot) \right\|_\infty \in (0, \infty)\,, \\ \kappa_2 &:= \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\| \nabla_{3,3}^B L(x, y, \cdot) \right\|_\infty < \infty\,. \end{aligned} \tag{3.32}$$

[8] E.g., $\mathcal{X} \subset \mathbb{R}^d$ closed. By definition of the Bouligand-derivative, $\mathcal{X}$ has to be a normed linear space.

Let P *and* $\mathrm{Q} \neq \mathrm{P}$ *be probability measures on* $\mathcal{X} \times \mathcal{Y}$, $\delta_1 > 0$, $\delta_2 > 0$,

$$\mathcal{N}_{\delta_1}(f_{L^\star,\mathrm{P},\lambda}) := \{f \in \mathcal{H} : \|f - f_{L^\star,\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta_1\}\,,$$

and $\lambda > \frac{1}{2}\kappa_2 \|k\|_\infty^3$. *Define* $G : (-\delta_2, \delta_2) \times \mathcal{N}_{\delta_1}(f_{L^\star,\mathrm{P},\lambda}) \to \mathcal{H}$,

$$G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_3^B L^\star(X, Y, f(X)) \cdot \Phi(X)\,, \tag{3.33}$$

and assume that $\nabla_2^B G(0, f_{L^\star,\mathrm{P},\lambda})$ *is strong. Then the Bouligand influence function* $\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P})$ *of* $T(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda}$ *exists, is bounded, and equals*

$$\begin{aligned} &S^{-1}\big(\mathbb{E}_\mathrm{P} \nabla_3^B L^\star(X, Y, f_{L^\star,\mathrm{P},\lambda}(X)) \cdot \Phi(X)\big) \\ -&S^{-1}\big(\mathbb{E}_\mathrm{Q} \nabla_3^B L^\star(X, Y, f_{L^\star,\mathrm{P},\lambda}(X)) \cdot \Phi(X)\big)\,, \end{aligned} \tag{3.34}$$

where $S := \nabla_2^B G(0, f_{L^\star,\mathrm{P},\lambda}) : \mathcal{H} \to \mathcal{H}$ *is given by*

$$S(\,\cdot\,) = 2\lambda\, \mathrm{id}_\mathcal{H}(\,\cdot\,) + \mathbb{E}_\mathrm{P} \nabla_{3,3}^B L^\star(X, Y, f_{L^\star,\mathrm{P},\lambda}(X)) \cdot \langle \Phi(X), \,\cdot\, \rangle_\mathcal{H} \Phi(X)\,.$$

Proof of Theorem 3.4.4. In Section 1.7.3 we have seen that $f_{L^\star,\mathrm{P},\lambda}$ exists and is unique. By definition of $L^\star$ it follows from (1.45) that $\nabla_3^B L(x,y,t) = \nabla_3^B L^\star(x,y,t)$. Therefore,

$$\begin{aligned} G(\varepsilon, f) &:= 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_3^B L^\star(X, Y, f(X)) \Phi(X) \\ &= 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_3^B L(X, Y, f(X)) \Phi(X)\,. \end{aligned}$$

Hence $G(\varepsilon, f)$ is the same as in Theorem 3.3.1. All conditions of Theorem 3.3.1 are fulfilled since we assumed that $\nabla_2^B G(0, f_{L^\star,\mathrm{P},\lambda})$ is strong. Hence the proof of Theorem 3.4.4 is identical to the proof of Theorem 3.3.1, which is based on the implicit function theorem for B-derivatives 3.2.2, and the assertion follows. □

Note that also in this case the Bouligand influence function of the SVM only depends on Q via the second term in (3.34). Recall from Subsection 3.2.2 that $(\kappa_1, \kappa_2) = (1, 0)$ for the ϵ-insensitive loss and $(\kappa_1, \kappa_2) = (\max\{1 - \tau, \tau\}, 0)$ for the pinball loss.

3.4.2 Numerical considerations

To illustrate our robustness result, we include two numerical examples. The first example treats some simulated data, based upon a Cauchy distribution, the second (real-life) example is an insurance data set with extreme values. For both examples we used R (R Development Core Team, 2009) for the numerical calculations.

Before we state our numerical results, we would like to draw attention to the estimation of the hyperparameters (such as γ for the Gaussian RBF kernel, the ϵ-value of the ϵ-insensitive loss or the c-value of the Huber loss or the regularization parameter λ). Since the quality of the estimator $\mathcal{R}_{L,\mathrm{D}}(f_{L,\mathrm{D},\lambda})$ and the accuracy of predictions $f_{L,\mathrm{D},\lambda}(x)$ for unseen $x \in \mathcal{X}$ not only depends on the data set used for the training of the SVM, but also on the hyperparameters, the choice of these is of crucial importance. Unfortunately, choosing optimal values for the hyperparameters usually requires computing $f_{L,\mathrm{D},\lambda}$ for many different combinations of the hyperparameters which means that one has to solve a series of convex problems instead of only one. A reasonable choice of the hyperparameters will depend on the criteria used to measure their quality. For regression problems, this criterium is usually a minimization of the empirical L-risk.

There exist of course various methods to obtain the optimal values of these parameters, some of which we will describe here in short, and none of which is optimal for all data sets and is applicable for sample sizes of any size. Most often the parameters are chosen in a data-dependent manner by methods such as random search, cross-validation, a grid search or through a training-validation SVM.

Optimization through a *grid search* is quite easy. After first determining the search space (being the space of all possible combinations of the hyperparameters), each search dimension is split up into parts. The intersections of these splits will form the grid with the trial points for which the objective function is calculated. The best performing point is then taken. A two stage grid search is also possible. In the first stage a rough grid covers a broad region of the space. The optimal point from this search is then in the second stage used as the center of a finer grid, and the best point of this last search is taken as the result. Given that the search range is large enough and the grid is fine enough, there is little danger that the algorithm will only find a local optimum instead of the global one. Clearly, the larger and the finer the grid is chosen, the more time consuming the method will become.

Another standard technique is (k-fold) *cross-validation.* This method is mostly used for relatively small to moderate-sized data sets. This method will randomly divide the data set in k equal-sized disjoint subsets, where each subset is once used as a validation set while the other $k-1$ subsets together form the training set. The combination of hyperparameters with the best performance will then be chosen. Although cross-validation is widely used in practice, there are however some disadvantages to the method, see, e.g., Schölkopf and Smola (2002). One of these is the apparent danger of overfitting, since the training set and the validation set are related to each other. Another is that the found hyperparameters depend on the number of folds (and thus the size of each subset) chosen, since often smaller training sets require a larger value of the regularization parameter λ, which in turn can lead to different values for the other hyperparameters.

A simple method specific for choosing the regularization parameter λ is the use of a training-validation SVM (Steinwart and Christmann, 2008b, Chapter 6.5). The idea of TV-SVMs is to use the training set to construct a couple of SVM decision functions and then use the decision function that performs best on some independent validation set. We need to remark here that the validation step not necessarily provides a unique regularization parameter. Steinwart and Christmann (2008b) show in Lemma 6.29 that for all interesting cases, a measurable TV-SVM will exist, and in Theorem 6.32 they provide the consistency of the method.

Other methods include the *Nelder-Mead algorithm* (Nelder and Mead, 1965), *heuristic choices* of the hyperparameters (Mattera and Haykin, 1999, Cherkassky and Ma, 2004), or *pattern search* (Momma and Bennett, 2002).

Numerical example for simulated data

For this first example we will try to predict the function

$$f(x) = 50\sin(x/20)\cos(x/10) + x$$

with an SVM. For this purpose, $n = 1000$ data points x_i from a uniform distribution $\mathcal{U}(-100, 100)$ have been generated. The corresponding output values y_i were generated by $y_i = f(x_i) + \varepsilon_i$, where ε_i were pseudo-random numbers from a Cauchy distribution. The SVM-regression was done by using the ϵ-insensitive loss function and the Gaussian RBF kernel as defined in Chapter 1. The computation was executed via the `R` function `svm` from the library `e1071`. Using the set of generated data points, the hyperparameters $(\lambda, \epsilon, \gamma)$ of the SVM have been determined by minimizing the $L^\star$-risk via a three dimensional grid search over $17 \times 12 \times 17 = 3468$ knots, where λ is the regularization parameter of the SVM, ϵ is the parameter of the ϵ-insensitive loss function, and γ is the parameter of the Gaussian RBF kernel. For each knot in the grid, an SVM was fitted to our 1000 data points. For each of these points the shifted loss was then calculated and used to determine the $L^\star$-risk for this particular SVM. The grid search resulted in the choice

$(\lambda, \epsilon, \gamma) = \left(2^{-9}/n, 2^{-8}, 2^{-4}\right)$ as parameters for the SVM with the smallest $L^{\star}$-risk. With these parameters, the optimal SVM was then determined.

Figure 3.3 shows the fitted SVM (in blue). The upper sub-plot clearly shows that there are extreme values in the data set which was intended because we simulated error terms from a Cauchy distribution, but the SVM fits nevertheless the pattern set by the majority of the data points quite well. Due to these extreme values, a relatively small systematic bias of the SVM would not be visible. Therefore, we zoomed in on the interesting part of the y-axis to get a better view of the plot, as shown in the lower sub-plot of Figure 3.3. In this graph we see that there is almost no bias and that the SVM neatly covers the true function (in red). Hence, this numerical example confirms our theoretical robustness result of a bounded Bouligand influence function: the SVM is a good approximation of the true function even for heavy-tailed distributions with large extreme values, if the sample size is large enough.

Numerical example for large fire insurance claims in Denmark

For the second example we used the data set ‘danish’ from the `R` package `evir`. This set consists of 2167 fire insurance claims over 1 million Danish Krone (DKK) during the period from Thursday, January 3rd, 1980 until Monday, December 31st, 1990. The claims are total figures, i.e., they include damage to buildings, furniture and personal property as well as loss of profits. The data were supplied by Mette Rytgaard of Copenhagen Re. These data form an irregular time series. The plot of the data shows that there really is a time effect which we will use for our SVM purposes. We have done both classical least squares regression as well as non-parametric conditional quantile regression by means of SVMs. Time was the only explanatory variable. The SVM-regression was done by using the pinball loss for different τ-values. We chose $\tau \in \{0.50, 0.75, 0.90, 0.99, 0.995\}$ since we are interested in big claims. We used the Gaussian RBF kernel. For the SVMs the computations were done with the function `kqr` from the package `kernlab` (Karatzoglou *et al.*, 2004) in `R`, for the least squares regression we used the standard `R` function `lm`. The optimal value of the hyperparameter γ of the kernel was determined by the `kqr` function itself. For each of the τ-values, an SVM was fitted to the data set.

Figure 3.4 shows the fitted curves for the Danish data set. The least squares fit is shown as a dotted line, the SVM-quantiles are drawn as solid

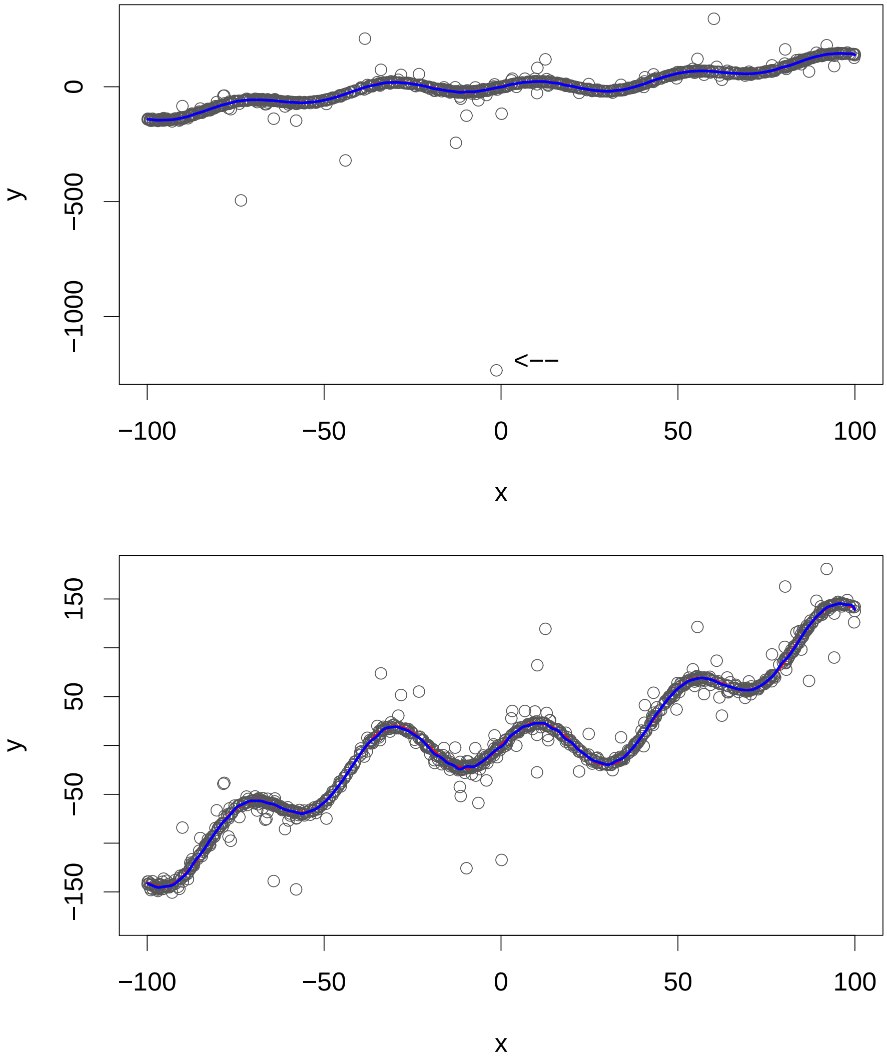

Figure 3.3: The upper sub-plot shows all data points, including all extreme values, as well as the true function (in red) and SVM (in blue). The difference between both functions is hardly visible. The lower sub-plot zooms in on the y-axis. It shows that the bias between the SVM and the true function is almost invisible in this example despite the extreme values.

curves. The upper sub-plot shows the data with 3 remarkable large extreme claims, these being the claims of over 100 million DKK. However, the SVM-quantiles appear to follow the bulk of the points, and do not seem to be much attracted by these extremes. This is in good agreement to our theoretical result stating that SVMs are robust. Due to the extremes, the conditional quantile curves fitted by SVMs and the LS-fit are hardly distinguishable in the upper plot. Therefore we zoomed in on the y-axis to get a better view. On the second sub-plot, we can clearly see all curves, but no longer the extreme data points. The third sub-plot zooms even further, showing only the 90%-quantile and lower quantiles, as well as the least squares regression line. We see that the LS fit lies above the 50% SVM-quantile curve (and in this case often even above the 75% SVM-quantile curve) because it is more attracted towards the extreme values and hence is less robust.

The Danish data set is well-known in the literature on extreme value theory and is therefore used as a benchmark data set in the `R` package `evir` developed by Alexander McNeil and Alec Stephenson. To demonstrate that the residuals of an SVM for non-parametric median regression based on the pinball loss function with $\tau = 0.5$ have a distribution close to an extreme value distribution, we fitted a generalized Pareto (GPD) distribution by the maximum likelihood method to the residuals, see Pickands (1975) and Hosking and Wallis (1987) for details. The computations were done with the function `gpd` from the `evir` package. Figure 3.5 shows that a GPD distribution, which is well-known to have heavy tails, actually offers a good fit for these residuals. The ML estimates for the two parameters of the GPD distribution are 0.498 (0.149) for the shape parameter and 7.824 (1.377) for the scale parameter.

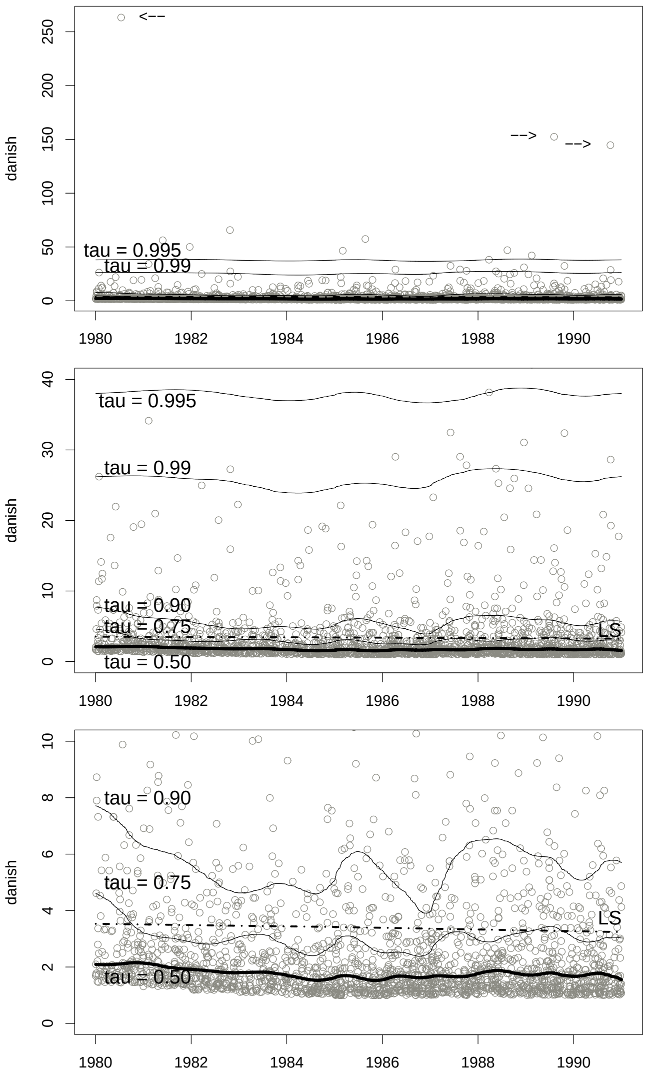

Figure 3.4: All data points, including the extreme values, are visible in the upper sub-plot. The SVM-quantiles seem to follow the mass of the data points and are not attracted to the extremes. The lower sub-plots zoom in on different scales of the y-axis. Here both the LS regression (dotted line) and the SVM-quantiles are distinguishable. There seems to be almost no attraction towards the extreme values for the SVM based quantile curves.

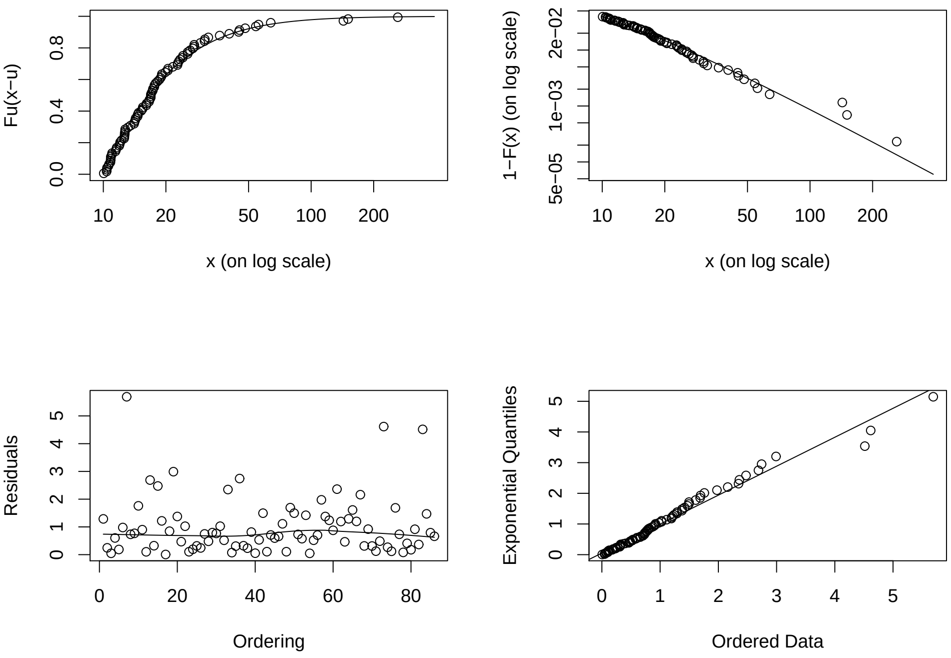

Figure 3.5: Plots to check whether the residuals of the median regression based on SVMs with the pinball loss function have an approximate generalized Pareto distribution. Upper left: excess distribution; Upper right: tail of the underlying distribution; Lower left: scatterplot of the residuals; Lower right: qqplot of the residuals.

Appendices

Appendix A

Mathematical Prerequisites

A.1 Topology

In this work, both metric spaces and Polish spaces are used. We will give their definitions here and will at the same time review the notions of continuity and convergence. For more information on topology and topological spaces, we refer to Kuratowski (1968), Willard (1970) or Dudley (2002).

Definition A.1.1. *Let X be a set. A subset τ of the power set 2^X of X is called a **topology** on X if it satisfies the following three conditions:*

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* $\bigcup_{i \in I} O_i \in \tau$.

*The pair (X, τ) is called a **topological space** and each $O \in \tau$ is called an **open set**.*

A special case of topological spaces are the metric spaces. For $d : X \times X \to [0, \infty)$ a *metric*, we call the pair (X, d) a *metric space*. If d is clear from the context, we omit it and simply call X a metric space.

The most trivial example of a metric space is the Euclidean space $\mathbb{R}^d$, $d \in \mathbb{N}$, equipped with the Euclidian distance

$$d_2(x, y) = \|x - y\|_2 := \Big(\sum_{i=1}^{d} |x_i - y_i| \Big)^{1/2}, \qquad x, y, \in \mathbb{R}^d .$$

A metric d is called *translation-invariant* if $d(x+a,y+a)=d(x,y)$ for all $x,y,a\in X$.

The *closed ball* with radius $\varepsilon>0$ and center $x\in X$ on some metric space (X,d) is defined as

$$B_d(x,\varepsilon):=\{y\in X: d(x,y)\leq\varepsilon\}\,,$$

while for the *open ball* $B_d^*(x,\varepsilon)$ a strict inequality holds. A subset $O\subset X$ is called *open* if for all $x\in O$ there exists an $\varepsilon>0$ such that $B_d^*(x,\varepsilon)\subset O$. The open sets in a metric space (X,d) form a topology on X, called the *metric topology* τ_d.

For (X,τ) a topological space, we call a set $A\subset X$ *closed* if $X\setminus A$ is open. The *closure* of a set A is defined by

$$\overline{A}:=\bigcap\Big\{C\subset X: C \text{ is closed and } A\subset C\Big\}\,.$$

A is said to be 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 indices $i_1,\dots,i_n\in I$ with $A\subset\bigcup_{j=1}^n O_{i_j}$. For $A\subset\mathbb{R}^d$ an easier equivalence holds: A is *compact* if and only if it is bounded and closed (Willard, 1970, Example 17.9(a)).

Moreover, A is called *dense* if $\overline{A}=X$. A topological space (X,τ) is called *separable* if there exists a countable and dense subset of X. $\mathbb{R}$ is separable, since $\mathbb{Q}$ is a countable and dense subset of $\mathbb{R}$.

A family $\mathcal{C}$ of sets is a *covering* of the space X if each point of X belongs to some member C of $\mathcal{C}$.

Let (X_1,τ_1) and (X_2,τ_2) be topological spaces and $x_0\in X_1$. 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 $f(O_1)\subset O_2$. The map f is called *continuous* if f is continuous at every $x\in X$. If f is a real function, an easier definition can be given. A function $f:X\to\mathbb{R}$ is continuous at x_0, if $\lim_{x\to x_0}f(x)=f(x_0)$ or in ε-δ-notation if for all $\varepsilon>0$ there exists a $G\subset X$ open with $x_0\in G$, and a $\delta>0$ such that for all $x\in G$ holds

$$|x-x_0|<\delta\Rightarrow|f(x)-f(x_0)|<\varepsilon\,.$$

The continuous image of a compact set is again compact. A *homeomorphism* or *topological isomorphism* is a continuous bijective function between two topological spaces that has a continuous inverse function.

A sequence $(x_n)_n$ is said to *converge* in a metric space (X, d) if there exists an element $x \in X$ such that for all $\varepsilon > 0$ there exists an $n_0 \geq 1$ such for all $n \geq n_0$ we have $d(x, x_n) \leq \varepsilon$. The *limit* x is unique and we write $\lim_{n\to\infty} x_n = x$ or $x_n \to x$ for $n \to \infty$. If for a sequence $(a_n)_n \subset \mathbb{R}$ holds that $a_n \to 0$, we call it a *null sequence.* A sequence $(x_n)_n$ is called a *Cauchy sequence* if for every $\varepsilon > 0$ there exists an $n_0 \geq 1$ such that, for all $m, n \geq n_0$, $d(x_m, x_n) \leq \varepsilon$. Trivially, every convergent sequence is also a Cauchy sequence, but the inverse is in general not true. Therefore, a metric space is called *complete* if and only if every Cauchy sequence converges. The metric d is then said to be a *complete metric* for X. A topological space (X, τ) is *(completely) metrizable* if it is homeomorphic to a (complete) metric space. This means that there exists at least one (complete) metric d for X which generates (or induces) the topology τ, i.e., for which τ equals the metric topology τ_d. Note that completeness is a property of a metric space, whereas complete metrizability is a property of a topological space. E.g., the open interval $(0, 1)$ is not complete since $(1/n)_n$ is a non-converging Cauchy sequence on $(0, 1)$, but it is completely metrizable since it is homeomorphic with the complete space $\mathbb{R}$.

A sequence of functions $(f_n)_{n\in\mathbb{N}}$, $f_n : X \to \mathbb{R}$ is said to be *pointwise convergent* to a function $f : X \to \mathbb{R}$, if and only if for each $x \in X$ we have that $f_n(x) \to f(x)$. The sequence is *uniformly convergent* if $\|f - f_n\|_\infty \to 0$.

A *basis* of a topology τ is any subset τ_1 of τ such that every open set can be written as a union of sets in τ_1. The set of open balls of a metric space is thus a basis of its topology.

The following definition was first introduced by Bourbaki.

Definition A.1.2. *A topological space (X, τ) is called a* ***Polish*** *space if τ has a countable basis and there exists a complete metric defining τ.*

Another definition is that the topological space (X, τ) is separable and completely metrizable. This means that the space has to be homeomorphic to a complete metric space that has a countable dense subset. Although Polish spaces are metrizable, they are not necessarily themselves metric spaces. Each Polish space admits many complete metrics giving rise to the same topology, but not one of these is singled out or distinguished. A Polish space with a distinguished complete metric is called a *Polish metric space.* For example, the Euclidean spaces $\mathbb{R}^d$ are Polish. Trivially, also all complete separable metric spaces are Polish.

A.2 Probability and Measure Theory

In this section we will state the necessary notions and results concerning measure and probability theory. More details on global measure theory and probability can be found in, e.g., Chow and Teicher (1988) and Billingsley (1995), for the specific parts on Polish spaces, we refer to Dudley (2002).

Definition A.2.1. *Let X be a non-empty set. 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) $A^C := X \setminus A \in \mathcal{A}$ 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 $\mathcal{A}$.

*We call $(X,\mathcal{A})$ a **measurable space** and the elements of $\mathcal{A}$ are called **measurable sets**.*

If $\mathcal{A}$ is clear from the context, or if its specific form is irrelevant, we just call X a measurable space.

It is easy to verify that the intersection of σ-algebras on X is once again a σ-algebra on X. This implies that for any $C \subset 2^X$, there exists a smallest σ-algebra that contains C. This σ-algebra will be denoted as $\sigma(C)$ and is called the σ-algebra *generated* by C. This means that $C \subset \sigma(C) \subset \mathcal{A}$ for all σ-algebras $\mathcal{A}$ on X with $C \subset \mathcal{A}$. An example of such a generated σ-algebra is the *Borel σ-algebra* $\mathcal{B}(\tau)$ of some topological space (X,τ). In this case $\mathcal{B}(\tau) := \mathcal{B}(X) := \sigma(X)$, and its elements are called *Borel sets.*

For $(X_1,\mathcal{A}_1)$ and $(X_2,\mathcal{A}_2)$ two measurable spaces, a function $f : X_1 \to X_2$ is called *measurable*, or $(\mathcal{A}_1,\mathcal{A}_2)$-measurable, if $f^{-1}\mathcal{A}_2 \subset \mathcal{A}_1$.

If $(f_n)_{n\in\mathbb{N}}$ is a sequence of measurable functions mapping from $(X,\mathcal{A})$ to $[-\infty,+\infty]$, 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, 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, meaning that $f_n(x) \to f(x)$ for all $x \in X$ and $f_n(x) \le f_{n+1}(x)$ for all $x \in X$ and $n \ge 1$. Finally, for f bounded, we can pick an increasing sequence (f_n) such that the convergence is even uniform, i.e., $\|f - f_n\|_\infty \to 0$.

Definition A.2.2. *Given some measurable space $(X,\mathcal{A})$, we call a function $\mu : \mathcal{A} \to [-\infty,+\infty]$ a **signed measure** if $\mu(\emptyset) = 0$ and*

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n)$$

for all sequences $(A_n)_{n\in\mathbb{N}}$ of mutually disjoint sets $A_n \in \mathcal{A}$.

A signed measure μ is called a ***measure*** *if $\mu(A) \geq 0$ for all $A \in \mathcal{A}$. A measure is said to be* ***finite*** *if, in addition, $\mu(X) < \infty$. Moreover, if $\mu(X) = 1$, then it is called a* ***probability measure*** *or a* ***distribution****. A measure is called* ***σ-finite*** *if $X = \cup_{n\in\mathbb{N}} A_n$ 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 *(finite, σ-finite) measure space* or a *probability space* if $(X, \mathcal{A})$ is a measurable space and μ is a (finite, σ-finite) measure, respectively probability measure, on $\mathcal{A}$. A probability measure will most often be written as P instead of μ.

An example of a σ-finite measure is the *counting measure* μ on $(\mathbb{Z}, 2^{\mathbb{Z}})$, where $\mu(A)$ equals the number of points in a set $A \in 2^{\mathbb{Z}}$. Another one is the *Lebesgue measure* on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, which is given by

$$\mu(\{x \in \mathbb{R}^d : a_i < x_i \leq b_i, i = 1, \ldots, d\}) = \prod_{i=1}^{d} (b_i - a_i)\,,$$

for all $a_i < b_i$, $i = 1, \ldots, d$. This means that, for bounded rectangles, the Lebesgue measure is nothing else than the ordinary volume. The *Dirac measure* δ_x for some measurable space $(X, \mathcal{A})$ and some $x \in X$ is defined as $\delta_x(A) := 1$ if $x \in A$ and $\delta_x(A) := 0$ if $x \notin A$.

The following theorem (Rademacher, 1919) describes the measure of the set of differentiable points of a Lipschitz continuous function f.

Theorem A.2.3 (Rademacher's theorem)**.** *Let $U \subset \mathbb{R}^n$ be open, and $f : U \to \mathbb{R}^m$ be a Lipschitz continuous function. Then f is Fréchet-differentiable almost everywhere (i.e., the points where f is not Fréchet-differentiable form a set of Lebesgue measure 0).*

Let $(X, \mathcal{A}, \mu)$ be a measure space, we call $N \in \mathcal{A}$ a *μ-zero set* or *μ-null set* if $\mu(N) = 0$. A property $\mathcal{P}(x)$ is said to hold *μ-almost surely* if $\mu(\{x \in X : \mathcal{P}(x) \text{ is false}\}) = 0$.

Now consider a probability space $(X, \mathcal{A}, \mathrm{P})$. In general, the subsets of P-zero set are not P-zero sets themselves, since it can be that they are not measurable. However, we can always add such sets to $\mathcal{A}$. If we define

$$\mathcal{A}_{\mathrm{P}} := \{A \cup B : A \in \mathcal{A}, \exists N \in \mathcal{A} \text{ with } \mathrm{P}(N) = 0 \text{ and } B \subset N\}\,,$$

then $\mathcal{A}_{\mathrm{P}}$ is a σ-algebra, called the P-*completion of* $\mathcal{A}$.

A measurable space $(X, \mathcal{A})$ is said to be P-*complete* if, for a probability measure $\mathrm{P} : \mathcal{A} \to [0, 1]$, $\mathcal{A} = \mathcal{A}_{\mathrm{P}}$. 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 $\mathcal{A}$.

For $(X, \mathcal{A}, \mathrm{P})$ a probability space, a sequence (f_n) of measurable functions $f_n : X \to \mathbb{R}$, is said to *converge in probability* P if for all $\varepsilon > 0$ and all $\delta > 0$ there exists an $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$\mathrm{P}(\{x \in X : |f(x) - f_n(x)| \geq \delta\}) \leq \varepsilon\,.$$

The sequence *converges* P-*almost surely* if $f_n(x) \to f(x)$ for P-almost all $x \in X$. It can be shown that convergence almost surely implies convergence in probability.

Let $(X, \mathcal{A}, \mu)$ be a measure space. A measurable function $f : X \to [-\infty, \infty]$ is called μ-*integrable* if

$$\int_X |f| d\mu < \infty\,.$$

The set of all μ-integrable functions is written as $\mathcal{L}_1(\mu)$.

The following theorem treats the continuity of the integral with respect to almost sure convergence (Dudley, 2002, p. 132).

Theorem A.2.4 (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*

$$\lim_{n\to\infty} \int_X f_n d\mu = \int_X \lim_{n\to\infty} f_n d\mu = \int_X f d\mu\,.$$

The next lemma, see, e.g., Bauer (2001, Theorem 23.8, p. 141), gives a formulation to compute expectations by using tail bounds.

Lemma A.2.5. *Let $(X, \mathcal{A}, \mu)$ be a finite measure space, and $f : X \to [0, \infty)$ be a measurable function. Let $\varphi : [0, \infty) \to [0, \infty)$ be a continuous function that is continuously differentiable op $(0, \infty)$ and satisfies $\varphi(0) = 0$. Then*

$$\int\limits_X \varphi \circ f d\mu = \int_0^\infty \varphi'(t) \mu(f \geq t) dt\,.$$

Let us also review some properties of σ-algebras and measures defined by a topology. We defined the Borel σ-algebra for a topological space (X, τ) already as $\mathcal{B}(X) := \sigma(\tau)$. A measure $\mu : \mathcal{B}(X) \to [0, \infty]$ is then called a *Borel measure*.

Let (X, τ) be a topological space and μ be a Borel measure on X. μ is said to be *regular* if for each $A \in \mathcal{B}(X)$ we have both *outer regularity*, i.e.,

$$\mu(A) = \inf\{\mu(O) : A \subset O, O \text{ open}\}\,,$$

and *inner regularity*, i.e.,

$$\mu(A) = \sup\{\mu(C) : C \subset A, C \text{ compact}\}\,.$$

The proof of the following theorem can be found in Dudley (2002, p. 225).

Theorem A.2.6 (Ulam's theorem)**.** *Every finite Borel measure on a Polish space is regular.*

Polish spaces are also important in 'splitting' probability measures, see, e.g. Dudley (2002, Section 10.2).

Lemma A.2.7 (Regular conditional distribution for Polish spaces)**.** *Let $(X, \mathcal{A})$ be a measurable space, Y be a Polish space with its Borel σ-algebra $\mathcal{B}(Y)$, and 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

$$P(A \times B) = \int_A \mathrm{P}(B|x) d\mathrm{P}_X(x)\,.$$

The map $\mathrm{P}(\,\cdot\,|x)$ is called a ***regular conditional probability*** *or* ***regular conditional distribution*** *of* P. P_X *is called the* ***marginal probability*** *or* ***marginal distribution****.*

Here, $\mathcal{A} \otimes \mathcal{B}(Y)$ denotes the *product σ-algebra* on the product space $X \times Y$. For a sequence $(X_n, \mathcal{A}_n)$ of measurable spaces, the product σ-algebra $\otimes_{n\in\mathbb{N}}\mathcal{A}_n$ on the product space $\times_{n\in\mathbb{N}} X_n$ is defined as the σ-algebra generated by the sets $A_n \times \times_{m\neq n} X_m$, $A_n \in \mathcal{A}$, $n \in \mathbb{N}$.

For a probability space $(X, \mathcal{A}, \mathrm{P})$ we define the *expectation* of a function $f \in L_1(\mathrm{P})$ as $\mathbb{E}_\mathrm{P} f := \int_X f d\mathrm{P}$. If P is a distribution on $X \times Y$, then $\mathbb{E}_\mathrm{P} f := \int_{X \times Y} f d\mathrm{P}$, and if in this case P can be split into a marginal distribution P_X on X and a conditional distribution $\mathrm{P}(\,\cdot\,|x)$ on Y, we write for $f \in L_1(\mathrm{P}_X)$ that $\mathbb{E}_{\mathrm{P}_X} f := \int_X f d\mathrm{P}_X$.

Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, we call a map $\xi : X \to Y$ a *random variable* if ξ is $(\mathcal{A}, \mathcal{B})$-measurable. The set $\sigma(\xi) := \{\xi^{-1}(B) : B \in \mathcal{B}\}$ is called the σ-algebra *generated* by ξ. Trivially, $\sigma(\xi)$ is a sub-σ-algebra of $\mathcal{A}$.

For $\xi : X \to Y$ and $\xi' : X' \to Y$ two random variables defined on the probability spaces $(X, \mathcal{A}, \mathrm{P})$ and $(X', \mathcal{A}', \mathrm{P}')$ respectively, and with $(Y, \mathcal{B})$ a measurable space, we say that ξ and ξ' are *identically distributed* if $\mathrm{P}(\xi \in B) = \mathrm{P}'(\xi' \in B)$ for all $B \in \mathcal{B}$. We will write $\mathrm{P}_\xi(B) := \mathrm{P}(\xi \in B)$.

If I is an index set, and $(A_i)_{i \in I}$ a family of sets for which $A_i \in \mathcal{A}$ for all $i \in I$, then the events A_i are said to be (stochastically) independent if for all distinct indices $i_1, \ldots, i_n \in I$ and for all $n \in \mathbb{N}$ holds that $\mathrm{P}\big(\bigcap_{j=1}^n A_{i_j}\big) = \prod_{j=1}^n \mathrm{P}(A_{i_j})$. The members of family of σ-algebras $(\mathcal{A}_i)_{i \in I}$ with $\mathcal{A}_i \subset \mathcal{A}$ are called independent if all families $(A_i)_{i \in I}$ of events for which $A_i \in \mathcal{A}_i$, are independent. For $(Y_i, \mathcal{B}_i)$, $i \in I$, measurable spaces, then the random variables $\xi_i : X \to Y_i$, $i \in I$, are *independent* if their generated σ-algebras $\sigma(\xi_i)$ are independent.

For $(X, \mathcal{A}, \mathrm{P})$ still a probability space, and for $\xi_i : X \to \mathbb{R}$, $i = 1, \ldots, n$, random variables holds that $\mathbb{E}_\mathrm{P} \prod_{i=1}^n \xi_i = \prod_{i=1}^n \mathbb{E}_\mathrm{P}(\xi_i)$. A similar result is valid for Hilbert spaces. If $\xi_1, \xi_2 : X \to H$ are independent random variables mapping into a separable Hilbert space, then

$$\mathbb{E}_\mathrm{P} \langle \xi_1, \xi_2 \rangle_H = \langle \mathbb{E}_\mathrm{P} \xi_1, \mathbb{E}_\mathrm{P} \xi_2 \rangle_H \,.$$

We will here also repeat the Borel-Cantelli lemma, see, e.g.,Dudley (2002, p. 262) or Billingsley (1995, Theorems 4.3 and 4.4). Recall that $\limsup A_n = \cap_{n \in \mathbb{N}} \cup_{i \geq n} A_i$ for a sequence of sets A_n, $n \in \mathbb{N}$.

Lemma A.2.8 (Borel-Cantelli). *Let $(X, \mathcal{A}, \mathrm{P})$ be a probability space and $(A_n)_{n \in \mathbb{N}}$ a sequence of sets with $A_n \in \mathcal{A}$. Then:*

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 $\mathrm{P}(\limsup A_n) = 1$.

A.3 Functional Analysis

In this section we will review the needed concepts from functional analysis. We refer the interested reader to, e.g., Dudley (2002) and Lax (2002).

A.3.1 Banach Spaces

Let E be a vector space and $\|\cdot\| : E \to [0,\infty)$ a *norm*, we then call $\big(E, \|\cdot\|\big)$ a *normed space*. If the metric associated with the norm is complete, the pair $\big(E, \|\cdot\|\big)$ is called a *Banach* space. If there is no confusion possible, we will write E instead of $\big(E, \|\cdot\|\big)$. To distinguish between norms, we will often add an index $\|\cdot\|_E$ for the norm of the normed space E.

We will denote the *closed unit ball* by $B_E := \{x \in E : \|x\|_E \leq 1\}$. A set $A \subset E$ is called *bounded* if $A \subset cB_E$ for some $c \in [0,\infty)$.

For E and F two vector spaces, a map $S : E \to F$ is called a *(linear) operator* if $S(\alpha x) = \alpha S(x)$ and $S(x+y) = S(x) + S(y)$ for all $\alpha \in \mathbb{R}$ and $x, y \in E$. We will often write Sx instead of $S(x)$. An operator $S : E \to F$ is *bounded* if the image SB_E of the unit ball is bounded under S. If E and F are normed spaces, then this is equivalent to saying that S is continuous, or that there exists a constant $c \in [0,\infty)$ such that for all $x \in E$ we have $\|Sx\|_E \leq c\,\|x\|_F$.

The space of all bounded (linear) operators mapping from E to F is written as $\mathcal{L}(E,F)$. If $E = F$, we will use $\mathcal{L}(E) := \mathcal{L}(E,E)$. If $S \in \mathcal{L}(E,F)$ satisfies $\|Sx\|_F = \|x\|_E$ for all $x \in E$, then S is called an *isometric embedding*. Obviously, S is injective in this case. If, in addition, S is also surjective, then S is called an *isometric isomorphism* and E and F are said to be *isometrically isomorphic*. An $S \in \mathcal{L}(E,F)$ is called *compact* if $\overline{SB_E}$ is a compact subset in F. The following result can be found in, e.g., Cheney (2001).

Theorem A.3.1 (Fredholm alternative)**.** *Let E be a Banach space and let $S : E \to E$ be a compact operator. Then* $\mathrm{id}_E + S$ *is surjective if and only if it is injective.*

A 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, due to the completeness of $\mathbb{R}$, dual spaces are always Banach spaces. For $x \in E$ and $x' \in E'$, the evaluation of x' at x is often written as a *dual pairing*, i.e., $\langle x', x\rangle_{E',E} := x'(x)$. 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.* For $S \in \mathcal{L}(E, F)$, the *adjoint operator* $S' : F' \to E'$ is defined by $\langle S'y', x \rangle_{E',E} := \langle y', Sx \rangle_{F',F}$ for all $x \in E$ and $y' \in F'$.

For the proof of Theorem 3.3.1 we will also need the following consequence of the open mapping theorem, see Lax (2002, p. 170) or Dudley (2002, p. 214).

Theorem A.3.2. *Let E and F be Banach spaces, $S : E \to F$ be a bounded, linear, and bijective operator. Then the inverse $S^{-1} : F \to E$ is a bounded linear operator.*

Next we will quickly review Banach space valued integration. For more details on this subject, we refer to Diestel and Uhl (1977, Chapter II). Let $(X, \mathcal{A})$ be a measurable space and E be a Banach space. A function $f : X \to E$ is called a *measurable step function* if there exist $x_1, \dots, x_n \in E$ and $A_1, \dots, A_n \in \mathcal{A}$ such that

$$f = \sum_{i=1}^{n} \mathbf{1}_{A_i} x_i \,. \tag{A.1}$$

We call $f : X \to E$ an *E-valued measurable function* if there exists a sequence (f_n) of measurable step functions $f_n : X \to E$ such that

$$\lim_{n \to \infty} \|f(x) - f_n(x)\|_E = 0$$

holds for all $x \in X$. The *integral* of a measurable step function $f : X \to E$ with representation (A.1) and a σ-finite measure μ on X is then defined as

$$\int_X f d\mu := \sum_{i=1}^{n} \mu(A_i) x_i \,.$$

Definition A.3.3. *Let $(X, \mathcal{A}, \mu)$ be a σ-finite measure space and E be a Banach space. An E-valued measurable function $f : X \to E$ is called **Bochner μ-integrable** if there exists a sequence (f_n) of E-valued measurable step functions $f_n : X \to E$ such that*

$$\lim_{n \to \infty} \int_X \|f_n - f\|_E \, d\mu = 0 \,.$$

Then the limit

$$\int_X f \, d\mu := \lim_{n \to \infty} \int_X f_n \, d\mu$$

*exists and is called the **Bochner integral** of f. If μ is a probability, the integral can also be written as $\mathbb{E}_\mu f$.*

It can easily be shown that this integral is linear. Furthermore, an E-valued measurable function $f : X \to E$ is Bochner μ-integrable if and only if $x \mapsto \|f(x)\|_E$ is μ-integrable. In this case

$$\Big\| \int_X f \, d\mu \Big\|_E \le \int_X \|f\|_E \, d\mu \, .$$

If $S : E \to F$ is a bounded linear operator, and $f : X \to E$ is Bochner μ-integrable, then the composition $S \circ f : X \to F$ will also be Bochner μ-integrable and the integral and S will commute:

$$S\Big(\int_X f \, d\mu \Big) = \int_X Sf \, d\mu \, .$$

The following theorem can be found in, e.g., Diestel and Uhl (1977, Theorem 3, p. 45)

Theorem A.3.4 (Dominated convergence theorem)**.** *Let $(X, \mathcal{A}, \mu)$ be a σ-finite measure space, E be a Banach space, and (f_n) a sequence of Bochner μ-integrable $f_n : X \to E$. If $\lim_{n\to\infty} f_n(x) = f(x)$ for μ-almost all $x \in X$ and if there exists a μ-integrable function $g : X \to \mathbb{R}$ with $\|f_n\| \le g$, then f is Bochner μ-integrable and*

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f \, d\mu \, .$$

The next result, see Diestel and Uhl (1977, Corollary 8, p. 48), shows that the Bochner integral is in some sense a convex combination.

Theorem A.3.5. *Let $(X, \mathcal{A}, \mu)$ be a finite measure space, E be a Banach space, and $f : X \to E$ be Bochner μ-integrable. Then, for each $A \in \mathcal{A}$ with $\mu(A) > 0$, we have*

$$\frac{1}{\mu(A)} \int_A f \, d\mu \in \overline{\mathrm{co}}(f(A)) \, .$$

Let us now take a look at some important Banach spaces. The supnorm of a function $f : X \to \mathbb{R}$ is defined as $\|f\|_\infty := \sup_{x\in X} |f(x)|$. Then the set $B(X) := \{f : X \to \mathbb{R} : \|f\|_\infty < \infty\}$ equipped with the supnorm is a Banach space. A function $f : X \to \mathbb{R}$ is called *bounded* if there exists an $M < \infty$ such that $\|f\|_\infty \le M$. A sequence of functions $f_n : X \to \mathbb{R}$, $n \in \mathbb{N}$, is *uniformly bounded* if there exists an $M < \infty$ such that, for all $n \in \mathbb{N}$, $\|f_n\|_\infty \le M$.

Given a measurable space $(X, \mathcal{A})$, $\mathcal{L}_0(X)$ denotes the set of all real-valued measurable functions f on X and $\mathcal{L}_\infty(X)$ the set of all bounded measurable functions, i.e., $\mathcal{L}_\infty(X) := \{f \in \mathcal{L}_0(X) : \|f\|_\infty < \infty\}$. $\mathcal{L}_0(X)$ is a vector space and $\mathcal{L}_\infty(X)$ becomes a Banach space when equipped with the norm $\|\cdot\|_\infty$. Let us now assume we have a measure μ on $\mathcal{A}$. For $p \in (0,\infty)$ and $f \in \mathcal{L}_0(X)$ we write $\|f\|_{\mathcal{L}_p(\mu)} := (\int_X |f|^p d\mu)^{1/p}$. To treat the case $p = \infty$, we call $N \in \mathcal{A}$ a local μ-zero set if $\mu(N \cap A) = 0$ for all $A \in \mathcal{A}$ with $\mu(A) < \infty$. Then $\|f\|_{\mathcal{L}_\infty(\mu)} := \inf\{a \geq 0 : \{x \in X : |f(x)| > a\}$ is a local μ-zero set$\}$. In both cases the *set of p-integrable functions* $\mathcal{L}_p(\mu) := \{f \in \mathcal{L}_0(X) : \|f\|_{\mathcal{L}_p(\mu)} < \infty\}$ is a vector space of functions, and for $p \in [1,\infty]$ all properties of a norm on $\mathcal{L}_p(\mu)$ are followed by the mapping $\|\cdot\|_{\mathcal{L}_p(\mu)}$. As usual, we call $f, f' \in \mathcal{L}_p(\mu)$ equivalent, written $f \sim f'$, 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 $x \in X$. The set of equivalence classes $L_p(\mu) := \{[f]_\sim : f \in \mathcal{L}_p(\mu)\}$, where $[f]_\sim := \{f' \in \mathcal{L}_p(\mu) : f \sim f'\}$, is a vector space and $\|[f]_\sim\|_{L_p(\mu)} := \|f\|_{\mathcal{L}_p(\mu)}$ is a complete norm on $L_p(\mu)$ for $p \in [1,\infty]$, i.e., $(L_p(\mu), \|\cdot\|_{L_p(\mu)})$ is a Banach space. It is common practice to identify the *Lebesgue spaces* $\mathcal{L}_p(\mu)$ and $L_p(\mu)$ and hence we often abbreviate both $\|\cdot\|_{\mathcal{L}_p(\mu)}$ and $\|\cdot\|_{L_p(\mu)}$ as $\|\cdot\|_p$. 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. For μ the counting measure on X, we write $\ell_p(X)$ instead of $\mathcal{L}_p(\mu)$.

The following result can, e.g., be found in Werner (2002, Theorem II.2.4) or Dudley (2002, p. 208).

Theorem A.3.6 (Riesz representation theorem)**.** *Let $(X, \mathcal{A}, \mu)$ be a σ-finite measure space and $1 \leq p < \infty$. Define q by $\frac{1}{p} + \frac{1}{q} = 1$. Then is $T : L_q(\mu) \to \big(L_p(\mu)\big)'$ defined by*

$$(Tg)(f) := \int_X fg \, d\mu$$

an isometric isomorphism.

A.3.2 Hilbert Spaces

A very important example of Banach spaces are Hilbert spaces. For $\langle \cdot\,, \cdot \rangle : H \times H \to \mathbb{R}$ an *inner product*, the pair $(H, \langle \cdot\,, \cdot \rangle)$ is called a *pre-Hilbert space*. To differentiate between different inner products, we will often write $\langle \cdot\,, \cdot \rangle_H$. If the inner product is clear from the context, H is called a pre-Hilbert space.

The *Cauchy-Schwarz inequality*

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle\,, \qquad x, y \in H\,,$$

can be used to show that $\|x\|_H := \sqrt{\langle x, x\rangle}$, $x \in H$, defines a norm on H. If this norm is complete, $(H, \langle\,\cdot\,,\,\cdot\,\rangle)$ is called a *Hilbert space.*

The following lemma shows the connection between the norm and the inner product.

Lemma A.3.7 (Parallellogram identity). *Let $(H, \langle\,\cdot\,,\,\cdot\,\rangle)$ be a Hilbert space. Then, for all $x, y \in H$, we have*

$$\begin{aligned} 4\langle x, y\rangle &= \|x+y\|_{\mathcal{H}}^2 - \|x-y\|_{\mathcal{H}}^2\,, \\ \|x+y\|_{\mathcal{H}}^2 + \|x-y\|_{\mathcal{H}}^2 &= 2\,\|x\|_{\mathcal{H}}^2 + 2\,\|y\|_{\mathcal{H}}^2\,. \end{aligned}$$

We refer to Werner (2002, Theorem II.2.5) or Dudley (2002, p. 174) for the following fact concerning the dual of a Hilbert space H. Note that for a given $x \in H$ the map $\langle\,\cdot\,, x\rangle : H \to \mathbb{R} : y \mapsto \langle y, x\rangle$ is a bounded linear functional, and therefore thus an element in H'.

Theorem A.3.8 (Fréchet-Riesz representation). *Let H be a Hilbert space and H' its dual. Then the mapping $\iota : H \to H'$ defined by $\iota x := \langle\,\cdot\,, x\rangle$ for all $x \in H$ is an isometric isomorphism.*

Some straightforward calculations allow us to transform the Bernstein inequality as given in Yurinsky (1995, Theorem 3.3.2) into the following version of Hoeffding's inequality.

Theorem A.3.9 (Hoeffding's inequality in Hilbert spaces). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, H be a separable Hilbert space, and $B > 0$. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to H$ be independent H-valued random variables satisfying $\|\xi_i\|_\infty \le B$ for all $i = 1, \ldots, n$. Then, for all $\tau > 0$, we have*

$$\mathrm{P}\Big(\big\|n^{-1}\sum_{i=1}^{n}(\xi_i - \mathbb{E}_{\mathrm{P}}\xi_i)\big\|_H \ge B\sqrt{\frac{2\tau}{n}} + B\sqrt{\frac{1}{n}} + \frac{4B\tau}{3n}\Big) \le e^{-\tau}\,.$$

A.3.3 Derivatives in Normed Spaces

Let us take a look at the following results from Averbukh and Smolyanov (1967, 1968), Fernholz (1983) and Rieder (1994) on various notions of differentiation to clarify the connections between these notions.

For every pair of normed real vector spaces (E, F) let a subset $\mathcal{S}(E, F)$ of the functions from E to F be given. The following conditions are imposed on this system $\mathcal{S}$, which will provide the (Landau) o remainder of the first-order Taylor approximation of an $\mathcal{S}$-differentiation:

i) $\varrho(0) = 0$, $\varrho \in \mathcal{S}(E, F)$,

ii) $\mathcal{S}(E, F)$ is a real vector subspace of all functions from E to F,

iii) $\mathcal{S}(E, F) \cap C^0(E, F) = \{0\}$ where $C^0(E, F)$ is the space of continuous linear mappings from E to F, and 0 stands for the zero operator and

iv) moreover, in the case where $E = \mathbb{R}$, it is required that

$$\mathcal{S}(\mathbb{R}, F) = \{\varrho : \mathbb{R} \to F \,|\, \lim_{t \to 0} \varrho(t)/t = 0\} .$$

If $\mathcal{S}$ fulfills (i) to (iv), then some mapping $T : E \to F$ is called $\mathcal{S}$-*differentiable* at x if there exists some $A \in C^0(E, F)$ and $\varrho \in \mathcal{S}(E, F)$ such that for all $h \in E$,

$$T(x + h) = T(x) + Ah + \varrho(h) .$$

The continuous linear mapping $\nabla^{\mathcal{S}} T(x) = A$ is called $\mathcal{S}$-*derivative* of T at x. The set of all functions $T : E \to F$ which are $\mathcal{S}$-differentiable at x is denoted by $\mathcal{D}_{\mathcal{S}}(E, F; x)$. From conditions (ii) and (iii) it is seen that the $\mathcal{S}$-derivative $\nabla^{\mathcal{S}} T(x)$ is uniquely defined. Condition (iv) ensures that $\mathcal{S}$-differentiability in case $E = \mathbb{R}$ coincides with the usual notion of differentiability. The function $T \mapsto \nabla^{\mathcal{S}} T(x)$ is a linear mapping from $\mathcal{D}_{\mathcal{S}}(E, F; x)$ to $C^0(E, F)$.

$\mathcal{S}$-differentiations may be constructed in a special way by means of coverings $\mathcal{C}$, whose elements are naturally assumed to be bounded sets C (so that $th \to 0$ uniformly for $h \in C$ as $t \to 0$). For every normed real vector space E let a covering $\mathcal{C}_E$ of E be given which consists of bounded subsets of E. If F is another normed real vector space, define

$$\mathcal{S}_{\mathcal{C}}(E, F) = \{\varrho : E \to F \,|\, \lim_{t \to 0} \sup_{h \in C} \frac{\|\varrho(th)\|}{t} = \varrho(0) = 0 \,, \forall\, C \in \mathcal{C}_E\} .$$

Then the class $\mathcal{S}_{\mathcal{C}}$ satisfies the conditions (i) to (iv). With E ranging through all normed real vector spaces, we can then define the following concepts of differentiation by varying the covering $\mathcal{C}_E$:

i) *Gâteaux*-differentiation corresponds to $\mathcal{C}_{GE} = \{C \subset E \,|\, C \text{ finite}\}$.

ii) For *Hadamard*-differentiation, $\mathcal{C}_{HE} = \{C \subset E \,|\, C \text{ compact}\}$, and

iii) *Fréchet*-differentiation uses $\mathcal{C}_{FE} = \{C \subset E \,|\, C \text{ bounded}\}$.

The three differentiations will be indicated by the corresponding authors' initials. From these definitions it is clear that ∇^F implies ∇^H which implies ∇^G. It can be shown that ∇^H is actually the weakest $\mathcal{S}$-derivative which fulfills the chain rule.

Since these definitions are hard to work with, we will also recall the definitions of the Gâteaux- and Fréchet-derivative by using limits. Let E and F be normed spaces, $U \subset E$ and $V \subset F$ be open sets, and $f : U \to V$ be a function. We say that f is *Gâteaux-differentiable* at $x_0 \in U$ if there exists a bounded linear operator $\nabla^G f(x_0) \in \mathcal{L}(E, F)$ such that

$$\lim_{t\to 0,\, t\neq 0} \frac{\left\| f(x_0 + tx) - f(x_0) - t\nabla^G f(x_0)(x) \right\|_F}{t} = 0\,, \ \ x \in E.$$

We say that f is *Fréchet-differentiable* at x_0 if there exists a bounded linear operator $\nabla^F f(x_0) \in \mathcal{L}(E, F)$ such that

$$\lim_{x\to 0,\, x\neq 0} \frac{\left\| f(x_0 + x) - f(x_0) - \nabla^F f(x_0)(x) \right\|_F}{\|x\|_E} = 0\,.$$

We call $\nabla^G f(x_0)$ the Gâteaux-derivative and $\nabla^F f(x_0)$ the Fréchet-derivative of f at x_0. The function f is called Gâteaux- (or Fréchet-) differentiable if f is Gâteaux- (or Fréchet-) differentiable for all $x_0 \in U$, respectively. Furthermore, f is called *continuously (Fréchet-) differentiable* if it is Fréchet-differentiable and the derivative $\nabla^F f : U \to \mathcal{L}(E, F)$ is continuous.

The next result, see, e.g., Akerkar (1999, Theorem 2.6), gives the Fréchet-derivative of a function that is defined on a product space.

Theorem A.3.10 (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 Fréchet-differentiable if and only if G is partially Fréchet-differentiable and the partial Fréchet-derivatives $\frac{\partial G}{\partial E_1}$ and $\frac{\partial G}{\partial E_2}$ are continuous. In this case, the Fréchet-derivative of G at $(x_1, x_2) \in U_1 \times U_2$ is given by*

$$\nabla^F G(x_1, x_2)(y_1, y_2) = \frac{\partial G}{\partial E_1}(x_1, x_2)y_1 + \frac{\partial G}{\partial E_2}(x_1, x_2)y_2\,, \quad (y_1, y_2) \in E_1 \times E_2\,.$$

We refer to Chapter 4 of Akerkar (1999) for the following implicit function theorem for Fréchet-derivatives.

Theorem A.3.11 (Implicit function theorem)**.** *Let E and F be Banach spaces, and let $G : E \times F \to F$ be a continuously Fréchet-differentiable* *function. Suppose that we have* $(x_0, y_0) \in E \times F$ *such that* $G(x_0, y_0) = 0$ *and* $\nabla_2^F G(x_0, y_0)$ *is invertible. Then there exists a* $\delta > 0$ *and a continuously Fréchet-differentiable function* $f : x_0 + \delta B_E \to y_0 + \delta B_F$ *such that for all* $x \in x_0 + \delta B_E$, $y \in y_0 + \delta B_F$ *we have*

$$G(x, y) = 0 \qquad \textit{if and only if} \qquad y = f(x)\,.$$

Moreover, the Fréchet-derivative of f *is given by*

$$\nabla^F f(x) = -\big(\nabla_2^F G(x, f(x))\big)^{-1} \nabla_1^F G(x, f(x))\,.$$

A.4 Convex Analysis

In this section we will discuss some necessary properties of convex functions. Let us therefore start with the definition of a convex set and a convex function.

A subset A of some Banach space E is called *convex* if, for all $x_1, x_2 \in A$ and for all $\alpha \in [0, 1]$ holds that $\alpha x_1 + (1 - \alpha)x_2 \in A$. In this case we call $f : A \to \mathbb{R} \cup \{\infty\}$ a *convex function* if, for all $x_1, x_2 \in A$ and for all $\alpha \in [0, 1]$, we have

$$f\big(\alpha x_1 + (1 - \alpha)x_2\big) \leq \alpha f(x_1) + (1 - \alpha)f(x_2)\,.$$

If, for all $x_1 \neq x_2$ the inequality is strict, f is called a *strictly convex function.* A function that is twice differentiable will be convex provided its Hessian matrix is positive semi-definite. The set A is called *absolute convex* if for all $x_1, x_2 \in A$ and for all α_1, α_2 satisfying $|\alpha_1| + |\alpha_2| \leq 1$ holds that $\alpha_1 x_1 + \alpha_2 x_2 \in A$.

Furthermore, f is called *concave* if $-f$ is convex. For $A \subset \mathbb{R}^d$, a function f is *affine* if f is convex, concave and finite; i.e., if there exists a vector $a \in \mathbb{R}^d$ and a constant $b \in \mathbb{R}$ such that $f(x) = a^T x + b$ for all $x \in A$.

The *convex hull* $\operatorname{co} A$ of $A \subset E$ is the smallest convex set containing A and can be characterized as

$$\operatorname{co} A = \{\sum_{i=1}^{n} a_i x_i : x_i \in A, a_i \in \mathbb{R}, n \in \mathbb{N}, \sum_{i=1}^{n} a_i = 1, a_i \geq 0\}\,.$$

The *absolute convex hull* is the intersection of all absolute convex sets containing A and is given by

$$\operatorname{aco} A = \{\sum_{i=1}^{n} a_i x_i : x_i \in A, a_i \in \mathbb{R}, n \in \mathbb{N}, \sum_{i=1}^{n} |a_i| \leq 1\}\,.$$

Clearly, $\mathrm{co}A \subset \mathrm{aco}A$.

Given two Banach spaces E and F and $A \subset E$, a function $f : A \to F$ is called *Lipschitz continuous* if there exists a constant $c \geq 0$ such that $\|f(x) - f(x')\|_F \leq c\,\|x - x'\|_E$ for all $x, x' \in A$. The smallest constant that fulfills this inequality is denoted by $|f|_1$. We will call this $|f|_1$ the *Lipschitz constant.*

In the next subsection we will review some continuity properties of convex functions, then we will give in introduction to subdifferential calculus in Subsection A.4.2 and Subsection A.4.3 describes the optimization of convex programs using Lagrange multipliers and duality.

A.4.1 Properties of Convex Functions

The following result on the continuity of convex functions can be found, e.g., in Rockafellar and Wets (2009).

Lemma A.4.1 (Continuity of convex functions)**.** *Let $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be a convex function with domain* $\mathrm{Dom} f := \{t \in \mathbb{R} : f(t) < \infty\}$*. Then f is continuous at all $t \in \mathrm{Int}\,\mathrm{Dom} f$.*

The next result is a consequence of Ekeland and Turnbull (1983, Proposition II.4.6).

Proposition A.4.2 (Uniqueness of minimizer)**.** *Let E be a Banach space and let $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex function. If f is continuous and $\lim_{\|x\|_E \to \infty} f(x) = \infty$, then f has a minimizer. Moreover, if f is strictly convex, then f has a unique minimizer in E.*

A.4.2 Some Facts on Subdifferentials

In this subsection we will state some important properties of the subdifferential of a convex function (see e.g., Phelps, 1993, Rockafellar and Wets, 2009). For the remainder of this subsection, E and F will denote $\mathbb{R}$-Banach spaces. Let us begin by recalling the definition of subdifferentials.

Definition A.4.3. *Let $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex function, and $w \in E$ with $f(w) < \infty$. Then the* ***subdifferential*** *of f at w is defined by*

$$\partial f(w) := \left\{ w' \in E' : \langle w', v - w \rangle \leq f(v) - f(w) \textit{ for all } v \in E \right\}.$$

The following proposition provides some elementary facts on the subdifferential, see Phelps (1993, Proposition 1.11).

Proposition A.4.4. *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 $c \geq 0$ and $\delta > 0$ are constants satisfying*

$$\big|f(v) - f(w)\big| \leq c\,\|v - w\|_E\,, \qquad v \in w + \delta B_E\,,$$

then we have $\|w'\|_E \leq c$ for all $w' \in \partial f(w)$.

This next proposition shows the extent to which the known rules of calculus carry over to subdifferentials.

Proposition A.4.5 (Subdifferential calculus)**.** *Let $f, g : E \to \mathbb{R} \cup \{\infty\}$ be convex functions, $\lambda \geq 0$, and $A : F \to E$ be a bounded linear operator. We then have:*

i) (Homogeneity) For all $w \in E$ with $f(x) < \infty$, we have

$$\partial(\lambda f)(w) = \lambda \partial f(w)\,.$$

ii) (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

$$\partial(f + g)(w) = \partial f(w) + \partial g(w)\,.$$

iii) (Chain rule) If there exists a $v_0 \in F$ 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.

iv) (Minima) The function f has a global minimum at $w \in E$ if and only if $0 \in \partial f(w)$.

v) (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 $\partial f(w) = \{f'(w)\}$.

vi) *(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.4.6 (Representation of subdifferential)**.** *Let $\tilde{L} : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ be a measurable function which is both convex and Lipschitz continuous with respect to its third argument,* P *be a distribution on $\mathcal{X} \times \mathcal{Y}$, and $p \in [1, \infty)$. Assume that $R : L_p(\mathrm{P}) \to \mathbb{R} \cup \{\pm\infty\}$ defined by*

$$R(f) := \int_{\mathcal{X}\times\mathcal{Y}} \tilde{L}(x, y, f(x, y))\, d\mathrm{P}(x, y)$$

exists for all $f \in L_p(\mathrm{P})$ and define p' by $\frac{1}{p} + \frac{1}{p'} = 1$. If $|R(f)| < \infty$ for at least one $f \in L_p(\mathrm{P})$, then, for all $f \in L_p(\mathrm{P})$, we have

$$\begin{aligned}\partial R(f) = \bigl\{h \in L_{p'}(\mathrm{P}) : h(x, y) \in \partial\tilde{L}(x, y, f(x, y))&\\ \textit{for } \mathrm{P}\textit{-almost all } (x, y)\bigr\}\,,&\end{aligned}$$

where $\partial\tilde{L}(x, y, t)$ denotes the subdifferential of $\tilde{L}(x, y, \,\cdot\,)$ at the point t.

Proof of Proposition A.4.6. Since $\tilde{L}$ is measurable, Lipschitz continuous, and finite, it is a continuous function with respect to its third argument. Thus it is a normal convex integrand by Proposition 2C of Rockafellar (1976). Then Corollary 3E of Rockafellar (1976) gives the assertion. □

A.4.3 Convex Progams, Lagrange Multipliers and Duality

Optimization theory will try to characterize the solutions of an optimization problem, which will typically be subject to some constraints. It will provide us with necessary and sufficient conditions for a function to be a solution to the problem. It will also try to develop effective algorithms to solve these problems. As shown in Section 1.2, the SVM problem can be converted into a form that suits this framework.

A number of classes of problems can be distinguished. However, for the case of support vector machines, it is sufficient to consider the optimization problem of maximizing (or minimizing) a convex function given a number of linear constraints.

Definition A.4.7. *A* ***convex program*** *(P) is the optimization problem*

$$\begin{array}{lll} \textit{minimize} & f(x)\,, & x \in A \\ \textit{subject to} & g_i(x) \leq 0\,, & i = 1, \ldots, m\,, \\ & h_i(x) = 0\,, & i = 1, \ldots, n\,, \end{array}$$

where $A \subset \mathbb{R}^d$ is a convex set, and the functions $f, g_1, \ldots, g_m, h_1, \ldots, h_n : A \to \mathbb{R}$ are finite convex functions.

The function f is known as the *objective function*, the functions $g_i(x) \leq 0$ define the *inequality constraints*, and the functions $h_i(x) = 0$ the *equality constraints*. For more details on convex problems we refer the interested reader to, e.g., Cristianini and Shawe-Taylor (2000, Chapter 5), and Steinwart and Christmann (2008b, Chapter A.6), a discussion on more general optimization problems can be found in, e.g., Gill *et al.* (1981, Chapter 3).

Remark that it suffices to consider the problem (P) since maximization problems can easily be converted to minimization problems by changing the sign of the function f. In the same way can the constraints always be written as given above.

If the objective function, the equality and the inequality constraints are all linear, the optimization problem is called a *linear program*. If the objective function is quadratic, while all constraints remain linear, it is called a *quadratic program*.

We call a vector z a *feasible solution* of the convex program (P) if $z \in A$ and z satisfies the constraints from Definition A.4.7. The set of all feasible solutions is called the *feasible region* and will be denoted as R. R is a (possibly empty) convex set. The objective function can also be rewritten as the convex function $f_0 : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f_0(x) := f(x)\mathbf{1}_R(x) + \infty\mathbf{1}_{R^C}(x)\,.$$

Thus minimizing f_0 over $\mathbb{R}^d$ is exactly the same as minimizing f over the feasible region R. The infimum of f_0 will be called the *optimal value* in (P), the points where the infimum is attained are known as the *optimal solutions* to (P), given that $R \neq \emptyset$.

A convex program is said to be *well-posed* in the sense of *Hadamard* if the optimal solution exists and is unique for all data sets, and it depends on the data in a smooth (or continuous) way.

An inequality constraint $g_i(x) \leq 0$ is said to be *active* (or thight) if the solution z satisfies $g_i(z) = 0$, otherwise it is called *inactive*. Equality constraints can be considered to be always active. Sometimes *slack variables* ξ_i, $i = 1, \dots, m$, are introduced to transform the inequality constraints into equality constraints:

$$g_i(x) \leq 0 \quad \Longleftrightarrow \quad g_i(x) + \xi_i = 0\,, \text{with } \xi_i \geq 0\,.$$

For active constraints, the slack variables will be zero, for inactive constraints they will give a measure of 'looseness' in the constraint.

One way to solve such a convex program is by using the Lagrange approach. The purpose of Lagrangian theory (1788) was to characterize the solution of an optimization problem with only equality constraints by introducing the Lagrange multipliers and the Lagrange function. This method was a generalization of the result of Fermat (1629), which gave the solution for an unconstrained optimization problem. Details on these methods can be found, e.g., in Vapnik (1998, Chapter 9.5). Later, in 1951, Kuhn and Tucker provided a more general result that was able to cope with both equality and inequality constraints.

We will first define the Lagrangian multipliers and the Lagrangian function, which contains information about both the objective function and the constraints, and then state the Kuhn-Tucker Theorem.

Definition A.4.8. *Given an optimization problem with objective function $f : A \to \mathbb{R}$, where $A \subset \mathbb{R}^d$, and equality constraints $h_i(x) = 0$, for $i = 1, \dots, n$, the* ***Lagrangian function****, or in short Lagrangian, is defined as*

$$\mathfrak{L}(x, \beta) := f(x) + \sum_{i=1}^{n} \beta_i h_i(x)$$

where $\beta = (\beta_1, \dots, \beta_n)$ and the $\beta_i \geq 0$ are called the ***Lagrange multipliers****. If, in addition, there are also inequality constraints $g_i(x) \leq 0$, $i = 1, \dots, m$, then the* ***generalized Lagrangian function*** *is defined as*

$$\mathfrak{L}(x, \alpha, \beta) := f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{i=1}^{n} \beta_i h_i(x)\,,$$

and the components of both $\alpha = (\alpha_1, \dots, \alpha_m)$, $\alpha_i \geq 0$, and $\beta = (\beta_1, \dots, \beta_n)$, $\beta_i \geq 0$, are the Lagrange multipliers.

Remark that, for inequalities of the form $g_i(x) \geq 0$ the sign of the middle summand will change, since in that case $-g_i(x) \leq 0$. Furthermore, if there are no equality constraints, we will simply write $\mathfrak{L}(x, \alpha)$ for the Lagrangian.

Theorem A.4.9 (Kuhn-Tucker)**.** *Given a convex program (P) with affine functions g_i, $i = 1, \ldots, m$ and h_i, $i = 1, \ldots, n$, then necessary and sufficient conditions for a point x^* to be an optimal solution are the existence of $\alpha^* = (\alpha_1^*, \ldots, \alpha_m^*)$ and $\beta^* = (\beta_1^*, \ldots, \beta_n^*)$ such that*

$$\begin{aligned}
\frac{\partial \mathfrak{L}(x^*, \alpha^*, \beta^*)}{\partial x} &= 0\,, \\
\frac{\partial \mathfrak{L}(x^*, \alpha^*, \beta^*)}{\partial \beta} &= 0\,, \\
\alpha_i^* g_i(x^*) &= 0\,, \qquad i = 1, \ldots, m\,, \\
g_i(x^*) &\leq 0\,, \qquad i = 1, \ldots, m\,, \\
\alpha_i &\geq 0\,, \qquad i = 1, \ldots, m\,.
\end{aligned}$$

The first condition will give us a set of new equations, the second one will return us the equality constraints. The third relation is known as the Karush-Kuhn-Tucker complementarity condition, which implies that for an active constraint the Lagrange multiplier will be $\alpha_i^* \geq 0$, while those of the inactive constraints need to be zero. The whole of these five conditions is often called the Karush-Kuhn-Tucker (KKT) conditions.

For the special case of the SVM as described in Subsection 1.2, the second relation will be superfluous, since there are no equality constraints. Also, the first condition will be split up, since the objective function, being the separating hyperplane, is given in terms of both the vector w and the real number b.

Definition A.4.10. *The **Lagrangian dual problem** of the primal problem from Definition A.4.7 is*

$$\begin{array}{ll}
\textit{maximize} & \theta(\alpha, \beta)\,, \\
\textit{subject to} & \alpha \geq 0\,,
\end{array}$$

where $\theta(\alpha, \beta) = \inf_{x \in A} \mathfrak{L}(x, \alpha, \beta)$.

The Lagrangian treatment of a convex problem passes via the dual description of the problem. Often this dual problem will be easier to treat than the primal problem, since it avoids handling the inequality constraints directly and tries to optimize the dual function over the Lagrange multipliers instead of optimizing over the vectors x in the feasible region.

To go from the primal to the dual program, we can set the derivatives of the Lagrangian with respect to the primal variables to zero, thus imposing stationarity, and then substitute these relations into the Lagrangian, which removes the dependence on the primal variables. This is precisely the same as computing the function

$$\theta(\alpha, \beta) = \inf_{x \in A} \mathfrak{L}(x, \alpha, \beta)\,.$$

The resulting function only contains the Lagrange multipliers as variables and will be maximized under simpler constraints, namely the remaining KKT conditions.

Appendix B

List of Symbols and Notations

Sets and Spaces

$\emptyset$	Empty set
$\mathbb{N}$	Set of positive integers
Q	Set of rational numbers
$\mathbb{R}$	Set of real numbers
(a, b)	Open interval
$[a, b]$	Closed interval
$\overline{A}$	Closure of a set A
$\lvert A \rvert$	Number of elements in a set A
co$A$	Convex hull of a set A
aco$A$	Absolute convex hull of a set A
$\mathcal{X}$	Input space (complete seperable metric space)
$\mathcal{Y}$	Output space (closed subset of $\mathbb{R}$)
H	Hilbert space
$\mathcal{H}_0$	Feature space
$\mathcal{H}$	Reproducing kernel Hilbert space
$C(X)$	Space of continuous functions $f : X \to \mathbb{R}$
$\mathcal{L}(E, F), \mathcal{L}(E)$	Space of bounded linear $S : E \to F$ or $S : E \to E$
$\mathcal{L}_0(X)$	Set of all measurable functions on X
$\mathcal{L}_\infty(X)$	Set of all bounded measurable functions on X
$\mathcal{L}_p(\mu)$	Set of p-integrable functions (w.r.t. μ)
$L_p(\mu)$	Set of equivalence classes of p-integrable functions
$\ell_p(X)$	$\mathcal{L}_p(\mu)$ with μ the counting measure
$\langle \cdot\,, \cdot \rangle$, $\langle \cdot\,, \cdot \rangle_H$	Inner product (in Hilbert space H)

B_E	Closed unit ball in space E
$\mathfrak{D}$	Set of discontinuity points of a function
$\mathcal{N}_\delta(x)$	δ-neighborhood of x

Functions and Operators

$\mathbf{1}_A(x)$	Indicator function: $\mathbf{1}_A(x) = 1$, $x \in A$; $\mathbf{1}_A(x) = 0$, $x \notin A$
id	Identity map: $x \mapsto x$
IF	Influence function
BIF	Bouligand influence function
ι	Fréchet-Riesz isomorphism: $x \mapsto \langle \cdot\,, x \rangle$
$f \ \sim_x \ F$	f approximates F in x
$f \ \approx \ g$	f strongly approximates g
$f \ \approx_x \ F$	f strongly approximates F in x
$F \ \approx_{(x,y)} \ G$	F strongly approximates G
S'	Adjoint operator of S
∇^F	Fréchet-derivative
∇^G	Gâteaux-derivative
∇^H	Hadamard-derivative
∇^B	Bouligand-derivative
$\partial f(x)$	Subdifferential of f at x

Norms

$\lVert \cdot \rVert_2$	Euclidean norm
$\lVert \cdot \rVert_p$	p-norm
$\lVert \cdot \rVert_{L_p}$	L_P-norm
$\lVert \cdot \rVert_\infty$	Supremum norm
$\lVert \cdot \rVert_E$	Norm of space E
$\lVert \cdot \rVert_\mathcal{H}$	Norm of RKHS $\mathcal{H}$
$\lVert \cdot \rVert_\mathcal{M}$	Norm of total variation

Measure theory and Probability

$(X, \mathcal{A})$	Measurable space with σ-algebra $\mathcal{A}$
$(X, \mathcal{A}, \mathrm{P})$	Probability space with distribution P
$\sigma(X)$	σ-algebra on a non-empty set X
$\mathcal{B}$, $\mathcal{B}(\tau)$	Borel σ-algebra on $\mathbb{R}$, or w.r.t. topology τ
μ	Unspecified (signed) measure
$\mathrm{P}, \bar{\mathrm{P}}, \mathrm{Q}$	Probability distributions
P_X	Marginal distribution
$\mathrm{P}(\,\cdot\,\vert x)$	Regular conditional distribution
D	Empirical distribution for the data set D
δ_x, $\delta_{\{x\}}$	Dirac measure at some point x
$\mathcal{M}_1$, $\mathcal{M}_1(Z)$	Set of all probability distributions on a measurable space

X, X_i	Input random variable
Y, Y_i	Output random variable
Z, Z_i	$Z = (X, Y)$, $Z_i = (X_i, Y_i)$
$\mathbb{E}_{\mathrm{P}}(X)$	Expectation of X w.r.t. P

Statistical Learning Theory

D	Training data set
n	Sample size
d	Dimension of the input vector X
sv	Set of support vectors
H_0	Separating hyperplane
H_1, H_2	Decision boundaries
γ_g	Geometrical margin
$\mathfrak{L}$, $\mathfrak{L}_P$, $\mathfrak{L}_D$	Lagrangian, primal and dual Lagrangian
α, β	Lagrange parameters
ξ	Slack variables
$f_{L,\mathrm{P},\lambda}$	SVM decision function w.r.t. P and L
$f_{L,\mathrm{D},\lambda}$	Empirical SVM decision function w.r.t. data set D
$f^*_{L,\mathrm{P}}$	Bayes decision function w.r.t. P and L
$T(\mathrm{P})$	Value of statistic T at P, often $T(P) = f_{L,\mathrm{P},\lambda}$
k	Kernel
k_{RBF}	Gaussian RBF kernel
γ	Width of Gaussian RBF kernel
Φ	Canonical feature map of RKHS $\mathcal{H}$
L	Loss function
$L^\star$	Shifted loss function
L_{c-log}	Logistic loss for classification
L_{hinge}	Hinge loss for classification
L_{LS}	Least squares loss
L_ϵ	ϵ-insenstive loss for regression
$L_{c-Huber}$	Huber loss for regression, $c > 0$
L_{r-log}	Logistic loss for regression
$L_{\tau-pin}$	Pinball loss for quantile regression, $\tau \in (0, 1)$
L_{L1}	$L1$-loss for regression
λ	Regularization parameter
$\mathcal{R}_{L,\mathrm{P}}(\,\cdot\,)$	L-risk w.r.t. P
$\mathcal{R}^*_{L,\mathrm{P}}$	Bayes risk
$\mathcal{R}_{L,\mathrm{D}}(\,\cdot\,)$	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. data set D

$\mathcal{C}_{L,\mathrm{Q}}(\,\cdot\,)$	Inner risk w.r.t. Q
$\mathcal{M}_{L^\star,\mathrm{Q}}(\varepsilon)$	Set of ε-approximate minimizers
$\mathcal{M}_{L^\star,\mathrm{Q}}(0^+)$	Set of exact minimizers
$\delta_{\max}(\varepsilon,\mathrm{Q})$	Self-calibration function

Abbreviations

BIF	Bouligand influence function
ERM	Empirical risk minimization
GPA	Generalized portrait algorithm
IF	Influence function
KKT	Karush-Kuhn-Tucker
k-NN	k-nearest neighbors
OLS	Ordinary least squares
RKHS	Reproducing kernel Hilbert space
RSS	Residual sum of squares
SVM	Support vector machine

Bibliography

Aizerman, M., Braverman, E., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, **25**, 821–837.

Akerkar, R. (1999). *Nonlinear Functional Analysis*. Narosa Publishing House, New Delhi.

Anlauf, J. K. and Biehl, M. (1989). The AdaTron: an Adaptive Perceptron Algorithm. *Europhysics Letters*, **10**(7), 687–692.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, **68**, 337–404.

Averbukh, V. I. and Smolyanov, O. G. (1967). The theory of differentiation in linear topological spaces. *Russian Mathematical Surveys*, **22**(6), 201–258.

Averbukh, V. I. and Smolyanov, O. G. (1968). The various definitions of the derivative in linear topological spaces. *Russian Mathematical Surveys*, **23**(6), 67–113.

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 Transactions on Information Theory*, **44**(2), 525–536.

Bauer, H. (2001). *Measure and Integration Theory*. De Gruyter, Berlin.

Bennett, K. P. and Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, **1**, 23–34.

Bergman, S. (1950). *The Kernel Function and the Conformal Mapping*, volume 5 of *Mathematical Surveys and Monographs*. AMS.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer, Boston.

Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons, New York, 3^{rd} edition.

Boser, B. E., Guyon, I., and Vapnik, V. (1992). A Training Algorithm for Optimal Margin Classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, pages 144–152.

Burges, C. J. C. (1998). A Ttorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, **2**(2), 121–167.

Cheney, W. (2001). *Analysis for Applied Mathematics*. Springer, New York.

Cherkassky, V. and Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, **17**, 113–126.

Chow, Y. S. and Teicher, H. (1988). *Probability Theory: Independence, Interchangeability, Martingales*. Springer-Verlag, New York, 2^{nd} edition.

Christmann, A. and Hable, R. (2011). Support Vector Machines for Additive Models: Consistency and Robustness. Accepted in *Computational Statistics and Data Analysis.*

Christmann, A. and Steinwart, I. (2004). On Robust Properties of Convex Risk Minimization Methods for Pattern Recognition. *Journal of Machine Learning Research*, **5**, 1007–1034.

Christmann, A. and Steinwart, I. (2007). Consistency and robustness of kernel based regression in convex minimization. *Bernoulli*, **13**(3), 799–819.

Christmann, A. and Steinwart, I. (2008). Consistency of kernel-based quantile regression. *Applied Stochastic Models in Business and Industry*, **24**, 171–183.

Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis*. Wiley & Sons, New York.

CORTES, C. AND VAPNIK, V. N. (1995). Support-vector networks. *Machine Learning*, **20**, 273–297.

COVER, T. M. (1965). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on Electronic Computers*, **14**(3), 326–334.

CRISTIANINI, N. AND SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.

DAVIES, P. L. (1993). Aspects of robust linear regression. *Annals of Statistics*, **21**, 1843–1899.

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 Transactions on Pattern Analysis and Machine Intelligence*, **4**, 154–157.

DIESTEL, J. AND UHL, J. J. (1977). *Vector Measures*. American Mathematical Society, Providence, RI.

DING, N. AND VISHWANATHAN, S. V. N. (2011). t-logistic regression. In *Advances in Neural Information Processing Systems 23*, Cambridge, Massachusetts. MIT Press. To appear.

DONTCHEV, A. L. AND HAGER, W. (1994). Implicit functions, Lipschitz maps, and stability in optimization. *Mathematics of Operations Research*, **19**(3), 753–768.

DUDA, R. O. AND HART, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley & Sons, New York.

DUDLEY, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press.

EKELAND, I. AND TURNBULL, T. (1983). *Infinite-dimensional Optimization and Convexity*. Chicago Lectures in Mathematics. The University of Chicago Press.

FERNHOLZ, L. T. (1983). *Von Mises Calculus for Statistical Functionals*, volume 19 of *Lecture Notes in Statistics*. Springer, New York.

Fisher, R. A. (1952). *Contributions to Mathematical Statistics*. Wiley, New York.

Gill, P. E., Murray, W., and Wright, M. H. (1981). *Practical Optimization*. Academic Press, London.

Guillory, A., Chastain, E., and Bilmes, J. (2009). Active Learning as Non-Convex Optimization. In *Twelfth International Conference on Artificial Intelligence and Statistics (AISTAT)*, Clearwater Beach, Florida.

Hable, R. (2011). Asymptotic Normality of Support Vector Machine Variants and other Regularized Kernel Methods. Submitted.

Hable, R. and Christmann, A. (2011). On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, **102**, 993–1007.

Hampel, F. R. (1968). Contributions to the theory of robust estimation. Unpublished Ph.D. thesis, Dept. of Statistics, University of California, Berkeley.

Hampel, F. R. (1974). The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, **69**, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: The Approach Based on Influence Functions*. Wiley & Sons, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

Hosking, J. and Wallis, J. (1987). Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, **29**(3), 339–349.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5^{th} Berkeley Symposium*, **1**, 221–233.

Huber, P. J. (1981). *Robust Statistics*. J. Wiley & Sons, New York.

IP, C. AND KYPARISIS, J. (1992). Local convergence of quasi-Newton methods for B-differentiable equations. *Mathematical Programming*, **56**, 71–89.

KARATZOGLOU, A., SMOLA, A., HORNIK, K., AND ZEILEIS, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, **11**(9), 1–20.

KOENKER, R. (2005). *Quantile Regression*. Cambridge University Press, New York.

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.

KURATOWSKI, K. (1968). *Topology*, volume 1. Academic Press, New York-London.

LAX, P. D. (2002). *Functional Analysis*. Wiley & Sons, New York.

MANGASARIAN, O. L. (1965). Linear and nonlinear separation of patterns by linear programming. *Operations Research*, **13**, 444–452.

MARONNA, R. A., MARTIN, R. D., AND YOHAI, V. J. (2006). *Robust Statistics. Theory and Methods*. Wiley & Sons, New York.

MASNADI-SHIRAZI, H. AND VASCONCELOS, N. (2009). On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost. In Y. B. D. Koller, D. Schuurmans and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages pages 1049–1056. MIT Press, Cambridge, Massachusetts.

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.

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.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

Phelps, R. R. (1993). *Convex Functions, Monotone Operators and Differentiability*. Lecture Notes in Mathematics 1364. Springer, Berlin.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, **3**(1), 119–131.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, **78**(9), 1481–1497.

Rademacher, H. (1919). Über partielle und totale Differenzierbarkeit von Funktionen mehrerer Variabeln und über die Transformation der Doppelintegrale. *Mathematische Annalen*, **79**, 340–359.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rieder, H. (1994). *Robust Asymptotic Statistics*. Springer, New York.

Robinson, S. M. (1987). Local structure of feasible sets in nonlinear programming, Part III: Stability and sensitivity. *Mathematical Programming Study*, **30**, 45–66.

Robinson, S. M. (1991). An implicit-function theorem for a class of nonsmooth functions. *Mathematics of Operations Research*, **16**(2), 292–309.

Rockafellar, R. T. (1976). Integral functionals, normal integrands and measurable selections. In J. P. Gossez, E. J. Lami Dozo, J. Mawhin, and L. Waelbroeck, editors, *Nonlinear Operators and the Calculus of Variations*, volume 543 of *Lecture Notes in Mathematics*, pages 157–207, Berlin. Springer.

Rockafellar, R. T. and Wets, R. J. B. (2009). *Variational Analysis*. Springer, Berlin, 3^{rd} edition.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**(6), 386–408.

Rosenblatt, R. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, D.C.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts.

Shawe-Taylor, J. and Cristianini, N. (1999). Margin Distribution and Soft Margin. In A. J. Smola, P. Bartlett, B. Schölkopf, and C. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 349–358. MIT Press.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M. (1998). Structural Risk Minimization over Data-Dependent Hierarchies. *IEEE Transactions on Information Theory*, **44**(5), 1926–1940.

Smith, F. W. (1968). Pattern Classifier Design by Linear Programming. *IEEE Transactions on Computers*, **C-17**(4), 367–372.

Steinwart, I. (2001). On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of Machine Learning Research*, **2**, 67–93.

Steinwart, I. (2002). Support vector machines are universally consistent. *Journal of Complexity*, **18**, 768–791.

Steinwart, I. (2005). Consistency of Support Vector Machines and Other Regularized Kernel Machines. *IEEE Transactions on Information Theory*, **51**(1), 128–142.

Steinwart, I. (2007). How to compare different loss functions. *Constrained Approximation*, **26**, 225–287.

Steinwart, I. and Christmann, A. (2008a). How SVMs can estimate quantiles and the median. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 305–312. MIT Press, Cambridge, Massachusetts.

Steinwart, I. and Christmann, A. (2008b). *Support Vector Machines*. Springer, New York.

Steinwart, I. and Christmann, A. (2009a). Fast Learning from Non-i.i.d. Observations. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1768–1776. MIT Press, Cambridge, Massachusetts.

Steinwart, I. and Christmann, A. (2009b). Sparsity of SVMs that use the epsilon-insensitive loss. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou., editors, *Advances in Neural Information Processing Systems 21*, pages 1569–1576. MIT Press, Cambridge, Massachusetts.

Steinwart, I., Hush, D., and Scovel, C. (2006). Function classes that approximate the Bayes risk. In G. Lugosi and H. U. Simon, editors, *COLT'06,* 19^{th} *Annual Conference on Learning Theory*, pages 79–93, New York. Springer.

Stone, C. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**(4), 595–620.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., and Vandewalle, J. (2002a). *Least Squares Support Vector Machines*. World Scientific, Singapore.

Suykens, J. A. K., De Brabanter, J., Lukas, L., and Vandewalle, J. (2002b). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, **48**, 85–105.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Non-parametric Quantile Estimation. *Journal of Machine Learning Research*, **7**, 1231–1264.

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.

Vapnik, V. N. (1979). *Estimation of Dependences Based on Empirical Data [in Russian]*. Nauka, Moscow. (English translation: Springer Verlag, 1982).

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1^{st} edition.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley & Sons, New York.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 2^{nd} edition.

Vapnik, V. N. and Chervonenkis, A. Y. (1964). A note on one class of perceptrons. *Automation and Remote Control*, **25**(1).

Vapnik, V. N. and Chervonenkis, A. Y. (1974). *Theory of Pattern Recognition [in Russian]*. Nauka, Moscow. (German translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie-Verlag, Berlin, 1979).

Vapnik, V. N. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, **24**, 774–780.

Wahba, G. (1990). Spline Models for Observational Data. In *CBMS-NSF Regional Conference Series in Applied Mathematics*, volume 59, Philadelphia, USA. SIAM: Society for Industrial and Applied Mathematics.

Werner, D. (2002). *Funktionalanalysis*. Springer, Berlin, 4^{th} edition.

Willard, S. (1970). *General Topology*. Addison-Wesley, Reading, Massachusetts.

Wu, Y. and Liu, Y. (2007). Robust Truncated Hinge Loss Support Vector Machines. *Journal of the American Statistical Association*, **102**(479), 974–983.

Xu, H., Caramanis, C., and Mannor, S. (2009). Robustness and Regularization of Support Vector Machines. *Journal of Machine Learning Research*, **10**, 1485–1510.

Yosida, K. (1974). *Functional Analysis*. Springer, Berlin, 6^{th} edition.

Yurinsky, V. (1995). *Sums and Gaussian Vectors*. Lecture Notes in Mathematics 1617. Springer, Berlin.

Zhang, T. (2004). Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**(1), 56–134.

Index