On Consistency and Stability of Support Vector Machines and Localized Support Vector Machines

Von der Universität Bayreuth
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Abhandlung

von

Hannes Köhler

aus Wiesbaden

1. Gutachter: Prof. Dr. Andreas Christmann
2. Gutachter: Prof. Dr. Ingo Steinwart
3. Gutachter: Prof. Dr. Yiming Ying

Tag der Einreichung: 30.09.2024
Tag des Kolloquiums: 28.02.2025

Abstract

In recent years, the demand for machine learning and artificial intelligence has grown rapidly. This has manifested itself in a drastic increase in the number of existing applications as well as in the pervasiveness of these applications. In these, different machine learning methods have shown enormous empirical success in accurately capturing relations between input and output variables that are far too complex to model them by hand or by classic statistical methods. The present work takes a more analytical approach by mathematically investigating what guarantees can be given for the behavior of one special type of machine learning methods, namely kernel-based minimizers of a regularized risk functional. These minimizers are also known as support vector machines (SVMs) in the literature. In recent years, SVMs have been investigated in much detail, but there still remain open questions.

The present work examines two properties of SVMs. First, SVMs are proven to exhibit different types of consistency—namely risk consistency, L_p-consistency and consistency with respect to the norm in the underlying reproducing kernel Hilbert space—under mild conditions. Surprising negative results occur when transitioning to so-called shifted loss functions, which in many cases helps to eliminate certain conditions regarding the underlying (and in practice unknown) probability measure. It is shown that this elimination is in general not possible for some of the results on consistency, but at the same time it is also shown that alternative and in a certain sense less restrictive conditions regarding the probability measure do in some cases suffice when using shifted loss functions. Secondly, total stability of SVMs is investigated, which is related to classic statistical robustness. Whereas the latter concept however only considers the effect of changes in the probability measure P (respectively an empirical probability measure D_n in applications) on the resulting SVM, total stability additionally takes into account the regularization parameter λ and the kernel k of the SVM and gives bounds on how much the resulting SVM can in the worst case scenario change based on simultaneous variations in the whole triple (P, λ, k) respectively $(\mathrm{D}_n, \lambda, k)$.

As SVMs can in practice suffer from their super-linear requirements regarding computation time as well as computer memory, localized SVMs are examined as well. The principal idea behind localized SVMs is to not learn a single global SVM on the whole input space, but to instead divide the input space into different regions and learn one local SVM in each of these regions and then plug them together to obtain a predictor on the whole input space. This approach reduces the number of data points used for computing each single of the local SVMs and hence—because of the super-linear computational requirements—reduces the overall computation time and space. Additionally, it can also yield advantages regarding the quality of the resulting predictions. Results on consistency as well as on total stability are transferred to localized SVMs. Notably, the consistency results also allow for regions that change as the size of the data set increases, and the total stability results also consider the effect of changes in the regions.

Kurzzusammenfassung

In den letzten Jahren ist die Nachfrage nach maschinellem Lernen und künstlicher Intelligenz rasant angestiegen. Dies hat sich in einer drastischen Zunahme der Zahl der bestehenden Anwendungen sowie in der Verbreitung dieser Anwendungen niedergeschlagen. Hierbei haben verschiedene Verfahren des maschinellen Lernens enorme empirische Erfolge beim akkuraten Erfassen von Beziehungen zwischen Eingabe- und Ausgabevariablen gezeigt, die deutlich zu komplex sind, um sie von Hand oder mittels klassischer statistischer Methoden zu modellieren. Die vorliegende Arbeit verfolgt einen analytischeren Ansatz, indem sie mathematisch untersucht, welche Garantien für das Verhalten einer speziellen Art von Verfahren des maschinellen Lernens gegeben werden können, nämlich kernbasierten Minimierern eines regularisiertes Risikofunktionals. Diese Minimierer sind in der Literatur auch als Support Vector Machines (SVMs) bekannt. In den letzten Jahren wurden SVMs detailliert untersucht, aber es gibt weiterhin offene Fragen.

Die vorliegende Arbeit untersucht zwei Eigenschaften von SVMs. Erstens wird bewiesen, dass SVMs unter schwachen Voraussetzungen verschiedene Arten der Konsistenz aufweisen – nämlich Risiko-Konsistenz, L_p-Konsistenz und Konsistenz bezüglich der Norm im zugrunde liegenden reproduzierenden Kern-Hilbertraum. Beim Übergang zu sogenannten geshifteten Verlustfunktionen, welche in vielen Fällen beim Eliminieren gewisser Voraussetzungen an das zugrunde liegende (und in der Praxis unbekannte) Wahrscheinlichkeitsmaß helfen, treten überraschende negative Resultate auf. Es wird gezeigt, dass dieses Eliminieren bei manchen der Konsistenzresultate im Allgemeinen nicht möglich ist, aber gleichzeitig wird auch gezeigt, dass in manchen Fällen alternative und in einem gewissen Sinn weniger restriktive Voraussetzungen bezüglich des Wahrscheinlichkeitsmaßes bei der Verwendung von geshifteten Verlustfunktionen genügen. Zweitens wird totale Stabilität von SVMs untersucht, welche verwandt mit klassicher statistischer Robustheit ist. Während letzteres Konzept jedoch nur den Effekt von Änderungen im Wahrscheinlichkeitsmaß P (beziehungsweise in einem empirischen Wahrscheinlichkeitsmaß D_n in Anwendungen) auf die resultierende SVM betrachtet, berücksichtigt totale Stabilität zusätzlich auch den Regularisierungsparameter λ und den Kern k der SVM und gibt Abschätzungen dafür, wie stark die resultierende SVM sich bei gleichzeitiger Änderung des gesamten Tripels (P, λ, k) beziehungsweise $(\mathrm{D}_n, \lambda, k)$ schlimmstenfalls ändern kann.

Da SVMs in der Praxis unter ihren superlinearen Anforderungen hinsichtlich Rechenzeit und Computerspeicher leiden können, werden zusätzlich lokalisierte SVMs untersucht. Die grundsätzliche Idee hinter lokalisierten SVMs besteht darin, nicht eine einzelne globale SVM auf dem kompletten Eingaberaum zu lernen, sondern den Eingaberaum stattdessen in verschiedene Regionen aufzuteilen und in jeder dieser Regionen eine lokale SVM zu lernen

und diese dann zusammenzufügen, um einen Prädiktor auf dem gesamten Eingaberaum zu erhalten. Dieser Ansatz verringert die Anzahl der Datenpunkte, die für die Berechnung jeder einzelnen lokalen SVM verwendet werden und reduziert somit – wegen der superlinearen Rechenanforderungen – die Gesamtrechenzeit sowie den Platzbedarf. Zusätzlich kann er auch Vorteile hinsichtlich der Qualität der resultierenden Vorhersagen liefern. Resultate zur Konsistenz sowie zur totalen Stabilität werden auf lokalisierte SVMs übertragen. Hierbei werden bei den Konsistenzresultaten insbesondere auch Regionen zugelassen, welche sich bei Zunahme der Größe des Datensatzes verändern, und bei den Resultaten zur totalen Stabilität wird insbesondere auch der Effekt von Änderungen in den Regionen betrachtet.

Contents

Chapter 1

Introduction

We live in *"The Age of Big Data"* as *The New York Times* already titled in 2012 (Lohr, 2012). Data has become available in vast amounts, be it data from digital sensors in manufacturing and research processes as well as in products for private use such as smartphones and cars, or data supplied by end-users on the web by uploading content such as images and by searching for specific products or answers on questions, or numerous further sources. Naturally, companies as well as public organizations want to make use of this "data flood" to "guide decisions, trim costs and lift sales" (Lohr, 2012)—or to just generally improve their performance. The question arises of how to make sense of the oftentimes overwhelming amount of data and how to identify and extract that part of all the available information that is relevant for the decision/prediction one tries to make—how to *learn* from data?

Because of the vast amount of potential influencing factors and the potentially complex relations, it quickly becomes impossible even for experts in their respective fields to develop accurate models by hand. This is where statistical machine learning comes into play, supported by the quickly increasing computational resources available. Informally, a characterizing property of machine learning is "to let the system learn by itself (...) instead of being programmed explicitly for the task" (Alpaydın, 2020, p. xix). A similar description is given by calling machine learning a "field of study that gives computers the ability to learn without being explicitly programmed" which Arthur Samuel is often accredited with (see for example Zhou, 2021, p. 22). Even though this exact quote can not be found in the often referenced seminal paper by Samuel (1959), it still captures the gist of that paper, which played its part in Samuel popularizing the expression machine learning. A slightly more formal definition is given by Mitchell (1997, p. 2): "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

These learning tasks can be manifold. Samuel (1959) looked at the game of checkers which—as a game—has the advantage of having very clearly defined rules and states. Later, the success of machine learning algorithms in the games of chess and Go, beating some of the world's leading players in both games after some of these algorithms had been given no information about the games apart from the rules, also reached a lot of attention in mainstream media, see for example McFarland (2016), Gibbs (2017). Machine learning has however also been successfully applied in numerous problems that are not restricted to the domains of games. A classic example is the recognition of handwritten letters, which is a typical task that many university students have to implement themselves in their machine learning courses. Nowadays, not only the recognition of written but also that of spoken language is well advanced and is used as an everyday helper in many smartphones or even whole homes in the form of voice assistants. Smartphones of course also use machine learning in many more applications, like for example face detection in the camera app. In medicine, machine learning helps to identify and predict diseases that a patient has respectively might get in the future based on the patient's blood levels, the results of imaging methods such as magnetic resonance imaging, and further characteristics of the patient. In cars, machine learning methods use the data stream coming from sensors and cameras for accident prevention systems and even for autonomous driving. This short list gives a first idea of how varied and omnipresent machine learning applications already are and that this will only further increase in the near future—not only in tasks which are as apparent to the end-user but also in such that are hidden in internal processes of companies.

The tasks can be formalized by introducing an input space $\mathcal{X}$ containing all potential explanatory variables (such as the signals coming from the different digital sensors in an autonomously driving car) as well as an output space $\mathcal{Y}$ containing the possible decisions that can be made based on the input variables (such as the different angles that the wheels of the car can get moved into) and an unknown probability measure P on $\mathcal{X} \times \mathcal{Y}$ describing combinations of values from $\mathcal{X}$ and $\mathcal{Y}$ which can occur (such as possible signals from the sensors and appropriate angles of the wheels that will not cause a crash). If we denote by (X, Y) a pair of random variables with distribution P, the *task* consists of predicting the value of Y based on that of X, that is finding a good predictor function $f\colon \mathcal{X} \to \mathcal{Y}$, while inflicting as few prior assumptions upon P as possible. In (supervised) machine learning, the *experience* used for this task consists of input-output-pairs $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$ for some $n \in \mathbb{N}$, of observations sampled from P, which together form a data set $D_n := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$.[1] Finally, as the relation between X and Y is usually not entirely deterministic but P allows Y to take different values for a given value $x \in \mathcal{X}$ of X, it is not always immediately obvious in which way one should aim to predict Y based on X. The exact goal of prediction gets specified by the *performance measure* one chooses and often consists of finding a function which estimates certain characteristics of the conditional distributions $\mathrm{P}(\cdot \mid X = x)$, $x \in \mathcal{X}$, of Y,[2] like for example the function of conditional means or that of conditional medians.

For approaching such machine learning tasks, a vast array of methods and corresponding algorithms exists. An overview of various popular approaches to the learning problem is for example given by the aforementioned books Mitchell (1997), Alpaydın (2020), Zhou (2021), but also by Devroye et al. (1996), Duda et al. (2001), Györfi et al. (2002), Clarke et al. (2009), Hastie et al. (2009), Shalev-Shwartz and Ben-David (2014) among others. We

[1] The data set is denoted as a tuple rather than a set as it is possible for the same input-output-pair to occur multiple times.

[2] These conditional distributions are known to uniquely exist whenever we have $\mathcal{Y} \subseteq \mathbb{R}$ closed (which will be assumed throughout this thesis), cf. Remark 2.0.3.

mostly focus on one special type of machine learning algorithms, namely that of computing *support vector machines (SVMs).* Whereas SVMs had originally only been proposed for classification problems using the so-called hinge loss function (cf. eq. (2.4)), for which reason some authors still use the term SVM to describe only this special type of *kernel-based regularized risk minimizers*, we use it in a broader sense allowing for arbitrary loss functions and thus notably also covering regression problems. SVMs as well as the necessary building blocks—some of them have already been mentioned here, such as loss functions, risks, kernels and regularization—are described in Section 2.1, see also Vapnik (1995, 1998), Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002), Cucker and Zhou (2007), Steinwart and Christmann (2008) for more detailed introductions. Additional references providing useful properties of SVMs can also be found in Section 2.1.

SVMs are an important and popular tool in machine learning mainly due to two reasons: First, they are known to possess many desirable theoretical properties such as universal consistency, statistical robustness and stability, and good learning rates (see for example the books mentioned in the previous paragraph). Secondly, they are the solutions of finite-dimensional convex programs (cf. Smola and Schölkopf, 2004) and empirically observe good performance (cf. Klambauer et al., 2017; Paoletti et al., 2019) if the data set is not too large. For large data sets, SVMs however suffer from their computational requirements growing at least quadratically in the size of the data set D_n, with regards to both time and memory, cf. Platt (1998), Joachims (1998), Thomann et al. (2017). For this reason, we do not only look at “regular” SVMs but also at so-called *localized SVMs*, which—instead of computing one SVM for the whole input space—divide the input space $\mathcal{X}$ into different regions,[3] compute a separate SVM in each of these regions, and then plug them together in order to obtain a global predictor on all of $\mathcal{X}$. By regionalizing the input space, the data set gets split into several smaller regional data sets—as the size of D_n increases, it is usually reasonable to also increase the number of regions, such that the regional data sets do only grow slowly in size—and the computational requirements are therefore greatly reduced for large data sets, see for example Thomann et al. (2017). Additionally, localizing SVMs can also yield improvements regarding the resulting predictions because the localization increases the flexibility of the method and can potentially separate regions in which $\mathrm{P}(\cdot \mid X = x)$ takes very different shapes. Section 2.2 defines these localized SVMs in more detail and reviews existing publications on them. That section additionally mentions some alternative approaches which can also reduce the computational requirements of SVMs, but which for the most part do not yield the additional improvement of the predictions that localized SVMs can yield.

This thesis is mainly concerned with two important theoretical properties of SVMs and localized SVMs. First, Chapter 3 takes a look at their *consistency*, i.e. at whether (localized) SVMs converge (in a suitable sense) to the “true” function one tries to estimate as the size n of the data set D_n tends to infinity. As this is a very fundamental property for any learning method, there already exist results for SVMs (see for example Christmann and Steinwart, 2007, Theorem 12) as well as for localized SVMs (see for example Hable, 2013, Theorem 1, and Dumpert and Christmann, 2018, Theorem 3.1). These results are in

[3] These regions are often chosen in such a way that they constitute a partition of the input space. However, most of the results from this thesis do not actually require the regions to be pairwise disjoint.

some sense generalized in Chapter 3. Additionally, all these results consider the same type of consistency, namely *risk consistency*, cf. Section 3.1. Whereas risk consistency is the most widely-used type of consistency in machine learning theory and the one that is aimed at by SVMs and localized SVMs by their definition, other types like *L_p-consistency* and *consistency with respect to the norm in a Hilbert space* (which are both also introduced in Section 3.1) can also be of interest as they compare the two functions—(localized) SVM and "true" function—more directly. For this reason, results on these types of consistency are derived as well. Notably and in addition to some further generalizations, the results for localized SVMs, contrary to those from Dumpert and Christmann (2018), also allow the underlying regions to change as the size n of D_n increases and, contrary to those from Hable (2013), do not assume any specific method for obtaining the regions. Note that this thesis does *not* investigate the rates of these types of convergence, but instead focuses on deriving consistency under mild conditions.

Afterwards, Chapter 4 investigates *total stability* of SVMs and localized SVMs. The expression "total stability" is based on the paper Christmann et al. (2018) and bears resemblance to the notion of statistical robustness, which is concerned with guaranteeing that small changes in the data (such as a small amount of outliers or slight changes in a potentially large part of the data set) will only lead to small changes in the resulting predictor. Such changes in the data may always occur in practice, for example due to measurement and rounding errors. Rounding inaccuracies can obviously lead to small changes in a large part of the data set, and it is hence apparent that this type of changes in the data is of great practical relevance. On the other hand, outliers constitute a common occurrence as well: "[A]ltogether, 1–10% gross errors in routine data seem to be more the rule rather than the exception" (Hampel et al., 1986, p. 28). Hence, it is obvious from the exemplary machine learning tasks given earlier in this introduction that robustness is a desirable property for machine learning methods such as (localized) SVMs: In an autonomous car, changes in a few pixels of the output of a camera (for example because of some dirt on the camera) should not lead to a completely different resulting angle of the wheels. Similarly, when using blood levels and other patient data for diagnosing diseases, rounding these values in a slightly different way should not yield a completely different result. There exists a multitude of publications on statistical robustness of SVMs and also some on that of localized SVMs (see Section 4.1). Our considerations, however, are more closely connected to those by Christmann et al. (2018) who additionally took into account that SVMs also depend on certain hyperparameters (cf. Section 2.1), which in turn are usually not predefined but instead depend on the data set D_n themselves. As changes in these hyperparameters can therefore be a consequence of changes in the data, one would hope that SVMs are stable with respect to them as well. Christmann et al. (2018) proved that this is indeed the case and used the expression "total stability" for this. Chapter 4 first considerably generalizes results from that paper, before then transferring them to localized SVMs. For these, the effect of changes in the underlying regionalization is considered as well.

Chapter 3 is partially taken from the peer-reviewed papers Köhler (2024a,b) that were published in *Neurocomputing* and *Journal of Machine Learning Research* respectively. Chapter 4 on the other hand is partially taken from the peer-reviewed paper Köhler and

Christmann (2022) that was published in *Journal of Machine Learning Research.* Both chapters do however also contain previously unpublished results.

Chapter 2

Support Vector Machines and Localized Support Vector Machines

This chapter gives a short introduction, first to support vector machines (SVMs) and then to localized SVMs. Both of these are examined regarding different properties in the subsequent chapters. Throughout this chapter, the following is assumed to hold true:

Assumption 2.0.1. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let $\mathcal{Y} \subseteq \mathbb{R}$ be closed.

Remark 2.0.2. Throughout this thesis we usually refrain from explicitly stating the σ-algebra $\mathcal{A}$ and instead just speak of the measurable space $\mathcal{X}$ to shorten the notation as long as the exact shape of $\mathcal{A}$ is not relevant or $\mathcal{A}$ is obvious from context. When speaking of probability measures on measurable spaces, we usually also omit explicitly stating the σ-algebra, thus for example speaking of probability measures on $\mathcal{X}$ instead of on $(\mathcal{X}, \mathcal{A})$. On Cartesian products of measurable spaces, we always assume the product σ-algebra if not specified differently. On measurable subsets $\tilde{\mathcal{X}} \subseteq \mathcal{X}$, we always assume the σ-algebra $\tilde{\mathcal{A}} := \{S \subseteq \tilde{\mathcal{X}} \mid S \in \mathcal{A}\}$ if not specified differently. A metric space (S, d_S)—we will often have $S \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$—is always assumed to be equipped with its Borel σ-algebra $\mathcal{B}_S$ if not specified differently. To once more shorten the notation, we usually do not explicitly state the metric d_S and instead just speak of the metric space S.

Remark 2.0.3. $\mathcal{Y} \subseteq \mathbb{R}$ being closed will be assumed in all main chapters and is known to imply that $\mathcal{Y}$ is a Polish space (Bauer, 2001, p. 157). If P is a probability measure on $\mathcal{X} \times \mathcal{Y}$, the (regular) conditional distributions $\mathrm{P}(\cdot \mid X = x)$, $x \in \mathcal{X}$, hence uniquely exist and it is possible to split P into a marginal distribution P^X on $\mathcal{X}$ and these conditional distributions (Dudley, 2004, Theorems 10.2.1 and 10.2.2).

With a slight abuse of notation, we sometimes just write $\mathrm{P}(\cdot \mid X)$ when discussing $\mathrm{P}(\cdot \mid X = x)$, $x \in \mathcal{X}$, in the remainder of this thesis.

2.1 Introduction to Support Vector Machines

SVMs have been an important method in machine learning for many years now. After important groundwork had been laid in the preceding decades (notably see Vapnik and

Lerner, 1963; Vapnik and Chervonenkis, 1964), SVMs as they are known today have been proposed in the 1990s by Boser et al. (1992), Cortes and Vapnik (1995) among others. Whereas the concept of SVMs had only been proposed for binary classification tasks (with the two classes being denoted as "-1" and "$+1$", i.e. with output space $\mathcal{Y} = \{-1, +1\}$) in these papers, Vapnik (1995), Drucker et al. (1996), Vapnik et al. (1996) among others generalized it to regression tasks, sometimes calling the resulting method *support vector regression* in order to differentiate it from SVMs for classification. We however call both of these SVMs—as it is the case in many publications nowadays—since the definition given in this section covers both cases.

In the following, Section 2.1.1 gives a definition of SVMs and Section 2.1.2 formally defines kernels and reproducing kernel Hilbert spaces, which are needed for SVMs. Section 2.1.3 then recalls some properties of the building blocks of SVMs that will be useful in later chapters. Further detailed introductions to SVMs can for example be found in Vapnik (1995, 1998), Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002), Cucker and Zhou (2007), Steinwart and Christmann (2008), with Sections 2.1.1 to 2.1.3 mostly using the definitions and notation used by Steinwart and Christmann (2008). Lastly, Section 2.1.4 recalls the concept of shifted loss functions and the advantages they can bring to SVMs.

2.1.1 Definition of Support Vector Machines

SVMs tackle the problem of (supervised) statistical machine learning described in the introduction. That is, the goal is to find some function $f\colon \mathcal{X} \to \mathbb{R}$ relating an input random variable X taking values in $\mathcal{X}$ to an output random variable Y taking values in $\mathcal{Y} \subseteq \mathbb{R}$ and predicting the value of Y based on that of X. To decide for a predictor f, the quality of predictions $f(x)$ has to be assessed. This can be done using *loss functions*.

Definition 2.1.1 (Loss Function)**.** A function $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called a loss function (or just loss) if it is measurable.

For a predictor f, the value $L(x, y, f(x))$ can then be interpreted as the loss or cost associated with predicting $f(x)$ while the true output belonging to x is y. In order to not only assess single predictions $f(x)$ but instead the whole predictor f, one has to look at the average loss this predictor produces, i.e. the expectation of L.

Definition 2.1.2 (Risk)**.** Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function, P be a probability measure on $\mathcal{X} \times \mathcal{Y}$, and $f\colon \mathcal{X} \to \mathbb{R}$ be a measurable function. Then,

$$\mathcal{R}_{L,\mathrm{P}}(f) := \int_{\mathcal{X}\times\mathcal{Y}} L(x, y, f(x)) \, \mathrm{dP}(x, y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} L(x, y, f(x)) \, \mathrm{dP}(y \,|\, x) \, \mathrm{dP}^X(x)$$

is called L-risk (or just risk) of f with respect to P.[4]

[4] Splitting P into marginal distribution P^X and conditional distribution $\mathrm{P}(\cdot \,|\, X)$ is possible because of Remark 2.0.3 and Assumption 2.0.1.

In practice, the true distribution P is usually unknown and the predictor f is learned on the basis of a training data set $D_n := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ of $n \in \mathbb{N}$ i.i.d. observations sampled from P. The associated empirical distribution is given by

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

with $\delta_{(x,y)}$ denoting the Dirac distribution in $(x, y) \in \mathcal{X} \times \mathcal{Y}$. This leads to the following important special case of Definition 2.1.2:

Remark 2.1.3 (Empirical Risk). Let D_n be the empirical distribution associated with the training data set D_n. Plugging D_n into Definition 2.1.2 yields the empirical L-risk (or just empirical risk)

$$\mathcal{R}_{L,\mathrm{D}_n}(f) = \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) \, \mathrm{d}\mathrm{D}_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, f(x_i)) \,.$$

The learning goal can be formalized as finding a function whose risk (i.e. expected loss/cost) with respect to the true distribution P underlying the data is as small as possible. Ideally, one would hope to find a measurable function that minimizes $\mathcal{R}_{L,\mathrm{P}}$. In practice, this is usually not possible because of P being unknown, all information about P stemming from the data set D_n. Still, it is apparent that this minimizer of the risk as well as the minimal risk itself both play an important role.

Definition 2.1.4 (Bayes Risk). Let $L \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a probability measure on $\mathcal{X} \times \mathcal{Y}$. Then,

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

is called Bayes L-risk (or just Bayes risk) with respect to P. Any measurable $f^*_{L,\mathrm{P}} \colon \mathcal{X} \to \mathbb{R}$ satisfying $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ is called a Bayes function.

For some loss functions, the Bayes functions correspond to easily interpretable characteristics of the conditional distribution $\mathrm{P}(\cdot \mid X)$. That is, the learning goals specified by these loss functions correspond to learning these characteristics of $\mathrm{P}(\cdot \mid X)$. Prime examples of this include the *least squares loss*

$$L_{\mathrm{LS}} \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty) \,, \; (x, y, t) \mapsto (y - t)^2 \,,$$

where the associated risk gets minimized by the function $x \mapsto \mathbb{E}\,[Y|X = x]$ of conditional means (cf. Györfi et al., 2002, Section 1.1), and the *(τ-)pinball loss*

$$L_{\tau\text{-pin}} \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty) \,, \; (x, y, t) \mapsto \begin{cases} (1 - \tau) \cdot (t - y) & , \text{ if } y < t \,, \\ \tau \cdot (y - t) & , \text{ if } y \geq t \,, \end{cases} \tag{2.1}$$

for $\tau \in (0, 1)$, where the associated risk gets minimized by the function of conditional τ-quantiles (cf. Koenker, 2005, Sections 1.3 and 1.4).[5] As quantiles are in general not unique, the latter is an example for Bayes functions also not always being unique.

[5] See also page 49 for further details and references on the pinball loss.

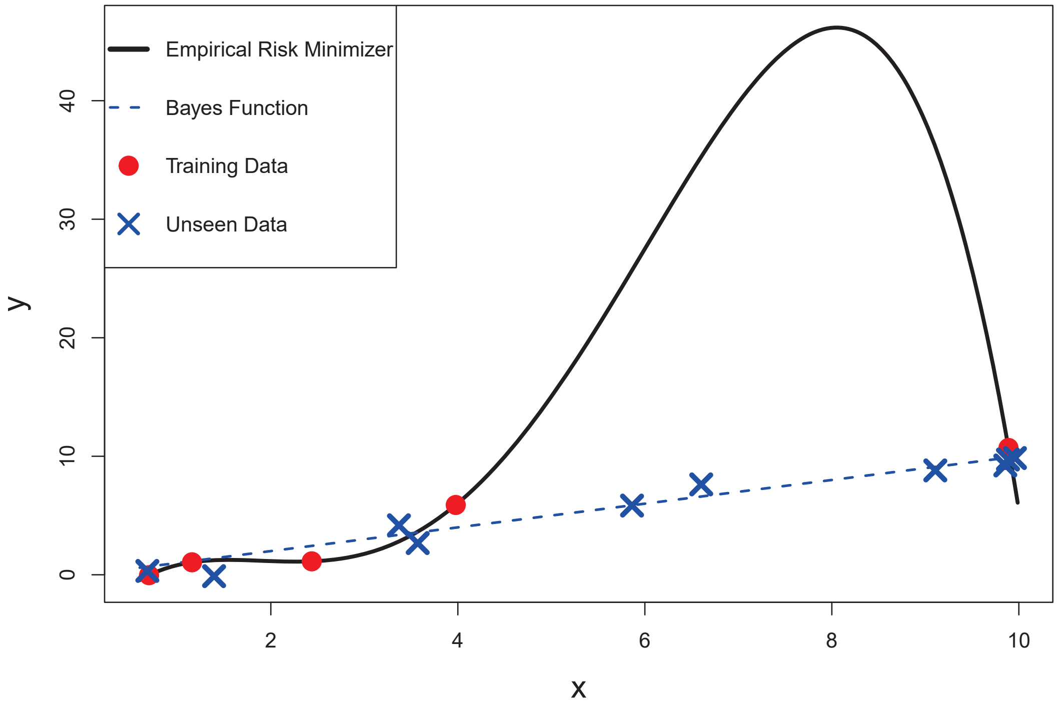

Figure 2.1.1: Example of overfitting for the least squares loss and data generated according to P^X being the uniform distribution on $(0, 10)$ and $\mathrm{P}(\cdot \,|\, X = x)$ being the normal distribution $\mathcal{N}(x, 1)$. The function represented by the solid line exhibits empirical risk 0 on the training data but does—in contrast to the Bayes function—not yield good predictions for unseen data coming from the same distribution. Here, the Bayes function is just the linear function defined by $f^*_{L,\mathrm{P}}(x) = x$ because of the underlying conditional distributions being symmetric about that function.

Now, trying to find a function whose risk (with respect to P) is as small as possible, ideally equal to the Bayes risk, based on the training data set D_n comes with the problem that not every function observing a small empirical risk on the finite data set D_n also offers good generalization to unseen data coming from P and will therefore yield good predictions for such unseen data (even though, by the law of large numbers, $\mathcal{R}_{L,\mathrm{D}_n}(f)$ of course converges to $\mathcal{R}_{L,\mathrm{P}}(f)$ for each single measurable f as $n \to \infty$). By just minimizing the empirical risk, one often runs into the problem of overfitting the function to the training data, for example just interpolating the points from D_n and thus obtaining empirical risk 0, but not accurately capturing the structure underlying the data, see for example Figure 2.1.1.

In order to circumvent overfitting, one can add a regularization term penalizing functions that are overly complex in a suitable sense, see also von Luxburg and Schölkopf (2011, Section 7) for more details on the idea behind regularization. For SVMs, the minimization is performed not over all measurable functions but only those from some *reproducing kernel Hilbert space* (*RKHS*) H over $\mathcal{X}$ and the penalty term is the squared H-norm of the function multiplied by some positive constant $\lambda > 0$, the so-called *regularization parameter*, controlling the amount of regularization. RKHSs as well as some of their properties are described in more detail in Section 2.1.2. For now, it suffices to know that RKHSs over $\mathcal{X}$ can be seen as special Hilbert spaces consisting of functions $f\colon \mathcal{X} \to \mathbb{R}$, and that RKHSs can be associated to a so-called *kernel* $k\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ on $\mathcal{X}$ and possess certain useful properties. RKHSs can be large enough to approximate every continuous function arbitrarily well and to contain functions separating arbitrary compact disjoint sets in the binary classification setting (cf. Steinwart and Christmann, 2008, Definition 4.52, Proposition 4.54, Corollary 4.58), but can still be handled well in practice because of not being too large and having a useful structure for representing their elements (see also Lemma 2.1.10(iv)). This yields the following definition of SVMs:

Definition 2.1.5 (Support Vector Machine)**.** Let $L\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to [0,\infty)$ be a loss function and P be a probability measure on $\mathcal{X} \times \mathcal{Y}$. Let k be a kernel on $\mathcal{X}$ with RKHS H and let $\lambda > 0$. Then,

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

is called support vector machine (SVM).

Remark 2.1.6 (Empirical Support Vector Machine)*.* In practice, the probability measure in Definition 2.1.5 is usually the empirical distribution D_n associated with the data set D_n. Plugging D_n into Definition 2.1.5 yields the empirical SVM

$$f_{L,\mathrm{D}_n,\lambda,k} = \arg\inf_{f\in H} \mathcal{R}_{L,\mathrm{D}_n}(f) + \lambda \left\|f\right\|_H^2 = \arg\inf_{f\in H} \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, f(x_i)) + \lambda \left\|f\right\|_H^2 \,.$$

In the considerations regarding consistency in Chapter 3, it is necessary in many results to think of the empirical SVM not as a fixed function depending on the observed data $D_n = ((x_1, y_1), \dots, (x_n, y_n))$ but instead as a random function depending on the random variables $(X_1, Y_1), \dots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} \mathrm{P}$, which the data points are realizations of. Since it will be clear from the context whether the SVM needs to be thought of as a random function, we will not introduce separate notation for this and instead also denote it by $f_{L,\mathrm{D}_n,\lambda,k}$.

Even though one usually needs empirical SVMs in practice, Definition 2.1.5 defines SVMs for general probability measures, most notably also for the true distribution underlying D_n, which is useful for theoretical considerations. To distinguish the latter from an empirical SVM, it is also called *theoretical SVM*. Furthermore, when contrasting SVMs with the localized SVMs introduced in Section 2.2, we sometimes also call them *global SVMs* or *regular SVMs* in order to emphasize the difference.

SVMs are known to possess many desirable theoretical properties under mild assumptions. These include existence and uniqueness as well as statistical robustness—making them the solutions of a well-posed problem in Hadamard's sense (Hable and Christmann, 2011)—, risk consistency and the existence of representation theorems, see for example the references given at the beginning of Section 2.1. Further active research is performed on learning rates of SVMs, that is, on how quickly the convergence of the risks of empirical SVMs to the Bayes risk takes place. Because of the no-free-lunch-theorem (Devroye, 1982), such learning rates cannot be derived without imposing assumptions on the unknown true distribution P. Learning rates under different conditions on P have for example been derived by Caponnetto and De Vito (2007), Steinwart et al. (2009), Mendelson and Neeman (2010), Eberts and Steinwart (2013), Hang and Steinwart (2017), Fischer and Steinwart (2020). Some additional references regarding consistency and robustness (and general stability) are given in Chapters 3 and 4 respectively, where new results on these properties are derived.

It is worth mentioning that there also exist several learning methods that are closely related to SVMs, such as *pairwise learning*. Pairwise learning uses loss functions depending on not only a single input-output tuple (x, y) and a prediction $f(x)$ but instead on pairs $((x, y), (x', y'))$ and either a prediction $f(x, x')$ or a pair of predictions $(f(x), f(x'))$. This approach is popular for different tasks. One such task is *distance metric learning* (Weinberger and Saul, 2009; Bellet and Habrard, 2015; Cao et al., 2016), where it is natural to look at pairs of instances because metrics work on pairs of instances as well, and where the learned metric can then be used to obtain predictions by learning methods such as k-nearest neighbors. Another important application of pairwise learning is that of *ranking* tasks, in which one aims at learning how different instances are ranked relatively to each other (Clémençon et al., 2008; Agarwal and Niyogi, 2009; Zhao et al., 2017). In principle, one could approach such tasks by trying to learn a real-valued function of scores belonging to the individual instances. As the precise values of such scores would however not contain any inherent meaning and as only the pairwise comparisons of such scores are of importance, it is also natural to consider pairwise learning for such tasks. See also Christmann and Zhou (2016), Ying and Zhou (2016), Huang and Wu (2021), Gensler and Christmann (2022) among others (with Ying and Zhou, 2016, combining pairwise learning with online learning) for general considerations on pairwise learning that do not focus on any specific task such as distance metric learning or ranking.

2.1.2 Kernels and Reproducing Kernel Hilbert Spaces

This section gives a short overview of RKHSs and the associated kernel functions. We however only state those properties that are necessary for later chapters, thus of course not giving an extensive introduction. For more detailed descriptions of RKHSs and kernels, see for example Aronszajn (1950), Berlinet and Thomas-Agnan (2004), Saitoh and Sawano (2016) as well as Steinwart and Christmann (2008, Chapter 4), from which most definitions and results from this section are taken. The results can for the most part however also be found in the other references.

We start by giving a definition of *kernels*, which are associated to RKHSs and which are useful for working with the functions contained in the RKHSs and describing them. As we are only interested in $\mathbb{R}$-valued and not in $\mathbb{C}$-valued kernels, the subsequent definition actually comprises two equivalent characterizations (Steinwart and Christmann, 2008, Definition 4.1 and Theorem 4.16), which will both be useful in later chapters.

Definition 2.1.7 (Kernel)**.** Let $\mathcal{X}$ be a non-empty set. Then, a function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a <u>kernel</u> on $\mathcal{X}$ if one of the following two equivalent conditions is satisfied:

(i) There exists an $\mathbb{R}$-Hilbert space H, called <u>feature space</u> of k, and a map $\Phi\colon \mathcal{X} \to H$, called <u>feature map</u> of k, such that

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_H \qquad \forall\, x, x' \in \mathcal{X}\,.$$

(ii) k is <u>symmetric</u>, i.e.

$$k(x, x') = k(x', x) \qquad \forall\, x, x' \in \mathcal{X}\,,$$

and <u>positive definite</u>, i.e.

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \qquad \forall\, m \in \mathbb{N}\,,\ \alpha_1, \ldots, \alpha_m \in \mathbb{R}\,,\ x_1, \ldots, x_m \in \mathcal{X}\,.$$

Looking at the first part of this definition, the evaluation of a kernel can be interpreted as first mapping the inputs into a potentially high-dimensional (or even infinite-dimensional) Hilbert space H via Φ and then computing the inner product of these H-valued representations of the inputs. By choosing a suitable kernel, these computations in H can be performed without ever explicitly moving to this potentially high-dimensional space, thus circumventing the drastic increase in computational demands coming from the high number of dimensions. Notably, this can even be done without knowing Φ, i.e. without knowing how the transformation to H is done. This is also known as the *kernel trick* and is described in detail by Schölkopf et al. (1998) but had already been used in early papers on SVMs such as Boser et al. (1992), Cortes and Vapnik (1995).

Now, by the Riesz-Fréchet representation theorem (see Dudley, 2004, Theorem 5.5.1) each element $h \in H$ can be one-to-one associated to a linear function $g\colon H \to \mathbb{R}$ by defining $g(h') := \langle h, h' \rangle_H$ for all $h' \in H$. Hence, it is also possible to associate each $h \in H$ to a function $f\colon \mathcal{X} \to \mathbb{R}$ by defining $f(x) := \langle h, \Phi(x) \rangle_H$ for all $x \in \mathcal{X}$, which is not necessarily linear in x because of possible non-linearities coming from Φ. As SVMs are elements of certain Hilbert spaces (cf. Definition 2.1.5), they can therefore be seen as not necessarily linear functions on $\mathcal{X}$ which can be obtained by only looking at linear functions on H—without even being required to explicitly perform computations in H because of the kernel trick. In contrast to the relation between elements from H and linear functions on H, the relation between elements from H and functions on $\mathcal{X}$ is however not one-to-one in general. On the one hand, not all functions on $\mathcal{X}$ can be represented by elements from H. This can be mitigated by choosing H large enough that most functions can at least be closely approximated.[6] On the other hand, it is also possible that multiple elements from H all represent the same function on $\mathcal{X}$. This gets circumvented by not choosing any possible combination of feature space and feature map, but instead the combination of

[6] In practice, H is not chosen freely but instead as the RKHS (cf. Definition 2.1.8) of a suitable kernel. However, popular kernels such as the Gaussian RBF kernels have RKHSs that are large enough to, for example, approximate all continuous functions arbitrarily well, cf. Example 2.1.12.

reproducing kernel Hilbert space (see Definition 2.1.8) and the so-called *canonical feature map* (see Remark 2.1.9). For reproducing kernel Hilbert spaces, we also give two different definitions in the following, which are equivalent by Steinwart and Christmann (2008, Lemma 4.19, Theorem 4.20, Theorem 4.21).

Definition 2.1.8 (Reproducing Kernel Hilbert Space)**.** Let $\mathcal{X}$ be a non-empty set and let H be an $\mathbb{R}$-Hilbert space consisting of functions mapping from $\mathcal{X}$ into $\mathbb{R}$. H is called a reproducing kernel Hilbert space (RKHS) over $\mathcal{X}$ if one of the following two equivalent conditions is satisfied:

(i) For all $x \in \mathcal{X}$, the Dirac functional

$$\delta_x \colon H \to \mathbb{R}\,,\; f \mapsto f(x)$$

is continuous.

(ii) There exists a kernel k on $\mathcal{X}$ satisfying $k(\cdot, x) \in H$ for all $x \in \mathcal{X}$ as well as the reproducing property

$$f(x) = \langle f, k(\cdot, x)\rangle_H$$

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

Remark 2.1.9 (Canonical Feature Map). In the situation of the second part of Definition 2.1.8,

$$\Phi \colon \mathcal{X} \to H\,,\; x \mapsto \Phi(x) \coloneqq k(\cdot, x)$$

is called canonical feature map of k. By Steinwart and Christmann (2008, Lemma 4.19), the RKHS H and Φ together form a possible combination of feature space and feature map of k. Whereas in general there exist many possible such combinations, the relation between kernel and RKHS is one-to-one (Steinwart and Christmann, 2008, Theorems 4.20 and 4.21) and the RKHS is in some sense the smallest feature space of the kernel (Steinwart and Christmann, 2008, Theorem 4.21). For these reasons, RKHS and canonical feature map form a canonical choice of feature space and feature map, which we always use if not specified differently, and H is also called the RKHS of k and k the (reproducing) kernel of H.

The subsequent lemma summarizes several interesting properties of kernels and their RKHSs that are useful for later chapters. For this, we define

$$||k||_\infty \coloneqq \sup_{x\in\mathcal{X}} \sqrt{k(x,x)} \tag{2.2}$$

for kernels k on $\mathcal{X}$ and call k bounded if $||k||_\infty < \infty$. Note that this definition of $||k||_\infty$ coincides with the square root of the general definition of the supremum norm when applied to kernels, cf. Cucker and Zhou (2007, p. 22).

Lemma 2.1.10. *Let k be a kernel on $\mathcal{X}$ with RKHS H.*

(i) If k is bounded, then

$$||f||_\infty \leq ||f||_H \, ||k||_\infty$$

for all $f \in H$.

(ii) Let Q be a probability measure on $\mathcal{X}$. If k is bounded and measurable, then $H \subseteq L_p(\mathrm{Q})$ for all $p \in [1, \infty]$.

(iii) If $\mathcal{X}$ is a separable topological space and k is continuous, then H is separable.

(iv) The set

$$H_{\mathrm{pre}} := \left\{ \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) \,\middle|\, m \in \mathbb{N},\ \alpha_1, \dots, \alpha_m \in \mathbb{R},\ x_1, \dots, x_m \in \mathcal{X} \right\}$$

is dense in H. For $f := \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) \in H_{\mathrm{pre}}$, we have

$$||f||_H^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \,.$$

Proof.

(i) See Steinwart and Christmann (2008, Lemma 4.23).

(ii) Follows from part (i) and Steinwart and Christmann (2008, Lemma 4.24) because $||f||_{L_p(\mathrm{Q})} \leq ||f||_\infty$ for all measurable functions and all $p \in [1, \infty]$.

(iii) See Steinwart and Christmann (2008, Lemma 4.33).

(iv) See Steinwart and Christmann (2008, Theorem 4.21). □

Also note that empirical SVMs based on convex loss functions (cf. Definition 2.1.13) always lie in H_{pre} from Lemma 2.1.10(iv), see Steinwart and Christmann (2008, Theorem 5.5).

The next lemma concerns RKHSs of multiples of kernels. This result and its proof are taken from the peer-reviewed paper Köhler and Christmann (2022) that was published in *Journal of Machine Learning Research*. We suppose that the result might be well-established, but we still give the proof because we did not explicitly find it in any literature preceding Köhler and Christmann (2022).

Lemma 2.1.11. *Let $\Omega \neq \emptyset$. Let $k\colon \Omega \times \Omega \to \mathbb{R}$ be a kernel with RKHS H. Let $\alpha > 0$ and define the kernel $\tilde{k}\colon \Omega \times \Omega \to \mathbb{R}$ by $\tilde{k} := \alpha k$. Then, $\tilde{H} := H$ equipped with the norm $||\cdot||_{\tilde{H}} := \frac{1}{\sqrt{\alpha}} ||\cdot||_H$ is the RKHS of $\tilde{k}$.*

Proof. Let $\Phi\colon \Omega \to H$, $x \mapsto k(\cdot, x)$ be the canonical feature map of k. By Remark 2.1.9, we obtain

$$\tilde{k}(x, x') = \alpha k(x, x') = \alpha \left\langle \Phi(x), \Phi(x') \right\rangle_H = \left\langle \sqrt{\alpha}\Phi(x), \sqrt{\alpha}\Phi(x') \right\rangle_H \qquad \forall\, x, x' \in \Omega \,,$$

and hence $\tilde{k}$ is indeed a kernel whose feature space and feature map can be chosen as H and $\tilde{\Phi} := \sqrt{\alpha}\Phi$ respectively.

Now, let $f \in H$ be arbitrary but fixed. Because H is the RKHS of k, the reproducing property yields that $g := \alpha^{-1/2} f \in H$ satisfies

$$\langle g, \tilde{\Phi}(x)\rangle_H = \left\langle \frac{1}{\sqrt{\alpha}} f, \sqrt{\alpha}\Phi(x) \right\rangle_H = \langle f, \Phi(x)\rangle_H = f(x) \qquad \forall\, x \in \Omega\,. \tag{2.3}$$

Hence, Steinwart and Christmann (2008, Theorem 4.21, eq. (4.10)) yields that $\tilde{H} := H$ equipped with a suitable norm is indeed the RKHS of $\tilde{k}$. To derive the norm, note that g is actually the only element of H satisfying (2.3) because for each $h \in H$ with $h \neq g$, there needs to exist at least one $x \in \Omega$ such that

$$f(x) = \sqrt{\alpha} \cdot g(x) \neq \sqrt{\alpha} \cdot h(x) = \sqrt{\alpha} \cdot \langle h, \Phi(x)\rangle_H = \langle h, \tilde{\Phi}(x)\rangle_H\,.$$

Steinwart and Christmann (2008, Theorem 4.21, eq. (4.11)) therefore yields

$$||f||_{\tilde{H}} = \inf\left\{ ||h||_H \,\middle|\, h \in H \text{ with } f = \langle h, \tilde{\Phi}(\cdot)\rangle_H \right\} = ||g||_H = \frac{1}{\sqrt{\alpha}}\, ||f||_H\,. \qquad \square$$

Even though there exist many useful kernels, the one that is probably the most popular choice for SVMs and related machine learning methods is the *Gaussian RBF (radial basis function) kernel*, for which reason this kernel will also appear in several examples throughout this thesis and also in some results in later chapters.

Example 2.1.12 (Gaussian RBF Kernel)**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$. For $\gamma \in (0, \infty)$, the Gaussian RBF kernel (or just Gaussian kernel) k_γ on $\mathcal{X}$ with bandwidth γ is defined by

$$k_\gamma \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}\,,\ (x, x') \mapsto \exp\left(-\frac{||x - x'||_2^2}{\gamma^2} \right)\,.$$

The RKHS of k_γ is large enough to approximate all continuous functions as well as all functions from $L_p(\mathrm{Q})$-spaces (where Q is a probability measure on $\mathcal{X}$) arbitrarily well, cf. Steinwart and Christmann (2008, Definition 4.52, Corollary 4.58, Theorem 4.63). The continuity of k_γ additionally implies the separability of its RKHS (see Lemma 2.1.10(iii)) and, if $\mathcal{X}$ and $\mathcal{Y}$ are both equipped with their respective Borel σ-algebras, also the measurability of k_γ.

Note that one does in practice usually not predetermine the exact kernel before learning an SVM. Instead, a typical course of action is to predetermine some *set of kernels* and then perform cross-validation to choose the kernel yielding the best prediction from this set. For example, one might decide on using a Gaussian RBF kernel and then choose a suitable bandwidth γ by performing cross-validation for all bandwidths from some grid of values. As cross-validation can be time-consuming, there exist approaches to speed it up by suitable approximations, see for example Liu et al. (2020). In addition, Ying and Zhou (2007), Xiang and Zhou (2009), Xiang (2013), Hu and Zhou (2021) (with the last one additionally considering distributed learning, see Section 2.2.1) among others directly include the choice of γ in their optimization problems[7] respectively choose γ depending on the size of the training data set and derive theoretical results such as learning rates by incorporating this flexibility of γ in their analyses. This approach is in parts motivated by Smale and Zhou (2003) who showed that fixed Gaussian RKHSs can—depending on $f^*_{L,\mathrm{P}}$—exhibit approximation errors that decay only logarithmically with respect to the regularization parameter λ. Lastly, Lanckriet et al. (2004), Micchelli and Pontil (2005), Ong et al. (2005), Wu et al. (2007), Ying and Campbell (2009), Cortes et al. (2010), Liu and Liao (2015), Lv et al. (2021) among others investigate learning the kernel for more general classes of kernels. The exact investigated framework differs between the mentioned papers, many however coincide in actually learning a convex combination (or sometimes a more general linear combination) of a set of base kernels. To achieve this, Ong et al. (2005) take the conceptually particularly interesting approach of introducing so-called hyper reproducing kernel Hilbert spaces—RKHSs whose elements are kernels again—and associated hyperkernels. These different approaches are also known as *multi-kernel learning*, and Gönen and Alpaydın (2011) give an extensive review of many of the approaches that existed at that point in time.

2.1.3 Some Additional Properties of Loss Functions and Risks

This section can be viewed as giving a collection of notation and properties regarding loss functions and risks that will be useful in later chapters. Most of these are taken from Steinwart and Christmann (2008).

Definition 2.1.13 (Properties of Loss Functions)**.** Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function.

(i) L is called <u>convex</u> if $L(x, y, \cdot)\colon \mathbb{R} \to [0, \infty)$ is convex for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

(ii) L is called <u>continuous</u> if $L(x, y, \cdot)\colon \mathbb{R} \to [0, \infty)$ is continuous for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

(iii) L is called <u>differentiable</u> if $L(x, y, \cdot)\colon \mathbb{R} \to [0, \infty)$ is differentiable for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. In this case, $L'(x, y, t_0)$ denotes the derivative of $L(x, y, \cdot)\colon \mathbb{R} \to [0, \infty)$ at $t_0 \in \mathbb{R}$ for $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

(iv) L is called <u>Lipschitz continuous</u> if there exists a constant $c \geq 0$ such that

$$\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |L(x, y, t_1) - L(x, y, t_2)| \leq c \cdot |t_1 - t_2| \qquad \forall\, t_1, t_2 \in \mathbb{R}\,.$$

The smallest such constant is denoted by $|L|_1$ and called <u>Lipschitz constant</u> of L.

In practice, one almost always uses convex—and hence also continuous—loss functions as this is needed to guarantee uniqueness of the SVM and also makes SVMs the solutions of convex programs, which makes their calculation computationally feasible because of

[7] In contrast to cross-validation, γ can not only be chosen from a discrete set of possible values in these optimization problems.

one being able to apply known results and methods for such convex programs, see for example Smola and Schölkopf (2004) (some theoretical results will however also hold true without the convexity). Lipschitz continuity will be of special importance in the stability results from Chapter 4, but is not satisfied by the popular least squares loss. On the other hand, whereas the least squares loss is differentiable, this is not satisfied by other popular choices such as the pinball loss or the ε-insensitive loss (cf. Figure 2.1.2). In such non-differentiable but still convex cases, the concept of a *subdifferential* will be useful. The following definition slightly simplifies the definitions found in books on convex analysis such as Rockafellar (1972, Chapter 23) because we only need subdifferentials with respect to one-dimensional arguments.

Definition 2.1.14 (Subdifferential)**.**

(i) Let $f\colon \mathbb{R} \to \mathbb{R}$ be convex. Then, the subdifferential of f at $r_0 \in \mathbb{R}$ is defined by

$$\partial f(r_0) := \left\{ s \in \mathbb{R} \,\middle|\, f(r) \geq f(r_0) + s \cdot (r - r_0) \text{ for all } r \in \mathbb{R} \right\} .$$

(ii) Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex loss function. Then, $\partial L(x, y, t_0)$ denotes the subdifferential of $L(x, y, \cdot)\colon \mathbb{R} \to [0, \infty)$ at $t_0 \in \mathbb{R}$ for $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

(iii) Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex loss function, and let $g\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $f\colon \mathcal{X} \to \mathbb{R}$ be functions. We say that g is from the subdifferential of L with respect to f if $g(x, y) \in \partial L(x, y, f(x))$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

The loss function controls the exact goal of prediction by specifying how different deviations between true output y and prediction $f(x)$ are penalized. The two loss functions already introduced in Section 2.1.1 (least squares loss and pinball loss) coincide in the structure of how this deviation is measured, namely by taking the difference between true output and prediction. Such loss functions are also called *distance-based*, are typically used in regression tasks and will be of special importance in Chapter 3.

Definition 2.1.15 (Distance-Based Loss Function)**.** A loss function $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called distance-based if there exists a representing function $\psi\colon \mathbb{R} \to [0, \infty)$ satisfying $\psi(0) = 0$ and $L(x, y, t) = \psi(y - t)$ for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$. If $\psi(r) = \psi(-r)$ for all $r \in \mathbb{R}$, then L is called symmetric.
Let $p \in (0, \infty)$. A distance-based loss $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ with representing function ψ is of

(i) upper growth type p if there is a constant $c > 0$ such that $\psi(r) \leq c\,(|r|^p + 1)$ for all $r \in \mathbb{R}$.

(ii) lower growth type p if there is a constant $c > 0$ such that $\psi(r) \geq c\,|r|^p - 1$ for all $r \in \mathbb{R}$.

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

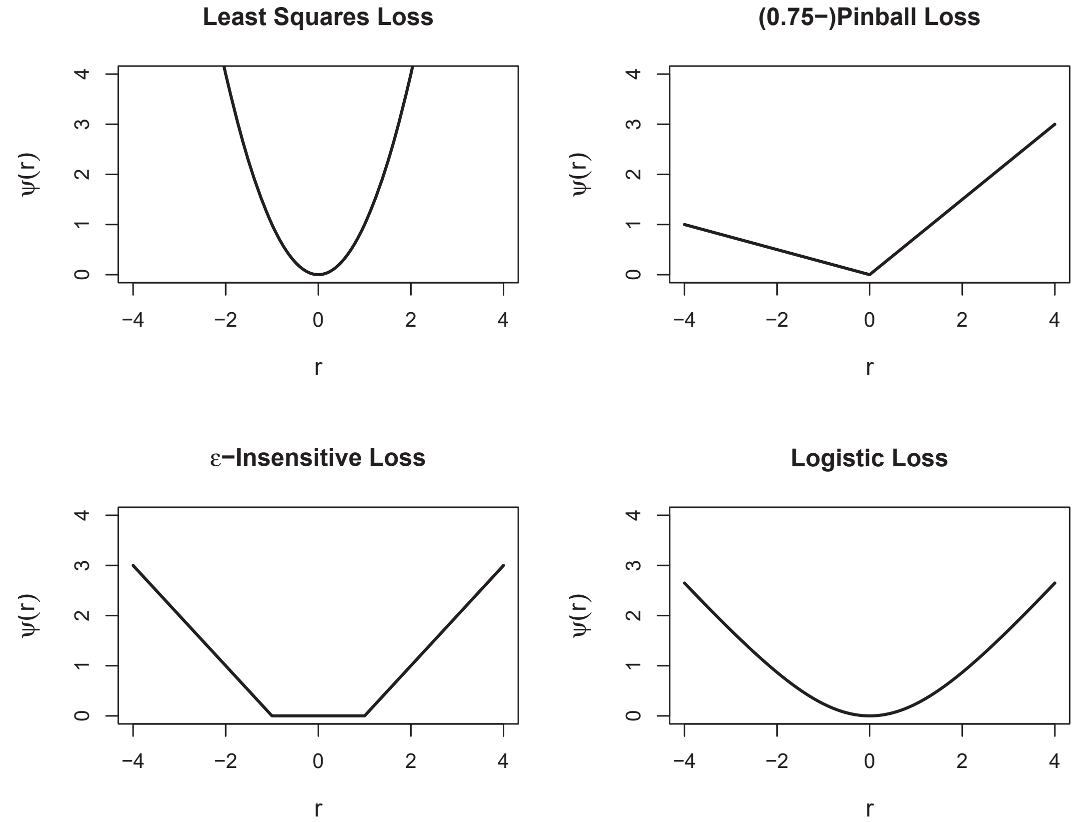

Figure 2.1.2: Representing function ψ for some popular distance-based loss functions.

Since the first argument does not matter in distance-based loss functions, we often ignore it and write $L\colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ and $L(y, t)$ instead.

Even though distance-based losses are typical for regression tasks, some of them, like the least squares loss, are also popular choices for classification tasks, see for example Györfi et al. (2002, Section 1.4). As an example of a distance-based loss, the least squares loss is of growth type 2 whereas many other common loss functions for regression tasks, like the pinball loss, *logistic loss*, *ε-insensitive loss* or *Huber loss*, are of growth type 1. For some of them, the associated representing functions are plotted exemplarily in Figure 2.1.2. The properties examined in later chapters show that the higher growth type sometimes leads to slightly more restrictive conditions regarding P (cf. Chapter 3 and also Remark 2.1.22) and sometimes even to results not being applicable at all (cf. Chapter 4) when using for example the least squares loss.

As many results from later chapters will specifically investigate such distance-based loss functions, some additional properties of them and their associated risks are needed. As a start, the subsequent lemma links properties of such loss functions to those of the respective representing functions. This lemma for the most part coincides with Steinwart and Christmann (2008, Lemma 2.33).

Lemma 2.1.16. *Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a distance-based loss function with representing function $\psi\colon \mathbb{R} \to [0, \infty)$.*

(i) L is convex if and only if ψ is convex.

(ii) L is continuous if and only if ψ is continuous.

(iii) L is Lipschitz continuous if and only if ψ is Lipschitz continuous. In that case, we have $|L|_1 = |\psi|_1$.

Proof. See Steinwart and Christmann (2008, Lemma 2.33). The equality of the two Lipschitz constants in (iii) easily follows because

$$\begin{aligned}&\sup_{y\in\mathcal{Y}} |L(y,t_1) - L(y,t_2)| = \sup_{y\in\mathcal{Y}} |\psi(y-t_1) - \psi(y-t_2)| \\ &\le \sup_{y\in\mathcal{Y}} |\psi|_1 \cdot |(y-t_1) - (y-t_2)| = |\psi|_1 \cdot |t_1 - t_2| \qquad \forall\, t_1, t_2 \in \mathbb{R}\end{aligned}$$

and

$$|\psi(r_1) - \psi(r_2)| = |L(0,-r_1) - L(0,-r_2)| \le |L|_1 \cdot |r_1 - r_2| \qquad \forall\, r_1, r_2 \in \mathbb{R}\,. \qquad \square$$

Additionally, the growth type of a distance-based loss function can be linked to its Lipschitz continuity:

Lemma 2.1.17. *Let $L\colon \mathcal{Y}\times\mathbb{R} \to [0,\infty)$ be a distance-based loss function with representing function $\psi\colon \mathbb{R} \to [0,\infty)$.*

(i) If L is Lipschitz continuous, then it is of upper growth type 1.

(ii) If L is convex and of upper growth type 1, then it is Lipschitz continuous.

(iii) If L is convex and there exist $r_- < 0$ and $r_+ > 0$ such that $\psi(r_-), \psi(r_+) > 0$, then L is of lower growth type 1.

(iv) If L is of lower growth type 1, then there exist $r_- < 0$ and $r_+ > 0$ such that $\psi(r_-), \psi(r_+) > 0$.

Proof.

(i)+(ii) See Steinwart and Christmann (2008, Lemma 2.36).

(iii) See Steinwart and Christmann (2008, Lemma 2.36) and additionally observe that $\lim_{|r|\to\infty} \psi(r) = \infty$ gets implied by the existence of $r_- < 0$ and $r_+ > 0$ such that $\psi(r_-), \psi(r_+) > 0$ because ψ is convex (which follows from L being convex, cf. Lemma 2.1.16) and $\psi(0) = 0$ by definition of distance-based losses.

(iv) Let $r_- := -\frac{2}{c}$ and $r_+ := +\frac{2}{c}$ with c being the constant from the definition of lower growth type 1. Then,

$$\psi(r_-) \ge c \cdot \left|-\frac{2}{c}\right| - 1 = 1 > 0\,,$$

and analogously $\psi(r_+) \ge 1 > 0$. $\square$

Remark 2.1.18. In practice and in most theoretical results, one almost always uses convex loss functions, see also the discussion subsequent to Definition 2.1.13. By parts (i) and (ii) of Lemma 2.1.17, Lipschitz continuity and upper growth type 1 are two properties that can then be used interchangeably if the loss is distance-based. Additionally, it is obvious that, for useful distance-based losses, there should exist $r_- < 0$ and $r_+ > 0$ such that $\psi(r_-), \psi(r_+) > 0$. By also considering parts (iii) and (iv), Lipschitz continuity can hence even be seen as equivalent to growth type 1 for the relevant distance-based loss functions.

Finally, the growth type can also be used to bound different useful quantities by each other. For this, we need two additional definitions.

Definition 2.1.19 (Nemitski Loss Function)**.** A loss function $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called Nemitski loss function if there exists a measurable function $b\colon \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ and an increasing function $h\colon [0, \infty) \to [0, \infty)$ such that

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

If there exists $p \in (0, \infty)$ and a constant $c > 0$ such that $h(|t|) = c|t|^p$ for all $t \in \mathbb{R}$, then L is called Nemitski loss function of order p.
If P is a probability measure on $\mathcal{X} \times \mathcal{Y}$ such that $\int_{\mathcal{X}\times\mathcal{Y}} b(x, y)\, \mathrm{dP}(x, y) < \infty$, then L is called P-integrable Nemitski loss function.

P-integrable Nemitski loss functions can for example be useful for guaranteeing the finiteness of $\mathcal{R}_{L,\mathrm{P}}(f)$ for all bounded functions f, like for example those from the RKHS of a bounded kernel, cf. Lemma 2.1.10(i). We furthermore need the following definition which can be thought of as describing the heaviness of the tails of the conditional distribution $\mathrm{P}(\cdot \,|\, X)$, averaged over the distribution P^X of X:

Definition 2.1.20 (Average p-th Moment)**.** Let P be a probability measure on $\mathcal{X} \times \mathcal{Y}$ and let $p \in (0, \infty)$. The average p-th moment of P is defined by

$$|\mathrm{P}|_p := \left(\int_{\mathcal{X}\times\mathcal{Y}} |y|^p \,\mathrm{dP}(x, y)\right)^{1/p} = \left(\int_{\mathcal{X}}\int_{\mathcal{Y}} |y|^p \,\mathrm{dP}(y \,|\, x)\,\mathrm{dP}^X(x)\right)^{1/p}.$$

Lemma 2.1.21. *Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a distance-based loss function, let* P *be a probability measure on $\mathcal{X} \times \mathcal{Y}$ and let $p \in (0, \infty)$.*

(i) If L is of upper growth type p and $|\mathrm{P}|_p < \infty$, then L is a P*-integrable Nemitski loss of order p.*

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

$$\mathcal{R}_{L,\mathrm{P}}(f) \le c_{L,p} \cdot \left(|\mathrm{P}|_p^p + ||f||_{L_p(\mathrm{P}^X)}^p + 1\right).$$

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

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

(iv) If L is of lower growth type p, there exists a constant $c_{L,p} > 0$ independent of P *such that, for all measurable $f\colon \mathcal{X} \to \mathbb{R}$,*

$$|\mathrm{P}|_p^p \le c_{L,p} \cdot \Big(\mathcal{R}_{L,\mathrm{P}}(f) + ||f||_{L_p(\mathrm{P}^X)}^p + 1\Big)$$

and

$$||f||_{L_p(\mathrm{P}^X)}^p \le c_{L,p} \cdot \Big(\mathcal{R}_{L,\mathrm{P}}(f) + |\mathrm{P}|_p^p + 1\Big)\,.$$

Proof. See Steinwart and Christmann (2008, Lemma 2.38). □

Remark 2.1.22 (Moment Condition)*.* Looking at the definition of SVMs, it is important that the RKHS H contains at least one function with finite risk. Lemma 2.1.21 is the reason why many results for distance-based loss functions of growth type $p \in (0, \infty)$ will impose the easier to interpret moment condition $|\mathrm{P}|_p < \infty$ upon P instead:

(i) If $|\mathrm{P}|_p < \infty$ and L is of upper growth type p, then part (ii) of Lemma 2.1.21 yields $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ and hence *any* RKHS H contains a function with finite risk. If on the other hand $|\mathrm{P}|_p = \infty$ and L is of lower growth type p, then part (iv) of Lemma 2.1.21 yields $\mathcal{R}_{L,\mathrm{P}}(0) = \infty$.
Therefore, $|\mathrm{P}|_p < \infty$ is equivalent to $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ if L is of growth type p.

(ii) If $p \ge 1$ and H is the RKHS of a bounded and measurable kernel (for example the Gaussian RBF kernel, cf. Example 2.1.12), then Lemma 2.1.10(ii) implies that $H \subseteq L_p(\mathrm{P}^X)$ and the preceding can be strengthened to the following:
If $|\mathrm{P}|_p < \infty$ and L is of upper growth type p, then part (ii) of Lemma 2.1.21 even yields $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for *all* $f \in H$. If on the other hand $|\mathrm{P}|_p = \infty$ and L is of lower growth type p, then part (iv) of the lemma even yields $\mathcal{R}_{L,\mathrm{P}}(f) = \infty$ for *all* $f \in H$.
Therefore, $|\mathrm{P}|_p < \infty$ is equivalent to the existence of an $f \in H$ with finite risk and also to *all* $f \in H$ having finite risk if L is of growth type p.

As a counterpart to distance-based loss functions, which are based on the difference between true output and prediction and are typically used in regression tasks, so-called *margin-based loss functions* constitute a second important type of loss functions, which are instead based on the *product* of true output and prediction and are typically used in (binary) classification tasks.

Definition 2.1.23 (Margin-Based Loss Function)**.** A loss function $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called margin-based if there exists a representing function $\varphi\colon \mathbb{R} \to [0, \infty)$ such that $L(x, y, t) = \varphi(yt)$ for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$.

Similarly to distance-based losses, the first argument can also be ignored in margin-based loss functions. Even though margin-based losses will, in contrast to distance-based ones, not play a special role in this thesis (i.e. there are no results only applicable to margin-based losses), we still include their definition because of the general importance of this subtype of loss functions for binary classification. As mentioned at the beginning of Section 2.1, SVMs had originally been developed only for such binary classification tasks with $\mathcal{Y} = \{-1, +1\}$. In these early publications, the *hinge loss*

$$L_{\text{hinge}} \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)\,, \; (x, y, t) \mapsto \max\{0, 1 - yt\} \tag{2.4}$$

had been used, which is still among the most popular choices for such tasks. Further note that the least squares loss is also often used for binary classification tasks and can actually also be interpreted as being margin-based because $L_{\text{LS}}(y, t) = (y - t)^2 = (1 - yt)^2$ holds true for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ for $\mathcal{Y} = \{-1, +1\}$.

Finally, we give some useful additional definitions and lemmas that are not only applicable to distance-based or margin-based loss functions. As an auxiliary tool for analyzing risks, it is sometimes useful to only look at the inner integral from Definition 2.1.2 and define a specific notation for this.

Definition 2.1.24 (Inner Risk)**.** Let $L \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function, P be a probability measure on $\mathcal{X} \times \mathcal{Y}$, $x \in \mathcal{X}$ and $t \in \mathbb{R}$. Then,

$$\mathcal{C}_{L,\mathrm{P}(\cdot \,|\, x),x}(t) \coloneqq \int_{\mathcal{Y}} L(x, y, t) \,\mathrm{dP}(y \,|\, x)$$

is called inner L-risk (or just inner risk) of t at x with respect to P.

Definition 2.1.25 (Inner Bayes Risk)**.** Let $L \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function, P be a probability measure on $\mathcal{X} \times \mathcal{Y}$ and $x \in \mathcal{X}$. Then,

$$\mathcal{C}^*_{L,\mathrm{P}(\cdot \,|\, x),x} \coloneqq \inf_{t \in \mathbb{R}} \mathcal{C}_{L,\mathrm{P}(\cdot \,|\, x),x}(t)$$

is called inner Bayes L-risk (or just inner Bayes risk) at x with respect to P.

The following lemma shows that using the expression "inner Bayes risk" is justified because integrating over $\mathcal{C}^*_{L,\mathrm{P}(\cdot \,|\, x),x}$ indeed yields the Bayes risk, which will be useful in some proofs.

Lemma 2.1.26. *Let $\mathcal{X}$ be a complete measurable space, $L \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function, and* P *be a probability measure on $\mathcal{X} \times \mathcal{Y}$. Then, $x \mapsto \mathcal{C}^*_{L,\mathrm{P}(\cdot \,|\, x),x}$ is measurable and we have*

$$\mathcal{R}^*_{L,\mathrm{P}} = \int_{\mathcal{X}} \mathcal{C}^*_{L,\mathrm{P}(\cdot \,|\, x),x} \,\mathrm{dP}^X(x)\,.$$

Proof. See Steinwart and Christmann (2008, Lemma 3.4). □

Furthermore, the property of convexity gets transferred from loss functions to risks as well as inner risks. For this, denote by $\mathcal{L}_0(\mathcal{X})$ the set of all measurable functions $f \colon \mathcal{X} \to \mathbb{R}$.

Lemma 2.1.27. *Let $L \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex loss function and* P *be a probability measure on $\mathcal{X} \times \mathcal{Y}$. Then, $\mathcal{R}_{L,\mathrm{P}} \colon \mathcal{L}_0(\mathcal{X}) \to [0, \infty]$ is convex and $\mathcal{C}_{L,\mathrm{P}(\cdot \,|\, x),x} \colon \mathbb{R} \to [0, \infty]$ is convex for all $x \in \mathcal{X}$.*

Proof. See Steinwart and Christmann (2008, Lemma 2.13) for the convexity of $\mathcal{R}_{L,\mathrm{P}}$. For $x \in \mathcal{X}$, the convexity of $\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}$ holds true because for all $t_1, t_2 \in \mathbb{R}$ and $\theta \in [0,1]$, we have

$$\begin{aligned}\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}\Big(\theta t_1 + (1-\theta)t_2\Big) &= \int_{\mathcal{Y}} L\Big(x,y,\theta t_1 + (1-\theta)t_2\Big)\,\mathrm{dP}(y\,|\,x) \\ &\leq \theta \int_{\mathcal{Y}} L(x,y,t_1)\,\mathrm{dP}(y\,|\,x) + (1-\theta)\int_{\mathcal{Y}} L(x,y,t_2)\,\mathrm{dP}(y\,|\,x) \\ &= \theta \cdot \mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}(t_1) + (1-\theta)\cdot \mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}(t_2)\end{aligned}$$

due to the convexity of L. □

2.1.4 Shifted Loss Functions

Looking at the definition of SVMs as minimizers of the regularized risk, it is apparent that the existence and uniqueness of SVMs can only be guaranteed if the RKHS H contains at least one function with finite risk (see also Steinwart and Christmann, 2008, Lemma 5.1 and Theorem 5.2) and that this assumption will also be required for more advanced results on SVMs. There are different ways to ensure such an element of H exists. For example, for distance-based loss functions of upper growth type $p \in (0,\infty)$, the moment condition $|\mathrm{P}|_p < \infty$ can be used, which does not only guarantee the existence of such an $f \in H$ but is even equivalent to it under slight additional assumptions, see Remark 2.1.22. However, this moment condition imposes an assumption on the probability measure P which excludes heavy-tailed distributions such as the Cauchy distribution and which can in general not even be verified because of P being unknown. To try to circumvent this problem, we use *shifted loss functions*, which have been applied in robust statistics for a long time, see for example Huber (1967) or Huber (1981, Chapter 3).

Definition 2.1.28 (Shifted Loss Function)**.** Let $L\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to [0,\infty)$ be a loss function. Then,

$$L^\star\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R} \to \mathbb{R}\,,\; (x,y,t) \mapsto L(x,y,t) - L(x,y,0)$$

is called the shifted loss function associated to L.

The corresponding risks, inner risks and SVMs as well as properties of shifted loss functions such as convexity or Lipschitz continuity can be defined analogously to how this was done for regular loss functions, for which reason we do not repeat these definitions here.

Even though this shift of the loss function does not seem to really change anything at first glance, Christmann et al. (2009) showed that it can indeed be used to eliminate the moment condition from many results on SVMs in the case of L being Lipschitz continuous (which is basically equivalent to L being of growth type 1 if it is distance-based, cf. Remark 2.1.18). They observed that

$$\mathcal{R}_{L,\mathrm{P}}(f) \leq |L|_1 \cdot ||f||_{L_1(\mathrm{P}^X)} + |L|_1 \cdot |\mathrm{P}|_1 \qquad \forall\, f \in \mathcal{L}_0(\mathcal{X})\,,$$

from which the finiteness of the risk can only be guaranteed under the moment condition $|\mathrm{P}|_1 < \infty$ when using L, while at the same time

$$|\mathcal{R}_{L^\star,\mathrm{P}}(f)| \leq |L|_1 \cdot ||f||_{L_1(\mathrm{P}^X)} \qquad \forall\, f \in \mathcal{L}_0(\mathcal{X})\,, \tag{2.5}$$

which means that the moment condition is not needed when using $L^\star$, since $\mathcal{R}_{L^\star,\mathrm{P}}(f)$ is finite for all $f \in L_1(\mathrm{P}^X)$ even if $|\mathrm{P}|_1 = \infty$. Also note that $L^\star$ and $\mathcal{R}_{L^\star,\mathrm{P}}$ can—in contrast to L and $\mathcal{R}_{L,\mathrm{P}}$—also take on negative values and it is therefore not only necessary to have $\mathcal{R}_{L^\star,\mathrm{P}}(f) < \infty$ for at least one $f \in H$ but also $\mathcal{R}_{L^\star,\mathrm{P}}(f) > -\infty$ for *all* $f \in H$ in order to guarantee existence and uniqueness of the SVM $f_{L^\star,\mathrm{P},\lambda,k}$ (see also Christmann et al., 2009, Theorems 5 and 6). As (2.5) however bounds the absolute value of $\mathcal{R}_{L^\star,\mathrm{P}}(f)$, one can conclude that using shifted loss functions looks promising when combining a Lipschitz continuous loss with a bounded and measurable kernel because the latter by Lemma 2.1.10(ii) guarantees that $H \subseteq L_1(\mathrm{P}^X)$ and hence that the right hand side of (2.5) is finite for all $f \in H$.

Indeed, Christmann et al. (2009) showed that many of the results on the desirable properties that regular SVMs possess can be transferred to SVMs using shifted loss functions without needing the moment condition that was required in the non-shifted case. In addition to the already mentioned existence and uniqueness, these results include a representer theorem as well as results on risk consistency and robustness. Moreover, the use of shifted loss functions is justified in that the loss gets shifted by a fixed amount (independently of the prediction that is plugged in), for which reason this shift does not change the learning goal, which gets reaffirmed by the following:

Lemma 2.1.29. *Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a loss function and let $L^\star$ be its shifted version. Let P be a probability measure on $\mathcal{X} \times \mathcal{Y}$, let H be an RKHS over $\mathcal{X}$ and let $\lambda > 0$. If $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, then*

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

Proof. See Christmann et al. (2009, p. 314). □

Finally, the following lemma tells us that certain properties of loss functions and the respective properties of the associated shifted loss functions can be used interchangeably, for which reason it will for example not matter which of L and $L^\star$ we require to be convex or Lipschitz continuous in the results from later chapters.

Lemma 2.1.30. *Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a loss function and let $L^\star$ be its shifted version.*

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

(ii) L is Lipschitz continuous if and only if $L^\star$ is Lipschitz continuous. Furthermore, we have $|L|_1 = |L^\star|_1$.

Proof. See Christmann et al. (2009, Proposition 2) for the one direction. The other direction follows analogously. □

Because of the sketched advantages, Chapter 4 will only investigate SVMs using shifted loss functions. Because of Lemma 2.1.29, the results from that chapter are however also valid for SVMs using regular loss functions whenever the risk of the zero function is finite. Chapter 3 takes a slightly different approach in that it investigates SVMs using regular loss functions and those using shifted ones separately, which is due to the surprising observation that the main advantage of shifted loss functions—eliminating the moment condition when working with Lipschitz continuous losses—does actually not come into effect completely for some of the results from that chapter.

2.2 Introduction to Localized Support Vector Machines

Whereas SVMs possess many desirable theoretical properties, as reviewed in Section 2.1.1, they suffer from their computational requirements growing at least quadratically in the size of the training data set, see for example Platt (1998), Joachims (1998), Thomann et al. (2017). One of the possible approaches to reduce this computational complexity is localization, which can additionally offer some advantages regarding the quality of prediction as well. Section 2.2.1 starts by giving a quick overview of some other existing approaches and then gives an informal introduction to the idea behind localization as well as stating the mentioned additional advantages. Afterwards, Section 2.2.2 formalizes the approach by mathematically defining localized SVMs and introducing some notation that is needed in later chapters.

Parts of Section 2.2.1 coincide with the introduction already given in the peer-reviewed paper Köhler (2024a, Section 3.1) that was published in *Neurocomputing*.

2.2.1 Localized Learning and Other Approaches to Deal with a Large Amount of Data

There exist different approaches to reduce the computational complexity of SVMs. A popular such approach is *online learning* by *stochastic gradient descent*, which tackles the high computational demands by only looking at a single training point respectively only a small batch of training points at a time and updating the learned predictor iteratively (Smale and Yao, 2006; Ying and Zhou, 2006; Dieuleveut and Bach, 2016; Lin and Rosasco, 2017; Ying and Zhou, 2017). Note that some of the referenced publications deviate further from the SVM approach by omitting the explicit regularization via $\lambda \, ||f||_H^2$, which gets justified by noting that the step size used by the gradient descent algorithm serves as a form of implicit regularization. Another approach is to find a suitable *low-rank approximation to the kernel matrix* by smaller matrices (and hence effectively reducing the size of the problem) by performing *column subsampling* (Williams and Seeger, 2000; Bach, 2013; El Alaoui and Mahoney, 2015; Rudi et al., 2015), for example using the *Nyström method.* Instead of approximating the kernel matrix, one can also approximate the whole kernel function by *random features* and obtain a low-dimensional feature representation, through which it is possible to efficiently find a good predictor (Rahimi and Recht, 2007; Sriperumbudur and

Szabó, 2015; Rudi and Rosasco, 2017; Liu et al., 2022; Mei et al., 2022). A popular type of random feature approximation is using *random Fourier features*, which can be applied to a large class of kernels (including the Gaussian RBF kernels). This type of random feature approximation makes use of Bochner's theorem, stating that each kernel on $\mathbb{R}^d$ that is shift-invariant (i.e. $k(x, y) = k(x + t, y + t)$ for all $x, y, t \in \mathbb{R}^d$) and continuous can be seen as the Fourier transform of a finite measure on $\mathbb{R}^d$ (Wendland, 2005, Theorem 6.6) respectively even of a probability measure if one scales the kernels properly, and approximates the Fourier transform—and hence also the kernel—by sampling from said probability measure and plugging this sampled data into the integrand of the Fourier transform. Further, Yang et al. (2012) compare the column subsampling approach (more specifically, the Nyström method) with the random features approach (more specifically, random Fourier features), and Rudi et al. (2017), Meanti et al. (2020) combine multiple of the mentioned approaches.

Closer to the localized approach are, however, methods that decompose the available data set into $m \in \mathbb{N}$ subsets and train m "small" SVMs on these subsets instead of a single "large" one on all of D_n, which can substantially reduce the training time as well as required storage space because of the aforementioned super-linear computational requirements of SVMs. This can for example be done by means of *distributed learning*, which randomly splits D_n into subsets, trains an SVM on each such subset, and then averages the resulting m SVMs in order to obtain the final predictor (Christmann et al., 2007; Zhang et al., 2015; Guo et al., 2017; Lin et al., 2017; Mücke and Blanchard, 2018; Sun and Wu, 2021; Liu and Shi, 2024).

In the localized approach, one also trains SVMs on subsets of D_n, but the split of D_n is now obtained in a spatial way—based on some regionalization of the input space $\mathcal{X}$—instead of randomly. Following early theoretical investigations of such localized approaches (Bottou and Vapnik, 1992; Vapnik and Bottou, 1993), different methods for obtaining the required regions have been examined. One such method is the use of decision trees, see for example Bennett and Blue (1998), Wu et al. (1999), Tibshirani and Hastie (2007), Chang et al. (2010). Among these, Chang et al. (2010) generate the tree by splitting the data in an axis-parallel way, whereas the other three articles all propose using an SVM for each decision in the tree. This difference is due to Bennett and Blue (1998), Wu et al. (1999), Tibshirani and Hastie (2007) mainly aiming at improving the accuracy of the predictor, whereas Chang et al. (2010) also want to reduce the training time. Cheng et al. (2010), Gu and Han (2013) on the other hand split the training data into clusters based on variants of k-means and then train an SVM on each resulting cluster, and other articles propose k-nearest neighbors (kNN) methods for obtaining the regionalization. Here, we have to differ between those approaches that measure distances for selecting the k nearest neighbors in the input space $\mathcal{X}$ (Zhang et al., 2006; Hable, 2013) and those that measure them in the feature space H (Blanzieri and Bryl, 2007; Blanzieri and Melgani, 2008; Segata and Blanzieri, 2010). One drawback of such kNN methods is their usually slow and computationally intensive prediction phase, which is due to them having to compute a new SVM in the k-neighborhood of each test point during the prediction phase. Even though each of these SVMs is only based on k data points instead of the whole training data set, this still considerably slows down the prediction phase if a large amount of predictions has to be made. Segata and Blanzieri (2010) mitigate this problem by slightly adapting the kNN approach to instead train an SVM on the k-neighborhood of each training point during the training phase and then use the SVM belonging to the single closest training point for predicting the output to a test point.

In comparison to distributed learning, all these localized approaches have the disadvantage that, no matter which method of regionalization is chosen, the process of regionalizing the input space clearly also takes some time for large data sets—albeit considerably less time than just training an SVM on the whole data set—, thus making the computational gain of such a localized approach in the training phase smaller than that of distributed learning. However, localization usually also results in a *significantly faster prediction phase*, that is a significantly faster evaluation of the resulting predictor for test samples—in comparison to not only regular SVMs but also to the distributed approach. Whereas one has to evaluate each of the m different SVMs (and then average the results) in distributed learning, it suffices to evaluate the one SVM belonging to the region of the test sample in localized learning (if the regions do not overlap). Comparing the localized approach to regular SVMs on the other hand, one obtains from the empirical representer theorem (see Steinwart and Christmann, 2008, Theorem 5.5) for SVMs that

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

for $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. Depending on L, different lower and upper bounds on the number of $\alpha_i \neq 0$ can be derived, see for example Steinwart (2003). For simplicity of the argument, consider a loss function for which $\alpha_i \neq 0$ for all $i \in \{1, \dots, n\}$ holds true almost surely, such as the logistic loss for classification, see Steinwart and Christmann (2008, Proposition 8.30). In this case, one has to evaluate n kernel functions in order to evaluate the SVM $f_{L,\mathrm{D}_n,\lambda,k}$. For the localized approach on the other hand—if we assume for simplicity that each of the m subsets that D_n is split into has approximately the same size, that is size n/m—one only has to evaluate those approximately n/m kernel functions belonging to training samples from the region of the test sample (if the regions do not overlap).

Several of the referenced publications on localized SVMs also include experimental analyses of how much the respective methods of localization reduce the computation time in comparison to regular SVMs. For example, Chang et al. (2010) compared their decision tree based localized SVMs *DTSVM* (among some other approaches) to regular SVMs on different medium-size data sets (ca. 10,000 to 500,000 samples, using about two thirds of them for training) and observed drastic reductions in training time, especially for the larger ones among these data sets. Depending on how the training of SVMs was implemented, the training time for the largest data set was reduced by a factor of almost 3,700 or even 5,800 (Chang et al., 2010, Figures 5 and 10). At the same time, *DTSVM* exhibited comparable or even increased test accuracy over regular SVMs (Chang et al., 2010, Figures 6 and 11) and also drastically reduced testing time (Chang et al., 2010, Table 4). Additionally, they also took a look at some large-size data sets (ca. 600,000 to 5,000,000 samples), for which they could not perform a comparison with regular SVMs because of them requiring an excessive amount of training time and memory, but showed that *DTSVM* is able to still perform well on such large-size data sets. Similarly, Segata and Blanzieri (2010) compared different variants of their kNN based localized SVMs with regular SVMs on different data sets containing ca. 50,000 to 1,000,000 data samples and also observed decreased training and testing times (by factors ranging up to ca. 100 for the variant *FaLK-SVM*, which was the one observing the highest test accuracy) as well as comparable or even increased test accuracy (Segata and Blanzieri, 2010, Tables 5–7). Gu and Han (2013) performed similar comparisons for their k-means based localized SVMs *CSVM* (as well as for some other approaches), however only for data sets containing only ca. 3,000 to 60,000 training samples—which were however high-dimensional, consisting of up to 784 features. Even for these comparatively small data sets, *CSVM* exhibits training times that are considerably lower than those of regular SVMs (called "kernel SVM" in their tables), by factors of up to almost 130, while at the same time yielding comparable test accuracy (Gu and Han, 2013, Tables 2 and 3). In all of these three publications, the library `LIBSVM` (Chang and Lin, 2011) was used for computing SVMs. Thomann et al. (2017) on the other hand used their own library `liquidSVM` (Steinwart and Thomann, 2017) and looked at large training data sets of up to almost 10,000,000 samples. They showed that localized SVMs based on the Voronoi partition approach that is built into `liquidSVM` can be computed in few hours and yield good test accuracy (as with the large-size data sets used by Chang et al. (2010), it was of course not feasible to also compute regular SVMs for such large data sets in order to compare them). Additionally, by using not only a single but instead multiple machines, they succeeded in obtaining good results in just a little over one day of combined training and testing time even for an enormous training data set consisting of 32,000,000 samples in 631 dimensions—whereas among the other data sets used in the four papers mentioned in this paragraph, there was none that had more than 54 dimensions and at the same time more than 240,000 samples.

Even though the exact training and testing time depends on the method chosen for localization as well as on the exact implementation and therefore differs between these publications, they all observed a drastic reduction compared to the computation time of regular SVMs. In addition to this computational gain, localizing the SVM approach can also yield advantages regarding the *quality of prediction*—compared to distributed learning as well as regular SVMs: Whereas the underlying true function, which one aims to estimate, can of course exhibit discontinuities, SVMs based on a continuous and bounded kernel such as the commonly used Gaussian RBF kernel from Example 2.1.12 are always continuous (and bounded) themselves (Steinwart and Christmann, 2008, Lemma 4.28). This can lead to SVMs not accurately modeling the true function near such discontinuities, but instead greatly oscillating and overshooting—an effect that is also known from Fourier series, where it is called the Gibbs phenomenon, cf. Hewitt and Hewitt (1979). Additionally, in global learning approaches like SVMs, the complexity of the predictor is usually controlled globally by a very small amount of hyperparameters—in the case of SVMs by the regularization parameter λ and potentially by hyperparameters of the kernel, for example the bandwidth γ of the Gaussian RBF kernel. Hence, an accurate prediction can be difficult for such global approaches if the complexity and variability of the true function, or that of the conditional distributions $\mathrm{P}(Y \mid X = x)$, greatly differ between different areas of the input space $\mathcal{X}$, even if the true function does not exhibit any discontinuities. Both of these problems can be overcome by the use of localized methods, as a good regionalization can split the input space into separate regions at (or at least close to) discontinuities and such that the

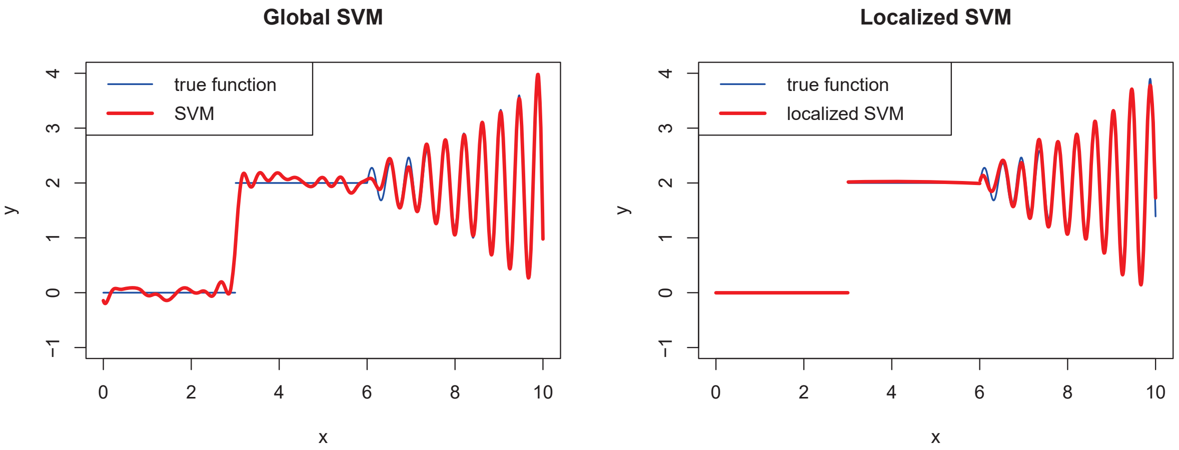

Figure 2.2.1: A global SVM (left plot) and a localized SVM (right plot; splits between the regions at $x = 3$ and $x = 6$) fitted to the same data which was generated according to the plotted true function and some normally distributed error. The global SVM (slightly) overshoots at the discontinuity at $x = 3$ and oscillates too much for $x \le 6$ because the underlying hyperparameters have to be chosen in a way that also allows for a reasonably good fit for $x > 6$, where the true function oscillates very quickly. The localized SVM does not exhibit these problems and yields a considerably better fit overall. [This is a minimally modified version of a figure that was first published in Köhler, 2024a.]

complexity and variability do not change too much throughout the individual regions.

To exemplify the increased quality of prediction that can be the result of localizing, we take a look at the following toy example:

Example 2.2.1. Let $\mathcal{X} = \mathbb{R}$ and $\mathrm{P}^X = \mathcal{U}(0, 10)$ be the uniform distribution on $(0, 10)$. Let $\mathcal{Y} = \mathbb{R}$ and let the output $y \in \mathcal{Y}$ be obtained by adding an $\mathcal{N}(0, \sigma^2)$-distributed error to

$$f(x) = \begin{cases} 0 & \text{, if } x \le 3 \\ 2 & \text{, if } 3 < x \le 6 \\ 2 + \sin(15(x-6)) \cdot \frac{x^4}{5000} & \text{, else} \,. \end{cases}$$

We computed a global SVM as well as a localized SVM—with fixed regions capturing the pattern in the data by splitting the regions at $x = 3$ and $x = 6$—for different values of σ and of the training set size n. The SVMs were all based on the 0.5-pinball loss function $L_{0.5\text{-pin}}$, for which reason the Bayes function is the function of conditional medians, which coincides with the underlying function f because of the symmetry of the additive error terms. Figure 2.2.1 exemplarily depicts the resulting predictors for $\sigma = 0.5$ and $n = 3{,}000$ and shows that the localized SVM does indeed yield a considerably better fit to the Bayes function f than the global SVM does.

We further computed such global and localized SVMs for all $\sigma \in \{0.01, 0.1, 0.5, 1\}$ for $n = 2{,}000$ as well as for $n = 8{,}000$ and estimated the associated risks based on 10,000

n	2,000		8,000	
σ	glob.	loc.	glob.	loc.
0.01	1.95	0.36	1.37	0.34
0.1	3.82	0.49	1.96	0.13
0.5	15.32	5.02	7.18	1.63
1	42.26	18.52	14.10	5.96

(a) Medians

n	2,000		8,000	
σ	glob.	loc.	glob.	loc.
0.01	0.15	0.02	0.04	0.02
0.1	0.34	0.08	0.08	0.02
0.5	1.50	1.04	0.53	0.37
1	5.30	3.95	1.63	1.35

(b) Median absolute deviations

Table 2.2.1: (a) Medians and (b) median absolute deviations of the excess risks of global and localized SVMs in the situation of Example 2.2.1 for different combinations of σ and n, based on 1,000 iterations of training a global respectively a localized SVM for each combination. To improve the readability and ease comparisons between the different combinations, all excess risks were multiplied by the factor 1,000.

test data points. We repeated this 1,000 times for each combination of σ and n in order to minimize the effect of chance and to also get an estimate of the variation in the risks. The computations were performed in R Statistical Software (R Core Team, 2022, v4.2.2) using the library `liquidSVM` (Steinwart and Thomann, 2017) for computing the SVMs. We computed the excess risks

$$\mathcal{R}_{L_{0.5\text{-pin}},\mathrm{P}}(g) - \mathcal{R}^*_{L_{0.5\text{-pin}},\mathrm{P}} = \mathcal{R}_{L_{0.5\text{-pin}},\mathrm{P}}(g) - \mathcal{R}_{L_{0.5\text{-pin}},\mathrm{P}}(f)$$

for g being the respective global SVM as well as for g being the respective localized SVM and collected the results in Table 2.2.1 (medians and median absolute deviations) and Figure 2.2.2 (box plots).[8] Whereas the exact values of the excess risks of course differ between the different combinations of σ and n—increasing as σ increases and decreasing as n increases (note that the scaling of the box plots differs between the different combinations of σ and n as the box plots for small σ and large n would be barely identifiable otherwise)—, the overall shape of the localized SVMs possessing lower excess risks than the global SVMs is always very similar, for which reason the results affirm the observations from Figure 2.2.1. Note that the box plots even show that for five of the eight examined combinations of σ and n (those with $\sigma \in \{0.01, 0.1\}$ and the combination of $\sigma = 0.5$ and $n = 8{,}000$), the *worst* excess risk of a localized SVM was still *better* than the *best* excess risk of a global SVM in these 1,000 repetitions.

The intuition of localized SVMs also being able to improve regular SVMs with regard to the quality of prediction gets affirmed mathematically by Blaschzyk and Steinwart (2022), who, in the case of using the hinge loss for classification, derive learning rates exceeding those known for regular SVMs. Whereas most of the papers on localized SVMs mentioned in the preceding paragraphs focus on the experimental analysis of a specific method of localization, Blaschzyk and Steinwart (2022) constitutes an example of a paper deriving theoretical results and additionally not requiring any special method of localization (instead

[8] In both displays, we multiplied all observed excess risks by the factor 1,000 in order to move to a scale that makes quick comparisons between the different values easier.

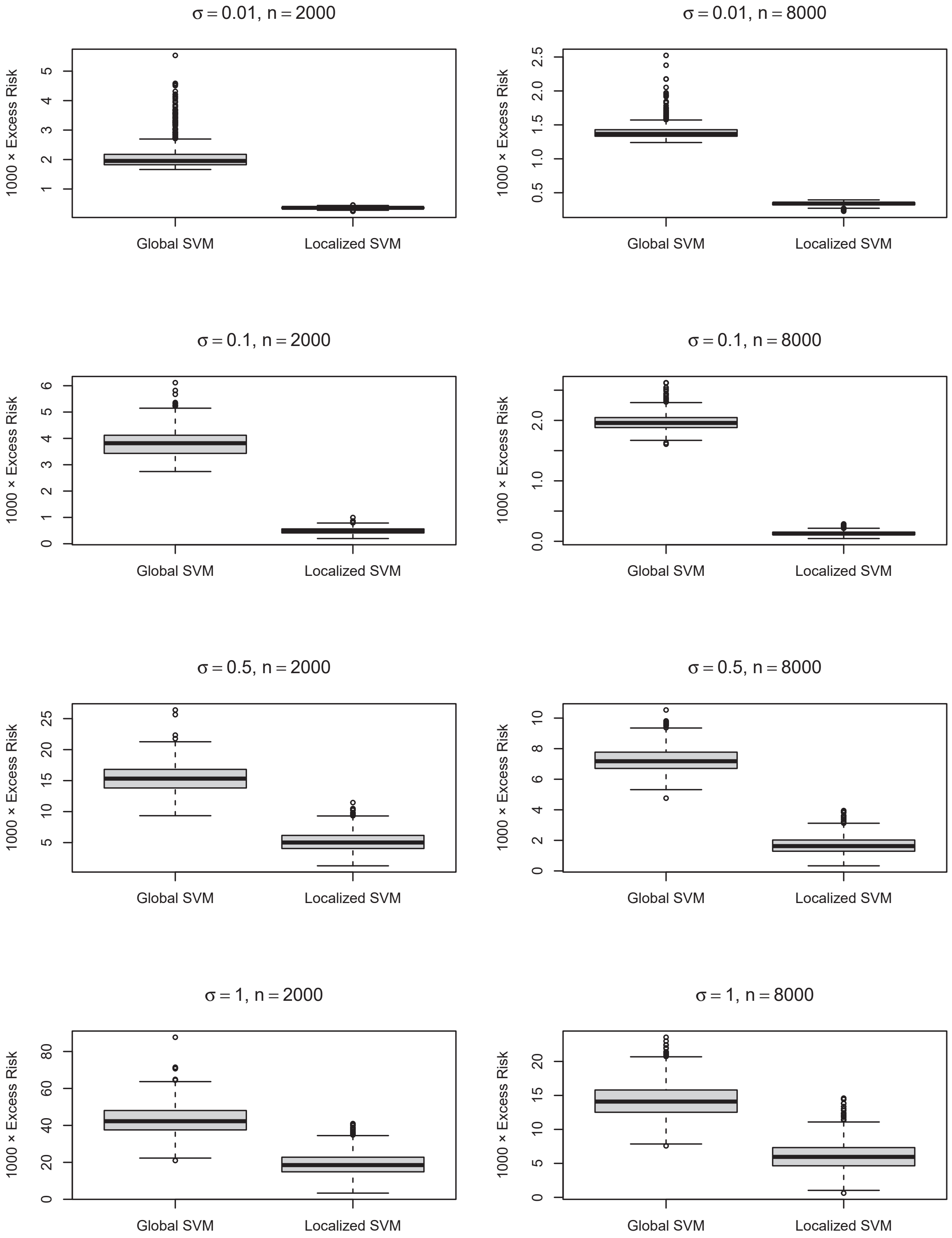

Figure 2.2.2: Box plots of the excess risks of global and localized SVMs in the situation of Example 2.2.1 for different combinations of σ and n. Each box plot is based on 1,000 iterations of training a global respectively a localized SVM. To improve the readability and ease comparisons between the different box plots, all excess risks were multiplied by the factor 1,000.

only requiring the resulting regionalization to satisfy certain conditions). There are several papers taking a similar approach and also deriving learning rates for such localized SVMs, with Thomann et al. (2017) also using the hinge loss, Meister and Steinwart (2016), Mücke (2019) investigating least squares regression, and Hamm and Steinwart (2022) considering both for a specific method of localization. Carratino et al. (2021) also examined least squares regression, but regionalized the feature space instead of the input space and combined localization with other techniques like column subsampling. Additionally, Dumpert and Christmann (2018), Dumpert (2020) proved that localized SVMs are risk consistent (which in some aspects gets considerably generalized in Section 3.4.3) as well as statistically robust.

2.2.2 Definition of Localized Support Vector Machines

To formally define localized SVMs, one first needs to split the input space $\mathcal{X}$ into different regions. The set of regions is called a *regionalization*:

Definition 2.2.2 (Regionalization)**.** A regionalization of $\mathcal{X}$ of size $A \in \mathbb{N}$ is a set $\boldsymbol{\mathcal{X}} := \{\mathcal{X}_1, \dots, \mathcal{X}_A\}$, where $\mathcal{X}_1, \dots, \mathcal{X}_A \subseteq \mathcal{X}$ are non-empty and measurable and satisfy $\bigcup_{a=1}^{A} \mathcal{X}_a = \mathcal{X}$. The sets $\mathcal{X}_a$, $a = 1, \dots, A$, are called regions. $\boldsymbol{\mathcal{X}}$ is called a partitioning regionalization if $\mathcal{X}_1, \dots, \mathcal{X}_A$ are pairwise disjoint.

Remark 2.2.3. Throughout this thesis, it is assumed that a regionalization is on hand and properties of the resulting localized SVMs are examined. How to obtain such a regionalization in a sensible way is of course important as well, but not the topic of this thesis, and we refer to the references given in Section 2.2.1 for more information on finding a good regionalization.

We impose different additional assumptions on the regionalizations in the sections deriving results for localized SVMs (Sections 3.4 and 4.4), all of them however being rather mild. Note that a regionalization does in general not need to be partitioning, but the regions can instead also overlap. The applied regionalizations being partitioning is required only in Section 4.4.2. For comparing two localized SVMs that are based on different localizations, the *combined regionalization* can be useful:

Definition 2.2.4 (Combined Regionalization)**.** The combined regionalization $\boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}}$ of two regionalizations $\boldsymbol{\mathcal{X}}_{\mathbf{1}} := \{\mathcal{X}_{1,1}, \dots, \mathcal{X}_{1,A_1}\}$ and $\boldsymbol{\mathcal{X}}_{\mathbf{2}} := \{\mathcal{X}_{2,1}, \dots, \mathcal{X}_{2,A_2}\}$ of $\mathcal{X}$ is defined as

$$\begin{aligned}\boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}} &:= \{\mathcal{X}^*_1, \dots, \mathcal{X}^*_B\} \\ &:= \{\mathcal{X}_{1,a_1} \cap \mathcal{X}_{2,a_2} \mid \mathcal{X}_{1,a_1} \in \boldsymbol{\mathcal{X}}_{\mathbf{1}},\, \mathcal{X}_{2,a_2} \in \boldsymbol{\mathcal{X}}_{\mathbf{2}}\} \setminus \{\emptyset\}\,.\end{aligned}$$

In order to define SVMs on the different regions, one first needs to find suitable measures on these regions: Given some region $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ and some probability measure P on $\mathcal{X} \times \mathcal{Y}$, it suggests itself to define a *local measure* on $\tilde{\mathcal{X}} \times \mathcal{Y}$ by

$$\mathrm{P}_{\tilde{\mathcal{X}}} := \begin{cases} \frac{1}{\mathrm{P}(\tilde{\mathcal{X}} \times \mathcal{Y})} \cdot \mathrm{P}|_{\tilde{\mathcal{X}} \times \mathcal{Y}} & , \text{if } \mathrm{P}(\tilde{\mathcal{X}} \times \mathcal{Y}) > 0\,, \\ 0 & , \text{else}\,. \end{cases} \tag{2.6}$$

Note that this obviously is a probability measure only if $\mathrm{P}(\tilde{\mathcal{X}} \times \mathcal{Y}) > 0$. For an empirical measure D_n associated to a data set $D_n := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, this definition of $(\mathrm{D}_n)_{\tilde{\mathcal{X}}}$ yields exactly the same measure as first choosing the subset $(D_n)_{\tilde{\mathcal{X}}} := D_n \cap (\tilde{\mathcal{X}} \times \mathcal{Y})$ and then constructing the associated empirical measure. Given a regionalization $\boldsymbol{\mathcal{X}} := \{\mathcal{X}_1, \ldots, \mathcal{X}_A\}$, further denote by $\mathbf{P}_{\boldsymbol{\mathcal{X}}} := (\mathrm{P}_{\mathcal{X}_1}, \ldots, \mathrm{P}_{\mathcal{X}_A})$ the vector of corresponding local measures. These local measures can be used to define *local SVMs*.

Definition 2.2.5 (Local Support Vector Machine)**.** Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ and P be a probability measure on $\mathcal{X} \times \mathcal{Y}$. Let $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ be non-empty and measurable. Let k be a kernel on $\tilde{\mathcal{X}}$ with RKHS H and let $\lambda > 0$ Then,

$$f_{L,\mathrm{P}_{\tilde{\mathcal{X}}},\lambda,k} := \begin{cases} \arg\inf_{f \in H} \mathcal{R}_{L,\mathrm{P}_{\tilde{\mathcal{X}}}}(f) + \lambda \left|\left|f\right|\right|_H^2 & \text{, if } \mathrm{P}(\tilde{\mathcal{X}} \times \mathcal{Y}) > 0\,, \\ 0 & \text{, else.} \end{cases}$$

is called local SVM on $\tilde{\mathcal{X}}$.

Note that a local SVM is just a regular SVM on the region $\tilde{\mathcal{X}}$ if $\mathrm{P}(\tilde{\mathcal{X}} \times \mathcal{Y}) > 0$ holds true. In some situations, it is required that all local SVMs are indeed SVMs, which leads to the following definition:

Definition 2.2.6 (Positive Probability Measure on Regionalization)**.** For a regionalization $\boldsymbol{\mathcal{X}} := \{\mathcal{X}_1, \ldots, \mathcal{X}_A\}$ of $\mathcal{X}$, a probability measure P on $\mathcal{X} \times \mathcal{Y}$ is called positive on $\boldsymbol{\mathcal{X}}$ if $\mathrm{P}(\mathcal{X}_a \times \mathcal{Y}) > 0$ for all $a \in \{1, \ldots, A\}$.

Now, all that remains to do, is to plug the local SVMs on the different regions together in a suitable way in order to obtain a global predictor. For this, they first need to be extended such that they are defined on all of $\mathcal{X}$. For $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ and a function $g\colon \tilde{\mathcal{X}} \to \mathbb{R}$, denote its zero-extension to $\mathcal{X}$ by

$$\hat{g}\colon \mathcal{X} \to \mathbb{R}\,,\; x \mapsto \begin{cases} g(x) & \text{, if } x \in \tilde{\mathcal{X}}\,, \\ 0 & \text{, else}\,. \end{cases}$$

Similarly, zero-extensions of kernels and probability measures will also be needed in later chapters and are denoted by the $(\hat{\cdot})$-notation as well: If $k\colon \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \to \mathbb{R}$ is a kernel on $\tilde{\mathcal{X}} \subseteq \mathcal{X}$, denote

$$\hat{k}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}\,,\; (x, x') \mapsto \begin{cases} k(x, x') & \text{, if } x, x' \in \tilde{\mathcal{X}}\,, \\ 0 & \text{, else}\,. \end{cases}$$

By Meister and Steinwart (2016, Lemma 2) this indeed defines a kernel on $\mathcal{X}$. If $\mathcal{X} \times \mathcal{Y}$ is equipped with the σ-algebra $\mathcal{F}$ and Q is a probability measure on $\tilde{\mathcal{X}} \times \mathcal{Y}$ for a measurable $\tilde{\mathcal{X}} \subseteq \mathcal{X}$, denote

$$\hat{\mathrm{Q}}\colon \mathcal{F} \to [0, 1]\,,\; S \mapsto \mathrm{Q}\left(S \cap \left(\tilde{\mathcal{X}} \times \mathcal{Y}\right)\right)\,.$$

Lastly, as the regions of a regionalization $\boldsymbol{\mathcal{X}} = \{\mathcal{X}_1, \ldots, \mathcal{X}_A\}$ are not necessarily pairwise disjoint, the influence of the different local SVMs on the resulting global predictor needs to be controlled pointwisely by means of weight functions w_a, $a = 1, \ldots, A$, associated with the regionalization. Throughout this thesis, we impose the following three standard assumptions on such a set of weight functions:

(W1) $w_a \colon \mathcal{X} \to [0,1]$ measurable for all $a \in \{1, \dots, A\}$.

(W2) $\sum_{a=1}^{A} w_a(x) = 1$ for all $x \in \mathcal{X}$.

(W3) $w_a(x) = 0$ for all $a \in \{1, \dots, A\}$ and $x \notin \mathcal{X}_a$.

Definition 2.2.7 (Localized Support Vector Machine)**.** Let $\boldsymbol{\mathcal{X}} = \{\mathcal{X}_1, \dots, \mathcal{X}_A\}$ be a regionalization of $\mathcal{X}$ and w_1,...,w_A be weight functions satisfying **(W1)**, **(W2)** and **(W3)**. Let $L \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a probability measure on $\mathcal{X} \times \mathcal{Y}$. For $a = 1, \dots, A$, let k_a be a kernel on $\mathcal{X}_a$ with RKHS H_a and let $\lambda_a > 0$. Denote $\boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_A)$ and $\boldsymbol{k} := (k_1, \dots, k_A)$. Then,

$$f_{L,\mathrm{P},\boldsymbol{\lambda},\boldsymbol{k},\boldsymbol{\mathcal{X}}} \colon \mathcal{X} \to \mathbb{R}\,,\; x \mapsto \sum_{a=1}^{A} w_a(x) \cdot \hat{f}_{L,\mathrm{P}_{\mathcal{X}_a},\lambda_a,k_a}(x)$$

is called localized support vector machine (localized SVM). The vector $\boldsymbol{k}$ is also called a vector of kernels on $\boldsymbol{\mathcal{X}}$.

This definition explicitly allows that the regularization parameters and kernels in the different regions differ from each other. This is integral for the increased capability of localized SVMs (compared to non-localized ones) to accurately learn a function whose complexity and variability differ between different areas of the input space, because the choice of regularization parameter and kernel (respectively of the hyperparameter(s) of the kernel) constitutes a principal mechanism for controlling the complexity of an SVM. As explained in Section 2.2.1, this increased capability is one of the main motivations behind the localized approach.

Remark 2.2.8. Note that **(W2)** and **(W3)** imply that $w_a \equiv \mathbb{1}_{\mathcal{X}_a}$ if $\boldsymbol{\mathcal{X}}$ is a partitioning regionalization. This leads to the following simplified definition of a localized SVM that is based on a partitioning regionalization $\boldsymbol{\mathcal{X}}$:

$$f_{L,\mathrm{P},\boldsymbol{\lambda},\boldsymbol{k},\boldsymbol{\mathcal{X}}} \colon \mathcal{X} \to \mathbb{R}\,,\; x \mapsto \sum_{a=1}^{A} \hat{f}_{L,\mathrm{P}_{\mathcal{X}_a},\lambda_a,k_a}(x)\,.$$

As it was the case for global SVMs, the definitions regarding localized SVMs can of course also be transferred to shifted loss functions completely analogously.

Chapter 3

Consistency

One of the most fundamental properties that learning methods should have is consistency: As the size of the underlying data set tends to infinity, the function resulting from the learning method should converge to the "true" function which one wishes to estimate, that is, the Bayes function from Definition 2.1.4. There exist different types of consistency that can be of interest and that are described in Section 3.1. Afterwards, Section 3.2 contains results on the relationship between these different types of consistency, that is, under what circumstances one type implies the other. Notably, these results are not only valid for (sequences of) SVMs respectively localized SVMs but instead for arbitrary sequences of functions, for which reason they can also be applied to other learning methods. In Sections 3.3 and 3.4, results on the different types of consistency are explicitly derived for SVMs and localized SVMs. The main focus of this chapter are distance-based loss functions (cf. Definition 2.1.15), but some results are also applicable to other types of loss functions. Also note: Whereas distance-based loss losses are typically used in regression tasks, some (like the least squares loss) are also applied in classification tasks, cf. Györfi et al. (2002, Section 1.4), which makes the results from this section applicable to an even wider array of learning tasks.

Throughout this chapter, the following standard assumptions are assumed to hold true:

Assumption 3.0.1. Let $\mathcal{X}$ be a complete separable metric space and let $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathcal{X}$ and $\mathcal{Y}$ be equipped with their respective Borel σ-algebras $\mathcal{B}_\mathcal{X}$ and $\mathcal{B}_\mathcal{Y}$. Let P be a probability measure on $\mathcal{X} \times \mathcal{Y}$. For all $n \in \mathbb{N}$, let the data set $D_n := ((x_1, y_1), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ consist of i.i.d. observations sampled from P, and let $\mathrm{D}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$ be the associated empirical distribution.

3.1 Types of Consistency

So far, the expression "learning method" has been used in a rather informal way. In order to describe different types of consistency, *measurable learning methods* have to be defined more formally.

Definition 3.1.1 (Measurable Learning Method)**.** A measurable learning method $\mathcal{L}$ on $\mathcal{X} \times \mathcal{Y}$ maps every data set $D_n \in (\mathcal{X} \times \mathcal{Y})^n$ to a function $f_{D_n} \colon \mathcal{X} \to \mathbb{R}$ and additionally satisfies that the map

$$(\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \to \mathbb{R}\,, \qquad (D_n, x) \mapsto f_{D_n}(x)$$

is measurable, for all $n \in \mathbb{N}$.

Measurable learning methods guarantee that the maps f_{D_n} are measurable for all fixed $D_n \in (\mathcal{X} \times \mathcal{Y})^n$. They furthermore guarantee that the convergence in probability in the subsequent three definitions is well-defined.[9] This follows directly from Steinwart and Christmann (2008, p. 205) for Definition 3.1.2 and in a similar way for Definitions 3.1.3 and 3.1.4. Note that SVMs constitute a measurable learning method under mild assumptions, cf. Steinwart and Christmann (2008, Lemma 6.23).

The most widely-used type of consistency in machine learning theory in general and for SVMs in particular—because this is what their definition aims at, cf. Definition 2.1.5—certainly is *risk consistency.*

Definition 3.1.2 (Risk Consistency)**.** Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function. Then, a measurable learning method $\mathcal{L}$ on $\mathcal{X} \times \mathcal{Y}$ is called $\underline{L\text{-risk consistent}}$ (or just $\underline{\text{risk consistent}}$) if

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_{D_n}) = \mathcal{R}^*_{L,\mathrm{P}} \qquad \text{in probability } \mathrm{P}^\infty.$$

As this is a very natural type of consistency to consider and the classic one aimed at in statistical learning theory (cf. Vapnik, 1995), results on risk consistency exist for many learning methods, see for example Steinwart (2005) (SVMs for classification), Zhang and Yu (2005) (boosting), Christmann and Steinwart (2007) (SVMs for regression; see also Section 3.3), Biau et al. (2008) (averaging classifiers such as random forests), Lin et al. (2022) (deep convolutional neural networks).

In addition to risk consistency, *L_p-consistency* can also be of interest as it compares the functions (the estimator f_{D_n} and the Bayes function $f^*_{L,\mathrm{P}}$) directly, weighted only based on the marginal distribution P^X, instead of additionally depending on the loss function and the conditional distribution of Y.

Definition 3.1.3 (L_p-Consistency)**.** Let $p \in [1, \infty]$ and $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function. Assume that $f^*_{L,\mathrm{P}}$ exists and is P^X-a.s. unique. Then, a measurable learning method $\mathcal{L}$ on $\mathcal{X} \times \mathcal{Y}$ is called $\underline{L_p(\mathrm{P}^X)\text{-consistent}}$ (or just $\underline{L_p\text{-consistent}}$) if

$$\lim_{n\to\infty} \left|\left| f_{D_n} - f^*_{L,\mathrm{P}} \right|\right|_{L_p(\mathrm{P}^X)} = 0 \qquad \text{in probability } \mathrm{P}^\infty.$$

At first glance, L_p-consistency looks like the slightly stronger one among these two types of consistency: Because one almost always uses continuous loss functions in practice (see also Section 2.1.3), $|f_{D_n}(x) - f^*_{L,\mathrm{P}}(x)|$ being small for $x \in \mathcal{X}$ immediately implies $|L(x, y, f_{D_n}(x)) - L(x, y, f^*_{L,\mathrm{P}}(x))|$ being small as well for all $y \in \mathcal{Y}$; the other direction does

[9] Similarly as it was the case for empirical SVMs (see considerations subsequent to Remark 2.1.6), one formally needs to denote f_{D_n} as a random function depending on the random variables underlying D_n (instead of on D_n itself) for these convergences. We refrain from doing so in order to simplify the notation.

not seem as straightforward. However, Section 3.2 shows that L_p- and risk consistency are actually equivalent under mild conditions.

Whereas risk consistency and L_p-consistency are the two main types of consistency investigated in the subsequent sections, a few results also consider *H-consistency* for an RKHS H.

Definition 3.1.4 (H-Consistency)**.** Let H be an RKHS and $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function. Assume that $f^*_{L,\mathrm{P}} \in H$ and that $f^*_{L,\mathrm{P}}$ is P^X-a.s. unique. Then, a measurable learning method $\mathcal{L}$ on $\mathcal{X} \times \mathcal{Y}$ is called $\underline{H\text{-consistent}}$ if

$$\lim_{n\to\infty} \left|\left| f_{D_n} - f^*_{L,\mathrm{P}} \right|\right|_H = 0 \qquad \text{in probability } \mathrm{P}^\infty.$$

Note that L_p- and H-consistency are defined in exactly the same way and the two definitions could hence be merged into a single one considering Banach spaces or even more general normed spaces. These two special cases are however important and yield different results in later sections, which is why they are defined separately.

Remark 3.1.5*.* For shifted loss functions, these three types of consistency can be defined analogously.

Remark 3.1.6*.* In general, $f^*_{L,\mathrm{P}}$ is neither guaranteed to exist nor to be unique, cf. Definition 2.1.4. As the predictors resulting from the learning method are compared directly to $f^*_{L,\mathrm{P}}$ in the definitions of L_p- and H-consistency, $f^*_{L,\mathrm{P}}$ is however assumed to exist and be P^X-a.s. unique in these definitions and all results dealing with these two types of consistency.

3.2 Connection between Different Types of Consistency

In this section, the connection between the three types of consistency described in Section 3.1 is examined. This is done in two parts, first for regular loss functions in Section 3.2.1 and then for shifted loss functions in Section 3.2.2. Whereas close connections are derived in both of these sections, the latter one also yields the interesting negative result that one of the relationships derived for regular losses—namely, L_p-consistency following from risk consistency—can surprisingly not be transferred to shifted loss functions in the generality one might expect based on the results from Section 3.2.1.

For the most part, this section already appeared in the peer-reviewed paper Köhler (2024b, Section 3) that was published in *Journal of Machine Learning Research.* It however also contains previously unpublished results, namely those connecting H-consistency to the other two types of consistency.

3.2.1 Connection for Regular Loss Functions

So far, there are no general results on L_p-consistency following from risk consistency, but only results regarding special loss functions: For the least squares loss, it has been known for many years that the excess risk of a function, i.e. the difference between its risk and the Bayes risk, corresponds to the squared $L_2(\mathrm{P}^X)$-norm of its deviation from the Bayes function, and risk consistency therefore implies L_2-consistency, cf. Cucker and Smale (2001, Proposition 1) or Cherkassky and Mulier (2007, pp. 26–28). Recently, this L_2-difference between a function and the Bayes function has also been bounded by the excess risk—by means of so-called comparison or self-calibration inequalities—in case of the asymmetric least squares loss by Farooq and Steinwart (2019) and in case of more general strongly convex loss functions under additional assumptions by Sheng et al. (2020). Additionally, Hable and Christmann (2014) showed that L_1-consistency follows from risk consistency in case of the pinball loss, and Steinwart and Christmann (2011), Xiang et al. (2012) derived self-calibration inequalities for this loss under additional assumptions. Tong and Ng (2019) did so for the ε-insensitive loss.

The subsequent theorem generalizes the aforementioned special cases to general convex, distance-based loss functions, with parts of the proof being closely inspired by the proof of Hable and Christmann (2014, Lemma A.1).

Theorem 3.2.1. *Let $L\colon \mathcal{Y}\times\mathbb{R}\to[0,\infty)$ be a convex, distance-based loss function of lower growth type $p\in[1,\infty)$. Assume that $f^*_{L,\mathrm{P}}$ exists and is P^X-a.s. unique, $f^*_{L,\mathrm{P}}\in L_p(\mathrm{P}^X)$ and $\mathcal{R}^*_{L,\mathrm{P}}<\infty$. Then, for every sequence $(f_n)_{n\in\mathbb{N}}\subseteq L_p(\mathrm{P}^X)$, we have*

$$\lim_{n\to\infty}\mathcal{R}_{L,\mathrm{P}}(f_n)=\mathcal{R}^*_{L,\mathrm{P}}\qquad\Rightarrow\qquad\lim_{n\to\infty}||f_n-f^*_{L,\mathrm{P}}||_{L_p(\mathrm{P}^X)}=0\,.$$

Proof. Let $g_n\colon\mathcal{X}\times\mathcal{Y}\to[0,\infty),(x,y)\mapsto L(y,f_n(x))$ for $n\in\mathbb{N}$, and $g^*\colon\mathcal{X}\times\mathcal{Y}\to[0,\infty),(x,y)\mapsto L(y,f^*_{L,\mathrm{P}}(x))$. Because of the convexity of L, we can apply Steinwart and Christmann (2008, Corollary 3.62)—where it is easy to see that we do not need the assumption of the sets $\mathcal{M}_{L,\mathrm{P}(\cdot\,|\,x),x}$ being singletons since we already know that $f^*_{L,\mathrm{P}}$ P^X-a.s. uniquely exists—, which yields that $f_n\xrightarrow{\mathrm{P}^X}f^*_{L,\mathrm{P}}$. Thus, because of the continuous mapping theorem and the continuity of L, we also have $g_n\xrightarrow{\mathrm{P}}g^*$. Since

$$\begin{aligned}\lim_{n\to\infty}\int|g_n|\,\mathrm{dP}=\lim_{n\to\infty}\int g_n\,\mathrm{dP}&=\lim_{n\to\infty}\mathcal{R}_{L,\mathrm{P}}(f_n)\\&=\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}})=\int g^*\,\mathrm{dP}=\int|g^*|\,\mathrm{dP}\,,\end{aligned}\tag{3.1}$$

the sequence $(|g_n|)_{n\in\mathbb{N}}$ is thus equi-integrable according to Bauer (2001, Theorem 21.7). That theorem can be applied because $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}})<\infty$, and hence $\mathcal{R}_{L,\mathrm{P}}(f_n)<\infty$ for n sufficiently large because of (3.1), and therefore $g^*\in L_1(\mathrm{P}^X)$ and $g_n\in L_1(\mathrm{P}^X)$ for n sufficiently large.

Because of L being of lower growth type p, there now exists a constant $c>0$ such that

$$\begin{aligned}|f_n(x)-f^*_{L,\mathrm{P}}(x)|^p&\leq\max\Big\{(2|y-f_n(x)|)^p\,,\,(2|y-f^*_{L,\mathrm{P}}(x)|)^p\Big\}\\&\leq 2^p\cdot\max\Big\{c^{-1}\Big(L(y,f_n(x))+1\Big)\,,\,c^{-1}\Big(L(y,f^*_{L,\mathrm{P}}(x))+1\Big)\Big\}\\&=\frac{2^p}{c}\cdot\Big(\max\left\{g_n(x,y),g^*(x,y)\right\}+1\Big)\\&\leq\frac{2^p}{c}\cdot\Big(g_n(x,y)+g^*(x,y)+1\Big)\qquad\forall\,(x,y,n)\in\mathcal{X}\times\mathcal{Y}\times\mathbb{N}\,,\end{aligned}\tag{3.2}$$

since g_n, $n \in \mathbb{N}$, and g^* are non-negative.

As $(|g_n|)_{n\in\mathbb{N}}$ is equi-integrable, and $g^* \in L_1(\mathrm{P}^X)$ and hence also equi-integrable (cf. Bauer, 2001, part 2 of the example on p. 122), every summand occurring on the right hand side of (3.2) is equi-integrable (as a sequence in n). By employing the example on p. 121 of Bauer (2001) as well as Corollary 21.3 from the same book, we hence obtain equi-integrability of the whole right hand side (as a sequence in n).

Thus, the sequence $(|f_n - f^*_{L,\mathrm{P}}|^p)_{n\in\mathbb{N}}$ is equi-integrable as well and L_p-convergence of f_n to $f^*_{L,\mathrm{P}}$, follows from Bauer (2001, Theorem 21.7). □

Remark 3.2.2. If L is of growth type p instead of only being of *lower* growth type p, the conditions $f^*_{L,\mathrm{P}} \in L_p(\mathrm{P}^X)$ and $\mathcal{R}^*_{L,\mathrm{P}} < \infty$ in Theorem 3.2.1 can also be replaced by the perhaps more intuitive and in this case equivalent moment condition $|\mathrm{P}|_p < \infty$. This equivalence can be obtained by combining Remark 2.1.22(i) with the observation that $\mathcal{R}^*_{L,\mathrm{P}} \leq \mathcal{R}_{L,\mathrm{P}}(0)$ and with Lemma 2.1.21(iv).

Notably, Theorem 3.2.1 strengthens Steinwart and Christmann (2008, Corollary 3.62), which stated that risk consistency implies weak consistency.

As suspected in Section 3.1, the opposite direction—risk consistency following from L_p-consistency—is generally the easier one. We formally state this implication in the subsequent Theorem 3.2.3. Hence, this theorem can be seen as the counterpart of Theorem 3.2.1, even though the conditions of the two theorems differ in some details. Notably, the function f^*, which the sequence is converging to, does not even necessarily need to be the Bayes function $f^*_{L,\mathrm{P}}$ here:

Theorem 3.2.3. *Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a continuous, distance-based loss function of upper growth type $p \in [1,\infty)$. Assume that $|\mathrm{P}|_p < \infty$. Then, for every sequence $(f_n)_{n\in\mathbb{N}} \subseteq L_p(\mathrm{P}^X)$ and every function $f^* \in L_p(\mathrm{P}^X)$, we have*

$$\lim_{n\to\infty} ||f_n - f^*||_{L_p(\mathrm{P}^X)} = 0 \qquad \Rightarrow \qquad \lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) = \mathcal{R}_{L,\mathrm{P}}(f^*)\,.$$

Proof. Since $||f_n - f^*||_{L_p(\mathrm{P}^X)} \to 0$, we also have $f_n \xrightarrow{\mathrm{P}^X} f^*$, and Bauer (2001, Theorem 21.7) yields equi-integrability of the sequence $(|f_n|^p)_{n\in\mathbb{N}}$. Let $g_n\colon \mathcal{X}\times\mathcal{Y} \to [0,\infty), (x,y) \mapsto L(y, f_n(x))$ for $n \in \mathbb{N}$, and $g^*\colon \mathcal{X}\times\mathcal{Y} \to [0,\infty), (x,y) \mapsto L(y, f^*(x))$. Because of L being of upper growth type p, there then exists a $c > 0$ such that

$$\begin{aligned} |g_n(x,y)| = g_n(x,y) = L(y, f_n(x)) &\leq c \cdot (|y - f_n(x)|^p + 1) \\ &\leq c \cdot (2^p \cdot (|y|^p + |f_n(x)|^p) + 1) \end{aligned} \tag{3.3}$$

for all $(x,y,n) \in \mathcal{X}\times\mathcal{Y}\times\mathbb{N}$.

Since every summand on the right hand side of (3.3) is equi-integrable (because $|\mathrm{P}|_p < \infty$), the whole right hand side is equi-integrable as well (as a sequence in n) by the example on p. 121 of Bauer (2001) and Corollary 21.3 from the same book. Hence, the sequence $(|g_n|)_{n\in\mathbb{N}}$ is equi-integrable as well.

Additionally, $g_n \xrightarrow{\mathrm{P}} g^*$ because of $f_n \xrightarrow{\mathrm{P}^X} f^*$ and the continuous mapping theorem in combination with the continuity of L, and thus, Bauer (2001, Theorem 21.7) yields

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) = \lim_{n\to\infty} \int g_n \,\mathrm{dP} = \lim_{n\to\infty} \int |g_n| \,\mathrm{dP} = \int |g^*| \,\mathrm{dP} = \int g^* \,\mathrm{dP} = \mathcal{R}_{L,\mathrm{P}}(f^*)\,. \quad \square$$

Together, Theorem 3.2.1 and Theorem 3.2.3 prove that L_p- and risk consistency are indeed equivalent under mild assumptions. Lastly, the subsequent corollary of Theorem 3.2.3 and Lemma 2.1.10 shows that H-consistency implies both of them.

Corollary 3.2.4. *Let H be the RKHS of a bounded and measurable kernel k on $\mathcal{X}$. Then, for every sequence $(f_n)_{n\in\mathbb{N}} \subseteq H$ and every function $f^* \in H$, the following hold true:*

(i) For all $p \in [1,\infty]$, we have

$$\lim_{n\to\infty} ||f_n - f^*||_H = 0 \qquad \Rightarrow \qquad \lim_{n\to\infty} ||f_n - f^*||_{L_p(\mathrm{P}^X)} = 0\,.$$

(ii) Let $L\colon \mathcal{Y}\times\mathbb{R} \to [0,\infty)$ be a continuous, distance-based loss function of upper growth type $p \in [1,\infty)$. Assume that $|\mathrm{P}|_p < \infty$. Then, we have

$$\lim_{n\to\infty} ||f_n - f^*||_H = 0 \qquad \Rightarrow \qquad \lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_n) = \mathcal{R}_{L,\mathrm{P}}(f^*)\,.$$

Proof.

(i) $H \subseteq L_p(\mathrm{P}^X)$ by Lemma 2.1.10(ii). Hence,

$$||f_n - f^*||_{L_p(\mathrm{P}^X)} \leq ||f_n - f^*||_\infty \leq ||f_n - f^*||_H \, ||k||_\infty$$

for all $n \in \mathbb{N}$ and $p \in [1,\infty]$ by Lemma 2.1.10(i), which yields the assertion because $||k||_\infty < \infty$ by assumption.

(ii) Follows directly from part (i), $H \subseteq L_p(\mathrm{P}^X)$, and Theorem 3.2.3. □

Remark 3.2.5. As can be seen from the preceding proof, H-consistency does actually not only imply L_p-consistency but even consistency with respect to the strong and nicely interpretable supremum norm $||\cdot||_\infty$. As the subsequent sections are however mainly concerned with L_p- and risk consistency, these are the types of consistency considered in the statement of Corollary 3.2.4.

3.2.2 Connection for Shifted Loss Functions

When looking at Theorem 3.2.1, it is obvious that the assumptions $f^*_{L,\mathrm{P}} \in L_p(\mathrm{P}^X)$ and $\mathcal{R}^*_{L,\mathrm{P}} < \infty$ are indeed necessary for the conclusion of the theorem and that one cannot hope to derive L_p- from risk consistency without them. Because these assumptions are equivalent to the moment condition $|\mathrm{P}|_p < \infty$ if L is of growth type p (cf. Remark 3.2.2), this however excludes heavy-tailed distributions such as the Cauchy distribution—even for $p = 1$. Analogously, Theorem 3.2.3 and part (ii) of Corollary 3.2.4 also require $|\mathrm{P}|_p < \infty$ and can therefore not be applied to such heavy-tailed distributions.

Based on the considerations from Section 2.1.4, it suggests itself to try to transfer the results from Section 3.2.1 to shifted loss functions in order to circumvent this problem and eliminate the moment condition and thus extend the applicability of these results in the case of using a Lipschitz continuous loss. As this Lipschitz continuity for distance-based losses all but coincides with the loss being of upper growth type 1 (cf. Remark 2.1.18), one might hope that the elimination of the moment condition is possible in the case $p = 1$.

When looking at the proof of Theorem 3.2.1, it is however easy to see that (3.1) does not hold true for shifted loss functions and the proof can thus not be transferred to the situation of this section. The following negative result shows that this is indeed not a failing of the specific proof we used, but that L_1-consistency does, somewhat surprisingly, actually not follow from $L^\star$-risk consistency in the generality one would have hoped for:

Proposition 3.2.6. *Let $\mathcal{Y} = \mathbb{R}$. Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex, distance-based and symmetric loss function of growth type 1, and let $L^\star$ be its shifted version. Then, even if $f^*_{L^\star,\mathrm{P}}$ is P^X-a.s. unique with $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}^X)$, a sequence $(f_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}^X)$ of functions satisfying*

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

*does in general **not** imply*

$$\lim_{n\to\infty} \left|\left| f_n - f^*_{L^\star,\mathrm{P}} \right|\right|_{L_1(\mathrm{P}^X)} = 0$$

without any additional assumptions besides Assumption 3.0.1 being imposed.

For proving Proposition 3.2.6, we need the following auxiliary lemma:

Lemma 3.2.7. *Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex and Lipschitz continuous loss function, and let $L^\star$ be its shifted version. If there exists a measurable function $f\colon \mathcal{X} \to \mathbb{R}$ satisfying $\mathcal{R}_{L^\star,\mathrm{P}}(f) = -\infty$, there also exists a measurable function $g\colon \mathcal{X} \to \mathbb{R}$ satisfying $\mathrm{P}^X(g \neq 0) > 0$ and $\mathcal{R}_{L^\star,\mathrm{P}}(g) \in (-\infty, 0]$.*

Proof. With the inner risk $\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}$, we have

$$\begin{aligned}\mathcal{R}_{L^\star,\mathrm{P}}(f) &= \int L^\star(x, y, f(x))\,\mathrm{dP}(x, y) = \int \mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(f(x))\,\mathrm{dP}^X(x) \\ &= \int \mathcal{C}^+_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(f(x))\,\mathrm{dP}^X(x) - \int \mathcal{C}^-_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(f(x))\,\mathrm{dP}^X(x) = -\infty\,,\end{aligned}$$

with $\mathcal{C}^+_{L^\star,\mathrm{P}(\cdot\,|\,x),x} := \max\{\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}\,,\,0\}$ and $\mathcal{C}^-_{L^\star,\mathrm{P}(\cdot\,|\,x),x} := \max\{-\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}\,,\,0\}$ denoting the positive and the negative part of $\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}$ respectively. From the definition of the integral, we hence obtain

$$\int \mathcal{C}^-_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(f(x))\,\mathrm{dP}^X(x) = \infty \tag{3.4}$$

and therefore the existence of $c \in (0, \infty)$ and $S \subseteq \mathcal{X}$ measurable such that $\mathrm{P}^X(S) > 0$ and $\mathcal{C}^-_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(f(x)) \geq c$ for all $x \in S$.

We further know that $|L|_1 > 0$ because it is clear from the definition of Lipschitz continuous loss functions that $|L|_1 = 0$ would imply $L(x, y, f(x)) = L(x, y, 0)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$

and hence $\mathcal{R}_{L^\star,\mathrm{P}}(f) = 0$, which contradicts our assumptions. Therefore, (3.4) directly implies that $|f(x)| \geq \frac{c}{|L|_1} > 0$ for all $x \in S$ because otherwise

$$\begin{aligned}\mathcal{C}^-_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(f(x)) &= \Big(\int L^\star(x,y,f(x))\,\mathrm{dP}^X(x)\Big)^- \leq \int \Big|L^\star(x,y,f(x))\Big|\,\mathrm{dP}^X(x)\\ &= \int \Big|L(x,y,f(x)) - L(x,y,0)\Big|\,\mathrm{dP}^X(x) \leq |L|_1 \cdot |f(x)| < c\,,\end{aligned}$$

which would form a contradiction to x coming from S.

Define

$$g(x) := \begin{cases} 0 & \text{, if } x \notin S\,,\\ \frac{c}{|L|_1}\cdot \operatorname{sign}(f(x)) & \text{, if } x \in S\,.\end{cases}$$

Then, $\mathrm{P}^X(g \neq 0) > 0$ and

$$\mathcal{R}_{L^\star,\mathrm{P}}(g) = \int_S \mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(g(x))\,\mathrm{dP}^X(x) + \underbrace{\int_{\mathcal{X}\setminus S} \mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(g(x))\,\mathrm{dP}^X(x)}_{=0}\,. \tag{3.5}$$

All that remains to investigate is the first integral on the right hand side. For all $x \in S$, we know that

$$\Big|\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(g(x))\Big| \leq \int |L(x,y,g(x)) - L(x,y,0)|\,\mathrm{dP}(y\,|\,x) \leq |L|_1 \cdot |g(x)| = c$$

and

$$\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(g(x)) \leq \max\Big\{\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(0), \mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(f(x))\Big\} = \mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(0) = 0$$

because $g(x)$ lies between 0 and $f(x)$, $\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}$ is convex (cf. Lemma 2.1.27, which is applicable because of $L^\star$ being convex), and additionally $\mathcal{C}_{L^\star,\mathrm{P}(\cdot\,|\,x),x}(f(x)) < 0$ by definition of S.

Plugging this into the right hand side of (3.5) yields $\mathcal{R}_{L^\star,\mathrm{P}}(g) \in [-c, 0]$ and hence the assertion. □

Proof of Proposition 3.2.6. We prove the statement by providing a counterexample.

Because of L being of lower growth type 1,

$$c_0 := \sup\{r \in [0,\infty)\,|\,\psi(r) = 0\}$$

is finite, where ψ denotes the representing function belonging to L, as introduced in Definition 2.1.15. Because of L being convex, distance-based, and symmetric, we have

$$L(y,t) = \psi(y-t) = 0 \qquad \Leftrightarrow \qquad y - t \in [-c_0, c_0]\,. \tag{3.6}$$

Assume without loss of generality that $c_0 \leq \frac{1}{2}$ (else just scale the subsequent example accordingly).

Choose $\mathcal{X} := (0,1)$, $\mathrm{P}^X := \mathcal{U}(0,1)$ and

$$\mathrm{P}(\cdot \mid X = x) := x \cdot \mathcal{U}(-1,1) + \frac{1-x}{2} \cdot \Big(\delta_{-a_x} + \delta_{a_x}\Big) \qquad \forall\, x \in \mathcal{X}\,, \tag{3.7}$$

where $\mathcal{U}(a,b)$ denotes the uniform distribution on (a,b), δ_z denotes the Dirac distribution in $z \in \mathbb{R}$ and $a_x > 1$ is a constant depending on x (and on L) that we will specify right after (3.15).[10] Further define

$$f_n \colon \mathcal{X} \to \mathbb{R}\,, \qquad x \mapsto \begin{cases} n & \text{, if } x \in \Big(0, \frac{1}{n}\Big)\,, \\ 0 & \text{, else}\,, \end{cases} \tag{3.8}$$

for $n \in \mathbb{N}$. As f_n is bounded and measurable for all $n \in \mathbb{N}$, we have $(f_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}^X)$. We now show that this example also possesses the remaining properties mentioned in the proposition, which consists of three main steps:

First, we show that $f^*_{L^\star,\mathrm{P}}$ is P^X-a.s. unique, more specifically $f^*_{L^\star,\mathrm{P}} \equiv 0$ P^X-a.s., and $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}^X)$:
Choose $f^* \equiv 0$. We show that $\mathcal{R}_{L^\star,\mathrm{P}}(f^*) < \mathcal{R}_{L^\star,\mathrm{P}}(f)$ for all measurable $f\colon \mathcal{X} \to \mathbb{R}$ satisfying $\mathrm{P}^X(f \neq 0) > 0$. As $\mathcal{R}_{L^\star,\mathrm{P}}(f^*) = 0$, the case $\mathcal{R}_{L^\star,\mathrm{P}}(f) = \infty$ is trivial. Furthermore, if there was an f satisfying $\mathcal{R}_{L^\star,\mathrm{P}}(f) = -\infty$ and thus contradicting our claim, there would by Lemma 3.2.7 (which is applicable by Lemma 2.1.17) also exist a measurable g with $\mathrm{P}^X(g \neq 0) > 0$ and $-\infty < \mathcal{R}_{L^\star,\mathrm{P}}(g) \leq 0 = \mathcal{R}_{L^\star,\mathrm{P}}(f^*)$, which would also contradict our claim. Hence, we can without loss of generality assume that $\mathcal{R}_{L^\star,\mathrm{P}}(f) \in \mathbb{R}$.
Since $f^* \equiv 0$, we have, for each $x \in \mathcal{X}$ and $y \geq 0$,

$$\begin{aligned} L^\star\left(-y, f^*(x)\right) + L^\star\left(y, f^*(x)\right) &= 2 \cdot L^\star(y, 0) \\ &= 2 \cdot L^\star\left(y,\, \frac{1}{2} \cdot (-f(x)) + \frac{1}{2} \cdot f(x)\right) \\ &\leq L^\star\left(y, -f(x)\right) + L^\star\left(y, f(x)\right) \\ &= L^\star\left(-y, f(x)\right) + L^\star\left(y, f(x)\right) \end{aligned} \tag{3.9}$$

because of L being distance-based, symmetric and convex.
Furthermore, by the definition of f, there exists $\varepsilon := (\varepsilon_1, \varepsilon_2)$ with $\varepsilon_1, \varepsilon_2 > 0$ such that $\mathrm{P}^X(\mathcal{X}_\varepsilon) > 0$, where $\mathcal{X}_\varepsilon := \{x \in \mathcal{X} : |f(x)| \geq \varepsilon_1 \text{ and } x \geq \varepsilon_2\}$. Now, specifically look at $x \in \mathcal{X}_\varepsilon$ and $y \in [c_0, c_0 + \min\{\frac{1}{2}, \frac{|f(x)|}{4}\}] \subseteq [0,1]$. First, only consider such x that satisfy $f(x) > 0$. We then obtain that

$$|-y - f(x)| = y + f(x) \geq c_0 + f(x) \quad \text{ and } \quad |\pm y - f^*(x)| = y \leq c_0 + \frac{f(x)}{4}\,, \tag{3.10}$$

and hence

$$L(-y, f(x)) \geq 4 \cdot L(-y, f^*(x)) = 2 \cdot \Big(L\left(y, f^*(x)\right) + L\left(-y, f^*(x)\right)\Big)$$

[10] For the sake of strictly adhering to the completeness assumption from Assumption 3.0.1, we can also choose $\mathcal{X}$ as $[0,1]$ or $\mathbb{R}$, and $\mathrm{P}(\cdot \mid X = x)$ as an arbitrary probability measure for $x \notin (0,1)$ without changing anything else.

because of (3.6) and the convexity, symmetry and distance-basedness of L. Thus,

$$\begin{aligned}
&\Big(L^{\star}\left(-y, f(x)\right) + L^{\star}\left(y, f(x)\right)\Big) - \Big(L^{\star}\left(-y, f^{*}(x)\right) + L^{\star}\left(y, f^{*}(x)\right)\Big) \\
&= \Big(L\left(-y, f(x)\right) + L\left(y, f(x)\right)\Big) - \Big(L\left(-y, f^{*}(x)\right) + L\left(y, f^{*}(x)\right)\Big) \\
&\geq \frac{1}{2} \cdot L\left(-y, f(x)\right) = \frac{1}{2} \cdot \psi(|-y - f(x)|) \geq \frac{1}{2} \cdot \psi(c_0 + f(x))\,,
\end{aligned}$$

where, in the last step, we again applied the convexity and symmetry of L, as well as (3.10).

By interchanging the roles of y and $-y$ in the preceding paragraph, we obtain an analogous inequality for the case that $f(x) < 0$. Combining these two cases yields that

$$\begin{aligned}
&\Big(L^{\star}\left(-y, f(x)\right) + L^{\star}\left(y, f(x)\right)\Big) - \Big(L^{\star}\left(-y, f^{*}(x)\right) + L^{\star}\left(y, f^{*}(x)\right)\Big) \\
&\geq \frac{1}{2} \cdot \psi(c_0 + |f(x)|) \qquad (3.11)
\end{aligned}$$

for all $x \in \mathcal{X}_{\varepsilon}$ and $y \in [c_0, c_0 + \min\{\frac{1}{2}, \frac{|f(x)|}{4}\}] \subseteq [0,1]$.

Because $\mathcal{R}_{L^{\star},\mathrm{P}}(f^{*}) = 0 \in \mathbb{R}$ by the definition of f^{*} and $\mathcal{R}_{L^{\star},\mathrm{P}}(f) \in \mathbb{R}$ by assumption, our considerations yield

$$\begin{aligned}
&\mathcal{R}_{L^{\star},\mathrm{P}}(f) - \mathcal{R}_{L^{\star},\mathrm{P}}(f^{*}) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} L^{\star}\left(y, f(x)\right) - L^{\star}\left(y, f^{*}(x)\right) \,\mathrm{dP}(y\,|\,x)\,\mathrm{dP}^{X}(x) \\
&= \int_{\mathcal{X}} \int_{[0,\infty)} \Big(L^{\star}\left(-y, f(x)\right) + L^{\star}\left(y, f(x)\right)\Big) \\
&\qquad\qquad - \Big(L^{\star}\left(-y, f^{*}(x)\right) + L^{\star}\left(y, f^{*}(x)\right)\Big) \,\mathrm{dP}(y\,|\,x)\,\mathrm{dP}^{X}(x) \\
&\overset{(3.9),(3.11)}{\geq} \int_{\mathcal{X}_{\varepsilon}} \int_{[c_0, c_0 + \min\{\frac{1}{2}, \frac{|f(x)|}{4}\}]} \frac{1}{2} \cdot \psi(c_0 + |f(x)|)\,\mathrm{dP}(y\,|\,x)\,\mathrm{dP}^{X}(x) \\
&= \int_{\mathcal{X}_{\varepsilon}} \frac{x}{2} \cdot \min\left\{\frac{1}{2}, \frac{|f(x)|}{4}\right\} \cdot \frac{1}{2} \cdot \psi(c_0 + |f(x)|)\,\mathrm{dP}^{X}(x) \\
&\geq \mathrm{P}^{X}(\mathcal{X}_{\varepsilon}) \cdot \frac{\varepsilon_2}{2} \cdot \min\left\{\frac{1}{2}, \frac{\varepsilon_1}{4}\right\} \cdot \frac{1}{2} \cdot \psi(c_0 + \varepsilon_1) \\
&\overset{(3.6)}{>} 0\,.
\end{aligned}$$

In the second step, we multiplied the integrand by 2 for $y = 0$, which does not change the value of the integral since $\mathrm{P}(Y = 0\,|\,X = x) = 0$ for all $x \in \mathcal{X}$. In the final steps, we additionally applied that $\mathrm{P}(\cdot\,|\,X = x)$ has Lebesgue density $\frac{x}{2}$ on $[c_0, c_0 + \min\{\frac{1}{2}, \frac{|f(x)|}{4}\}] \subseteq [0,1]$, respectively the definition of $\mathcal{X}_{\varepsilon}$.

Hence, $f^{*}_{L^{\star},\mathrm{P}} \equiv 0$ P^{X}-a.s. and thus also $f^{*}_{L^{\star},\mathrm{P}} \in L_1(\mathrm{P}^{X})$.

Next, we show that $\lim_{n\to\infty} \mathcal{R}_{L^{\star},\mathrm{P}}(f_n) = \mathcal{R}^{*}_{L^{\star},\mathrm{P}}$:

Recall the definition of f_n, $n \in \mathbb{N}$, from (3.8). For all $n \in \mathbb{N}$, we have $f^{*}_{L^{\star},\mathrm{P}}, f_n \in L_1(\mathrm{P}^{X})$

and therefore $\mathcal{R}^*_{L^\star,\mathrm{P}} = \mathcal{R}_{L^\star,\mathrm{P}}(f^*_{L^\star,\mathrm{P}}) \in \mathbb{R}$ and $\mathcal{R}_{L^\star,\mathrm{P}}(f_n) \in \mathbb{R}$ by (2.5). Hence, we can write

$$\begin{aligned}
&\mathcal{R}_{L^\star,\mathrm{P}}(f_n) - \mathcal{R}^*_{L^\star,\mathrm{P}} \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} L^\star\left(y, f_n(x)\right) - L^\star\left(y, f^*_{L^\star,\mathrm{P}}(x)\right) \mathrm{dP}(y\,|\,x)\, \mathrm{dP}^X(x) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} L\left(y, f_n(x)\right) - L\left(y, f^*_{L^\star,\mathrm{P}}(x)\right) \mathrm{dP}(y\,|\,x)\, \mathrm{dP}^X(x) \\
&= \int_0^{1/n} \int_{-1}^{1} \frac{x}{2} \cdot \Big(L\left(y,n\right) - L\left(y,0\right)\Big)\, \mathrm{d}y\, \mathrm{d}x \\
&\quad + \int_0^{1/n} \frac{1-x}{2} \cdot \bigg(\Big(L\left(-a_x,n\right) + L\left(a_x,n\right)\Big) - \Big(L\left(-a_x,0\right) + L\left(a_x,0\right)\Big)\bigg)\, \mathrm{d}x\,, \qquad (3.12)
\end{aligned}$$

where we applied the definition of f_n, $f^*_{L^\star,\mathrm{P}}$, and P in the last step. We will now analyze the two integrals on the right hand side separately and show that they both converge to 0 as $n \to \infty$, starting with the first one:

$$\begin{aligned}
&\left| \int_0^{1/n} \int_{-1}^{1} \frac{x}{2} \cdot \Big(L\left(y,n\right) - L\left(y,0\right)\Big)\, \mathrm{d}y\, \mathrm{d}x \right| \\
&\leq \int_0^{1/n} \int_{-1}^{1} \frac{x}{2} \cdot |L|_1 \cdot |n-0|\, \mathrm{d}y\, \mathrm{d}x = \frac{|L|_1}{2n} \xrightarrow{n\to\infty} 0
\end{aligned}$$

with L being Lipschitz continuous by Lemma 2.1.17.
As for the second integral on the right hand side of (3.12):
We take a look at the subdifferential $\partial\psi$ of the representing function ψ (cf. Definition 2.1.15) of L. Because of the symmetry of L, we will without loss of generality only investigate $\partial\psi(r)$ for $r \in [0,\infty)$. Define

$$z(r) \coloneqq \sup \partial\psi(r) \in [0,\infty) \qquad \forall\, r \in [0,\infty)\,,$$

where $z(r) < \infty$ will follow from (3.13) and $z(r) \geq 0$ follows from ψ being monotonically increasing on $[0,\infty)$ because of L being distance-based and convex. Furthermore, let c_L be the constant from the definition of the upper growth type 1 of L, that is

$$\psi(r) \leq c_L \cdot (|r|+1) \qquad \forall\, r \in \mathbb{R}\,.$$

Assume there was an $r_0 \in [0,\infty)$ such that $z(r_0) > c_L$. Then, by the definition of the subdifferential, we would obtain

$$c_L \cdot (r+1) \geq \psi(r) \geq \psi(r_0) + z(r_0) \cdot (r - r_0) \qquad \forall\, r \in [0,\infty)$$

and hence

$$r \leq \frac{\psi(r_0) - z(r_0)r_0 - c_L}{c_L - z(r_0)} \qquad \forall\, r \in [0,\infty)\,,$$

which is a contradiction because the right hand side is a constant in $\mathbb{R}$ that is independent of r. Hence, z is bounded by c_L. Because of ψ being monotonically increasing on $[0,\infty)$ and convex, we additionally obtain that z is monotonically increasing on $[0,\infty)$ and

$$\tilde{c}_L \coloneqq \lim_{r\to\infty} z(r) = \sup_{r\in[0,\infty)} z(r) \leq c_L \qquad (3.13)$$

exists.
We can therefore, for each $x \in (0,1)$, choose $r_x \in [0, \infty)$ such that

$$0 \leq \tilde{c}_L - z(r_x) \leq x \tag{3.14}$$

and

$$\psi(r_x) + z(r_x) \cdot (r - r_x) \leq \psi(r) \leq \psi(r_x) + \tilde{c}_L \cdot (r - r_x) \qquad \forall\, r \in [r_x, \infty)\,. \tag{3.15}$$

Now choose a_x in the definition of $\mathrm{P}(\cdot \,|\, X = x)$ in (3.7) as $a_x := r_x + \frac{1}{x}$ for all $x \in (0,1)$. Please note that $a_x > 1$ for all $x \in (0,1)$. We obtain

$$\begin{aligned}
&L(-a_x, n) + L(a_x, n) \\
&= \psi\left(|-a_x - n|\right) + \psi\left(|a_x - n|\right) \\
&= \psi\left(r_x + \frac{1}{x} + n\right) + \psi\left(r_x + \frac{1}{x} - n\right) \\
&\in \left[2 \cdot \psi(r_x) + z(r_x) \cdot \left(\frac{1}{x} + n + \frac{1}{x} - n\right)\,,\, 2 \cdot \psi(r_x) + \tilde{c}_L \cdot \left(\frac{1}{x} + n + \frac{1}{x} - n\right)\right] \\
&= \left[2 \cdot \left(\psi(r_x) + \frac{z(r_x)}{x}\right)\,,\, 2 \cdot \left(\psi(r_x) + \frac{\tilde{c}_L}{x}\right)\right] \qquad \forall\, n \in \mathbb{N}\,,\ x \in \left(0, \frac{1}{n}\right),
\end{aligned}$$

where we applied the symmetry of L as well as (3.15) combined with the fact that $\frac{1}{x} + n \geq 0$ and $\frac{1}{x} - n \geq 0$. Analogously, we obtain

$$\begin{aligned}
&L(-a_x, 0) + L(a_x, 0) \\
&= 2 \cdot \psi\left(r_x + \frac{1}{x}\right) \\
&\in \left[2 \cdot \left(\psi(r_x) + \frac{z(r_x)}{x}\right)\,,\, 2 \cdot \left(\psi(r_x) + \frac{\tilde{c}_L}{x}\right)\right] \qquad \forall\, x \in \left(0, \frac{1}{n}\right).
\end{aligned}$$

Plugging these results into the second integral on the right hand side of (3.12) finally yields

$$\begin{aligned}
&\left| \int_0^{1/n} \frac{1-x}{2} \cdot \Big(\big(L(-a_x, n) + L(a_x, n) \big) - \big(L(-a_x, 0) + L(a_x, 0) \big) \Big) \,\mathrm{d}x \right| \\
&\leq \int_0^{1/n} \frac{1-x}{2} \cdot \left(2 \cdot \left(\psi(r_x) + \frac{\tilde{c}_L}{x}\right) - 2 \cdot \left(\psi(r_x) + \frac{z(r_x)}{x}\right) \right) \mathrm{d}x \\
&= \int_0^{1/n} \frac{1-x}{2} \cdot \frac{2}{x} \cdot \left(\tilde{c}_L - z(r_x)\right) \mathrm{d}x \\
&\overset{(3.14)}{\leq} \int_0^{1/n} (1-x) \,\mathrm{d}x = \frac{1}{n} - \frac{1}{2n^2} \xrightarrow{n \to \infty} 0\,,
\end{aligned}$$

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

Finally and as a last step, we have to show that $\lim_{n\to\infty} \left|\left| f_n - f^*_{L^\star, \mathrm{P}} \right|\right|_{L_1(\mathrm{P}^X)} \neq 0$:

$$\lim_{n\to\infty} \left|\left| f_n - f^*_{L^\star, \mathrm{P}} \right|\right|_{L_1(\mathrm{P}^X)} = \lim_{n\to\infty} \int_0^{1/n} |n - 0| \,\mathrm{d}x = \lim_{n\to\infty} 1 \neq 0\,. \qquad \square$$

Note that in the situation of Proposition 3.2.6, risk consistency does also not imply L_p-consistency for any $p > 1$ since L_p-consistency for $p > 1$ would imply L_1-consistency.

We now take a special look at the τ-pinball loss for $\tau \in (0,1)$, cf. (2.1), which is convex and distance-based with growth type 1, but not symmetric for $\tau \neq 0.5$. The pinball loss can be used for quantile regression, i.e. for estimating the conditional quantiles

$$\begin{aligned} F^*_{\tau,\mathrm{P}} \colon\ & \mathcal{X} \to 2^{\mathbb{R}}, \\ & x \mapsto \{t^* \mid \mathrm{P}((-\infty, t^*]|x) \geq \tau \text{ and } \mathrm{P}([t^*, \infty)|x) \geq 1 - \tau\}, \end{aligned}$$

see also Koenker and Bassett (1978), Koenker (2005), Takeuchi et al. (2006), Steinwart and Christmann (2011) for more details on quantile regression.

If one assumes these conditional quantiles $F^*_{\tau,\mathrm{P}}(x)$ to P^X-a.s. be singletons, it is possible to denote them by the P^X-a.s. unique quantile function $f^*_{\tau,\mathrm{P}} \colon \mathcal{X} \to \mathbb{R}$ defined by $\{f^*_{\tau,\mathrm{P}}(x)\} = F^*_{\tau,\mathrm{P}}(x)$ for all $x \in \mathcal{X}$. Recall that this $f^*_{\tau,\mathrm{P}}$ is the up to P^X-zero sets only measurable function satisfying

$$\mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}}) = \mathcal{R}^*_{L_{\tau\text{-pin}},\mathrm{P}} \tag{3.16}$$

if $\mathcal{R}^*_{L_{\tau\text{-pin}},\mathrm{P}}$ is finite (see also Steinwart and Christmann, 2008, Proposition 3.9 and Lemma 3.12), and similarly, that $f^*_{\tau,\mathrm{P}}$ satisfies

$$\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}}) = \mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} \tag{3.17}$$

and is the up to P^X-zero sets only measurable function doing so if $\mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}}$ is finite. This ties our assumption of the conditional quantiles P^X-a.s. being singletons to Remark 3.1.6 about the required P^X-a.s. uniqueness of the Bayes function and yields $f^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} \equiv f^*_{\tau,\mathrm{P}}$ P^X-a.s.

As non-symmetric loss functions are not covered by Proposition 3.2.6 and as the pinball loss is the probably most popular among these, we specifically investigate the behavior of this loss function and obtain the following analogous result to Proposition 3.2.6:

Proposition 3.2.8. *Let $\mathcal{Y} = \mathbb{R}$. Let $\tau \in (0,1)$ and let $L^\star_{\tau\text{-}pin}$ be the shifted version of the τ-pinball loss.*[11] *Then, even if $f^*_{\tau,\mathrm{P}}$ is P^X-a.s. unique with $f^*_{\tau,\mathrm{P}} \in L_1(\mathrm{P}^X)$, a sequence $(f_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}^X)$ of functions satisfying*

$$\lim_{n\to\infty} \mathcal{R}_{L^\star_{\tau\text{-}pin},\mathrm{P}}(f_n) = \mathcal{R}^*_{L^\star_{\tau\text{-}pin},\mathrm{P}}$$

[11] It can easily be seen that this shifted pinball loss function is, for $\tau \in (0,1)$,

$$\begin{aligned} L^\star_{\tau\text{-pin}} \colon\ & \mathcal{Y} \times \mathbb{R} \to \mathbb{R} \\ & (y,t) \mapsto L_{\tau\text{-pin}}(y,t) - L_{\tau\text{-pin}}(y,0) = \begin{cases} (1-\tau)\cdot t & \text{, if } y < \min\{0,t\}, \\ (1-\tau)\cdot t - y & \text{, if } 0 \leq y < t, \\ y - \tau\cdot t & \text{, if } t \leq y < 0, \\ -\tau\cdot t & \text{, if } y \geq \max\{0,t\}. \end{cases} \end{aligned}$$

*does in general **not** imply*

$$\lim_{n\to\infty} \left|\left|f_n - f^*_{\tau,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)} = 0$$

without any additional assumptions besides Assumption 3.0.1 being imposed.

Proof. Similarly to Proposition 3.2.6, we prove the statement by providing a counterexample:

Choose $\mathcal{X} := (0,1)$, $\mathcal{Y} := \mathbb{R}$, $\mathrm{P}^X := \mathcal{U}(0,1)$, and

$$\begin{aligned}\mathrm{P}(\cdot\,|\,X=x) = x\cdot\Big(\tau\cdot\mathcal{U}((-1,0)) + (1-\tau)\cdot\mathcal{U}((0,1))\Big)&\\ + (1-x)\cdot\Big(\tau\cdot\delta_{-1/x} + (1-\tau)\cdot\delta_{1/x}\Big) \qquad &\forall\, x\in\mathcal{X}\,,\end{aligned}$$

where $\mathcal{U}(a,b)$ denotes the uniform distribution on (a,b) and δ_z denotes the Dirac distribution in $z\in\mathbb{R}$.[12] From this definition, we immediately obtain that $f^*_{\tau,\mathrm{P}} \equiv 0 \in L_1(\mathrm{P}^X)$.

Further define

$$f_n\colon \mathcal{X}\to\mathbb{R}\,, \qquad x\mapsto \begin{cases} n & \text{, if } x\in\left(0,\frac{1}{n}\right)\,,\\ 0 & \text{, else}\,,\end{cases}$$

for all $n\in\mathbb{N}$. As f_n is bounded and measurable for all $n\in\mathbb{N}$, we have $(f_n)_{n\in\mathbb{N}}\subseteq L_1(\mathrm{P}^X)$.

Because of the occurring risks both being finite, cf. (2.5), and $\mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} = \mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}})$, cf. (3.16), we can for all $n\in\mathbb{N}$ write

$$\begin{aligned}&\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}}\\ &= \int_{(0,1)}\int_{\mathbb{R}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{d}\mathrm{P}(y\,|\,x)\,\mathrm{d}\mathrm{P}^X(x)\,. \qquad (3.18)\end{aligned}$$

For P^X-almost all $x\in\mathcal{X}$, we can now further analyze the inner integral, applying that $f_n(x)\geq f^*_{\tau,\mathrm{P}}(x)$, by

$$\begin{aligned}&\int_{\mathbb{R}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{d}\mathrm{P}(y\,|\,x)\\ &= \int_{\mathbb{R}} L_{\tau\text{-pin}}(y, f_n(x)) - L_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{d}\mathrm{P}(y\,|\,x)\\ &= \int_{\left(-\infty, f^*_{\tau,\mathrm{P}}(x)\right)} (1-\tau)\cdot\Big(f_n(x) - f^*_{\tau,\mathrm{P}}(x)\Big)\,\mathrm{d}\mathrm{P}(y\,|\,x)\\ &\quad + \int_{\left[f^*_{\tau,\mathrm{P}}(x), f_n(x)\right)} (-\tau)\cdot\Big(f_n(x) - f^*_{\tau,\mathrm{P}}(x)\Big) + (f_n(x) - y)\,\mathrm{d}\mathrm{P}(y\,|\,x)\\ &\quad + \int_{[f_n(x),\infty)} (-\tau)\cdot\Big(f_n(x) - f^*_{\tau,\mathrm{P}}(x)\Big)\,\mathrm{d}\mathrm{P}(y\,|\,x)\\ &= \int_{\left[f^*_{\tau,\mathrm{P}}(x), f_n(x)\right)} (f_n(x) - y)\,\mathrm{d}\mathrm{P}(y\,|\,x)\,. \qquad (3.19)\end{aligned}$$

[12] For the sake of strictly adhering to the completeness assumption from Assumption 3.0.1, we can also choose $\mathcal{X}$ as $[0,1]$ or $\mathbb{R}$, and $\mathrm{P}(\cdot\,|\,X=x)$ as an arbitrary probability measure for $x\notin(0,1)$ without changing anything else.

In the last step, we employed that, for P^X-almost all $x \in \mathcal{X}$, we know from the definition of P that $\mathrm{P}(\{f^*_{\tau,\mathrm{P}}(x)\} \,|\, x) = 0$ and therefore $\mathrm{P}((-\infty, f^*_{\tau,\mathrm{P}}(x)) \,|\, x) = \tau$ and $\mathrm{P}([f^*_{\tau,\mathrm{P}}(x), \infty) \,|\, x) = 1 - \tau$ by the definition of $f^*_{\tau,\mathrm{P}}$.

Plugging (3.19) and the definition of f_n and $f^*_{\tau,\mathrm{P}}$ into (3.18), we obtain

$$\begin{aligned}\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f_n) - \mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} &= \int_{\left(0,\frac{1}{n}\right)} \int_{[0,n)} (n-y) \,\mathrm{d}\mathrm{P}(y \,|\, x) \,\mathrm{d}\mathrm{P}^X(x) \\ &= \int_0^{\frac{1}{n}} \int_0^1 (n-y) \cdot x \cdot (1-\tau) \,\mathrm{d}y \,\mathrm{d}x \\ &= (1-\tau) \cdot \frac{2n-1}{4n^2} \to 0\,, \qquad n \to \infty\,.\end{aligned}$$

On the other hand,

$$\left|\left|f_n - f^*_{\tau,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)} = \int_0^{\frac{1}{n}} |n - 0| \,\mathrm{d}x = 1 \not\to 0\,, \qquad n \to \infty\,,$$

which completes the proof. □

As the preceding results allow for arbitrary sequences of functions in $L_1(\mathrm{P}^X)$, we might still hope to deduce L_1-consistency following from $L^\star$-risk consistency by restricting ourselves to smaller function spaces with more structure like *Sobolev spaces*. However, the subsequent corollary shows that Proposition 3.2.6 and Proposition 3.2.8 can even be strengthened to sequences of functions from Sobolev spaces. Here, we assume that $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$, and we denote by $W^{m,q}(\mathcal{X})$ the *Sobolev space* consisting of all functions from $L_q(\mathcal{X})$ whose weak derivatives (cf. Adams and Fournier, 2003, Paragraph 1.62) up to order m are also in $L_q(\mathcal{X})$, cf. Adams and Fournier (2003, Definition 3.2). Here, as usual, $L_q(\mathcal{X})$ denotes the L_q-space with respect to the Lebesgue measure on $\mathcal{X}$.

Corollary 3.2.9. *Let $d \in \mathbb{N}$, $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{Y} = \mathbb{R}$. Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based and symmetric loss function of growth type 1, or the τ-pinball loss for some $\tau \in (0,1)$. Let $L^\star$ be its shifted version. Let $m \in \mathbb{N}$ and $1 \leq q \leq \infty$. Then, even if $f^*_{L^\star,\mathrm{P}}$ is P^X-a.s. unique with $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}^X)$, a sequence $(f_n)_{n\in\mathbb{N}} \subseteq W^{m,q}(\mathcal{X}) \cap L_1(\mathrm{P}^X)$ of functions satisfying*

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

*does in general **not** imply*

$$\lim_{n\to\infty} \left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)} = 0$$

without any additional assumptions besides Assumption 3.0.1 being imposed.

Proof. The assertion follows directly from the proof of Proposition 3.2.6 respectively Proposition 3.2.8 by changing the functions f_n, $n \in \mathbb{N}$, to

$$f_n\colon \mathcal{X} \to \mathbb{R}\,, \qquad x \mapsto \begin{cases} n \cdot (1-nx)^m & \text{, if } x \in \left(0, \frac{1}{n}\right)\,, \\ 0 & \text{, else}\,. \end{cases}$$

Since, for all $n \in \mathbb{N}$, f_n is bounded, measurable and m times weakly differentiable, we obtain $(f_n)_{n\in\mathbb{N}} \subseteq W^{m,\infty}(\mathcal{X}) \cap L_1(\mathrm{P}^X) \subseteq W^{m,q}(\mathcal{X}) \cap L_1(\mathrm{P}^X)$.[13]

If we denote the functions from the mentioned proofs by g_n, $n \in \mathbb{N}$, we have $f^*_{L^\star,\mathrm{P}}(x) \leq f_n(x) \leq g_n(x)$ for P^X-almost all $x \in \mathcal{X}$ because $f^*_{L^\star,\mathrm{P}} = 0$ P^X-a.s. (with $f^*_{L^\star,\mathrm{P}} = f^*_{\tau,\mathrm{P}}$ P^X-a.s. in the situation of $L^\star = L^\star_{\tau\text{-pin}}$ by the considerations prior to Proposition 3.2.8). It is easy to see that the convexity of L and the definition of $f^*_{L^\star,\mathrm{P}}$ as a minimizer of $\mathcal{R}_{L^\star,\mathrm{P}}$ therefore implies $\mathcal{R}_{L^\star,\mathrm{P}}(f_n) - \mathcal{R}^*_{L^\star,\mathrm{P}} \leq \mathcal{R}_{L^\star,\mathrm{P}}(g_n) - \mathcal{R}^*_{L^\star,\mathrm{P}}$, which then yields $\lim_{n\to\infty} \mathcal{R}_{L^\star,\mathrm{P}}(f_n) = \mathcal{R}^*_{L^\star,\mathrm{P}}$.

At the same time, we obtain

$$\left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)} = \int_0^{1/n} |n \cdot (1-nx)^m - 0| \,\mathrm{d}x = \frac{1}{m+1} \not\to 0\,, \qquad n \to \infty\,,$$

which completes the proof. □

The preceding results show that it is not possible to get rid of the moment condition from Theorem 3.2.1 (cf. Remark 3.2.2) just by transferring it to shifted loss functions. It might, however, still be possible to circumvent this moment condition by instead imposing some different and less restrictive conditions. For the pinball loss, i.e. for performing quantile regression, we are indeed able to derive such an alternative and in many cases less restrictive condition regarding P. To be more specific, the conditional distribution $\mathrm{P}(\cdot\,|\,X)$ is, in some sense, not allowed to be too heteroscedastic and it has to be continuous in the conditional quantiles $f^*_{\tau,\mathrm{P}}(x)$, $x \in \mathcal{X}$, which gets formalized by (3.20) and (3.21) in the subsequent theorem. Condition (3.20) gets visualized in Figure 3.2.1.

Theorem 3.2.10. *Let $\tau \in (0,1)$ and $L^\star_{\tau\text{-}pin}$ be the shifted version of the τ-pinball loss. Assume that $f^*_{\tau,\mathrm{P}}$ exists and is P^X-a.s. unique, $f^*_{\tau,\mathrm{P}} \in L_1(\mathrm{P}^X)$, and P additionally satisfies at least one of the following conditions:*

(i) $|\mathrm{P}|_1 < \infty$.

(ii) There exist $c_1, c_2 > 0$ such that

$$\mathrm{P}\Big((f^*_{\tau,\mathrm{P}}(X)-c_1, f^*_{\tau,\mathrm{P}}(X))\,\Big|\,X\Big) \geq c_2 \quad \textit{and} \quad \mathrm{P}\Big((f^*_{\tau,\mathrm{P}}(X), f^*_{\tau,\mathrm{P}}(X)+c_1)\,\Big|\,X\Big) \geq c_2 \qquad (3.20)$$

P^X*-a.s., as well as*

$$\mathrm{P}(f^*_{\tau,\mathrm{P}}(X)\,|\,X) = 0 \qquad (3.21)$$

P^X*-a.s.*

Then, for every sequence $(f_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}^X)$, we have

$$\lim_{n\to\infty} \mathcal{R}_{L^\star_{\tau\text{-}pin},\mathrm{P}}(f_n) = \mathcal{R}^*_{L^\star_{\tau\text{-}pin},\mathrm{P}} \qquad \Rightarrow \qquad \lim_{n\to\infty} ||f_n - f^*_{\tau,\mathrm{P}}||_{L_1(\mathrm{P}^X)} = 0\,.$$

[13] If $\mathcal{X}$ is not chosen as $(0,1)$ but instead as $\mathbb{R}$ in the proofs of Proposition 3.2.6 and Proposition 3.2.8 in order to strictly adhere to the assumptions, it is obviously possible to extend the functions f_n, $n \in \mathbb{N}$, in such a way that they are still in $W^{m,\infty}(\mathcal{X}) \cap L_1(\mathrm{P}^X)$.

Proof. By (2.5), both $\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f_n)$, $n \in \mathbb{N}$, and $\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}})$ are finite.

If condition (i) is satisfied, we further obtain as in Remark 3.2.2 that $\mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(0)$ and $\mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(f_n)$, for $n \in \mathbb{N}$, are finite, and therefore also $\mathcal{R}^*_{L_{\tau\text{-pin}},\mathrm{P}}$. As $\mathcal{R}^*_{L_{\tau\text{-pin}},\mathrm{P}} = \mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}})$ and $\mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} = \mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}})$ by (3.16) and (3.17), we hence obtain

$$\mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(f_n) = \mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f_n) + \mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(0) \qquad \forall\, n \in \mathbb{N}$$

and

$$\mathcal{R}^*_{L_{\tau\text{-pin}},\mathrm{P}} = \mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}}) = \mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}}) + \mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(0) = \mathcal{R}^*_{L^\star_{\tau\text{-pin}},\mathrm{P}} + \mathcal{R}_{L_{\tau\text{-pin}},\mathrm{P}}(0)\,.$$

Theorem 3.2.1 and Remark 3.2.2 then yield the assertion because of $L_{\tau\text{-pin}}$ being of growth type 1. Thus, it is only left to show that condition (ii) yields the assertion as well:

Because of the finiteness of $\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f_n)$, $n \in \mathbb{N}$, and $\mathcal{R}_{L^\star_{\tau\text{-pin}},\mathrm{P}}(f^*_{\tau,\mathrm{P}})$, the assumed risk consistency implies that the P-integral of $L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))$ converges to 0 as $n \to \infty$. We will now begin by fixing an $x \in \mathcal{X}$ and further analyzing the inner integral with respect to $\mathrm{P}(\cdot\,|\,x)$:

First, we look at the case that $f_n(x) \geq f^*_{\tau,\mathrm{P}}(x)$. In this case, repeating the considerations from (3.19), where we can apply (3.21) in the last step, yields for P^X-almost all such x that

$$\begin{aligned}
&\int_{\mathcal{Y}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{dP}(y|x)\\
&= \int_{\left[f^*_{\tau,\mathrm{P}}(x), f_n(x)\right)} (f_n(x) - y)\,\mathrm{dP}(y|x)\\
&\geq \int_{\left[f^*_{\tau,\mathrm{P}}(x), \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right)} (f_n(x) - y)\,\mathrm{dP}(y|x)\\
&\geq \left(f_n(x) - \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right)\cdot \mathrm{P}\left(\left(f^*_{\tau,\mathrm{P}}(x), \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right)\middle|\, x\right)\\
&= \frac{f_n(x) - f^*_{\tau,\mathrm{P}}(x)}{2}\cdot \mathrm{P}\left(\left(f^*_{\tau,\mathrm{P}}(x), \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right)\middle|\, x\right)\,.
\end{aligned}$$

If on the other hand $f_n(x) < f^*_{\tau,\mathrm{P}}(x)$, we analogously obtain for P^X-almost all such x:

$$\begin{aligned}
&\int_{\mathcal{Y}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{dP}(y|x)\\
&\geq \frac{f^*_{\tau,\mathrm{P}}(x) - f_n(x)}{2}\cdot \mathrm{P}\left(\left(\frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}, f^*_{\tau,\mathrm{P}}(x)\right)\middle|\, x\right)\,.
\end{aligned}$$

In summary,

$$\int_{\mathcal{Y}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x))\,\mathrm{dP}(y|X) \geq \frac{|f_n(X) - f^*_{\tau,\mathrm{P}}(X)|}{2}\cdot \mathrm{P}\left(J_{X,n}\middle|\, X\right) \tag{3.22}$$

P^X-a.s., where

$$J_{x,n} := \left(\min\left\{f^*_{\tau,\mathrm{P}}(x), \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right\}, \max\left\{f^*_{\tau,\mathrm{P}}(x), \frac{f_n(x)+f^*_{\tau,\mathrm{P}}(x)}{2}\right\}\right)$$

for all $x \in \mathcal{X}$.

Additionally, Christmann et al. (2009, Corollary 31) yields $f_n \xrightarrow{\mathrm{P}^X} f^*_{\tau,\mathrm{P}}$, i.e.

$$\lim_{n\to\infty} \mathrm{P}^X(|f_n(X) - f^*_{\tau,\mathrm{P}}(X)| > \varepsilon) = 0 \qquad \forall \varepsilon > 0\,. \tag{3.23}$$

Now, let $\varepsilon > 0$ be an arbitrary positive number (without loss of generality $\varepsilon < 2c_1$). $\mathcal{X}$ can be partitioned as $\mathcal{X} = \dot{\bigcup}_{i=1}^{3} \mathcal{X}_{i,\varepsilon}$, where

$$\begin{aligned}
\mathcal{X}_{1,\varepsilon} &:= \Big\{x \in \mathcal{X} : |f_n(x) - f^*_{\tau,\mathrm{P}}(x)| \le \varepsilon\Big\}\,,\\
\mathcal{X}_{2,\varepsilon} &:= \Big\{x \in \mathcal{X} : \varepsilon < |f_n(x) - f^*_{\tau,\mathrm{P}}(x)| \le 2 \cdot c_1\Big\}\,,\\
\mathcal{X}_{3,\varepsilon} &:= \mathcal{X}_3 := \Big\{x \in \mathcal{X} : |f_n(x) - f^*_{\tau,\mathrm{P}}(x)| > 2 \cdot c_1\Big\}\,,
\end{aligned}$$

such that

$$||f_n - f^*_{\tau,\mathrm{P}}||_{L_1(\mathrm{P}^X)} = \sum_{i=1}^{3} \int_{\mathcal{X}_{i,\varepsilon}} |f_n(x) - f^*_{\tau,\mathrm{P}}(x)| \,\mathrm{dP}^X(x)\,. \tag{3.24}$$

The three summands can now be analyzed separately:

$$\begin{aligned}
&\int_{\mathcal{X}_{1,\varepsilon}} |f_n(x) - f^*_{\tau,\mathrm{P}}(x)| \,\mathrm{dP}^X(x) \le \varepsilon,\\
&\int_{\mathcal{X}_{2,\varepsilon}} |f_n(x) - f^*_{\tau,\mathrm{P}}(x)| \,\mathrm{dP}^X(x) \le 2 \cdot c_1 \cdot \mathrm{P}^X(\mathcal{X}_{2,\varepsilon}) \xrightarrow{(3.23)} 0\,, \qquad n \to \infty\,,
\end{aligned}$$

and

$$\begin{aligned}
&\int_{\mathcal{X}_{3,\varepsilon}} |f_n(x) - f^*_{\tau,\mathrm{P}}(x)| \,\mathrm{dP}^X(x)\\
&= \int_{\mathcal{X}_3} \left(\frac{|f_n(x) - f^*_{\tau,\mathrm{P}}(x)|}{2} \cdot \mathrm{P}(J_{x,n}|x)\right) \cdot \frac{2}{\mathrm{P}(J_{x,n}|x)} \,\mathrm{dP}^X(x)\\
&\overset{(3.20),(3.22)}{\le} \frac{2}{c_2} \cdot \int_{\mathcal{X}_3} \int_{\mathcal{Y}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x)) \,\mathrm{dP}(y|x)\,\mathrm{dP}^X(x)\\
&\to 0\,, \qquad n \to \infty\,,
\end{aligned}$$

with the last convergence holding true because

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x)) \,\mathrm{dP}(y|x)\,\mathrm{dP}^X(x) \to 0\,, \qquad n \to \infty\,,$$

by assumption and

$$\int_{\mathcal{Y}} L^\star_{\tau\text{-pin}}(y, f_n(x)) - L^\star_{\tau\text{-pin}}(y, f^*_{\tau,\mathrm{P}}(x)) \,\mathrm{dP}(y|X) \ge 0$$

P^X-a.s. by (3.22).

Plugging these results into (3.24) yields the assertion. □

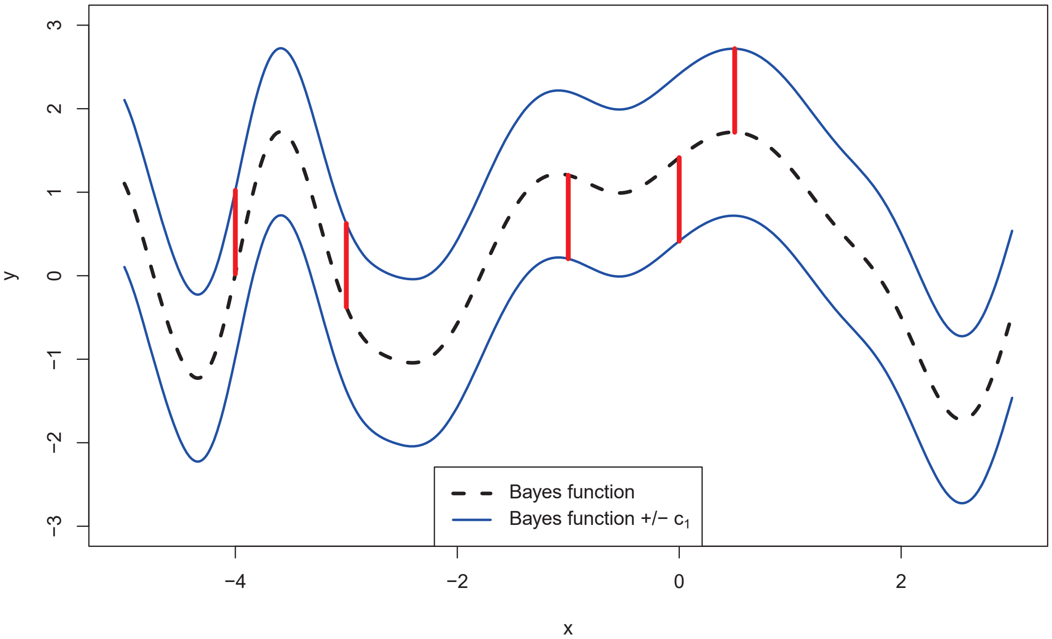

Figure 3.2.1: Visualization of (3.20). Each vertical slice between $f^*_{\tau,\mathrm{P}} - c_1$ and $f^*_{\tau,\mathrm{P}}$ as well as between $f^*_{\tau,\mathrm{P}}$ and $f^*_{\tau,\mathrm{P}} + c_1$ needs to have a conditional probability (given x) of at least c_2. The solid vertical lines depict some examples of such slices whose conditional probability needs to be at least c_2. [This is a minimally modified version of a figure that was first published in Köhler, 2024b.]

Even though it was not possible to get rid of the moment condition (i) without imposing the new conditions (ii), this still substantially expands the applicability of the theorem since there are many cases in which (ii) (whose first part is visualized in Figure 3.2.1) is satisfied even though (i) is not:

Example 3.2.11. Assume that $\tau \in (0, 1)$ and that we have an underlying homoscedastic regression model like

$$Y = f(X) + \varepsilon\,,$$

where $f\colon \mathcal{X} \to \mathcal{Y}$ is an arbitrary measurable function and ε is a continuous random variable whose distribution does not depend on the value of X. Whenever ε has a unique τ-quantile $q_\tau \in \mathbb{R}$, (ii) from Theorem 3.2.10 holds true with $f^*_{\tau,\mathrm{P}} = f + q_\tau$. For example, ε can follow a Cauchy distribution with location and scale parameters which are fixed independently of the value of X. In this case, the moment condition (i) does not hold true, but Theorem 3.2.10 does still yield L_1-consistency following from risk consistency.

Example 3.2.12. The independence of ε from X in Example 3.2.11 is not even strictly necessary. Assume the more general heteroscedastic model

$$Y = f(X) + \varepsilon_X\,,$$

where the distribution of ε_X is now allowed to depend on the value x of X. If, for example, there exist $C > 0$ and $c_1 > 0$ such that ε_x has a unique τ-quantile $q_{x,\tau} \in \mathbb{R}$ and Lebesgue density greater than C on $(q_{x,\tau} - c_1, q_{x,\tau} + c_1)$ for P^X-almost all $x \in \mathcal{X}$, condition (ii) from Theorem 3.2.10 is still satisfied.

For example, this situation is on hand if $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$, $\mathcal{Y} = \mathbb{R}$, and ε_x follows a Cauchy distribution with location parameter $\cos(||x||_2)$ and scale parameter $2+\sin(||x||_2)$ for all $x \in \mathcal{X}$. More generally, the same also holds true for different choices of location and scale parameters, as long as they are bounded from above and from below (in the case of the scale parameter we mean bounded away from zero by bounded from below).

We saw that L_1-consistency can not be obtained from risk consistency without imposing some different, albeit in some sense weaker, condition regarding P in exchange for omitting the moment condition. It is, however, indeed possible to just omit the moment condition in the reverse statement (Theorem 3.2.3) when transferring this to shifted loss functions in the case of having a convex loss function of upper growth type 1, which again hints at this direction being the easier one as it was suspected in Section 3.1.

Theorem 3.2.13. *Let $L\colon \mathcal{Y} \to \mathbb{R}$ be a convex, distance-based loss function of upper growth type 1, and let $L^\star$ be its shifted version. Then, for every sequence $(f_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}^X)$ and every function $f^* \in L_1(\mathrm{P}^X)$, we have*

$$\lim_{n\to\infty} ||f_n - f^*||_{L_1(\mathrm{P}^X)} = 0 \qquad \Rightarrow \qquad \lim_{n\to\infty} \mathcal{R}_{L^\star,\mathrm{P}}(f_n) = \mathcal{R}_{L^\star,\mathrm{P}}(f^*)\,.$$

Proof. We know from (2.5) that all risks appearing in this result are finite. L additionally being Lipschitz continuous (cf. Lemma 2.1.17) yields

$$\begin{aligned} |\mathcal{R}_{L^\star,\mathrm{P}}(f_n) - \mathcal{R}_{L^\star,\mathrm{P}}(f^*)| &\leq \int |L^\star(y, f_n(x)) - L^\star(y, f^*(x))| \,\mathrm{d}\mathrm{P}(x,y) \\ &= \int |L(y, f_n(x)) - L(y, f^*(x))| \,\mathrm{d}\mathrm{P}(x,y) \\ &\leq |L|_1 \cdot \int |f_n(x) - f^*(x)| \,\mathrm{d}\mathrm{P}(x,y) \\ &= |L|_1 \cdot ||f_n - f^*||_{L_1(\mathrm{P}^X)} \to 0 \qquad n \to \infty\,. \end{aligned}$$ □

Lastly, it is obvious that the first part of Corollary 3.2.4—H-consistency implying L_p-consistency (for any $p \in [1,\infty]$, not only for $p = 1$)—also holds true in the situation of the present section as the loss function did not come into play in that statement. The second part of Corollary 3.2.4—H-consistency implying risk consistency—does depend on the loss function but can be transferred to shifted loss functions seamlessly:

Corollary 3.2.14. *Let $L\colon \mathcal{Y} \to \mathbb{R}$ be a convex, distance-based loss function of upper growth type 1, and let $L^\star$ be its shifted version. Let H be the RKHS of a bounded and measurable kernel k on $\mathcal{X}$. Then, for every sequence $(f_n)_{n\in\mathbb{N}} \subseteq H$ and every function $f^* \in H$, we have*

$$\lim_{n\to\infty} ||f_n - f^*||_H = 0 \qquad \Rightarrow \qquad \lim_{n\to\infty} \mathcal{R}_{L^\star,\mathrm{P}}(f_n) = \mathcal{R}_{L^\star,\mathrm{P}}(f^*)\,.$$

Proof. The assertion follows directly from Corollary 3.2.4(i) and Theorem 3.2.13, which can be applied because $H \subseteq L_1(\mathrm{P}^X)$ by Lemma 2.1.10(ii). □

3.3 Consistency of Support Vector Machines

In this section, the results from Section 3.2 are used as an aid for deriving new results on different types of consistency of SVMs. First, SVMs using regular loss functions are considered. For them, Section 3.3.1 examines L_p-consistency. The results from that section are then employed in Section 3.3.2 to in some sense (which gets specified in Sections 3.3.1 and 3.3.2) improve existing results on risk consistency of such SVMs, and results on their H-consistency are derived in Section 3.3.3. Finally, SVMs based on shifted loss functions are investigated in Section 3.3.4 (L_p-consistency) and Section 3.3.5 (H-consistency). Risk consistency of SVMs based on shifted loss functions is not investigated in a separate section since the results from the preceding Section 3.2 would not yield any improvement over existing results in this case. This is explained in slightly more detail at the end of Section 3.3.4.

Sections 3.3.1, 3.3.2 and 3.3.4 are mostly taken from the peer-reviewed paper Köhler (2024b, Section 4) that was published in *Journal of Machine Learning Research*. Sections 3.3.3 and 3.3.5 consist of previously unpublished results.

3.3.1 L_p-Consistency Using Regular Loss Functions

Whereas SVMs based on distance-based losses are known to be risk consistent under mild assumptions (cf. Christmann and Steinwart, 2007, Theorem 12), there are no general results on their L_p-consistency so far, but instead only corollaries for special loss functions based on the results mentioned at the beginning of Section 3.2.1.

Since the conditions required by Christmann and Steinwart (2007, Theorem 12) also imply the validity of Theorem 3.2.1, L_p-consistency of such SVMs would now directly follow under these conditions. However, by some more thorough investigations, we are even able to slightly relax the conditions on the sequence $(\lambda_n)_{n\in\mathbb{N}}$ of regularization parameters, namely only requiring it to satisfy $\lambda_n^{p^*} n \to \infty$ (as $n \to \infty$) for $p^* = \max\{p+1, p(p+1)/2\}$ instead of for $p^* = \max\{2p, p^2\}$, which is required by Christmann and Steinwart (2007, Theorem 12). This relaxation gets apparent in the following example:

Example 3.3.1. The popular least squares loss function is of growth type $p = 2$. Hence in this case $\max\{p+1, p(p+1)/2\} = 3 < 4 = \max\{2p, p^2\}$. Thus, the subsequent Theorem 3.3.2 yields L_p-consistency (and Corollary 3.3.5 will yield risk consistency) of SVMs using the least squares loss under the condition that $\lambda_n^3 n \to \infty$ as $n \to \infty$, which is for example satisfied if $\lambda_n \propto n^{-1/4}$. On the other hand, Christmann and Steinwart (2007, Theorem 12) guarantees risk consistency of such SVMs only if $\lambda_n^4 n \to \infty$ as $n \to \infty$, which is not satisfied for $\lambda_n \propto n^{-1/4}$. Thus, our new results allow for slightly faster convergence of the regularization parameter to 0 and one therefore becomes more flexible in choosing the regularization parameters while still being guaranteed consistency.

It should be noted that such a relaxation takes place whenever $p > 1$ holds true. The case $p = 1$ is the only one, in which $\max\{p+1, p(p+1)/2\} = \max\{2p, p^2\}$.

Theorem 3.3.2. *Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex, distance-based loss function of growth type $p \in [1, \infty)$. Let $H \subseteq L_p(\mathrm{P}^X)$ dense and separable be the RKHS of a bounded* *and measurable kernel k on $\mathcal{X}$. Assume that $f^*_{L,\mathrm{P}}$ exists and is P^X-a.s. unique and $|\mathrm{P}|_p < \infty$. Define $p^* := \max\{p+1, p(p+1)/2\}$. If the sequence $(\lambda_n)_{n\in\mathbb{N}}$ satisfies $\lambda_n > 0$ for all $n \in \mathbb{N}$ as well as $\lambda_n \to 0$ and $\lambda_n^{p^*} n \to \infty$ for $n \to \infty$, then*

$$\lim_{n\to\infty} ||f_{L,\mathrm{D}_n,\lambda_n,k} - f^*_{L,\mathrm{P}}||_{L_p(\mathrm{P}^X)} = 0 \qquad \textit{in probability } \mathrm{P}^\infty.$$

We outsource the following statement, which is needed in the proof, into a separate lemma as this will be needed again in the proofs of Propositions 3.3.6 and 3.3.7.

Lemma 3.3.3. *Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss function of upper growth type $p \in [1,\infty)$. Let k be a bounded and measurable kernel on $\mathcal{X}$ with separable RKHS H. Assume that $|\mathrm{P}|_p < \infty$. Define $p^* := \max\{p+1, p(p+1)/2\}$. If the sequence $(\lambda_n)_{n\in\mathbb{N}}$ is bounded and satisfies $\lambda_n > 0$ for all $n \in \mathbb{N}$ as well as $\lambda_n^{p^*} n \to \infty$ for $n \to \infty$, then*

$$\lim_{n\to\infty} ||f_{L,\mathrm{D}_n,\lambda_n,k} - f_{L,\mathrm{P},\lambda_n,k}||_H = 0 \qquad \textit{in probability } \mathrm{P}^\infty.$$

Proof. Start by noting that applying Lemma 2.1.10(i), Steinwart and Christmann (2008, eq. (5.4)), and Lemma 2.1.21(ii) yields

$$||f_{L,\mathrm{P},\lambda_n,k}||_\infty \leq ||k||_\infty \cdot ||f_{L,\mathrm{P},\lambda_n,k}||_H \leq ||k||_\infty \cdot \mathcal{R}_{L,\mathrm{P}}(0)^{1/2} \cdot \lambda_n^{-1/2} \leq c_{p,L,\mathrm{P},k} \cdot \lambda_n^{-1/2} \quad (3.25)$$

for all $n \in \mathbb{N}$, with $c_{p,L,\mathrm{P},k} \in (0,\infty)$ denoting a constant depending only on p, L, P and k, but not on λ_n.

We know from Steinwart and Christmann (2008, Corollary 5.11) that there exist functions $h_n\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, $n \in \mathbb{N}$, such that

$$||f_{L,\mathrm{D}_n,\lambda_n,k} - f_{L,\mathrm{P},\lambda_n,k}||_H \leq \frac{1}{\lambda_n} \cdot ||\mathbb{E}_{\mathrm{D}_n}[h_n\Phi] - \mathbb{E}_{\mathrm{P}}[h_n\Phi]||_H \qquad \forall\, n \in \mathbb{N}\,, \quad (3.26)$$

and, for $s := p/(p-1)$,

$$\begin{aligned} ||h_n||_{L_s(\mathrm{P})} &\leq 8^p \cdot c_L \cdot \left(1 + |\mathrm{P}|_p^{p-1} + ||f_{L,\mathrm{P},\lambda_n,k}||_\infty^{p-1}\right) \\ &\leq 8^p \cdot c_L \cdot \left(1 + |\mathrm{P}|_p^{p-1} + c_{p,L,\mathrm{P},k}^{p-1} \cdot \lambda_n^{-(p-1)/2}\right) \\ &\leq \tilde{c}_{p,L,\mathrm{P},k} \cdot \lambda_n^{-(p-1)/2} \qquad\qquad \forall\, n \in \mathbb{N}\,, \end{aligned} \quad (3.27)$$

where we employed (3.25) in the second and the boundedness of $(\lambda_n)_{n\in\mathbb{N}}$ in the third step, and where $c_L \in (0,\infty)$ and $\tilde{c}_{p,L,\mathrm{P},k} \in (0,\infty)$ denote constants depending only on L respectively p, L, P and k.

Now, we can apply Steinwart and Christmann (2008, Lemma 9.2) with $q := p/(p-1)$ if $p > 1$ and $q := 2$ if $p = 1$, which leads to $q^* := \min\{1/2, 1-1/q\} = \min\{1/2, 1/p\} = (p+1)/(2p^*)$, to the functions $h_n\Phi$, $n \in \mathbb{N}$: First of all, with the help of (3.27) we obtain

$$||h_n\Phi||_q := \left(\mathbb{E}_{\mathrm{P}}[||h_n\Phi||_H^q]\right)^{1/q} \leq ||k||_\infty \cdot ||h_n||_{L_q(\mathrm{P})} \leq ||k||_\infty \cdot \tilde{c}_{p,L,\mathrm{P},k} \cdot \lambda_n^{-(p-1)/2} < \infty$$

for all $n \in \mathbb{N}$. We employed that, for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$||h_n(x,y)\Phi(x)||_H^q = |h_n(x,y)|^q \cdot ||\Phi(x)||_H^q = |h_n(x,y)|^q \cdot k(x,x)^{q/2} \leq |h_n(x,y)|^q \cdot ||k||_\infty^q$$

by the reproducing property. Hence, we obtain for all $\varepsilon > 0$, by combining this Lemma 9.2 with (3.26),

$$\begin{aligned}
&\mathrm{P}^n \left(D_n \in (\mathcal{X} \times \mathcal{Y})^n : ||f_{L,\mathrm{D}_n,\lambda_n,k} - f_{L,\mathrm{P},\lambda_n,k}||_H \geq \varepsilon\right) \\
&\leq \mathrm{P}^n \left(D_n \in (\mathcal{X} \times \mathcal{Y})^n : ||\mathbb{E}_{\mathrm{D}_n}[h_n\Phi] - \mathbb{E}_{\mathrm{P}}[h_n\Phi]||_H \geq \lambda_n \cdot \varepsilon\right) \\
&\leq c_q \cdot \left(\frac{||h_n\Phi||_q}{\lambda_n \varepsilon n^{q^*}}\right)^q \leq \hat{c}_{p,L,\mathrm{P},k} \cdot \left(\frac{1}{\lambda_n^{(p+1)/2} \varepsilon n^{q^*}}\right)^q \to 0\,, \qquad n \to \infty\,,
\end{aligned}$$

with $c_q \in (0,\infty)$ and $\hat{c}_{p,L,\mathrm{P},k} \in (0,\infty)$ denoting constants depending only on q (that is, only on p) respectively p, L, P and k, and with the convergence in the last step holding true because

$$\lambda_n^{(p+1)/2} n^{q^*} = \left(\lambda_n^{(p+1)/(2q^*)} n\right)^{q^*} = \left(\lambda_n^{p^*} n\right)^{q^*} \to \infty\,, \qquad n \to \infty\,,$$

by the assumptions on $(\lambda_n)_{n\in\mathbb{N}}$. This completes the proof. □

Proof of Theorem 3.3.2. We can split up the difference, which we have to investigate, as

$$\begin{aligned}
\left|\left|f_{L,\mathrm{D}_n,\lambda_n,k} - f^*_{L,\mathrm{P}}\right|\right|_{L_p(\mathrm{P}^X)} &\leq ||f_{L,\mathrm{D}_n,\lambda_n,k} - f_{L,\mathrm{P},\lambda_n,k}||_{L_p(\mathrm{P}^X)} + \left|\left|f_{L,\mathrm{P},\lambda_n,k} - f^*_{L,\mathrm{P}}\right|\right|_{L_p(\mathrm{P}^X)} \\
&\leq ||k||_\infty \, ||f_{L,\mathrm{D}_n,\lambda_n,k} - f_{L,\mathrm{P},\lambda_n,k}||_H + \left|\left|f_{L,\mathrm{P},\lambda_n,k} - f^*_{L,\mathrm{P}}\right|\right|_{L_p(\mathrm{P}^X)}
\end{aligned} \tag{3.28}$$

by Lemma 2.1.10(i). The first summand on the right hand side converges to 0 in probability as $n \to \infty$ by Lemma 3.3.3.

As for the second summand: First of all, Lemma 2.1.21(i) yields that L is a P-integrable Nemitski loss of order p. Hence, we know from Steinwart and Christmann (2008, Theorem 5.31) that

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

and Steinwart and Christmann (2008, Lemma 5.15) (with $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}} < \infty$ by Remark 3.2.2) then yields

$$\lim_{n\to\infty} \lambda_n \, ||f_{L,\mathrm{P},\lambda_n,k}||_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda_n,k}) - \mathcal{R}^*_{L,\mathrm{P}} = 0$$

because $\lambda_n \to 0$ as $n \to \infty$. Since $\lambda_n \, ||f_{L,\mathrm{P},\lambda_n,k}||_H^2$ is non-negative and $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda_n,k}) \geq \mathcal{R}^*_{L,\mathrm{P}}$ by the definition of $\mathcal{R}^*_{L,\mathrm{P}}$, we obtain

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda_n,k}) = \mathcal{R}^*_{L,\mathrm{P}}\,.$$

Hence, Theorem 3.2.1, whose conditions are satisfied because of the considerations from Remark 3.2.2, yields convergence to 0 (as $n \to \infty$) of the second summand on the right hand side of (3.28), which completes the proof. □

Remark 3.3.4. The conditions on H in Theorem 3.3.2 can be difficult to check directly. Because of $\mathcal{X}$ being separable by Assumption 3.0.1, the separability of H however immediately follows whenever k is continuous (cf. Lemma 2.1.10(iii)) and it suffices to verify this continuity instead. For example, the commonly used Gaussian RBF kernel (among many other kernels) satisfies this continuity and its RKHS additionally is dense in $L_p(\mathrm{P}^X)$ (cf. Example 2.1.12), for which reason this RKHS satisfies both conditions from Theorem 3.3.2.

3.3.2 Risk Consistency Using Regular Loss Functions

As we successfully slightly reduced the conditions regarding $(\lambda_n)_{n\in\mathbb{N}}$ compared to the referenced result on risk consistency in Section 3.3.1, we can now transfer this slight relaxation back from L_p-consistency to risk consistency by using Theorem 3.2.3:

Corollary 3.3.5. *Let $L\colon \mathcal{Y}\times\mathbb{R}\to[0,\infty)$ be a convex, distance-based loss function of growth type $p\in[1,\infty)$. Let $H\subseteq L_p(\mathrm{P}^X)$ dense and separable be the RKHS of a bounded and measurable kernel k on $\mathcal{X}$. Assume that $f^*_{L,\mathrm{P}}$ exists and is P^X-a.s. unique and $|\mathrm{P}|_p<\infty$. Define $p^*:=\max\{p+1,p(p+1)/2\}$. If the sequence $(\lambda_n)_{n\in\mathbb{N}}$ satisfies $\lambda_n>0$ for all $n\in\mathbb{N}$ as well as $\lambda_n\to 0$ and $\lambda_n^{p^*}n\to\infty$ for $n\to\infty$, then*

$$\lim_{n\to\infty}\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\lambda_n,k})=\mathcal{R}^*_{L,\mathrm{P}}\qquad \textit{in probability } \mathrm{P}^\infty.$$

Proof. The assertion follows directly from Theorem 3.3.2 and Theorem 3.2.3. □

Alas, the slight relaxation of the mentioned condition regarding the regularization parameters also comes along with an additional condition compared to Christmann and Steinwart (2007, Theorem 12): Corollary 3.3.5 requires $f^*_{L,\mathrm{P}}$ to P^X-a.s. uniquely exist. Thus, Corollary 3.3.5 pays for the slight relaxation in one condition by introducing this new additional condition and should therefore not be seen as a replacement of Theorem 12 from Christmann and Steinwart (2007) but as an addition instead.

3.3.3 H-Consistency Using Regular Loss Functions

If the Bayes function $f^*_{L,\mathrm{P}}$ is contained in H, it is even possible to strengthen the L_p- and risk consistency results from the preceding two sections to the stronger (cf. Corollary 3.2.4) H-consistency. Alas—in contrast to the condition $f^*_{L,\mathrm{P}}\in L_p(\mathrm{P}^X)$ that is needed for L_p-consistency but that could be replaced by the moment condition $|\mathrm{P}|_p<\infty$ in Theorem 3.3.2 based on Remark 3.2.2—the generally also not verifiable condition $f^*_{L,\mathrm{P}}\in H$ can not easily be replaced by a more intuitive alternative condition here.

Proposition 3.3.6. *Let $L\colon \mathcal{Y}\times\mathbb{R}\to[0,\infty)$ be a convex, distance-based loss function of upper growth type $p\in[1,\infty)$. Let H be the separable RKHS of a bounded and measurable kernel k on $\mathcal{X}$. Assume that $f^*_{L,\mathrm{P}}$ exists and is P^X-a.s. unique, $f^*_{L,\mathrm{P}}\in H$ and $|\mathrm{P}|_p<\infty$. Define $p^*:=\max\{p+1,p(p+1)/2\}$. If the sequence $(\lambda_n)_{n\in\mathbb{N}}$ satisfies $\lambda_n>0$ for all $n\in\mathbb{N}$ as well as $\lambda_n\to 0$ and $\lambda_n^{p^*}n\to\infty$ for $n\to\infty$, then*

$$\lim_{n\to\infty}\left|\left|f_{L,\mathrm{D}_n,\lambda_n,k}-f^*_{L,\mathrm{P}}\right|\right|_H=0\qquad \textit{in probability } \mathrm{P}^\infty.$$

Proof. By applying the triangle inequality, we obtain

$$\left|\left|f_{L,\mathrm{D}_n,\lambda_n,k}-f^*_{L,\mathrm{P}}\right|\right|_H\leq||f_{L,\mathrm{D}_n,\lambda_n,k}-f_{L,\mathrm{P},\lambda_n,k}||_H+\left|\left|f_{L,\mathrm{P},\lambda_n,k}-f^*_{L,\mathrm{P}}\right|\right|_H\,. \tag{3.29}$$

Lemma 3.3.3 then yields that the first summand on the right hand side converges to 0 in probability as $n\to\infty$.

Thus, only the second summand remains to be examined. By Lemma 2.1.21(i), L is a P-integrable Nemitski loss. Hence, Steinwart and Christmann (2008, Corollary 5.19) is applicable and yields (since $f^*_{L,\mathrm{P}} \in H$) that $\lambda \mapsto f_{L,\mathrm{P},\lambda,k}$ defines a continuous mapping from $[0,\infty]$ to H (by extending the definition of an SVM to the cases $\lambda = 0$ and $\lambda = \infty$). By noting that $f^*_{L,\mathrm{P}} \in H$ results in $f^*_{L,\mathrm{P}} = f_{L,\mathrm{P},0}$, this continuity yields, for all $\varepsilon > 0$,

$$\left|\left|f_{L,\mathrm{P},\lambda_n,k} - f^*_{L,\mathrm{P}}\right|\right|_H = ||f_{L,\mathrm{P},\lambda_n,k} - f_{L,\mathrm{P},0}||_H \leq \varepsilon$$

for n sufficiently large, which then yields the assertion. □

Notably and by Corollary 3.2.4, this result strengthens Theorem 3.3.2 to L_q-consistency for *any* $q \in [1,\infty]$—also for q exceeding the growth type p of the loss—if the additional assumption $f^*_{L,\mathrm{P}} \in H$ is satisfied. Furthermore, Proposition 3.3.6 offers the additional advantage of being transferable to Lipschitz continuous but not necessarily distance-based losses rather easily (cf. Proposition 3.3.7), thus being applicable to an even larger class of loss functions than the results on L_p- and risk consistency from Sections 3.3.1 and 3.3.2. This way, the subsequent Proposition 3.3.7 can also be applied for margin-based loss functions (cf. Definition 2.1.23) which are usually used in classification tasks. Most of the commonly used margin-based losses, like the hinge loss or the logistic loss, are Lipschitz continuous and satisfy $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ (when actually being used in classification tasks), which is required by the proposition.

Proposition 3.3.7. *Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, Lipschitz continuous loss function. Let H be the separable RKHS of a bounded and measurable kernel k on $\mathcal{X}$. Assume that $f^*_{L,\mathrm{P}}$ exists and is P^X-a.s. unique, $f^*_{L,\mathrm{P}} \in H$ and $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$. If the sequence $(\lambda_n)_{n\in\mathbb{N}}$ satisfies $\lambda_n > 0$ for all $n \in \mathbb{N}$ as well as $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$ for $n \to \infty$, then*

$$\lim_{n\to\infty} \left|\left|f_{L,\mathrm{D}_n,\lambda_n,k} - f^*_{L,\mathrm{P}}\right|\right|_H = 0 \qquad \textit{in probability } \mathrm{P}^\infty.$$

Proof. The proof works in exactly the same way as that of Proposition 3.3.6, with only small changes. Again, the triangle inequality yields

$$\left|\left|f_{L,\mathrm{D}_n,\lambda_n,k} - f^*_{L,\mathrm{P}}\right|\right|_H \leq ||f_{L,\mathrm{D}_n,\lambda_n,k} - f_{L,\mathrm{P},\lambda_n,k}||_H + \left|\left|f_{L,\mathrm{P},\lambda_n,k} - f^*_{L,\mathrm{P}}\right|\right|_H . \tag{3.30}$$

The first summand on the right hand side converging to 0 in probability can be shown similarly to Lemma 3.3.3. By Steinwart and Christmann (2008, Corollary 5.10), there exist functions $h_n\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, $n \in \mathbb{N}$, such that

$$||f_{L,\mathrm{D}_n,\lambda_n,k} - f_{L,\mathrm{P},\lambda_n,k}||_H \leq \frac{1}{\lambda_n} \cdot ||\mathbb{E}_{\mathrm{D}_n}[h_n\Phi] - \mathbb{E}_{\mathrm{P}}[h_n\Phi]||_H \qquad \forall\, n \in \mathbb{N}$$

and

$$||h_n||_\infty \leq |L|_1 \qquad \forall\, n \in \mathbb{N}\,.$$

Now, the convergence can be proven analogously to Lemma 3.3.3 by applying Steinwart and Christmann (2008, Lemma 9.2), where only the case $q \coloneqq 2$ (and hence $q^* = 1/2$) needs to be considered this time. This yields, for all $\varepsilon > 0$,

$$\mathrm{P}^n(D_n \in (\mathcal{X} \times \mathcal{Y})^n : ||f_{L,\mathrm{D}_n,\lambda_n,k} - f_{L,\mathrm{P},\lambda_n,k}||_H \geq \varepsilon) \leq c_2 \cdot \left(\frac{||k||_\infty \, |L|_1}{\lambda_n \varepsilon n^{1/2}} \right)^2$$
$$\to 0\,, \qquad n \to \infty\,,$$

with $c_2 \in (0, \infty)$ denoting the constant from the applied Lemma 9.2, and with the convergence holding true by the assumptions on $(\lambda_n)_{n\in\mathbb{N}}$.

Hence, only the second summand on the right hand side of (3.30) remains to be investigated. Here, L being a P-integrable Nemitski loss can now be seen from the fact that

$$L(x,y,t) \leq L(x,y,0) + |L|_1 \cdot |t| \qquad \forall\,(x,y,t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$$

by definition, and $L(\cdot,\cdot,0) \in L_1(\mathrm{P})$ (since $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ and $L \geq 0$). The rest follows the same way as in the proof of Proposition 3.3.6. □

3.3.4 L_p-Consistency Using Shifted Loss Functions

If one uses a Lipschitz continuous loss function (i.e. one of growth type 1 if it is distance-based, cf. Remark 2.1.18), it was explained in Section 2.1.4 that switching to the corresponding shifted loss function eliminates the moment condition $|\mathrm{P}|_1 < \infty$ in many results on SVMs.

The natural hope that Theorem 3.3.2 can be transferred to the shifted case similarly, thus also ridding it of the moment condition, might have already decreased because of the negative results from Section 3.2.2. As SVMs are always contained in some RKHS H, one might however still hope that counterexamples like the ones from the proofs of these results can not occur in such RKHSs because of the additional structure they possess compared to $L_1(\mathrm{P}^X)$.[14] Alas, Sobolev spaces like the ones considered in Corollary 3.2.9 are also RKHSs if one chooses an appropriate kernel like for example the ones found in Wendland (2005), which are classic examples of kernels with compact support. Hence, we obtain the following:

Corollary 3.3.8. *Let $d \in \mathbb{N}$, $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{Y} = \mathbb{R}$. Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based and symmetric loss function of growth type 1, or the τ-pinball loss for some $\tau \in (0,1)$. Let $L^\star$ be its shifted version. Then, even if H is the RKHS of a bounded and measurable kernel k on $\mathcal{X}$ and $f^*_{L^\star,\mathrm{P}}$ is P^X-a.s. unique with $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}^X)$, a sequence $(f_n)_{n\in\mathbb{N}} \subseteq H$ of functions satisfying*

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

*does in general **not** imply*

$$\lim_{n\to\infty} \left|\left| f_n - f^*_{L^\star,\mathrm{P}} \right|\right|_{L_1(\mathrm{P}^X)} = 0$$

without any additional assumptions besides Assumption 3.0.1 being imposed.

[14] By Lemma 2.1.21(ii), the associated kernel k being bounded and measurable implies that $H \subseteq L_1(\mathrm{P}^X)$.

Proof. There exist different kernels whose RKHS is $W^{2,2}(\mathcal{X})$. Examples of such kernels can be found in Berlinet and Thomas-Agnan (2004, Chapter 7), Saitoh and Sawano (2016, Theorem 1.11) among others. For this proof, we will however use the kernel $k_{1,1}$ defined by $k_{1,1}(x,x') := \phi_{1,1}(||x-x'||_2)$ with $\phi_{1,1}$ as in Wendland (2005, Definition 9.11), that is, $\phi_{1,1}(r) \propto (1-r)_+^3(3r+1)$ (cf. Wendland, 2005, Table 9.1). By Wendland (2005, Theorem 10.35), the RKHS of $k_{1,1}$ is indeed $W^{2,2}(\mathcal{X})$. Additionally, $k_{1,1}$ is bounded by $\phi_{1,1}(0) < \infty$ and because of its continuity also measurable. Applying Corollary 3.2.9 yields the assertion. □

As the (probably) most commonly used RKHSs for computing SVMs are those of the Gaussian RBF kernels, cf. Example 2.1.12, we also want to take a special look at these. After proving in Corollary 3.3.8 that RKHSs, in which L_1-consistency does not follow from risk consistency, do in fact exist, we see in the subsequent Corollary 3.3.9 that this phenomenon can not only occur for kernels whose RKHS is a Sobolev space but also for the Gaussian RBF kernel.

Corollary 3.3.9. *Let $d \in \mathbb{N}$, $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{Y} = \mathbb{R}$. Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based and symmetric loss function of growth type 1, or the τ-pinball loss for some $\tau \in (0,1)$. Let $L^\star$ be its shifted version. Let $\gamma \in (0,\infty)$ and H_γ be the RKHS of the Gaussian RBF kernel k_γ on $\mathcal{X}$. Then, even if $f^*_{L^\star,\mathrm{P}}$ is P^X-a.s. unique with $f^*_{L^\star,\mathrm{P}} \in L_1(\mathrm{P}^X)$, a sequence $(f_n)_{n\in\mathbb{N}} \subseteq H_\gamma$ of functions satisfying*

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

does in general ***not*** *imply*

$$\lim_{n\to\infty} \left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)} = 0$$

without any additional assumptions besides Assumption 3.0.1 being imposed.

Proof. Denote, for some $m \in \mathbb{N}$, the functions from the proof of Corollary 3.2.9 by g_n, $n \in \mathbb{N}$, that is,

$$g_n\colon \mathcal{X} \to \mathbb{R}\,, \qquad x \mapsto \begin{cases} n\cdot(1-nx)^m & \text{, if } x \in \left(0,\frac{1}{n}\right)\,, \\ 0 & \text{, else}\,. \end{cases}$$

Because $(g_n)_{n\in\mathbb{N}} \subseteq L_1(\mathrm{P}^X)$, there exists by Steinwart and Christmann (2008, Theorem 4.63) a sequence $(f_n)_{n\in\mathbb{N}} \subseteq H_\gamma$ such that

$$||f_n - g_n||_\infty \leq \frac{1}{n}$$

for all $n \in \mathbb{N}$.

Since both f_n and g_n are bounded, we obtain from (2.5) that, for all $n \in \mathbb{N}$, $\mathcal{R}_{L^\star,\mathrm{P}}(f_n) \in \mathbb{R}$ and $\mathcal{R}_{L^\star,\mathrm{P}}(g_n) \in \mathbb{R}$. Hence,

$$\begin{aligned}
&|\mathcal{R}_{L^\star,\mathrm{P}}(f_n) - \mathcal{R}_{L^\star,\mathrm{P}}(g_n)| \leq \int_{\mathcal{X}\times\mathcal{Y}} |L^\star(y, f_n(x)) - L^\star(y, g_n(x))| \,\mathrm{dP}(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} |L(y, f_n(x)) - L(y, g_n(x))| \,\mathrm{dP}(x,y) \leq |L|_1 \cdot \int_{\mathcal{X}\times\mathcal{Y}} |f_n(x) - g_n(x)| \,\mathrm{dP}(x,y) \\
&\leq |L|_1 \cdot \frac{1}{n} \to 0\,, \qquad n \to \infty\,.
\end{aligned}$$

with L being Lipschitz continuous by Lemma 2.1.17(ii). The risk consistency of $(g_n)_{n\in\mathbb{N}}$ shown in the proof of Corollary 3.2.9 then yields risk consistency of $(f_n)_{n\in\mathbb{N}}$.

On the other hand,

$$\lim_{n\to\infty} ||f_n - g_n||_{L_1(\mathrm{P}^X)} = \lim_{n\to\infty} \int_{\mathcal{X}} |f_n(x) - g_n(x)| \,\mathrm{dP}^X(x) \leq \lim_{n\to\infty} \frac{1}{n} = 0$$

combined with

$$\lim_{n\to\infty} \left|\left|g_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)} = \frac{1}{m+1}\,,$$

which is known from the proof of Corollary 3.2.9, yields

$$\lim_{n\to\infty} \left|\left|f_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)} \geq \lim_{n\to\infty} \left(\left|\left|g_n - f^*_{L^\star,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)} - ||f_n - g_n||_{L_1(\mathrm{P}^X)}\right) = \frac{1}{m+1}$$

and thus $(f_n)_{n\in\mathbb{N}}$ not being L_1-consistent. □

The previous results show that L_1-consistency of SVMs using shifted loss functions does in general not follow from their risk consistency, with the latter being known from Christmann et al. (2009, Theorem 8). Note that it might still be possible for such SVMs to be L_1-consistent for different reasons though.

At least in the special case of the shifted pinball loss, we found some alternative conditions to replace—and in many situations weaken—the moment condition from Theorem 3.3.2 in Theorem 3.2.10. With this, we can now at least deduce L_1-consistency of SVMs using this shifted pinball loss without needing to impose the moment condition:

Corollary 3.3.10. *Let $\tau \in (0,1)$ and $L^\star_{\tau\text{-}pin}$ be the shifted τ-pinball loss. Let $H \subseteq L_1(\mathrm{P}^X)$ dense and separable be the RKHS of a bounded and measurable kernel k on $\mathcal{X}$. Assume that $f^*_{\tau,\mathrm{P}}$ exists and is P^X-a.s. unique, $f^*_{\tau,\mathrm{P}} \in L_1(\mathrm{P}^X)$ and P additionally satisfies at least one of the additional conditions (i) and (ii) from Theorem 3.2.10. If the sequence $(\lambda_n)_{n\in\mathbb{N}}$ satisfies $\lambda_n > 0$ for all $n \in \mathbb{N}$ as well as $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$ for $n \to \infty$, then*

$$\lim_{n\to\infty} ||f_{L^\star_{\tau\text{-}pin},\mathrm{D}_n,\lambda_n} - f^*_{\tau,\mathrm{P}}||_{L_1(\mathrm{P}^X)} = 0 \qquad \textit{in probability } \mathrm{P}^\infty.$$

Proof. Christmann et al. (2009, Theorem 8) yields

$$\lim_{n\to\infty} \mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star_{\tau\text{-pin}},\mathrm{D}_n,\lambda_n}) = \mathcal{R}_{L^\star,\mathrm{P}}(f^*_{\tau,\mathrm{P}})$$

in probability P^∞. The assertion follows directly from Theorem 3.2.10. □

It would now be possible to use Corollary 3.3.10 to derive a result on risk consistency of SVMs which are based on the shifted pinball loss, similarly to what we did in the non-shifted case in Section 3.3.2, where we used Theorem 3.3.2 on L_p-consistency to derive Corollary 3.3.5 on risk consistency. In the latter result, we however only achieved an actual improvement (over already existing results) regarding the conditions on the regularization parameters if the loss function is of growth type $p > 1$. Similarly, a result on risk consistency which is based on Corollary 3.3.10 would offer no benefit over Theorem 8 from Christmann et al. (2009) because of the pinball loss being of growth type 1.

3.3.5 H-Consistency Using Shifted Loss Functions

In contrast to the results on L_p-consistency, it is indeed possible to transfer Proposition 3.3.7 on H-consistency of SVMs based on Lipschitz continuous losses to the shifted case and eliminate the condition $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ (which corresponds to the moment condition $|\mathrm{P}|_1 < \infty$ if the loss is additionally distance-based, cf. Remarks 2.1.18 and 2.1.22(i)) without needing to add a different condition regarding P instead.

Proposition 3.3.11. *Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a Lipschitz continuous and convex loss function, and let $L^\star$ be its shifted version. Let H be the separable RKHS of a bounded and measurable kernel k on $\mathcal{X}$. Assume that $f^*_{L^\star,\mathrm{P}}$ is P^X-a.s. unique. If $f^*_{L^\star,\mathrm{P}} \in H$ and the sequence $(\lambda_n)_{n\in\mathbb{N}}$ satisfies $\lambda_n > 0$ for all $n \in \mathbb{N}$ as well as $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$ for $n \to \infty$, then*

$$\lim_{n\to\infty} \left|\left| f_{L^\star,\mathrm{D}_n,\lambda_n,k} - f^*_{L^\star,\mathrm{P}} \right|\right|_H = 0 \qquad \textit{in probability } \mathrm{P}^\infty.$$

The proof needs the subsequent auxiliary result, which transfers Steinwart and Christmann (2008, Corollary 5.19) to shifted loss functions. In this, the definition of SVMs from Definition 2.1.5 is extended analogously to the cases $\lambda = 0$ and $\lambda = \infty$.

Lemma 3.3.12. *Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex and Lipschitz continuous loss function, and let $L^\star$ be its shifted version. Let H be the RKHS of a bounded and measurable kernel on $\mathcal{X}$. Define $\mathcal{R}^*_{L^\star,\mathrm{P},H} := \inf_{f\in H} \mathcal{R}_{L,\mathrm{P}}(f)$. Then, the following statements hold true:*

(i) If $\mathcal{R}^*_{L^\star,\mathrm{P},H} > -\infty$, $\lambda \mapsto f_{L^\star,\mathrm{P},\lambda,k}$ is a continuous mapping from $(0,\infty]$ to H.

(ii) If $\mathcal{R}^*_{L^\star,\mathrm{P},H} > -\infty$, $\lambda \mapsto \mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{P},\lambda,k})$ is a continuous mapping from $(0,\infty]$ to $[0,\infty)$.

(iii) If there exists an $f^* \in H$ minimizing $\mathcal{R}_{L^\star,\mathrm{P}}$ in H, then $\mathcal{R}^*_{L^\star,\mathrm{P},H} > -\infty$ and the mappings from (i) and (ii) are also defined and continuous at 0.

Proof. We omit stating the proof in full detail, as the result just transfers Steinwart and Christmann (2008, Corollary 5.19) from regular to shifted loss functions, which does not change the proof apart from some minor details. The main change is that the condition that L has to be a P-integrable Nemitski loss is dropped. This is possible because this condition was only necessary for the risk functional to be continuous and we are only examining Lipschitz continuous losses, for which the continuity of the risk functional follows from Christmann et al. (2009, Lemma 29). □

Proof of Proposition 3.3.11. By applying the triangle inequality, we obtain

$$\left|\left|f_{L^\star,\mathrm{D}_n,\lambda_n,k} - f^*_{L^\star,\mathrm{P}}\right|\right|_H \leq ||f_{L^\star,\mathrm{D}_n,\lambda_n,k} - f_{L^\star,\mathrm{P},\lambda_n,k}||_H + \left|\left|f_{L^\star,\mathrm{P},\lambda_n,k} - f^*_{L^\star,\mathrm{P}}\right|\right|_H . \tag{3.31}$$

The first summand on the right hand side converging to 0 in probability can be shown exactly the same way as in the proof of Proposition 3.3.7 by just replacing Steinwart and Christmann (2008, Corollary 5.10) with Christmann et al. (2009, Theorem 7).

As for the second summand, we obtain from Lemma 3.3.12 (since $f^*_{L^\star,\mathrm{P}} \in H$) that $\lambda \mapsto f_{L^\star,\mathrm{P},\lambda,k}$ defines a continuous mapping from $[0,\infty]$ to H. By noting that $f^*_{L,\mathrm{P}} \in H$ results in $f^*_{L^\star,\mathrm{P}} = f_{L^\star,\mathrm{P},0}$, this continuity yields, for all $\varepsilon > 0$,

$$\left|\left|f_{L^\star,\mathrm{P},\lambda_n,k} - f^*_{L^\star,\mathrm{P}}\right|\right|_H = ||f_{L^\star,\mathrm{P},\lambda_n,k} - f_{L^\star,\mathrm{P},0}||_H \leq \varepsilon$$

for n sufficiently large, which then yields the assertion. □

Recall that H-consistency implies L_p-consistency for all $p \in [1,\infty]$ by Corollary 3.2.4. Hence, the preceding result—similarly to Corollary 3.3.10, but for arbitrary distance-based losses of growth type 1 instead of only the pinball loss—yields an alternative (albeit not verifiable) condition to replace the moment condition $|\mathrm{P}|_1 < \infty$ in Theorem 3.3.2 and still obtain L_p-consistency, namely $f^*_{L^\star,\mathrm{P}} \in H$.

Since $0 \in H$ for any RKHS H and the Sobolev space $W^{2,2}(\mathcal{X})$ is separable (cf. Adams and Fournier, 2003, Theorem 3.6), Proposition 3.3.11 is actually applicable to the situation of the counterexample used to prove Corollary 3.3.8 (see also proof of Corollary 3.2.9). Thus, this is an example for which L_1-consistency does not directly follow from risk consistency, but SVMs are L_1-consistent nonetheless.

3.4 Consistency of Localized Support Vector Machines

In this section, consistency results from Section 3.3 are transferred to localized SVMs, showing that they are L_p- and risk consistent under analogous assumptions. This is done for L_p-consistency in Section 3.4.2 and for risk consistency in Section 3.4.3. Notably, the regionalizations underlying the different localized SVMs are allowed to change with n in all results. Before stating the results, some additional definitions and assumptions are stated in Section 3.4.1.

Remark 3.4.1. Some of the cases that were investigated for global SVMs are not explicitly examined for localized SVMs, which is due to different reasons:

- In contrast to global SVMs, there is generally not an obvious associated RKHS which contains a specific localized SVM and which therefore suggests itself to be used for H-consistency.
- For shifted loss functions, proving L_p-consistency without the moment condition would yield similar problems as in the non-localized case. Risk consistency on the other hand has already been derived by Dumpert and Christmann (2018), even if only for regionalizations that do not change as the size of the data set D_n increases.

This section is for the most part taken from the peer-reviewed paper Köhler (2024a, Sections 3.2 and 4) that was published in *Neurocomputing.*

3.4.1 Prerequisites

As consistency of localized SVMs is investigated in this section and as the regionalization is allowed to change with n, localized SVMs $f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}$, $n \in \mathbb{N}$, are required. For $n \in \mathbb{N}$, the regionalization $\boldsymbol{\mathcal{X}_n} := \{\mathcal{X}_{n,1}, \ldots, \mathcal{X}_{n,A_n}\}$ is assumed to be of size $A_n \in \mathbb{N}$. For $n \in \mathbb{N}$ and $a \in \{1, \ldots, A_n\}$, denote by $\mathrm{P}_{n,a} := \mathrm{P}_{\mathcal{X}_{n,a}}$ and $\mathrm{D}_{n,a} := (\mathrm{D}_n)_{\mathcal{X}_{n,a}}$ (as the empirical measure associated to $D_{n,a} := (D_n)_{\mathcal{X}_{n,a}}$) local measures on $\mathcal{X}_{n,a}$ as defined in eq. (2.6), and $d_{n,a} := |D_{n,a}|$. Additionally, denote

$$I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}} := \{a \in \{1, \ldots, A_n\} \,|\, \mathrm{P}(\mathcal{X}_{n,a} \times \mathcal{Y}) > 0\}$$

and $\tilde{A}_n := |I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}|$ for all $n \in \mathbb{N}$.

Further denoting $\boldsymbol{\mathcal{X}_n}(x) := \{\tilde{\mathcal{X}} \in \boldsymbol{\mathcal{X}_n} \,|\, x \in \tilde{\mathcal{X}}\}$ for all $x \in \mathcal{X}$ and $n \in \mathbb{N}$, the regionalizations are assumed to satisfy the following three conditions:

(R1) $\mathcal{X}_{n,a}$ complete (as a metric space) for all $n \in \mathbb{N}$ and $a \in \{1, \ldots, A_n\}$.

(R2) $\exists\, s_{\max} \in \mathbb{N}$ such that $|\boldsymbol{\mathcal{X}_n}(x)| \leq s_{\max}$ for all $x \in \mathcal{X}$ and $n \in \mathbb{N}$.

(R3) The sequence $(\boldsymbol{\mathcal{X}_n})_{n\in\mathbb{N}}$ is stochastically independent of the sequence $(D_n)_{n\in\mathbb{N}}$ of training data sets.

Condition **(R3)** might seem restrictive at first glance because it seemingly constitutes a restriction to only using regionalizations whose construction does not take the observed data into account. However, one can easily circumvent this restriction by randomly partitioning the whole data set into not only the usual three parts—namely a training data set D_n, a validation data set and a test data set—but four parts instead, where the fourth part is a regionalization data set. This way, the regionalizations can be chosen data-dependently without violating **(R3)**. By putting only a relatively small part of the available data into the regionalization data set—which can be sufficient because one reason for regionalizing is to just reduce the subsequent training time of the SVMs, for which no "perfect" regionalization is necessary—, this procedure does not substantially reduce the amount of data available for training, validating and testing.

Further note that, for every $n \in \mathbb{N}$, the regions need not necessarily be pairwise disjoint but can instead also overlap—as long as **(R2)** is satisfied, that is, as long as the number of regions overlapping does not exceed some global constant $s_{\max}$ in any point $x \in \mathcal{X}$. If the regionalization does not change with n, then **(R2)** is trivially satisfied for $s_{\max} = A_1$.

Remark 3.4.2. By Dunford and Schwartz (1957, Lemma I.6.4 and Theorem I.6.12), any subset of a separable metric space is a separable metric space again if it is equipped with the metric of the original space. Hence, Assumption 3.0.1 being satisfied for $\mathcal{X}$ implies it also being satisfied for the regions $\mathcal{X}_{n,a}$, $n \in \mathbb{N}$ and $a \in \{1, \ldots, A_n\}$.

In the consistency results, we further need the concept of families of kernels of type $\boldsymbol{\beta}$, which is introduced in the following definition.

Definition 3.4.3 (Family of Kernels of Type $\boldsymbol{\beta}$)**.** Let I be an index set such that $0 \in I$. For kernels $k^{(r)}$ and constants $\beta^{(r)} \in (0, \infty)$, $r \in I$, we say that $\boldsymbol{k} := (k^{(r)})_{r\in I}$ is a family of kernels of type $\boldsymbol{\beta} := (\beta^{(r)})_{r\in I}$ if, for all $r \in I$,

(i) $H^{(r)} \supseteq H^{(0)}$, where $H^{(r)}$ and $H^{(0)}$ are the RKHSs associated with $k^{(r)}$ and $k^{(0)}$ respectively, and

(ii) $||f||_{H^{(r)}} \leq \beta^{(r)} \cdot ||f||_{H^{(0)}}$ for all $f \in H^{(0)}$.

Remark 3.4.4. By Saitoh and Sawano (2016, Theorem 2.17) (see also Aronszajn, 1950, Part I.7, and Berlinet and Thomas-Agnan, 2004, Section 4.5, for related considerations), condition (i) from Definition 3.4.3 already implies that there exists some $\beta^{(r)} \in (0, \infty)$ such that (ii) is satisfied as well. Hence, every family of kernels satisfying (i) will also be a family of kernels of type $\boldsymbol{\beta}$ for suitable $\boldsymbol{\beta}$. Furthermore, the same theorem also yields that the two conditions from Definition 3.4.3 are equivalent to

(iii) $(\beta^{(r)})^2 \cdot k^{(r)} - k^{(0)}$ is a kernel,

for which reason families of kernels of type $\boldsymbol{\beta}$ are equivalently characterized by (iii) holding true for all $r \in I$.

Example 3.4.5. Let $d \in \mathbb{N}$, $\mathcal{X} \subseteq \mathbb{R}^d$ non-empty and I be an index set such that $0 \in I$. For $r \in I$, define $k^{(r)}$ as the Gaussian RBF kernel on $\mathcal{X}$ with bandwidth $\gamma^{(r)} \in (0, \infty)$, cf. Example 2.1.12. By Steinwart and Christmann (2008, Proposition 4.46), the conditions from Definition 3.4.3 are satisfied with $\beta^{(r)} := (\gamma^{(0)}/\gamma^{(r)})^{d/2}$ if $\gamma^{(0)} \geq \sup_{r \in I\setminus\{0\}} \gamma^{(r)}$.

Hence, every family $(k^{(r)})_{r\in J}$, $0 \notin J$, of Gaussian RBF kernels with bounded bandwidth can be turned into a family of kernels of type $\boldsymbol{\beta} = ((\gamma^{(0)}/\gamma^{(r)})^{d/2})_{r \in I}$, $I := J \cup \{0\}$, by choosing $k^{(0)}$ as the Gaussian RBF kernel with bandwidth $\gamma^{(0)} = \sup_{r\in J} \gamma^{(r)}$.

We introduced these families of kernels of type $\boldsymbol{\beta}$ since all kernels $k_{n,a}$, $n \in \mathbb{N}$, $a \in \{1, \ldots, A_n\}$, used in the local SVMs will be required to come from the union of $\ell \in \mathbb{N}$ such families $\boldsymbol{k^{(1)}}, \ldots, \boldsymbol{k^{(\ell)}}$. To be more specific, $\boldsymbol{k^{(j)}}$, $j = 1, \ldots, \ell$, will consist of kernels on $\mathcal{X}$ and each $k_{n,a}$ will be the restriction of such a kernel to $\mathcal{X}_{n,a} \times \mathcal{X}_{n,a}$. That is, we will have $k_{n,a} = k^{(j_0,r_0)}|_{\mathcal{X}_{n,a}\times\mathcal{X}_{n,a}}$ for some $j_0 \in \{1, \ldots, \ell\}$ and $r_0 \in I^{(j_0)}$, where $I^{(j_0)}$ denotes the index set of the j_0-th family. Based on this, we introduce the additional notation $\beta_{n,a} := \beta^{(j_0,r_0)}$ and $k_{n,a}^{(0)} := k^{(j_0,0)}|_{\mathcal{X}_{n,a}\times\mathcal{X}_{n,a}}$ (in case of ambiguity regarding j_0 and r_0, any of the options may be chosen), which will be needed later on.

Note that the concept of families of kernels of type $\boldsymbol{\beta}$ also allows for infinite index sets (see also Example 3.4.5). This will lead to the kernels $k_{n,a}$, $n \in \mathbb{N}$, $a \in \{1, \ldots, A_n\}$, being allowed to be chosen from a possibly infinite set of kernels.

Using this, the following assumptions are needed in the main results from the subsequent sections, for which reason they are stated here in order to be able to shorten the formulation of the results themselves:

Assumption 3.4.6.

- Let $L \colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex, distance-based loss function of growth type $p \in [1, \infty)$.
- Let, for $n \in \mathbb{N}$, $\boldsymbol{\mathcal{X}_n} := \{\mathcal{X}_{n,1}, \ldots, \mathcal{X}_{n,A_n}\}$ be a regionalization of $\mathcal{X}$ of size $A_n \in \mathbb{N}$ such that the regionalizations satisfy **(R1)**, **(R2)**, **(R3)**, and let the weight functions $w_{n,a}$, $n \in \mathbb{N}$ and $a \in \{1, \ldots, A_n\}$, satisfy **(W1)**, **(W2)**, **(W3)**.

- Let P be a probability measure on $\mathcal{X}\times\mathcal{Y}$ that satisfies $|\mathrm{P}|_p < \infty$ as well as $\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} |\mathrm{P}_{n,a}|_p < \infty$.
- For $n \in \mathbb{N}$, let $\boldsymbol{\lambda_n} \coloneqq (\lambda_{n,1},\dots,\lambda_{n,A_n}) \in (0,\infty)^{A_n}$.
- Let $\ell \in \mathbb{N}$ and let, for $j = 1,\dots,\ell$, $\boldsymbol{k^{(j)}} \coloneqq (k^{(j,r)})_{r\in I^{(j)}}$ be a family of uniformly bounded and measurable kernels of type $\boldsymbol{\beta^{(j)}} \coloneqq (\beta^{(j,r)})_{r\in I^{(j)}}$ on $\mathcal{X}$ with separable RKHSs $(H^{(j,r)})_{r\in I^{(j)}}$ such that $H^{(j,0)} \subseteq L_p(\mathrm{P}^X)$ dense. For all $n \in \mathbb{N}$, let $\boldsymbol{k_n} \coloneqq (k_{n,1},\dots,k_{n,A_n})$ such that for all $a \in \{1,\dots,A_n\}$
$$k_{n,a} \in \left\{k^{(j,r)}|_{\mathcal{X}_{n,a}\times\mathcal{X}_{n,a}} \,:\, j \in \{1,\dots,\ell\}, r \in I^{(j)}\right\} .$$

These relatively weak assumptions can for the most part be ensured to hold true as they only concern entities that can be chosen by the person computing the localized SVMs.[15] Notably, the conditions that—directly or indirectly—concern the regionalization and therefore also the way this regionalization is obtained, can be ensured for most of the methods mentioned in Section 2.2.1, with the exception of some of the kNN methods, which do not satisfy **(R2)**.

Remark 3.4.7. The condition $\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} |\mathrm{P}_{n,a}|_p < \infty$ is disadvantageous in that it requires knowledge about all regionalizations $\boldsymbol{\mathcal{X}_n}$, $n \in \mathbb{N}$. Because
$$|\mathrm{P}_{n,a}|_p^p = \int_{\mathcal{X}_{n,a}} |\mathrm{P}(\cdot\,|\,x)|_p^p \,\mathrm{d}\mathrm{P}_{n,a}^X(x) \le \sup_{x\in\mathcal{X}} |\mathrm{P}(\cdot\,|\,x)|_p^p$$
for all $n \in \mathbb{N}$ and $a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}$ (and analogously also $|\mathrm{P}|_p^p \le \sup_{x\in\mathcal{X}} |\mathrm{P}(\cdot\,|\,x)|_p^p$), it however suffices if $\sup_{x\in\mathcal{X}} |\mathrm{P}(\cdot\,|\,x)|_p < \infty$.

On the other hand, even though the finiteness of $|\mathrm{P}|_p$ does already imply the finiteness of $|\mathrm{P}_{n,a}|_p$ for all $n \in \mathbb{N}$ and $a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}$ because
$$\begin{aligned}|\mathrm{P}_{n,a}|_p^p &= \int_{\mathcal{X}_{n,a}} |\mathrm{P}(\cdot\,|\,x)|_p^p \,\mathrm{d}\mathrm{P}_{n,a}^X(x) = \frac{1}{\mathrm{P}^X(\mathcal{X}_{n,a})} \cdot \int_{\mathcal{X}_{n,a}} |\mathrm{P}(\cdot\,|\,x)|_p^p \,\mathrm{d}\mathrm{P}^X(x)\\ &\le \frac{1}{\mathrm{P}^X(\mathcal{X}_{n,a})} \cdot \int_{\mathcal{X}} |\mathrm{P}(\cdot\,|\,x)|_p^p \,\mathrm{d}\mathrm{P}^X(x) = \frac{1}{\mathrm{P}^X(\mathcal{X}_{n,a})} \cdot |\mathrm{P}|_p^p\,,\end{aligned}$$
$|\mathrm{P}|_p$ being finite is not sufficient to guarantee $\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} |\mathrm{P}_{n,a}|_p < \infty$, as can be seen from the following example:

Let $\mathrm{P}^X \coloneqq \mathcal{U}(0,1)$ and $\mathrm{P}(\cdot\,|\,X = x) \coloneqq \mathcal{U}(0, x^{-1/2})$ for all $x \in (0,1)$, where $\mathcal{U}(a,b)$ denotes the uniform distribution on (a,b). Then, we have
$$|\mathrm{P}|_1 = \int_0^1 \int_0^{\frac{1}{\sqrt{x}}} y\sqrt{x} \,\mathrm{d}y \,\mathrm{d}x = 1 < \infty\,,$$
but for $\mathcal{X}_{n,1} \coloneqq (0, \frac{1}{n})$, $n \in \mathbb{N}$, we obtain
$$|\mathrm{P}_{n,1}|_1 = \int_0^{\frac{1}{n}} \int_0^{\frac{1}{\sqrt{x}}} y\sqrt{x} \,\mathrm{d}y \cdot n \,\mathrm{d}x = \sqrt{n}\,,$$

[15] The most notable exception to this is the moment condition, which is however necessary for the existence of functions with finite risks in the respective RKHSs and for even having a P^X-a.s. unique Bayes function lying in $L_p(\mathrm{P}^X)$, cf. Remark 2.1.22.

which yields $\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} |\mathrm{P}_{n,a}|_p = \infty$.

Hence, the condition $\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} |\mathrm{P}_{n,a}|_p < \infty$ is not superfluous in itself and can not just be erased without adding a replacement like $\sup_{x\in\mathcal{X}} |\mathrm{P}(\cdot\,|\,x)|_p < \infty$.

3.4.2 L_p-Consistency Using Regular Loss Functions

The subsequent theorem shows that localized SVMs are indeed L_p-consistent under Assumption 3.4.6.

Theorem 3.4.8. *Let Assumption 3.4.6 be satisfied. Assume that $f^*_{L,\mathrm{P}}$ exists and is P^X-a.s. unique. Define $p^*_1 := \max\{p+1, p(p+1)/2\}$. Further choose $p^*_2 := \max\{2(p-1)/p, p-1\}$ if $p > 1$ and $p^*_2 \in (0,\infty)$ arbitrary if $p = 1$. If the regularization parameters satisfy $\lambda_{n,a} \in (0, C)$ for all $n \in \mathbb{N}$ and $a \in \{1,\dots,A_n\}$ for some $C \in (0,\infty)$, as well as*

$$\max_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \beta^2_{n,a}\lambda_{n,a} \to 0 \tag{3.32}$$

and

$$\min_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \frac{\lambda^{p^*_1}_{n,a} d_{n,a}}{\tilde{A}^{p^*_2}_n} \to \infty \tag{3.33}$$

as $n \to \infty$, then

$$\lim_{n\to\infty} \left|\left| f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f^*_{L,\mathrm{P}} \right|\right|_{L_p(\mathrm{P}^X)} = 0 \qquad \text{in probability } \mathrm{P}^\infty.$$

Remark 3.4.9. The conditions (3.32) and (3.33) are closely connected to the ones from the non-localized case (see Theorem 3.3.2), requiring that the regularization parameters tend to 0, but not too fast. (3.32) however additionally needs to take changes in the kernels into account by including $\beta_{n,a}$, and (3.33) additionally states that the number of regions must not grow too fast.

In some special cases, we can slightly simplify these two conditions:

If one only allows for a finite amount of kernels to choose from (instead of a finite amount of families of kernels of type $\boldsymbol{\beta}$), it is obviously possible to view each of these kernels as its own family of kernels with index set $I^{(j)} = \{0\}$ and $\beta^{(j,0)} = 1$ for all $j \in \{1,\dots,\ell\}$, and thus simplify (3.32) by eliminating $\beta_{n,a}$ from it.

Additionally, if the regionalization $\boldsymbol{\mathcal{X}_n}$ does not change with n, then $\tilde{A}_n$ is constant and we can erase it from (3.33).

Hence, if both of these hold true (finite amount of kernels and constant regionalization), the conditions regarding the regularization parameters are indeed exactly the same as in Theorem 3.3.2 on L_p-consistency of non-localized SVMs, with the only difference being that the conditions obviously need to hold true for each region now instead of only globally.

For proving Theorem 3.4.8, we need to investigate the difference between $f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}$ and $f^*_{L,\mathrm{P}}$. Analogously to the non-localized case, we do this by plugging in the theoretical localized SVM $f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}$ as an intermediate step. The two subsequent auxiliary results examine the difference between $f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}$ and $f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}$ and that between $f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}$ and $f^*_{L,\mathrm{P}}$ respectively.

As the assumptions needed for these lemmas are slightly weaker than those needed in the theorems from this section, Assumption 3.4.6 is *not* assumed to hold true in these lemmas, but we will instead explicitly list the required assumptions in the lemmas.

Lemma 3.4.10. *Let $L\colon \mathcal{Y}\times\mathbb{R}\to[0,\infty)$ be a convex, distance-based loss function of upper growth type $p\in[1,\infty)$. Let the regionalizations $\boldsymbol{\mathcal{X}_n}$, $n\in\mathbb{N}$, satisfy* ***(R1)**, **(R3)**, and let the weight functions $w_{n,a}$, $n\in\mathbb{N}$ and $a\in\{1,\dots,A_n\}$, satisfy **(W1)**, **(W2)**, **(W3)**. Assume $\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}}|\mathrm{P}_{n,a}|_p<\infty$. Let, for all $n\in\mathbb{N}$ and $a\in\{1,\dots,A_n\}$, $k_{n,a}$ be a bounded and measurable kernel on $\mathcal{X}_{n,a}$ with separable RKHS $H_{n,a}$, such that $\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}}||k_{n,a}||_\infty<\infty$. Define $p_1^*:=\max\{p+1,p(p+1)/2\}$. Further choose $p_2^*:=\max\{2(p-1)/p,p-1\}$ if $p>1$ and $p_2^*\in(0,\infty)$ arbitrary if $p=1$. If the regularization parameters satisfy $\lambda_{n,a}\in(0,C)$ for all $n\in\mathbb{N}$ and $a\in\{1,\dots,m_n\}$ for some $C\in(0,\infty)$, as well as*

$$\min_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}}\frac{\lambda_{n,a}^{p_1^*}d_{n,a}}{\tilde{A}_n^{p_2^*}}\to\infty \tag{3.34}$$

as $n\to\infty$, then

$$\lim_{n\to\infty}||f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}-f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}||_{L_\infty(\mathrm{P}^X)}=0\qquad\textit{in probability }\mathrm{P}^\infty.$$

Proof. To shorten the notation, denote $f_{\mathrm{P},n,a}:=f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}$ and $f_{\mathrm{D}_n,n,a}:=f_{L,\mathrm{D}_{n,a},\lambda_{n,a},k_{n,a}}$ for all $n\in\mathbb{N}$ and $a\in\{1,\dots,A_n\}$, as well as $\kappa:=\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}}||k_{n,a}||_\infty$ and $\rho:=\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}}|\mathrm{P}_{n,a}|_p$ throughout this proof.

Because applying **(W1)** and **(W2)** yields

$$\begin{aligned}&|f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}(x)-f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}(x)|=\left|\sum_{a=1}^{A_n}w_{n,a}(x)\cdot\left(\hat{f}_{\mathrm{D}_n,n,a}(x)-\hat{f}_{\mathrm{P},n,a}(x)\right)\right|\\&\le\sum_{a=1}^{A_n}w_{n,a}(x)\cdot\left|\hat{f}_{\mathrm{D}_n,n,a}(x)-\hat{f}_{\mathrm{P},n,a}(x)\right|\le\max_{a\in\{1,\dots,A_n\}}\left|\hat{f}_{\mathrm{D}_n,n,a}(x)-\hat{f}_{\mathrm{P},n,a}(x)\right|\end{aligned}$$

for all $n\in\mathbb{N}$ and all $x\in\mathcal{X}$, we obtain

$$\begin{aligned}&||f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}-f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}||_{L_\infty(\mathrm{P}^X)}\le\max_{a\in\{1,\dots,A_n\}}\left|\left|\hat{f}_{\mathrm{D}_n,n,a}-\hat{f}_{\mathrm{P},n,a}\right|\right|_{L_\infty(\mathrm{P}^X)}\\&=\max_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}}||f_{\mathrm{D}_n,n,a}-f_{\mathrm{P},n,a}||_{L_\infty(\mathrm{P}^X_{n,a})}\le\kappa\cdot\max_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}}||f_{\mathrm{D}_n,n,a}-f_{\mathrm{P},n,a}||_{H_{n,a}}\end{aligned}\tag{3.35}$$

for all $n\in\mathbb{N}$, with the last inequality holding true because of Lemma 2.1.10(i). Hence, we start by fixing an $n\in\mathbb{N}$ and an $a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}$ and investigating the corresponding difference on the right hand side of (3.35).

First, note that employing Lemma 2.1.10(i), Steinwart and Christmann (2008, eq. (5.4)) and Lemma 2.1.21(ii) yields

$$||f_{\mathrm{P},n,a}||_\infty\le||k_{n,a}||_\infty\cdot||f_{\mathrm{P},n,a}||_{H_{n,a}}\le||k_{n,a}||_\infty\cdot\mathcal{R}_{\mathrm{P}_{n,a}}(0)^{1/2}\cdot\lambda_{n,a}^{-1/2}\le c_{p,L,\rho,\kappa}\cdot\lambda_{n,a}^{-1/2}\tag{3.36}$$

with $c_{p,L,\rho,\kappa}\in(0,\infty)$ denoting a constant depending only on p, L, ρ and κ, but not on $\lambda_{n,a}$.

Assume now without loss of generality that $d_{n,a} > 0$ (which by (3.34) has to be satisfied for n sufficiently large), i.e. that $f_{\mathrm{D}_n,n,a}$ is indeed an empirical SVM and not just defined as the zero function. We know from Steinwart and Christmann (2008, Corollary 5.11) that there exists a function $h_{n,a} \colon \mathcal{X}_{n,a} \times \mathcal{Y} \to \mathbb{R}$ such that

$$||f_{\mathrm{D}_n,n,a} - f_{\mathrm{P},n,a}||_{H_{n,a}} \leq \frac{1}{\lambda_{n,a}} \cdot \left|\left|\mathbb{E}_{\mathrm{D}_{n,a}}\left[h_{n,a}\Phi_{n,a}\right] - \mathbb{E}_{\mathrm{P}_{n,a}}\left[h_{n,a}\Phi_{n,a}\right]\right|\right|_{H_{n,a}} \tag{3.37}$$

and, for $s := p/(p-1)$,

$$\begin{aligned} ||h_{n,a}||_{L_s(\mathrm{P}_{n,a})} &\leq 8^p \cdot c_L \cdot \left(1 + |\mathrm{P}_{n,a}|_p^{p-1} + ||f_{\mathrm{P},n,a}||_\infty^{p-1}\right) \\ &\leq 8^p \cdot c_L \cdot \left(1 + \rho^{p-1} + c_{p,L,\rho,\kappa}^{p-1} \cdot \lambda_{n,a}^{-(p-1)/2}\right) \\ &\leq \tilde{c}_{p,L,\rho,\kappa} \cdot \lambda_{n,a}^{-(p-1)/2}, \end{aligned} \tag{3.38}$$

where we employed (3.36) in the second and $\lambda_{n,a} \leq C$ in the third step, and where $c_L \in (0,\infty)$ and $\tilde{c}_{p,L,\rho,\kappa} \in (0,\infty)$ denote constants depending only on L respectively p, L, ρ and κ.

Assume without loss of generality that $p_2^* \leq 1$ if $p = 1$. Then, we can apply Steinwart and Christmann (2008, Lemma 9.2) with $q := p/(p-1)$ if $p > 1$ and $q := 2/p_2^*$ if $p = 1$, which leads to $q^* := \min\{1/2, 1 - 1/q\} = \min\{1/2, 1/p\} = (p+1)/(2p_1^*)$, to the function $h_{n,a}\Phi_{n,a}$: First of all, with the help of (3.38) we obtain

$$\begin{aligned} ||h_{n,a}\Phi_{n,a}||_q &:= \left(\mathbb{E}_{\mathrm{P}_{n,a}}\left[||h_{n,a}\Phi_{n,a}||_{H_{n,a}}^q\right]\right)^{1/q} \\ &\leq ||k_{n,a}||_\infty \cdot ||h_{n,a}||_{L_q(\mathrm{P}_{n,a})} \leq \kappa \cdot \tilde{c}_{p,L,\rho,\kappa} \cdot \lambda_{n,a}^{-(p-1)/2} < \infty\,, \end{aligned}$$

where we employed that, for all $(x,y) \in \mathcal{X}_{n,a} \times \mathcal{Y}$,

$$\begin{aligned} ||h_{n,a}(x,y)\Phi_{n,a}(x)||_{H_{n,a}}^q &= |h_{n,a}(x,y)|^q \cdot ||\Phi_{n,a}(x)||_{H_{n,a}}^q \\ &= |h_{n,a}(x,y)|^q \cdot k_{n,a}(x,x)^{q/2} \leq |h_{n,a}(x,y)|^q \, ||k_{n,a}||_\infty^q \end{aligned}$$

by the reproducing property. Hence, we obtain for all $\varepsilon > 0$, by combining this Lemma 9.2 with (3.37),

$$\begin{aligned} &\mathrm{P}_{n,a}^{d_{n,a}}\left(D_{n,a} \in (\mathcal{X}_{n,a} \times \mathcal{Y})^{d_{n,a}} : ||f_{\mathrm{D}_n,n,a} - f_{\mathrm{P},n,a}||_{H_{n,a}} \geq \frac{\varepsilon}{\kappa}\right) \\ &\leq \mathrm{P}_{n,a}^{d_{n,a}}\left(D_{n,a} \in (\mathcal{X}_{n,a} \times \mathcal{Y})^{d_{n,a}} : \left|\left|\mathbb{E}_{\mathrm{D}_{n,a}}\left[h_{n,a}\Phi n, a\right] - \mathbb{E}_{\mathrm{P}_{n,a}}\left[h_{n,a}\Phi_{n,a}\right]\right|\right|_{H_{n,a}} \geq \frac{\lambda_{n,a}\varepsilon}{\kappa}\right) \\ &\leq c_q \cdot \left(\frac{\kappa\, ||h_{n,a}\Phi_{n,a}||_q}{\lambda_{n,a}\varepsilon d_{n,a}^{q^*}}\right)^q \leq c_{q,p,L,\rho,\kappa} \cdot \left(\frac{1}{\lambda_{n,a}^{(p+1)/2}\varepsilon d_{n,a}^{q^*}}\right)^q \end{aligned}$$

with $c_q \in (0,\infty)$ and $c_{q,p,L,\mathrm{P},k} \in (0,\infty)$ denoting constants depending only on q (which means only on p in the case $p > 1$) respectively q, p, L, ρ and κ.

With this, we can now return to investigating the whole global predictors with the help of (3.35): For all $\varepsilon > 0$ and $n \in \mathbb{N}$, we have

$$\begin{aligned}
&\mathrm{P}^n \Big(D_n \in (\mathcal{X} \times \mathcal{Y})^n : ||f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}||_{L_\infty(\mathrm{P}^X)} \geq \varepsilon \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\Big| |D_{n,1}| = d_{n,1}, \ldots, |D_{n,A_n}| = d_{n,A_n}\Big) \\
&\leq \mathrm{P}^n \left(D_n \in (\mathcal{X} \times \mathcal{Y})^n : \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} ||f_{\mathrm{D}_n,n,a} - f_{\mathrm{P},n,a}||_{H_{n,a}} \geq \frac{\varepsilon}{\kappa}\right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\Big| |D_{n,1}| = d_{n,1}, \ldots, |D_{n,A_n}| = d_{n,A_n}\Big) \\
&\leq \sum_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \mathrm{P}_{n,a}^{d_{n,a}} \left(D_{n,a} \in (\mathcal{X}_{n,a} \times \mathcal{Y})^{d_{n,a}} : ||f_{\mathrm{D}_n,n,a} - f_{\mathrm{P},n,a}||_{H_{n,a}} \geq \frac{\varepsilon}{\kappa}\right) \\
&\leq c_{q,p,L,\rho,\kappa} \cdot \tilde{A}_n \cdot \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\frac{1}{\lambda_{n,a}^{(p+1)/2} \varepsilon d_{n,a}^{q^*}}\right)^q , \qquad (3.39)
\end{aligned}$$

and it remains to further investigate the right hand side:

If $p > 1$, we obtain $(qq^*)^{-1} = ((p-1)/p) \cdot \max\{2, p\} = p_2^*$. If $p = 1$, we analogously obtain $(qq^*)^{-1} = (p_2^*/2) \cdot 2 = p_2^*$. Thus, we have

$$\begin{aligned}
\tilde{A}_n \cdot \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\frac{1}{\lambda_{n,a}^{(p+1)/2} d_{n,a}^{q^*}}\right)^q &= \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\frac{\tilde{A}_n^{1/(qq^*)}}{\lambda_{n,a}^{(p+1)/(2q^*)} d_{n,a}}\right)^{qq^*} \\
&= \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\frac{\tilde{A}_n^{p_2^*}}{\lambda_{n,a}^{p_1^*} d_{n,a}}\right)^{qq^*} \qquad \to 0 , \qquad n \to \infty ,
\end{aligned}$$

by assumption. Hence, the whole right hand side of (3.39) converges to 0, which completes the proof. □

Lemma 3.4.11. *Let $L \colon \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex, distance-based loss function of upper growth type $p \in [1, \infty)$. Let the regionalizations $\boldsymbol{\mathcal{X}_n}$, $n \in \mathbb{N}$, satisfy **(R1)**, **(R2)**, and let the weight functions $w_{n,a}$, $n \in \mathbb{N}$ and $a \in \{1, \ldots, A_n\}$, satisfy **(W1)**, **(W2)**, **(W3)**. Assume that $|\mathrm{P}|_p < \infty$. Let $\ell \in \mathbb{N}$ and let, for $j = 1, \ldots, \ell$, $\boldsymbol{k}^{(j)} := (k^{(j,r)})_{r \in I^{(j)}}$ be a family of measurable kernels of type $\boldsymbol{\beta}^{(j)} := (\beta^{(j,r)})_{r \in I^{(j)}}$ on $\mathcal{X}$ with RKHSs $(H^{(j,r)})_{r \in I^{(j)}}$ such that $H^{(j,0)} \subseteq L_p(\mathrm{P}^X)$ dense. Let, for all $n \in \mathbb{N}$ and $a \in \{1, \ldots, A_n\}$,*

$$k_{n,a} \in \{k^{(j,r)}|_{\mathcal{X}_{n,a} \times \mathcal{X}_{n,a}} : j \in \{1, \ldots, \ell\}, r \in I^{(j)}\} .$$

If the regularization parameters satisfy $\lambda_{n,a} > 0$ for all $n \in \mathbb{N}$ and $a \in \{1, \ldots, A_n\}$ as well as $\max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \beta_{n,a}^2 \lambda_{n,a} \to 0$ as $n \to \infty$, then

$$\lim_{n \to \infty} \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) = \mathcal{R}_{L,\mathrm{P}}^* .$$

Proof. First, we show that all risks appearing in the assertion are finite: Lemma 2.1.21(ii) yields $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ as well as $\mathcal{R}_{L,\mathrm{P}_{n,a}}(0) < \infty$ for all $n \in \mathbb{N}$ and $a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}$ (with the latter holding true because $|\mathrm{P}_{n,a}|_p < \infty$ by Remark 3.4.7). Since $\mathcal{R}^*_{L,\mathrm{P}} \leq \mathcal{R}_{L,\mathrm{P}}(0)$ by definition, we obtain the finiteness of $\mathcal{R}^*_{L,\mathrm{P}}$. Furthermore,

$$\begin{aligned}
\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) &= \int_{\mathcal{X}\times\mathcal{Y}} L(y, f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}(x))\,\mathrm{dP}(x,y) \\
&\leq \int_{\mathcal{X}\times\mathcal{Y}} \sum_{a=1}^{A_n} w_{n,a}(x)\cdot L(y, \hat{f}_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}(x))\,\mathrm{dP}(x,y) \\
&\leq \sum_{a=1}^{A_n} \int_{\mathcal{X}_{n,a}\times\mathcal{Y}} L(y, f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}(x))\,\mathrm{dP}(x,y) \\
&= \sum_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \mathrm{P}(\mathcal{X}_{n,a}\times\mathcal{Y})\cdot\mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}})\,,
\end{aligned}$$

where we applied **(W1)**, **(W2)** and the convexity of L in the second and its non-negativity as well as **(W1)** and **(W3)** in the third step. In the last step, we employed that $\mathcal{X}_{n,a}\times\mathcal{Y}$ is a P-zero set for $a \notin I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}$, leading to the corresponding P-integrals being 0. Since $\mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}) \leq \mathcal{R}_{L,\mathrm{P}_{n,a}}(0)$ for all $a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}$ by the definition of $f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}$, and since we already saw that $\mathcal{R}_{L,\mathrm{P}_{n,a}}(0) < \infty$, the finiteness of $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}})$ follows for all $n \in \mathbb{N}$.

With the inner risk $\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}$ and the minimal inner risk $\mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x}$, we can now write

$$\begin{aligned}
&\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) - \mathcal{R}^*_{L,\mathrm{P}} \\
&= \int_{\mathcal{X}} \Big(\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x}\Big)\,\mathrm{dP}^X(x) \\
&\leq \int_{\mathcal{X}} \sum_{a=1}^{A_n} w_{n,a}(x)\cdot\Big(\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}(\hat{f}_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x}\Big)\,\mathrm{dP}^X(x) \\
&\leq \sum_{a=1}^{A_n} \int_{\mathcal{X}_{n,a}} \Big(\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}(f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x}\Big)\,\mathrm{dP}^X(x) \\
&= \sum_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\mathrm{P}(\mathcal{X}_{n,a}\times\mathcal{Y})\cdot\mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}) - \int_{\mathcal{X}_{n,a}} \mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x}\,\mathrm{dP}^X(x)\right)\,, \qquad (3.40)
\end{aligned}$$

where we applied Lemma 2.1.26 in the first, **(W1)**, **(W2)** and the convexity of L in the second, and **(W1)**, **(W3)** and $\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}(f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}) - \mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x} \geq 0$ for all $x \in \mathcal{X}$ (by the definition of $\mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x}$) in the third step. In the final step, we once more used that $\mathrm{P}(\mathcal{X}_{n,a}\times\mathcal{Y}) = 0$ for $a \notin I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}$.

If we define $\tilde{\lambda}_n := \max_{a \in I_{\boldsymbol{\mathcal{X}}_{\boldsymbol{n}},\mathrm{P}}} \beta_{n,a}^2 \lambda_{n,a}$ as well as $\tilde{k}_{n,a} \in \{k^{(j,r)} : j \in \{1,\dots,\ell\}, r \in I^{(j)}\}$ such that $\tilde{k}_{n,a}\big|_{\mathcal{X}_{n,a} \times \mathcal{X}_{n,a}} = k_{n,a}$ and analogously $\tilde{k}^{(0)}_{n,a} \in \{k^{(j,0)} : j \in \{1,\dots,\ell\}\}$ such that $\tilde{k}^{(0)}_{n,a}\big|_{\mathcal{X}_{n,a} \times \mathcal{X}_{n,a}} = k^{(0)}_{n,a}$, we can further analyze the right hand side of (3.40) by noting that, for all $n \in \mathbb{N}$ and $a \in I_{\boldsymbol{\mathcal{X}}_{\boldsymbol{n}},\mathrm{P}}$,

$$
\begin{aligned}
&\mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}) \\
&\quad \leq \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}) + \lambda_{n,a} \cdot \left|\left| f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}} \right|\right|^2_{H_{n,a}} \\
&\quad \leq \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P}_{n,a},\beta^2_{n,a}\lambda_{n,a},k^{(0)}_{n,a}}) + \lambda_{n,a} \cdot \left|\left| f_{L,\mathrm{P}_{n,a},\beta^2_{n,a}\lambda_{n,a},k^{(0)}_{n,a}} \right|\right|^2_{H_{n,a}} \\
&\quad \leq \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P}_{n,a},\beta^2_{n,a}\lambda_{n,a},k^{(0)}_{n,a}}) + \beta^2_{n,a} \cdot \lambda_{n,a} \cdot \left|\left| f_{L,\mathrm{P}_{n,a},\beta^2_{n,a}\lambda_{n,a},k^{(0)}_{n,a}} \right|\right|^2_{H^{(0)}_{n,a}} \\
&\quad \leq \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}}|_{\mathcal{X}_{n,a}}) + \beta^2_{n,a} \cdot \lambda_{n,a} \cdot \left|\left| f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}}|_{\mathcal{X}_{n,a}} \right|\right|^2_{H^{(0)}_{n,a}} \\
&\quad \leq \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}}|_{\mathcal{X}_{n,a}}) + \tilde{\lambda}_n \cdot \left|\left| f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}}|_{\mathcal{X}_{n,a}} \right|\right|^2_{H^{(0)}_{n,a}} \\
&\quad \leq \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}}|_{\mathcal{X}_{n,a}}) + \tilde{\lambda}_n \cdot \left|\left| f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}} \right|\right|^2_{\tilde{H}^{(0)}_{n,a}} .
\end{aligned}
$$

Here, we employed the definition of $f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}$ respectively $f_{L,\mathrm{P}_{n,a},\beta^2_{n,a}\lambda_{n,a},k^{(0)}_{n,a}}$ as the minimizers of the respective regularized risks (combined with the fact that $f_{L,\mathrm{P}_{n,a},\beta^2_{n,a}\lambda_{n,a},k^{(0)}_{n,a}} \in H^{(0)}_{n,a} \subseteq H_{n,a}$ and that $f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}}|_{\mathcal{X}_{n,a}} \in H^{(0)}_{n,a}$ by Berlinet and Thomas-Agnan, 2004, Theorem 6) in the second and in the fourth step, and again Berlinet and Thomas-Agnan (2004, Theorem 6) in the last step. Furthermore, the third step holds true because

$$
||f||_{H_{n,a}} = \min_{\substack{g \in \tilde{H}_{n,a}: \\ g|_{\mathcal{X}_{n,a}} = f}} ||g||_{\tilde{H}_{n,a}} \leq \min_{\substack{g \in \tilde{H}^{(0)}_{n,a}: \\ g|_{\mathcal{X}_{n,a}} = f}} ||g||_{\tilde{H}_{n,a}} \leq \beta_{n,a} \cdot \min_{\substack{g \in \tilde{H}^{(0)}_{n,a}: \\ g|_{\mathcal{X}_{n,a}} = f}} ||g||_{\tilde{H}^{(0)}_{n,a}} = \beta_{n,a} \cdot ||f||_{H^{(0)}_{n,a}}
$$

for all $f \in H^{(0)}_{n,a}$, where we once more applied Berlinet and Thomas-Agnan (2004, Theorem 6) and that $\tilde{H}^{(0)}_{n,a} \subseteq \tilde{H}_{n,a}$.

Plugging this into the right hand side of (3.40), we obtain

$$\begin{aligned}
&\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq \sum_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \Bigg(\mathrm{P}(\mathcal{X}_{n,a} \times \mathcal{Y}) \cdot \Bigg(\mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}} | \mathcal{X}_{n,a}) + \tilde{\lambda}_n \cdot \left\| f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}} \right\|^2_{\tilde{H}^{(0)}_{n,a}} \Bigg) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad - \int_{\mathcal{X}_{n,a}} \mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x} \, \mathrm{dP}^X(x) \Bigg) \\
&= \sum_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \Bigg(\mathrm{P}(\mathcal{X}_{n,a} \times \mathcal{Y}) \cdot \tilde{\lambda}_n \cdot \left\| f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}} \right\|^2_{\tilde{H}^{(0)}_{n,a}} \\
&\qquad\qquad\qquad + \int_{\mathcal{X}_{n,a}} \Big(\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}(f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,a}}(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x} \Big) \, \mathrm{dP}^X(x) \Bigg) \\
&\leq \sum_{j=1}^{\ell} \sum_{a=1}^{A_n} \Bigg(\mathrm{P}(\mathcal{X}_{n,a} \times \mathcal{Y}) \cdot \tilde{\lambda}_n \cdot \left\| f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}} \right\|^2_{H^{(j,0)}} \\
&\qquad\qquad\qquad + \int_{\mathcal{X}_{n,a}} \Big(\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x} \Big) \, \mathrm{dP}^X(x) \Bigg) \\
&\leq \sum_{j=1}^{\ell} s_{\max} \cdot \left(\tilde{\lambda}_n \cdot \left\| f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}} \right\|^2_{H^{(j,0)}} + \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}) - \mathcal{R}^*_{L,\mathrm{P}} \right) ,
\end{aligned} \tag{3.41}$$

with the third step holding true because of the summands being non-negative and the final step employing that, for all $j \in \{1, \dots, l\}$,

$$\begin{aligned}
&\sum_{a=1}^{A_n} \int_{\mathcal{X}_{n,a}} \Big(\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x} \Big) \, \mathrm{dP}^X(x) \\
&= \int_{\mathcal{X}} \sum_{a=1}^{A_n} \mathbb{1}_{\mathcal{X}_{n,a}}(x) \cdot \Big(\mathcal{C}_{L,\mathrm{P}(\cdot\,|\,x),x}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot\,|\,x),x} \Big) \, \mathrm{dP}^X(x) \\
&\leq s_{\max} \cdot \Big(\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}) - \mathcal{R}^*_{L,\mathrm{P}} \Big)
\end{aligned}$$

by **(R2)**, and analogously $\sum_{a=1}^{A_n} \mathrm{P}(\mathcal{X}_{n,a} \times \mathcal{Y}) \leq s_{\max}$.

Now, by Lemma 2.1.21(i), L is a P-integrable Nemitski loss of order p. Hence, for all $j \in \{1, \dots, l\}$, we know from Steinwart and Christmann (2008, Theorem 5.31) that

$$\mathcal{R}^*_{L,\mathrm{P},H^{(j,0)}} := \inf_{f \in H^{(j,0)}} \mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}^*_{L,\mathrm{P}} < \infty$$

and Steinwart and Christmann (2008, Lemma 5.15) then yields that

$$\lim_{n \to \infty} \tilde{\lambda}_n \left\| f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}} \right\|^2_{H^{(j,0)}} + \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}) - \mathcal{R}^*_{L,\mathrm{P}} = 0$$

because $\tilde{\lambda}_n \to 0$ as $n \to \infty$. Thus, the whole right hand side of (3.41) converges to 0 as $n \to \infty$ and we obtain the assertion because $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) - \mathcal{R}^*_{L,\mathrm{P}} \geq 0$ by the definition of $\mathcal{R}^*_{L,\mathrm{P}}$. □

Proof of Theorem 3.4.8. We can split up the difference, which we wish to investigate, as

$$\begin{aligned}
&\left|\left|f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f^*_{L,\mathrm{P}}\right|\right|_{L_p(\mathrm{P}^X)} \\
&\leq \left|\left|f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}\right|\right|_{L_p(\mathrm{P}^X)} + \left|\left|f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f^*_{L,\mathrm{P}}\right|\right|_{L_p(\mathrm{P}^X)} . \qquad (3.42)
\end{aligned}$$

Because $||f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}||_{L_p(\mathrm{P}^X)} \leq ||f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}||_{L_\infty(\mathrm{P}^X)}$, we know from Lemma 3.4.10 that the first summand on the right hand side converges to 0 in probability as $n \to \infty$.

Thus, only the second summand remains to be examined: From Lemma 3.4.11, we obtain

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) = \mathcal{R}^*_{L,\mathrm{P}} .$$

We further know for all $n \in \mathbb{N}$ that $f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} \in L_p(\mathrm{P}^X)$ because

$$\begin{aligned}
&||f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}||_{L_p(\mathrm{P}^X)} \leq ||f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}||_{L_\infty(\mathrm{P}^X)} \leq \max_{a\in\{1,\dots,A_n\}} \left|\left|\hat{f}_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}\right|\right|_{L_\infty(\mathrm{P}^X)} \\
&\leq \max_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left|\left|f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}\right|\right|_{L_\infty(\mathrm{P}^X_{n,a})} \leq \max_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} ||k_{n,a}||_\infty \left|\left|f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}\right|\right|_{H_{n,a}} < \infty
\end{aligned}$$

by **(W1)**, **(W2)** and Lemma 2.1.10(i), similarly to (3.35). Employing Theorem 3.2.1 and Remark 3.2.2 then yields convergence to 0 (as $n \to \infty$) of the second summand on the right hand side of (3.42), which completes the proof. □

Example 3.4.12. If $p = 2$, like for the popular least squares loss, we have $p_1^* = 3$ and $p_2^* = 1$ in Theorem 3.4.8 and condition (3.33) therefore becomes

$$\min_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \frac{\lambda_{n,a}^3 d_{n,a}}{\tilde{A}_n} \to \infty .$$

If $p = 1$, like for the pinball loss or the ε-insensitive loss, we have $p_1^* = 2$ and p_2^* can be chosen arbitrarily small. Hence, condition (3.33) relaxes even further in this case, becoming

$$\min_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \frac{\lambda_{n,a}^2 d_{n,a}}{\tilde{A}_n^\delta} \to \infty$$

for an arbitrarily small $\delta > 0$.

In the subsequent example, we empirically examine the convergence postulated in Theorem 3.4.8 for some simulated data. As this example also considers the case $A_n = 1$ for the different investigated sample sizes n, it also covers the L_p-consistency of *non-localized* SVMs that was stated in Theorem 3.3.2.

Example 3.4.13. We used R Statistical Software (R Core Team, 2022, v4.2.2) to perform median regression (that is, we used the 0.5-pinball loss function in our SVMs) on synthetic data generated according to the regression problem “Friedman 1” from the library `mlbench` (Leisch and Dimitriadou, 2021) as described by Friedman (1991). Here, the input space $\mathcal{X}$

is 10-dimensional and each component of the input is uniformly distributed on $[0,1]$, with however only 5 of these components actually influencing the output y via the function

$$f(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5\,,$$

from which the value of y is obtained by adding an $\mathcal{N}(0,1)$-distributed error, which yields that $f^*_{L,\mathrm{P}}$ is P^X-a.s. unique and coincides with f. The remaining 5 components of the input data are hence just adding noise.

We proceeded by generating a regionalization data set of size 10,000, based on which we used a k-means approach to partition $\mathcal{X}$ into 3, 5, 10, 20, 40 and 100 regions. For each of these regionalization choices, we then used `liquidSVM` (Steinwart and Thomann, 2017) with the 0.5-pinball loss function to compute corresponding localized SVMs for different training set sizes ranging from $n = 600$ to $n = 2{,}000{,}000$. Additionally, we did the same computations for a regular SVM (i.e. one based on a single global region). We used fixed Gaussian RBF kernels not changing with n. By Example 3.4.12, Theorem 3.4.8 then guarantees convergence of $\left|\left|f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f^*_{L,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)}$ to 0 whenever $\lambda_{n,i}$ tends to 0 slower than $d_{n,i}^{-1/2}$ (because $\tilde{A}_n$ is constant for each fixed regionalization). For this reason, we chose some constant $c_i > 0$ on each region $\mathcal{X}_i$ (for each regionalization) and then used $\lambda_{n,i} = c_i \cdot d_{n,i}^{-1/3}$. To empirically verify the postulated convergence, we estimated the resulting values of $\left|\left|f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f^*_{L,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)}$ based on 1,000,000 test data points generated according to “Friedman 1” without the random errors in order to obtain evaluations of the Bayes function $f^*_{L,\mathrm{P}}$.

We repeated this whole procedure 30 times and collected the means of the estimated values of $\left|\left|f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f^*_{L,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)}$ for the different combinations of n and the number of regions in Table 3.4.1.[16] We also added the estimated standard errors of the means based on these 30 repetitions in that table.

It can be seen that—no matter the number of regions—the convergence with respect to n does indeed seem to take place. Looking at the table row-wise instead of column-wise, one further notices that the number of regions yielding the best results slowly increased as we increased n, starting from 1 (i.e. using a global SVM) for $n = 600$ and reaching the maximum examined amount of regions, namely 100, for $n = 2{,}000{,}000$. Looking at how small the estimated values of the standard errors of the means are, one can conclude that this pattern was not just due to chance but really exists and hence supports the idea of increasing the number of regions as n increases in practice.

This idea also gets supported by Table 3.4.2, where we collected the corresponding computation times (training and testing times combined).[17] The table shows that computation

[16] The missing values in the table are due to the corresponding localized SVMs not having been computed—either because of having too few data points in the different regions (depicted by “–”) or because of having so many data points in a single region that the calculations become exceedingly memory-intensive (depicted by “+”).

[17] Because of us not computing a new regionalization for each n but instead using regionalizations that are fixed independently of n, the computation times do *not* include the time needed for computing the regionalization, which was however negligible for the simple k-means approach that was used. The time needed for assigning training and test points to the different regions on the other hand is included in the stated computation times.

#Reg. n	1	3	5	10	20	40	100
600	**1.154** *0.025*	1.227 *0.028*	1.318 *0.020*	1.590 *0.009*	1.826 *0.010*	–	–
2,000	0.967 *0.024*	**0.952** *0.030*	0.954 *0.019*	1.084 *0.009*	1.227 *0.005*	1.403 *0.004*	–
6,000	0.846 *0.026*	0.799 *0.032*	**0.764** *0.020*	0.812 *0.009*	0.874 *0.005*	0.981 *0.003*	1.160 *0.002*
20,000	0.744 *0.027*	0.689 *0.032*	**0.640** *0.019*	0.641 *0.009*	0.657 *0.004*	0.703 *0.002*	0.800 *0.002*
60,000	0.663 *0.028*	0.612 *0.031*	0.559 *0.018*	0.544 *0.009*	**0.541** *0.004*	0.557 *0.002*	0.611 *0.002*
200,000	+	+	0.483 *0.016*	0.463 *0.008*	0.453 *0.004*	**0.453** *0.002*	0.476 *0.002*
600,000	+	+	+	0.402 *0.008*	0.394 *0.004*	**0.386** *0.002*	0.392 *0.002*
2,000,000	+	+	+	+	+	0.331 *0.002*	**0.326** *0.002*

Table 3.4.1: Means and estimated standard errors of the means (in small font) of $\left\|f_{L,\mathrm{D}_n,\boldsymbol{\lambda}_n,\boldsymbol{k}_n,\boldsymbol{\mathcal{X}}_n} - f^*_{L,\mathrm{P}}\right\|_{L_1(\mathrm{P}^X)}$ for the regression problem "Friedman 1" with different training set sizes n and different amounts of underlying regions (based on 30 repetitions for each combination). Bold font highlights the minimum value for each n.

times indeed drastically decrease as the number of regions increases, also recall the analyses of computation times referenced in Section 2.2.1.

3.4.3 Risk Consistency Using Regular Loss Functions

To our knowledge, the only existing results which explicitly examine risk consistency of localized SVMs are those by Hable (2013, Theorem 1) and Dumpert and Christmann (2018, Theorem 3.1), both of which are in certain aspects considerably less general than the subsequent Theorem 3.4.14: Dumpert and Christmann (2018) only considered Lipschitz continuous (shifted) loss functions, whereas we take a look at distance-based loss functions, thus covering a different subset of all loss functions, notably also including the popular but not Lipschitz continuous least squares loss. Additionally, Dumpert and Christmann (2018) assumed a fixed regionalization and fixed kernels on the different regions, which stay the same independently of the size n of the underlying data set. We however also allow for regionalizations which change with n (cf. Section 3.4.1), since the regionalization is oftentimes not predefined in practice but instead might change when new data points are added to the data set—for example, becoming finer when n grows. We also allow for kernels that change with n and that are chosen from an possibly infinite set of kernels—for example, Gaussian kernels whose bandwidth decreases as n increases (cf. Example 3.4.5). Thus, we significantly generalize the investigations by Dumpert and Christmann (2018) in these aspects. Hable (2013) on the other hand only allowed for a bounded output space $\mathcal{Y}$ and only considered the special case of the regionalization stemming from some k-

#Reg. n	1	3	5	10	20	40	100
600	10	7	6	5	6	–	–
2,000	28	13	10	8	7	8	–
6,000	86	32	22	14	11	10	13
20,000	340	95	61	34	21	16	16
60,000	1,486	509	178	94	51	32	26
200,000	+	+	1,077	377	177	101	56
600,000	+	+	+	2,429	834	395	184
2,000,000	+	+	+	+	+	3,898	1,142

Table 3.4.2: Means of the computation times (training plus testing) in seconds based on for the regression problem "Friedman 1" with different training set sizes n and different amounts of underlying regions (based on 30 repetitions for each combination).

nearest neighbor method. Whereas this approach implicitly also allows for regionalizations which change with n, this makes our Theorem 3.4.14 applicable to a much wider array of localization methods—even though the k-nearest neighbor approach described by Hable (2013) is not one of them because it can lead to condition **(R2)** from Section 3.4.1 being violated, thus making our result and that of Hable (2013) applicable to different situations.

Apart from that, the oracle inequalities by Meister and Steinwart (2016), Thomann et al. (2017), Mücke (2019), Blaschzyk and Steinwart (2022) of course also imply risk consistency if the different parameters in these results are chosen accurately. However, these oracle inequalities are only valid for the least squares respectively the hinge loss, whereas we aim at deriving a much more general result which is applicable for the considerably larger class of convex, distance-based loss functions (even though this class does not contain the hinge loss). Additionally, these oracle inequalities require stricter conditions than our consistency results, like for example $\mathcal{X}$ being contained in a ball of fixed radius, $\mathcal{Y}$ being bounded, the kernels all being Gaussian kernels, and also additional requirements regarding the regionalization.

In the subsequent theorem, we derive such a general result on risk consistency of localized SVMs. Condition (3.43) in that theorem is slightly more restrictive and complicated than its counterpart (3.33) in the result on L_p-consistency. However, the additional factor $\lambda_{n,b}^{p_3^*}$ can be eliminated from (3.43) in several important special cases, thus weakening and simplifying this condition again: If the loss function is of growth type $p = 1$, one directly obtains $p_3^* = 0$, and if the regionalizations underlying the localized SVMs partition $\mathcal{X}$ or $f_{L,\mathrm{P}}^*$ exists and is P^X-a.s. unique, the special cases (i) and (ii) of the theorem also yield similar relaxations.

Theorem 3.4.14. *Let Assumption 3.4.6 be satisfied. Define* $p_1^* := \max\{p+1, p(p+1)/2\}$ *and* $p_3^* := \max\{p-1, p(p-1)/2\}$*. Further choose* $p_2^* := \max\{2(p-1)/p, p-1\}$ *if* $p > 1$ *and* $p_2^* \in (0,\infty)$ *arbitrary if* $p = 1$*. If the regularization parameters satisfy* $\lambda_{n,a} \in (0, C)$ *for all* $n \in \mathbb{N}$ *and* $a \in \{1, \dots, A_n\}$ *for some* $C \in (0,\infty)$*, as well as*

$$\max_{a \in I_{\boldsymbol{\mathcal{X}}_{\boldsymbol{n}},\mathrm{P}}} \beta_{n,a}^2 \lambda_{n,a} \to 0$$

and

$$\min_{a,b\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \frac{\lambda_{n,b}^{p_3^*}\lambda_{n,a}^{p_1^*}d_{n,a}}{\tilde{A}_n^{p_2^*}} \to \infty \tag{3.43}$$

as $n \to \infty$, then

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) = \mathcal{R}_{L,\mathrm{P}}^* \qquad \textit{in probability } \mathrm{P}^\infty.$$

If some additional conditions are satisfied, it is possible to slightly relax assumption (3.43) *regarding the regularization parameters:*

(i) If, for all $n \in \mathbb{N}$, the regionalization $\boldsymbol{\mathcal{X}_n}$ is a partition of $\mathcal{X}$, then it suffices if (3.43) *is satisfied for $p_1^* := \max\{2p, p^2\}$ and $p_3^* := 0$.*

(ii) If $f_{L,\mathrm{P}}^*$ exists and is P^X-a.s. unique, then it suffices if (3.43) *is satisfied for $p_3^* := 0$.*

For proving this theorem, the following lemma is useful. Because it does not need all of Assumption 3.4.6, the required assumptions are explicitly stated in the lemma instead.

Lemma 3.4.15. *Let $L\colon \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss function of upper growth type $p \in [1,\infty)$. Let the regionalizations $\boldsymbol{\mathcal{X}_n}$, $n \in \mathbb{N}$, satisfy **(R1)**, **(R3)**, and let the weight functions $w_{n,a}$, $n \in \mathbb{N}$ and $a \in \{1,\dots,A_n\}$, satisfy **(W1)**, **(W2)**, **(W3)**. Assume that $|\mathrm{P}|_p < \infty$ and $\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} |\mathrm{P}_{n,a}|_p < \infty$. Let, for all $n \in \mathbb{N}$ and $a \in \{1,\dots,A_n\}$, $k_{n,a}$ be a bounded and measurable kernel on $\mathcal{X}_{n,a}$ with separable RKHS $H_{n,a}$, such that $\sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} ||k_{n,a}||_\infty < \infty$. Define $p_1^* := \max\{p+1, p(p+1)/2\}$ and $p_3^* := \max\{p-1, p(p-1)/2\}$. Further choose $p_2^* := \max\{2(p-1)/p, p-1\}$ if $p > 1$ and $p_2^* \in (0,\infty)$ arbitrary if $p = 1$. If the regularization parameters satisfy $\lambda_{n,a} \in (0, C)$ for all $n \in \mathbb{N}$ and $a \in \{1,\dots,A_n\}$ for some $C \in (0,\infty)$, as well as*

$$\min_{a,b\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \frac{\lambda_{n,b}^{p_3^*}\lambda_{n,a}^{p_1^*}d_{n,a}}{\tilde{A}_n^{p_2^*}} \to \infty \tag{3.44}$$

as $n \to \infty$, then

$$\lim_{n\to\infty} |\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) - \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}})| = 0 \qquad \textit{in probability } \mathrm{P}^\infty.$$

If additionally, the regionalizations $\boldsymbol{\mathcal{X}_n}$, $n \in \mathbb{N}$, are partitions of $\mathcal{X}$, then it suffices if (3.44) *is satisfied for $p_1^* := \max\{2p, p^2\}$ and $p_3^* := 0$.*

Proof. Assume, for all $n \in \mathbb{N}$ and $a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}$, without loss of generality that $d_{n,a} > 0$ (which by (3.44) has to be satisfied for n sufficiently large), such that the respective local empirical SVM $f_{L,\mathrm{D}_{n,a},\lambda_{n,a},k_{n,a}}$ is indeed an empirical SVM and not just defined as the zero function. To shorten the notation, we denote $f_{\mathrm{P},n} := f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}$, $f_{\mathrm{D}_n,n} := f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}$, $f_{\mathrm{P},n,a} := f_{L,\mathrm{P}_{n,a},\lambda_{n,a},k_{n,a}}$ and $f_{\mathrm{D}_n,n,a} := f_{L,\mathrm{D}_{n,a},\lambda_{n,a},k_{n,a}}$ for all $n \in \mathbb{N}$ and $a \in \{1,\dots,A_n\}$, as well as $\kappa := \sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} ||k_{n,a}||_\infty$, $\rho := |\mathrm{P}|_p \vee \sup_{n\in\mathbb{N},a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} |\mathrm{P}_{n,a}|_p$ and $\tilde{\lambda}_n := \min_{a\in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \lambda_{n,a}$ throughout this proof. Additionally, note that Lemma 3.4.10 is applicable in the situation of this lemma (in the base case as well as in the special case of the regionalizations being partitions of $\mathcal{X}$) as (3.44) in combination with $\lambda_{n,b} \in (0, C)$ for all $n \in \mathbb{N}$ and $b \in \{1, \dots, A_n\}$ implies the validity of (3.34).

We start by proving the main assertion before turning our attention to the special case of the regionalizations being partitions of $\mathcal{X}$ afterwards.

By applying Lemma 2.1.21(iii) with $q \coloneqq p$, we obtain

$$\begin{aligned}
&|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},n})| \\
&\leq c_{p,L} \cdot \left(|\mathrm{P}|_p^{p-1} + ||f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)}^{p-1} + ||f_{\mathrm{D}_n,n}||_{L_p(\mathrm{P}^X)}^{p-1} + 1 \right) \cdot ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)} \,, \qquad (3.45)
\end{aligned}$$

where $c_{p,L} \in (0, \infty)$ denotes a constant only depending on p and L.

We can further analyze the right hand side of this inequality by noting that

$$\begin{aligned}
||f_{\mathrm{P},n}||_{L_\infty(\mathrm{P}^X)} &\leq \max_{a \in \{1,\dots,A_n\}} \left|\left| \hat{f}_{\mathrm{P},n,a} \right|\right|_{L_\infty(\mathrm{P}^X)} \\
&= \max_{a \in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} ||f_{\mathrm{P},n,a}||_{L_\infty(\mathrm{P}^X_{n,a})} \leq \max_{a \in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} c_{p,L,\rho,\kappa} \cdot \lambda_{n,a}^{-1/2} \,,
\end{aligned}$$

with the first inequality following from **(W1)** and **(W2)**, similarly to (3.35), and the last one analogously to (3.36), with $c_{p,L,\rho,\kappa} \in (0, \infty)$ denoting a constant depending only on p, L, ρ and κ. Hence,

$$||f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)}^{p-1} \leq ||f_{\mathrm{P},n}||_{L_\infty(\mathrm{P}^X)}^{p-1} \leq \max_{a \in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} c_{p,L,\rho,\kappa}^{p-1} \cdot \lambda_{n,a}^{-(p-1)/2} = c_{p,L,\rho,\kappa}^{p-1} \cdot \tilde{\lambda}_n^{-(p-1)/2} \,. \qquad (3.46)$$

Similarly, we obtain

$$\begin{aligned}
||f_{\mathrm{D}_n,n}||_{L_p(\mathrm{P}^X)}^{p-1} &\leq \left(||f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)} + ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)} \right)^{p-1} \\
&\leq 2^{p-1} \cdot c_{p,L,\rho,\kappa}^{p-1} \cdot \tilde{\lambda}_n^{-(p-1)/2} + 2^{p-1} \cdot ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)}^{p-1} \,, \qquad (3.47)
\end{aligned}$$

where we applied (3.46) in the last step.

Plugging (3.46) and (3.47) into (3.45) then yields

$$\begin{aligned}
&|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},n})| \\
&\leq c_{p,L} \cdot \left(\rho^{p-1} + (2^{p-1} + 1) \cdot c_{p,L,\rho,\kappa}^{p-1} \cdot \tilde{\lambda}_n^{-(p-1)/2} + 2^{p-1} \cdot ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)}^{p-1} + 1 \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)} \\
&= c_{p,L} \cdot \Bigg(\left(\rho^{p-1} \tilde{\lambda}_n^{(p-1)/2} + (2^{p-1} + 1) \cdot c_{p,L,\rho,\kappa}^{p-1} + \tilde{\lambda}_n^{(p-1)/2} \right) \\
&\qquad\qquad\qquad \cdot \tilde{\lambda}_n^{-(p-1)/2} \cdot ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)} + 2^{p-1} \cdot ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)}^{p} \Bigg) \\
&\leq \tilde{c}_{p,L,\rho,\kappa} \cdot \left(\tilde{\lambda}_n^{-(p-1)/2} \cdot ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_\infty(\mathrm{P}^X)} + ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_\infty(\mathrm{P}^X)}^{p} \right) \,,
\end{aligned}$$

where we employed $\tilde{\lambda}_n \leq C$ and $||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_p(\mathrm{P}^X)} \leq ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_\infty(\mathrm{P}^X)}$ in the last step.

We know from Lemma 3.4.10 that the second summand on the right hand side converges to 0 in probability as $n \to \infty$. Hence, we only need to further investigate the first summand.

For this, we can proceed in exactly the same way as in the proof of Lemma 3.4.10 and only need to additionally consider the factor $\tilde{\lambda}_n^{-(p-1)/2}$. By doing this, we obtain for all $\varepsilon > 0$

$$
\begin{aligned}
&\mathrm{P}^n \Big(D_n \in (\mathcal{X} \times \mathcal{Y})^n : \tilde{\lambda}_n^{-(p-1)/2} \cdot ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||_{L_\infty(\mathrm{P}^X)} \geq \varepsilon \\
&\qquad\qquad\qquad\qquad\qquad\qquad \Big| \, |D_{n,1}| = d_{n,1}, \ldots, |D_{n,A_n}| = d_{n,A_n} \Big) \\
&\leq \mathrm{P}^n \left(D_n \in (\mathcal{X} \times \mathcal{Y})^n : \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} ||f_{\mathrm{D}_n,n,a} - f_{\mathrm{P},n,a}||_{H_{n,a}} \geq \frac{\varepsilon \tilde{\lambda}_n^{(p-1)/2}}{\kappa} \right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad \left. \Big| \, |D_{n,1}| = d_{n,1}, \ldots, |D_{n,A_n}| = d_{n,A_n} \right) \\
&\leq \sum_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \mathrm{P}_{n,a}^{d_{n,a}} \left(D_{n,a} \in (\mathcal{X}_{n,a} \times \mathcal{Y})^{d_{n,a}} : ||f_{\mathrm{D}_n,n,a} - f_{\mathrm{P},n,a}||_{H_{n,a}} \geq \frac{\varepsilon \tilde{\lambda}_n^{(p-1)/2}}{\kappa} \right) \\
&\leq c_{q,p,L,\rho,\kappa} \cdot \tilde{A}_n \cdot \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\frac{1}{\tilde{\lambda}_n^{(p-1)/2} \lambda_{n,a}^{(p+1)/2} \varepsilon d_{n,a}^{q^*}} \right)^q ,
\end{aligned}
\tag{3.48}
$$

analogously to (3.39), with $c_{q,p,L,\rho,\kappa} \in (0, \infty)$ denoting a constant depending only on q, p, L, ρ and κ. Here, as in the proof of Lemma 3.4.10, $q := p/(p-1)$ if $p > 1$, $q := 2/p_2^*$ if $p = 1$, and $q^* := \min\{1/2, 1 - 1/q\} = \min\{1/2, 1/p\} = (p+1)/(2p_1^*) = (p-1)/(2p_3^*)$.

Because $(qq^*)^{-1} = p_2^*$ (cf. proof of Lemma 3.4.10), we furthermore obtain

$$
\tilde{A}_n \cdot \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\frac{1}{\tilde{\lambda}_n^{(p-1)/2} \lambda_{n,a}^{(p+1)/2} d_{n,a}^{q^*}} \right)^q = \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\frac{\tilde{A}_n^{p_2^*}}{\tilde{\lambda}_n^{p_3^*} \lambda_{n,a}^{p_1^*} d_{n,a}} \right)^{qq^*} \to 0 \, , \qquad n \to \infty \, ,
$$

by assumption. Hence, the whole right hand side of (3.48) converges to 0, which yields the main assertion.

As for the special case of the regionalizations being partitions of $\mathcal{X}$: If $\boldsymbol{\mathcal{X}_n}$ is a partition of $\mathcal{X}$, then the conditions **(W2)** and **(W3)** imply that $w_{n,a} = \mathbb{1}_{\mathcal{X}_{n,a}}$ for all $a \in \{1, \ldots, A_n\}$. Hence, we obtain

$$
\begin{aligned}
&|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},n})| \\
&= \left| \int_{\mathcal{X} \times \mathcal{Y}} L\left(y, \sum_{a=1}^{A_n} \mathbb{1}_{\mathcal{X}_{n,a}}(x) \hat{f}_{\mathrm{D}_n,n,a}(x)\right) \, \mathrm{dP}(x,y) \right. \\
&\qquad\qquad\qquad\qquad \left. - \int_{\mathcal{X} \times \mathcal{Y}} L\left(y, \sum_{a=1}^{A_n} \mathbb{1}_{\mathcal{X}_{n,a}}(x) \hat{f}_{\mathrm{P},n,a}(x)\right) \, \mathrm{dP}(x,y) \right| \\
&= \left| \sum_{a=1}^{A_n} \left(\int_{\mathcal{X}_{n,a} \times \mathcal{Y}} L(y, f_{\mathrm{D}_n,n,a}(x)) \, \mathrm{dP}(x,y) - \int_{\mathcal{X}_{n,a} \times \mathcal{Y}} L(y, f_{\mathrm{P},n,a}(x)) \, \mathrm{dP}(x,y) \right) \right| \\
&\leq \sum_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \mathrm{P}(\mathcal{X}_{n,a} \times \mathcal{Y}) \cdot \left| \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{\mathrm{D}_n,n,a}) - \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{\mathrm{P},n,a}) \right| \\
&\leq \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left| \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{\mathrm{D}_n,n,a}) - \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{\mathrm{P},n,a}) \right|
\end{aligned}
\tag{3.49}
$$

in this case. In the third step, we applied that $\mathcal{X}_{n,a} \times \mathcal{Y}$ is a P-zero set for all $a \notin I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}$, leading to the corresponding P-integrals being 0.

The argument of the maximum on the right hand side of (3.49) can, for each $a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}$, be examined in the same way as we previously examined the difference on the left hand side for proving the main assertion. A difference appears in (3.46), where we now have

$$||f_{\mathrm{P},n,a}||^{p-1}_{L_p(\mathrm{P}^X_{n,a})} \leq ||f_{\mathrm{P},n,a}||^{p-1}_{L_\infty(\mathrm{P}^X_{n,a})} \leq c^{p-1}_{p,L,\rho,\kappa} \cdot \lambda_{n,a}^{-(p-1)/2} .$$

That is, we can omit the final step of bounding this with the help of $\tilde{\lambda}_n$ because we are now not interested in $\max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} ||f_{\mathrm{P},n,a}||_{L_\infty(\mathrm{P}^X_{n,a})}$ but only in $||f_{\mathrm{P},n,a}||_{L_\infty(\mathrm{P}^X_{n,a})}$ for a specific a.

By applying this to the subsequent steps of our proof, we obtain

$$\begin{aligned}
&|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},n})| \\
&\leq \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left|\mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{\mathrm{D}_n,n,a}) - \mathcal{R}_{L,\mathrm{P}_{n,a}}(f_{\mathrm{P},n,a})\right| \\
&\leq \tilde{c}_{p,L,\rho,\kappa} \cdot \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\lambda_{n,a}^{-(p-1)/2} \cdot ||f_{\mathrm{D}_n,n,a} - f_{\mathrm{P},n,a}||_{L_\infty(\mathrm{P}^X_{n,a})} + ||f_{\mathrm{D}_n,n,a} - f_{\mathrm{P},n,a}||^p_{L_\infty(\mathrm{P}^X_{n,a})}\right) \\
&\leq \tilde{c}_{p,L,\rho,\kappa} \cdot \left(\max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\lambda_{n,a}^{-(p-1)/2} \cdot ||f_{\mathrm{D}_n,n,a} - f_{\mathrm{P},n}||_{L_\infty(\mathrm{P}^X_{n,a})}\right) + ||f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}||^p_{L_\infty(\mathrm{P}^X)}\right) ,
\end{aligned}$$

where the second summand on the right hand side converges to 0 in probability by Lemma 3.4.10.

As for the first summand, we can derive

$$\begin{aligned}
&\mathrm{P}^n\Bigg(D_n \in (\mathcal{X} \times \mathcal{Y})^n : \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\lambda_{n,a}^{-(p-1)/2} \cdot ||f_{\mathrm{D}_n,n,a} - f_{\mathrm{P},n,a}||_{L_\infty(\mathrm{P}^X_{n,a})}\right) \geq \varepsilon \\
&\qquad\qquad\qquad\qquad\qquad\Bigg|\, |D_{n,1}| = d_{n,1}, \dots, |D_{n,A_n}| = d_{n,A_n}\Bigg) \\
&\leq c_{q,p,L,\rho,\kappa} \cdot \tilde{A}_n \cdot \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\frac{1}{\lambda^p_{n,a} \varepsilon d^{q^*}_{n,a}}\right)^q ,
\end{aligned}$$

analogously to (3.48). Finally, we obtain convergence to 0 of the right hand side, and thus the assertion, because

$$\tilde{A}_n \cdot \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\frac{1}{\lambda^p_{n,a} d^{q^*}_{n,a}}\right)^q = \max_{a \in I_{\boldsymbol{\mathcal{X}_n},\mathrm{P}}} \left(\frac{\tilde{A}_n^{p_2^*}}{\lambda^{p_1^*}_{n,a} d_{n,a}}\right)^{qq^*} \to 0 , \qquad n \to \infty ,$$

by assumption, where we applied that $p/q^* = p_1^*$ since $p_1^* = \max\{2p, p^2\}$ now. □

Proof of Theorem 3.4.14. We start by proving the main assertion and the special case (i): We can split up the difference, which we wish to investigate, as

$$\begin{aligned}
&|\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) - \mathcal{R}^*_{L,\mathrm{P}}| \\
&\leq |\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) - \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}})| + |\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) - \mathcal{R}^*_{L,\mathrm{P}}| . \quad (3.50)
\end{aligned}$$

The assertions then follow directly by applying Lemma 3.4.15 to the first and Lemma 3.4.11 to the second summand on the right hand side.

As for the special case (ii): If $f^*_{L,\mathrm{P}}$ P^X-a.s. uniquely exists, the assertion follows directly from Theorem 3.4.8 and Theorem 3.2.3, which is applicable because $f^*_{L,\mathrm{P}} \in L_p(\mathrm{P}^X)$ (cf. Remark 3.2.2) and $f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} \in L_p(\mathrm{P}^X)$ for all $n \in \mathbb{N}$ (cf. proof of Theorem 3.4.8). □

If $p = 1$, the cases (i) and (ii) of Theorem 3.4.14 can be ignored since they do not yield an actual relaxation because $p_3^* = 0$ then also holds true in the general case. Furthermore, the possible relaxations mentioned in Remark 3.4.9 are obviously also valid for Theorem 3.4.14.

The subsequent example transfers Example 3.4.13 to now examine risk consistency instead of L_p-consistency. As it also considers the case $A_n = 1$ for the different investigated sample sizes n, it also covers the risk consistency of *non-localized* SVMs that was stated in Corollary 3.3.5.

Example 3.4.16. We look at the regression problem "Friedman 1" the same way as we did in Example 3.4.13, notably also choosing the regularization parameters as $\lambda_{n,i} = c_i \cdot d_{n,i}^{-1/3}$ for constants c_i. Theorem 3.4.14 yields that $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) - \mathcal{R}^*_{L,\mathrm{P}}$ converges to 0 because the special case (ii) of that theorem tells us that condition (3.43) coincides with (3.33) in the situation of this example, and the latter condition was explained to be satisfied for this choice of $\lambda_{n,i}$ in Example 3.4.13. Table 3.4.3 shows the resulting values of the excess risk $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) - \mathcal{R}^*_{L,\mathrm{P}}$, from which it can be seen that the postulated convergence does indeed take place and that it does so considerably faster than that of $\left|\left|f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}} - f^*_{L,\mathrm{P}}\right|\right|_{L_1(\mathrm{P}^X)}$ in Example 3.4.13.[18]

[18] The missing values in the table are due to the corresponding localized SVMs not having been computed—either because of having too few data points in the different regions (depicted by "–") or because of having so many data points in a single region that the calculations become exceedingly memory-intensive (depicted by "+").

#Reg. n	1	3	5	10	20	40	100
600	**0.316** *0.010*	0.344 *0.012*	0.381 *0.008*	0.498 *0.004*	0.604 *0.004*	–	–
2,000	0.244 *0.009*	0.238 *0.012*	**0.237** *0.008*	0.286 *0.004*	0.343 *0.002*	0.417 *0.002*	–
6,000	0.200 *0.009*	0.184 *0.011*	**0.170** *0.007*	0.185 *0.003*	0.206 *0.002*	0.245 *0.001*	0.314 *0.001*
20,000	0.163 *0.009*	0.147 *0.011*	0.129 *0.006*	**0.129** *0.003*	0.133 *0.002*	0.145 *0.001*	0.176 *0.001*
60,000	0.136 *0.009*	0.122 *0.010*	0.104 *0.006*	0.100 *0.003*	**0.098** *0.001*	0.101 *0.001*	0.114 *0.001*
200,000	+	+	0.082 *0.005*	0.077 *0.002*	0.074 *0.001*	**0.072** *0.001*	0.076 *0.001*
600,000	+	+	+	0.061 *0.002*	0.059 *0.001*	0.056 *0.001*	**0.055** *0.000*
2,000,000	+	+	+	+	+	0.044 *0.001*	**0.041** *0.000*

Table 3.4.3: Means and estimated standard errors of the means (in small font) of the excess risk $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\boldsymbol{\lambda_n},\boldsymbol{k_n},\boldsymbol{\mathcal{X}_n}}) - \mathcal{R}^*_{L,\mathrm{P}}$ for the regression problem "Friedman 1" with different training set sizes n and different amounts of underlying regions (based on 30 repetitions for each combination). Bold font highlights the minimum value for each n.

Chapter 4

Total Stability

This chapter concerns itself with total stability of SVMs and localized SVMs. That is, the influence that simultaneous slight changes in probability measure, (vector of) regularization parameter(s), (vector of) kernel(s) and—in the case of localized SVMs—regionalization have on the resulting predictor is investigated.[19] Naturally, one would hope that such small changes only lead to small changes in the predictor, such that small random errors in the available data do not completely skew the predictor.

The notion of total stability is formally defined in Section 4.1. Afterwards, Section 4.2 introduces ways to measure the differences between probability measures, regularization parameters, kernels and regionalizations, which are needed to derive results on total stability of SVMs in Section 4.3 and on that of localized SVMs in Section 4.4.

Throughout this chapter, the following standard assumptions are assumed to hold true:

Assumption 4.0.1. Let $\mathcal{X}$ be a complete separable metric space and let $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathcal{X}$ and $\mathcal{Y}$ be equipped with their respective Borel σ-algebras $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$.

Note that the results from this chapter only consider SVMs and localized SVMs using shifted loss functions. Because of Lemma 2.1.29, they can however be transferred to regular loss functions whenever $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$.[20]

4.1 Introduction to Total Stability

Total stability can in some sense be seen as an extension to the considerably more well-known concept of statistical robustness. Roughly speaking, the latter concerns itself with

[19] Total stability thus considers the effect of changes in all components influencing the predictor except for the loss function. It is explained in Section 4.1, why it makes sense to treat the loss function differently than probability measure, regularization parameter(s), kernel(s) and regionalization and *not* include it in the notion of total stability.

[20] Transferring it to localized SVMs, Lemma 2.1.29 strictly speaking does not demand $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, but instead that the risks with respect to the local measures on all regions with positive probability are all finite. If $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ is such a region with $\mathrm{P}(\tilde{\mathcal{X}} \times \mathcal{Y}) > 0$, it however follows from the definition of the local measures that $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ already implies $\mathcal{R}_{L,\mathrm{P}_{\tilde{\mathcal{X}}}}(0) < \infty$. (If on the other hand $\mathrm{P}(\tilde{\mathcal{X}} \times \mathcal{Y}) = 0$, then $f_{L^\star,\mathrm{P}_{\tilde{\mathcal{X}}},\lambda,k} = f_{L,\mathrm{P}_{\tilde{\mathcal{X}}},\lambda,k}$ is immediately obvious from the definition of local SVMs.)

the effect that changes in the probability measure (respectively the data) have on an estimator and the estimator is called statistically robust if small changes in the probability measure only lead to small changes in the (distribution of the) estimator. In practice, such changes in the data for example occur because of measurement or rounding errors and can therefore consist either of few extreme outliers or of slight changes in a large part of the data. Different notions of statistical robustness take different approaches, for example, to how these changes are quantified, but they all examine the same fundamental concept. We refer to Huber (1964), Hampel (1971), Huber (1981), Hampel et al. (1986), Rousseeuw and Leroy (1987), Maronna et al. (2006) among others for details on robust statistics.

When applying predictors like SVMs, we of course hope that such changes in the data do not skew the predictions by too much, for which reason we would like SVMs to be robust. Indeed, several papers derived results regarding different notions of statistical robustness. These include Bousquet and Elisseeff (2002), Christmann and Steinwart (2004, 2007), Christmann and Van Messem (2008), Hable and Christmann (2011), Sheng et al. (2020), Eckstein et al. (2023). Additionally, Dumpert and Christmann (2018), Dumpert (2020) transferred some of these results to localized SVMs. Similar considerations for related machine learning methods can for example be found in Poggio et al. (2004), Mukherjee et al. (2006) and the references cited therein.

While classic statistical robustness only considers the effect of small changes in the probability measure (respectively the data set in practice) on the resulting predictor, this is not the only component influencing the resulting (localized) SVM. Instead, loss function, regularization parameter(s), kernel(s) and—for localized SVMs—regionalization also play a role.

Among these, the loss function stands out as it usually is entirely predefined and not chosen depending on the data in practice. Instead the loss function is chosen depending on the property of the underlying distribution one wishes to estimate. If one tries to estimate the function of conditional means, one might choose the least squares loss, whereas the τ-pinball loss is the obvious choice for estimating the function of conditional τ-quantiles, and so on. Hence, the choice of the loss function describes the goal of the prediction and it is usually entirely intended that changes in the loss function can lead to considerable changes in the predictor. For this reason, the effect of such changes in the loss function is not included in the notion of total stability that is considered here.

On the other hand, regularization parameter(s), kernel(s) and regionalization are usually not predefined but instead chosen data-dependently. To be more specific, the type of kernel is often predefined (for example, a Gaussian RBF kernel), but not its hyperparameter(s). Regularization parameter and hyperparameter(s) of the kernel are often chosen from some predefined grid by means of cross-validation. Thus, they might change if the data changes and one would hope that (localized) SVMs are not only stable/robust with respect to the changes in the data itself but also with respect to the changes in regularization parameter(s) and kernel(s) that might result from them or from slightly changing the grid used in the cross-validation scheme.

Similar considerations suggest themselves regarding the regionalization: As described in Section 2.2.1, there are different methods for obtaining a regionalization in practice, like for example tree based methods. All of them have in common that the resulting regionalization depends on the data that is used for generating it. Hence, changes in the data might also lead to changes in the regionalization and it seems sensible to investigate whether localized SVMs are also stable with respect to such changes.

In the following, these considerations are transferred into formal definitions of total stability of SVMs and localized SVMs. Note that these definitions are highly adapted to the results presented in this thesis. They could of course also be generalized to different norms on the left hand side of the bound as well as to slightly different kinds of bounds, as long as the principal idea of total stability is adhered to: bounding the difference between the predictors by a weighted sum of the differences in the components influencing the predictors.

For SVMs, the previous considerations lead to the notion of total stability needing to consider simultaneous changes in the whole triple (P, λ, k) consisting of probability measure, regularization parameter and kernel. Because of this focus on the triple (P, λ, k), we will usually omit the loss function from the notation for SVMs and localized SVMs in this chapter.

Definition 4.1.1 (Total Stability of SVMs)**.** Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a loss function and let $L^\star$ be its shifted version. Let $\mathscr{A}$ be a set of assumptions. We call SVMs based on $L^\star$ totally sup-stable (under the assumptions $\mathscr{A}$ and with respect to d_1, d_2 and d_3) if, for all $\mathrm{P}_1, \mathrm{P}_2, \lambda_1, \lambda_2, k_1, k_2$ satisfying $\mathscr{A}$,

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty \leq c_1 \cdot d_1(\mathrm{P}_1, \mathrm{P}_2) + c_2 \cdot d_2(\lambda_1, \lambda_2) + c_3 \cdot d_3(k_1, k_2) \tag{4.1}$$

where, for $\ell = 1, 2, 3$, $c_\ell \in (0, \infty)$ can be written as

$$c_\ell = \max \{c_{\ell,\lambda_1}, c_{\ell,\lambda_2}\} \cdot \max \{c_{\ell,k_1}, c_{\ell,k_2}\}$$

for constants c_{ℓ,λ_j} and c_{ℓ,k_j} depending on λ_j and k_j respectively, $j = 1, 2$. That is, each c_ℓ is allowed to depend on each of $\lambda_1, \lambda_2, k_1, k_2$ individually but not on their differences.

For $p \in [1, \infty)$, we call SVMs based on $L^\star$ totally L_p-stable (under the assumptions $\mathscr{A}$ and with respect to d_1, d_2 and d_3) if (4.1) holds true with $||\cdot||_\infty$ replaced by $||\cdot||_{L_p(\mathrm{P}_i^X)}$, $i = 1, 2$, and with $d_3 := d_{3,\mathrm{P}_i^X}$ being allowed to depend on P_i^X.

Note that even though d_1, d_2 and d_3 are not required to be metrics, those that are used in the results on total stability and introduced in Sections 4.2.1 to 4.2.3 do actually satisfy the axioms of a metric.

Similarly to the previous definition, total stability of localized SVMs considers changes in the whole quadruple $(\mathrm{P}, \boldsymbol{\lambda}, \boldsymbol{k}, \boldsymbol{\mathcal{X}})$ consisting of probability measure, vector of regularization parameters, vector of kernels, and regionalization. Alas, Section 4.4.2 will explain that it is not possible to derive meaningful bounds on $||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}}||_\infty$ if the two regionalizations do actually differ. Hence, we first give a definition of *regionalization-subtotal stability* of localized SVMs, i.e. stability with respect to changes in only the triple $(\mathrm{P}, \boldsymbol{\lambda}, \boldsymbol{k})$, while the regionalization stays the same (as do the associated weight functions).

For the subsequent definitions, recall the notations $\mathbf{P}_{\boldsymbol{\mathcal{X}}}$ and $\boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}}$ from Section 2.2.2, denoting the vector of local measures on $\boldsymbol{\mathcal{X}}$ associated with P respectively the combined regionalization of two regionalizations $\boldsymbol{\mathcal{X}_1}$ and $\boldsymbol{\mathcal{X}_2}$.

Definition 4.1.2 (Regionalization-Subtotal Stability of Localized SVMs)**.** Let $L\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R}\to[0,\infty)$ be a loss function and let $L^\star$ be its shifted version. Let $\mathscr{A}$ be a set of assumptions. We call localized SVMs based on $L^\star$ regionalization-subtotally sup-stable (under the assumptions $\mathscr{A}$ and with respect to d_1, d_2 and d_3) if, for all $\mathrm{P}_1, \mathrm{P}_2, \boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{k_1}, \boldsymbol{k_2}, \boldsymbol{\mathcal{X}}$ satisfying $\mathscr{A}$,

$$||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}}||_\infty \le c_1\cdot d_1(\mathbf{P}_{\mathbf{1},\boldsymbol{\mathcal{X}}}, \mathbf{P}_{\mathbf{2},\boldsymbol{\mathcal{X}}}) + c_2\cdot d_2(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}) + c_3\cdot d_3(\boldsymbol{k_1}, \boldsymbol{k_2})\,, \tag{4.2}$$

where, for $\ell = 1,2,3$, $c_\ell\in(0,\infty)$ can be written as

$$c_\ell = \max\left\{c_{\ell,\boldsymbol{\lambda_1}}, c_{\ell,\boldsymbol{\lambda_2}}\right\}\cdot\max\left\{c_{\ell,\boldsymbol{k_1}}, c_{\ell,\boldsymbol{k_2}}\right\} \tag{4.3}$$

for constants $c_{\ell,\boldsymbol{\lambda_j}}$ and $c_{\ell,\boldsymbol{k_j}}$ depending on $\boldsymbol{\lambda_j}$ and $\boldsymbol{k_j}$ respectively, $j=1,2$. That is, each c_ℓ is allowed to depend on each of $\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{k_1}, \boldsymbol{k_2}$ individually but not on their differences.

For $p\in[1,\infty)$, we call localized SVMs based on $L^\star$ regionalization-subtotally L_p-stable (under the assumptions $\mathscr{A}$ and with respect to d_1, d_2 and d_3) if (4.2) holds true with $||\cdot||_\infty$ replaced by $||\cdot||_{L_p(\mathrm{P}_i^X)}$, $i=1,2$, with $d_3 := d_{3,\mathrm{P}_i^X}$ being allowed to depend on P_i^X, and with each of c_1, c_2, c_3 being allowed to additionally depend on $\mathrm{P}_i^X(\boldsymbol{\mathcal{X}}) := (\mathrm{P}_i^X(\mathcal{X}_1),\dots,\mathrm{P}_i^X(\mathcal{X}_A))$ via a factor $c_{\ell,\mathrm{P}_i^X(\boldsymbol{\mathcal{X}})}$ that is multiplied to (4.3).

Definition 4.1.3 (Total Stability of Localized SVMs)**.** Let $L\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R}\to[0,\infty)$ be a loss function and let $L^\star$ be its shifted version. Let $\mathscr{A}$ be a set of assumptions. For $p\in[1,\infty)$, we call localized SVMs based on $L^\star$ totally L_p-stable (under the assumptions $\mathscr{A}$ and with respect to d_1, d_2, d_3 and d_4) if, for all $\mathrm{P}_1, \mathrm{P}_2, \boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{k_1}, \boldsymbol{k_2}, \boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}$ satisfying $\mathscr{A}$,

$$\begin{aligned}||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}}||_{L_p(\mathrm{P}_i^X)} &\le c_1\cdot d_1(\mathbf{P}_{\mathbf{1},\boldsymbol{\mathcal{X}^*_{1,2}}}, \mathbf{P}_{\mathbf{2},\boldsymbol{\mathcal{X}^*_{1,2}}}) + c_2\cdot d_2(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2})\\ &\quad + c_3\cdot d_3(\boldsymbol{k_1}, \boldsymbol{k_2}) + c_4\cdot d_4(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2})\end{aligned}$$

for $i=1,2$ and where, for $\ell=1,2,3,4$, $c_\ell\in(0,\infty)$ can be written as

$$c_\ell = \max\{c_{\ell,\mathrm{P}_i^X(\boldsymbol{\mathcal{X}_1})}, c_{\ell,\mathrm{P}_i^X(\boldsymbol{\mathcal{X}_2})}\}\cdot\max\left\{c_{\ell,\boldsymbol{\lambda_1}}, c_{\ell,\boldsymbol{\lambda_2}}\right\}\cdot\max\left\{c_{\ell,\boldsymbol{k_1}}, c_{\ell,\boldsymbol{k_2}}\right\}$$

for constants $c_{\ell,\mathrm{P}_i^X(\boldsymbol{\mathcal{X}_j})}$, $c_{\ell,\boldsymbol{\lambda_j}}$ and $c_{\ell,\boldsymbol{k_j}}$ depending on $\mathrm{P}_i^X(\boldsymbol{\mathcal{X}_j}) := (\mathrm{P}_i^X(\mathcal{X}_{j,1}),\dots,\mathrm{P}_i^X(\mathcal{X}_{j,A_j}))$, $\boldsymbol{\lambda_j}$ and $\boldsymbol{k_j}$ respectively, $j=1,2$. That is, each c_ℓ is allowed to depend on each of $\mathrm{P}_i^X(\boldsymbol{\mathcal{X}_1}), \mathrm{P}_i^X(\boldsymbol{\mathcal{X}_2}), \boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{k_1}, \boldsymbol{k_2}$ individually but not on their differences. Additionally, $d_2 := d_{2,\boldsymbol{\mathcal{X}^*_{1,2}}}$, $d_3 := d_{3,\mathrm{P}_i^X,\boldsymbol{\mathcal{X}^*_{1,2}}}$ and $d_4 := d_{4,\mathrm{P}_i^X}$ are allowed to depend on P_i^X and/or $\boldsymbol{\mathcal{X}^*_{1,2}}$.

Suitable distance measures d_1, d_2, d_3 and d_4 are described in Section 4.2.

So far, there are—to our knowledge—no results on total stability of localized SVMs. Results on total stability of non-localized SVMs have already been derived by Christmann et al. (2018). These are however considerably generalized by those derived in Section 4.3. Namely, Theorems 2.7 and 2.10 from Christmann et al. (2018) both show total sup-stability (in the case of the latter, an application of the property $||f||_\infty \le ||k||_\infty\,||f||_H$, see Lemma 2.1.10(i) from this thesis, is needed to actually obtain sup-stability), but compared to Theorem 4.3.19 they impose additional assumptions regarding both the loss function L and either the regularization parameters λ_1 and λ_2 (Theorem 2.7) or the kernels k_1

and k_2 (Theorem 2.10). The generalization that Theorem 4.3.19 achieves is explained in more detail in Section 4.3.4 and comes with a small drawback regarding the distance d_3 for measuring the difference between the kernels, cf. Remark 4.3.21. Additionally, Section 4.3 contains results with respect to two different choices of d_1, namely the total variation distance and the Wasserstein distance, whereas Christmann et al. (2018) only considered the former. This further substantially adds to the list of comparisons in Definition 4.1.1 for which one can obtain meaningful bounds, cf. Section 4.2.1.

4.2 Measuring Differences between the Components Influencing (Localized) Support Vector Machines

This section describes the different ways in which the distance measures d_j, $j = 1, \dots, 4$, in the definitions of total stability from Section 4.1 are chosen in the results on total stability from Sections 4.3 and 4.4.

4.2.1 Differences between Probability Measures

There exist many different metrics for quantifying the difference between two probability measures. This section does not give a complete overview of such metrics. Instead, it focuses on those two that are used in the results from Sections 4.3 and 4.4, namely the *total variation distance* and the *Wasserstein distance*, and gives some examples of when these distances are useful and when they are not. A more extensive overview of metrics for probability measures is given by Rachev (1991), Gibbs and Su (2002), see also Zolotarev (1976) for some theoretical considerations underlying such metrics.

We start by giving a definition of the total variation distance, see also Tierney (1996, p. 61).

Definition 4.2.1 (Total Variation Distance)**.** Let $\mathrm{P}_1, \mathrm{P}_2$ be probability measures on some measurable space $(\Omega, \mathcal{A})$. The total variation distance between P_1 and P_2 is given by

$$\begin{aligned} &d_{\mathrm{tv}}(\mathrm{P}_1, \mathrm{P}_2) \\ &:= \sup \left\{ \sum_{i=1}^{n} |\mathrm{P}_1(S_i) - \mathrm{P}_2(S_i)| \;\middle|\; n \in \mathbb{N}\,,\ S_1, \dots, S_n \text{ measurable partition of } \Omega \right\} \\ &= 2 \cdot \sup_{S \in \mathcal{A}} |\mathrm{P}_1(S) - \mathrm{P}_2(S)|\,. \end{aligned}$$

Note that some authors define the total variation distance as half of that from Definition 4.2.1, see for example Tsybakov (2009, Definition 2.4). Further, the total variation distance does actually define a metric on the space of probability measures on $(\Omega, \mathcal{A})$, cf. Tsybakov (2009, p. 84). In our case, the measurable space usually is $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X}\times\mathcal{Y}})$.

There are several situations in which two probability measures are similar with respect to the total variation distance, and results on total stability using this metric will therefore yield meaningful bounds. One such situation is the comparison of two empirical distributions, where one is obtained from the other by perturbing (for example, because of measurement/reading errors) or adding (because of obtaining new observations) a small amount of data points:[21]

Example 4.2.2. Let $n \in \mathbb{N}$ and let D_1 and D_2 be the empirical distributions belonging to data sets D_1 and D_2 with $D_1 \coloneqq (\omega_1, \dots, \omega_n) \in \Omega^n$.

If $D_2 \coloneqq (\tilde{\omega}_1, \dots, \tilde{\omega}_n) \in \Omega^n$ is of the same size as D_1 but differs from D_1 in at most $\ell \in \mathbb{N}$ data points, which means that we have (after potentially reordering the data sets) $(\omega_1, \dots, \omega_{n-\ell}) = (\tilde{\omega}_1, \dots, \tilde{\omega}_{n-\ell})$, then

$$d_{\mathrm{tv}}(\mathrm{D}_1, \mathrm{D}_2) \leq \frac{2\ell}{n}\,.$$

Similarly, if D_2 was obtained by adding $m \in \mathbb{N}$ new data points to D_1, such that now $D_2 \coloneqq (\omega_1, \dots, \omega_n, \omega_{n+1}, \dots, \omega_{n+m}) \in \Omega^{n+m}$, then

$$d_{\mathrm{tv}}(\mathrm{D}_1, \mathrm{D}_2) \leq \frac{2m}{n+m}\,.$$

Similarly, this metric can be useful for comparing probability measures that have densities with respect to the same measure:

Example 4.2.3. Suppose that P_n, $n \in \mathbb{N}$, and P are probability measures with densities f_n, $n \in \mathbb{N}$, and f with respect to some measure μ on the measurable space $(\Omega, \mathcal{A})$. If the densities satisfy $f_n \to f$ μ-almost everywhere as $n \to \infty$, then Scheffé's Theorem (cf. Billingsley, 1995, Theorem 16.12) yields

$$d_{\mathrm{tv}}(\mathrm{P}_n, \mathrm{P}) = 2 \cdot \sup_{S \in \mathcal{A}} |\mathrm{P}_n(S) - \mathrm{P}(S)| \leq 2 \cdot \int_\Omega |f_n - f| \,\mathrm{d}\mu \to 0\,, \qquad n \to \infty\,.$$

On the other hand, the subsequent two examples show cases in which the total variation distance is not useful.

Example 4.2.4. Suppose that P is a probability measure with Lebesgue density on the measurable space $(\Omega, \mathcal{A})$ and that D_n is the empirical distribution belonging to a data set $D_n \coloneqq (\omega_1, \dots, \omega_n) \in \Omega^n$ drawn from P. Then, we have

$$d_{\mathrm{tv}}(\mathrm{P}, \mathrm{D}_n) = 2 \cdot |\mathrm{P}(\mathrm{supp}(\mathrm{D}_n)) - \mathrm{D}_n(\mathrm{supp}(\mathrm{D}_n))| = 2 \cdot |0 - 1| = 2$$

because of the finiteness of the support $\mathrm{supp}(\mathrm{D}_n)$ of D_n and the Lebesgue continuity of P. As this holds true regardless of the size n of the data set, $d_{\mathrm{tv}}(\mathrm{P}, \mathrm{D}_n)$ does not converge to 0 as $n \to \infty$ in such cases.

Example 4.2.2 showed that the total variation distance yields meaningful bounds for two of the three main types of changes in the data, in which one would naturally hope that the resulting predictors do not change by much: (arbitrarily large) measurement/reading errors in a small amount of data points, and adding a small amount of new data points. For the third such slight change of the data, namely slight changes in a large amount of the data points (for example, because of general imprecision in measuring the data), it is however easy to see from the subsequent example that d_{tv} is not as useful.

[21] Examples 4.2.2 to 4.2.4 are taken from the peer-reviewed paper Köhler and Christmann (2022).

Example 4.2.5. Let $n \in \mathbb{N}$ and let D_1 be the empirical distribution belonging to a data set $D_1 := (\omega_1, \dots, \omega_n) \in \Omega^n$. Assume that $D_2 := (\tilde{\omega}_1, \dots, \tilde{\omega}_n) \in \Omega^n$ is obtained from D_1 by just slightly perturbing **all** n data points, for example by adding the same small $\varepsilon > 0$ to all ω_i, $i = 1, \dots, n$, and let D_2 be the associated empirical distribution. Then,

$$d_{\mathrm{tv}}(\mathrm{D}_1, \mathrm{D}_2) \geq 2 \cdot |\mathrm{D}_1(\mathrm{supp}(\mathrm{D}_1)) - \mathrm{D}_2(\mathrm{supp}(\mathrm{D}_1))| = 2 \cdot |1 - 0| = 2$$

if not by chance ω_i coincides with $\tilde{\omega}_j$ for some i, j. This holds true no matter how small the perturbations in the data points are. Similarly, if not all but only some ratio $q \in (0, 1)$ of the data points is perturbed, one still obtains $d_{\mathrm{tv}}(\mathrm{D}_1, \mathrm{D}_2) \geq 2q$, no matter how small these perturbations are.

To conclude, total stability results using the total variation distance are useful for bounding the influence of slight changes in the underlying distribution (cf. Example 4.2.3) or that of perturbing or adding a small amount of data points (cf. Example 4.2.2).

They do however not yield useful bounds in the situations from Examples 4.2.4 and 4.2.5, i.e. for comparing an empirical distribution with an underlying distribution with Lebesgue density and for bounding the influence of slight changes in all data points (or at least in a significant portion of them). The former of these two negative examples is not the primary focus of considerations on total stability anyway, as total stability is mainly about the effect of small changes in the data respectively the underlying distribution and not about comparing an empirical with a theoretical SVM. The latter example, i.e. having slight measurement errors in many/all data points, however constitutes a typical situation occurring in practical problems. One would hope that such slight errors do also only lead to slight changes in the resulting predictor, but Example 4.2.5 shows that total stability results using the total variation distance can not help here.

At this point, the Wasserstein distance comes into play, which will also be used in stability results in Sections 4.3 and 4.4 and which yields meaningful bounds in this situation, cf. Example 4.2.8—and actually even in the situation from Example 4.2.4, cf. Example 4.2.10.

Definition 4.2.6 (Wasserstein Distance)**.** Let $d \in \mathbb{N}$ and let $\Omega \subseteq \mathbb{R}^d$ be separable, complete and equipped with its Borel σ-algebra. Let $\mathrm{P}_1, \mathrm{P}_2$ be probability measures on Ω. The Wasserstein distance (of order 1) between P_1 and P_2 is given by

$$d_{\mathrm{W}}(\mathrm{P}_1, \mathrm{P}_2) := \inf_{\nu} \int_{\Omega \times \Omega} ||\omega - \omega'||_1 \, \mathrm{d}\nu(\omega, \omega') ,$$

where the infimum is taken over all joint distributions ν on $\Omega \times \Omega$ with marginal distributions P_1 and P_2.[22]

Note that the Wasserstein distance can also be defined for other metrics than the one induced by $||\cdot||_1$ and for higher orders, cf. Villani (2009, Definition 6.1), but we only need this special case from Definition 4.2.6. Also note that the Wasserstein distance is also known by several other names in the literature, for example Kantorovich-Rubinstein distance or earth mover's distance, see also Villani (2009, pp. 118–119).

[22] There always exists at least one such distribution ν as it is possible to choose ν as the product measure $\mathrm{P}_1 \otimes \mathrm{P}_2$.

Further note that d_{W} defines a metric if one only allows for probability measures P which satisfy

$$\int_{\Omega} ||x - x_0||_1 \, \mathrm{dP}(x) < \infty \tag{4.4}$$

for some (and hence every) $x_0 \in \Omega$, cf. Dudley (2004, Lemma 11.8.3 and subsequent Remark). This subspace of Borel probability measures is also called *Wasserstein space (of order 1)* and denoted by $\mathcal{W}_1(\Omega)$. For probability measures P_1 and P_2 from this Wasserstein space, the Kantorovich-Rubinstein theorem yields the sometimes useful dual representation

$$d_{\mathrm{W}}(\mathrm{P}_1, \mathrm{P}_2) = \sup \left\{ \int_{\Omega} f \, \mathrm{dP}_1 - \int_{\Omega} f \, \mathrm{dP}_2 \,\middle|\, |f|_1 \leq 1 \right\} , \tag{4.5}$$

where $|f|_1$ denotes the Lipschitz constant of f, cf. Villani (2009, Remark 6.5).

More detailed introductions to Wasserstein distances including their interpretation as measuring the optimal transport cost between two measures can for example be found in Rachev and Rüschendorf (1998), Villani (2009), Panaretos and Zemel (2020).

Remark 4.2.7. Using the dual representation of the Wasserstein distance, it can be observed that both the total variation and the Wasserstein distance share the same structure of being special cases of so-called *integral probability metrics* (see Müller, 1997), which means that they can both be written as

$$d_{\bullet}(\mathrm{P}_1, \mathrm{P}_2) = \sup_{f \in \mathscr{F}_{\bullet}} \left| \int f \, \mathrm{dP}_1 - \int f \, \mathrm{dP}_2 \right| ,$$

where $d_{\bullet}$ denotes either of d_{tv} and d_{W} and $\mathscr{F}_{\bullet}$ is a suitable class of functions (see Müller, 1997, Section 5.2 for d_{tv}). Further examples of such integral probability metrics include the *Kolmogorov distance* as well as the *Dudley distance* (also known as *bounded Lipschitz distance*) and the *maximum mean discrepancy* (*MMD*). Among these, the Dudley distance possesses the nice property of metrizing weak convergence (Dudley, 2004, Theorem 11.3.3) and it additionally bears close resemblance to the Wasserstein distance because of the similarity of the associated function classes,

$$\mathscr{F}_{\mathrm{Dudley}} = \{f \,;\, |f|_1 + ||f||_{\infty} \leq 1\} \subseteq \{f \,;\, |f|_1 \leq 1\} = \mathscr{F}_{\mathrm{W}} \,.$$

Thus, d_{Dudley} can be bounded by d_{W} and convergence with respect to d_{W} hence implies weak convergence as well. Because the reverse does in general not hold true, d_{W} does in general however not metrize weak convergence but instead metrizes a slightly stronger statement, see Villani (2009, Definition 6.8 and Theorem 6.9).

MMD (see Gretton et al., 2006; Muandet et al., 2017) on the other hand is particularly interesting because it is often referred to in a way that seems very different from the definition of integral probability metrics, namely as the distance between the *kernel mean embeddings* of two probability distributions. These kernel mean embeddings can be seen as embeddings of the distributions into an RKHS and therefore transfer the idea of using RKHSs for representing single data points (cf. Section 2.1.2) to using them for representing whole distributions, see Sriperumbudur et al. (2011) for the relation between kernels that are typically used for SVMs and kernels that are typically used for such embeddings. Because of its representation as an integral probability metric (Muandet et al., 2017, Section 3.5), MMD however also has a close connection to the total variation as well as the Wasserstein distance. This connection gets especially apparent in Sriperumbudur et al. (2010, Theorem 21), where it is shown that both can be used to upper bound MMD. Lastly, Sriperumbudur et al. (2010, Theorems 23 and 24) give conditions under which MMD metrizes weak convergence, with the former theorem being further generalized by Sriperumbudur (2016, Theorem 3.2).

A major advantage that the Wasserstein distance offers over the total variation distance lies in the fact that it is also useful in the situation of Example 4.2.5. In the following, the situations from Examples 4.2.2 and 4.2.5 are combined into a single example because the Wasserstein distance is able to handle them both similarly well, yielding bounds of the same structure for both:

Example 4.2.8. Let $n \in \mathbb{N}$ and let D_1 and D_2 be the empirical distributions belonging to data sets D_1 and D_2 with $D_1 \coloneqq (\omega_1, \dots, \omega_n) \in \Omega^n$.

If $D_2 \coloneqq (\tilde{\omega}_1, \dots, \tilde{\omega}_n) \in \Omega^n$ is of the same size as D_1, then

$$d_{\mathrm{W}}(\mathrm{D}_1, \mathrm{D}_2) \leq \int_{\Omega \times \Omega} ||\omega - \tilde{\omega}||_1 \,\mathrm{d}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{(\omega_i, \tilde{\omega}_i)}\right)(\omega, \tilde{\omega}) = \frac{1}{n}\sum_{i=1}^{n} ||\omega_i - \tilde{\omega}_i||_1 \,.$$

Hence, the Wasserstein distance is small no matter whether D_2 was obtained from D_1 by changing few data points greatly (first scenario in Example 4.2.2) or by changing many data points slightly (scenario in Example 4.2.5).

Similarly, if D_2 was obtained by adding $m \in \mathbb{N}$ new data points to D_1, such that now $D_2 \coloneqq (\omega_1, \dots, \omega_n, \omega_{n+1}, \dots, \omega_{n+m}) \in (\mathcal{X} \times \mathcal{Y})^{n+m}$ (second scenario in Example 4.2.2), then

$$\begin{aligned} d_{\mathrm{W}}(\mathrm{D}_1, \mathrm{D}_2) &\leq \inf_{\mu} \int_{\Omega \times \Omega} ||\omega - \tilde{\omega}||_1 \,\mathrm{d}\left(\frac{1}{n+m}\sum_{i=1}^{n} \delta_{(\omega_i, \omega_i)} + \frac{m}{n+m}\mu\right)(\omega, \tilde{\omega}) \\ &= \frac{m}{n+m} \cdot \inf_{\mu} \int_{\Omega \times \Omega} ||\omega - \tilde{\omega}||_1 \,\mathrm{d}\mu(\omega, \tilde{\omega}) \\ &= \frac{m}{n+m} \cdot d_{\mathrm{W}}(\mathrm{D}_1, \tilde{\mathrm{D}}_2)\,, \end{aligned}$$

where $\tilde{\mathrm{D}}_2$ is the empirical distribution associated with $(\omega_{n+1}, \dots, \omega_{n+m})$ and where the infimum is taken over all joint distributions μ with marginal distributions D_1 and $\tilde{\mathrm{D}}_2$.

In both scenarios from Example 4.2.2, the bound derived for the total variation distance was stronger than that derived for the Wasserstein distance in Example 4.2.8 if the data points that differ between D_1 and D_2 differ by much (first scenario) respectively if the added data points in D_2 are very different from those already present in D_1 (second scenario). This is due to the total variation distance only being influenced by the number of points differing between D_1 and D_2 but not by how large those individual differences are. Nevertheless, the bounds derived for the Wasserstein distance can also be useful as they also become smaller if the number of differing/added points decreases. Additionally, the Wasserstein distance shows the great advantage of also yielding a nice bound in the scenario from Example 4.2.5.

Furthermore, the Wasserstein distance can also be used for the comparison between two theoretical distributions if they have densities with respect to the same measure, i.e. in the scenario from Example 4.2.3:

Example 4.2.9. Suppose that P_n, $n \in \mathbb{N}$, and P are probability measures from the Wasserstein space $\mathcal{W}_1(\Omega)$ with densities f_n, $n \in \mathbb{N}$, and f with respect to some measure μ on the measurable space $(\Omega, \mathcal{B}_\Omega)$. Assume that the densities satisfy $f_n \to f$ μ-almost everywhere as $n \to \infty$.

Defining g_n by $g_n(\omega) := ||\omega||_1 \cdot (f_n(\omega) - f(\omega))$ for $\omega \in \Omega$, one obtains that $g_n^+ \to 0$ μ-almost everywhere and that $0 \leq g_n^+(\omega) \leq ||\omega||_1 \cdot f_n(\omega)$ for all ω and n. As $\mathrm{P}_n \in \mathcal{W}_1(\Omega)$, $\omega \mapsto ||\omega||_1 \cdot f_n(\omega)$ is μ-integrable, and hence the dominated convergence theorem yields $\int g_n^+ \,\mathrm{d}\mu \to 0$ as $n \to \infty$. Analogously, $\int g_n^- \,\mathrm{d}\mu \to 0$ follows as well and one therefore obtains

$$\int_\Omega ||\omega||_1 \,\mathrm{dP}_n(\omega) - \int_\Omega ||\omega||_1 \,\mathrm{dP}(\omega) = \int_\Omega g_n(\omega)\,\mathrm{d}\mu(\omega) \to 0\,, \qquad n \to \infty\,.$$

By replacing $||\cdot||_1$ with an arbitrary bounded and continuous function, one can show in a similar way that P_n converges weakly to P. Together, these two facts show that the probability measures satisfy Villani (2009, Definition 6.8(i)), for which reason Villani (2009, Theorem 6.9) yields that

$$d_{\mathrm{W}}(\mathrm{P}_n, \mathrm{P}) \to 0\,, \qquad n \to \infty\,.$$

Lastly—even though this is not the main focus of results on total stability, as explained earlier—, the Wasserstein distance can even be useful for comparing an empirical distribution with an underlying theoretical distribution with Lebesgue density, i.e. in the scenario for which Example 4.2.4 showed that the total variation distance can not be used.

Example 4.2.10. Suppose that $\mathrm{P} \in \mathcal{W}_1(\Omega)$ is a probability measure with Lebesgue density and that D_n is the empirical distribution belonging to a data set $D_n := (\omega_1, \ldots, \omega_n) \in \Omega^n$ drawn from P. Then Panaretos and Zemel (2020, Proposition 2.2.6) yields

$$d_{\mathrm{W}}(\mathrm{P}, \mathrm{D}_n) \to 0\,, \qquad n \to \infty\,,$$

almost surely.

If P even lies in a Wasserstein space of some slightly higher order,[23] then Dereich et al. (2013, Theorem 1) and Fournier and Guillin (2015, Theorems 1 and 2) even bound the rate of convergence by $n^{-1/d}$.[24] Alas, one observes the curse of dimensionality as it is in general not possible to derive a better rate, cf. Dereich et al. (2013, Theorem 2) and Fournier and Guillin (2015, p. 709). See also Weed and Bach (2019) for related considerations in compact metric spaces.

[23] This means that $||x - x_0||_1$ is replaced by $||x - x_0||_1^q$ in condition (4.4) for some suitably chosen q slightly greater than 1.

[24] To be more precise, this rate is derived for $d \geq 3$ and the rates for $d = 1, 2$ differ slightly.

Remark 4.2.11. If Ω is assumed to be bounded, the Wasserstein distance can actually be bounded by some multiple of the total variation distance, cf. Gibbs and Su (2002, Theorem 4). Further, it is obvious that $\mathcal{W}_1(\Omega)$ contains all (Borel) probability measures on Ω in this case, for which reason the observations from Examples 4.2.9 and 4.2.10 then hold true without needing to explicitly require the probability measures to be in $\mathcal{W}_1(\Omega)$.

To conclude, for measuring the difference between two probability measures P_1 and P_2, the results on total stability of SVMs will use either

$$d_1(\mathrm{P}_1, \mathrm{P}_2) = d_{\mathrm{tv}}(\mathrm{P}_1, \mathrm{P}_2)$$

or

$$d_1(\mathrm{P}_1, \mathrm{P}_2) = d_{\mathrm{W}}(\mathrm{P}_1, \mathrm{P}_2)$$

in Definition 4.1.1. For comparing two vectors of local measures on a regionalization $\boldsymbol{\mathcal{X}} = \{\mathcal{X}_1, \dots, \mathcal{X}_A\}$, the results will use

$$d_1(\mathbf{P}_{\mathbf{1},\boldsymbol{\mathcal{X}}}, \mathbf{P}_{\mathbf{2},\boldsymbol{\mathcal{X}}}) = \max_{a\in\{1,\dots,A\}} d_\bullet(\mathrm{P}_{1,\mathcal{X}_a}, \mathrm{P}_{1,\mathcal{X}_a})$$

for sup-stability and

$$d_1(\mathbf{P}_{\mathbf{1},\boldsymbol{\mathcal{X}}}, \mathbf{P}_{\mathbf{2},\boldsymbol{\mathcal{X}}}) = \sum_{a=1}^{A} d_\bullet(\mathrm{P}_{1,\mathcal{X}_a}, \mathrm{P}_{1,\mathcal{X}_a}) \,,$$

for L_p-stability in Definitions 4.1.2 and 4.1.3, with $d_\bullet$ denoting either of d_{tv} and d_{W}.

4.2.2 Differences between Regularization Parameters

As the regularization parameters are just real numbers, it suggests itself to measure the difference between two regularization parameters λ_1 and λ_2 by just calculating the absolute value of their difference and hence choose

$$d_2(\lambda_1, \lambda_2) = |\lambda_1 - \lambda_2|$$

in Definition 4.1.1. For vectors $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ underlying localized SVMs that are both based on the same regionalization and that are hence of the same length, the results will use

$$d_2(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}) = ||\boldsymbol{\lambda_1} - \boldsymbol{\lambda_2}||_\infty$$

for sup-stability and

$$d_2(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}) = ||\boldsymbol{\lambda_1} - \boldsymbol{\lambda_2}||_1 \,,$$

for L_p-stability in Definition 4.1.3. If the regionalizations differ, Section 4.4.2 yields that this can be reduced to the case of still comparing vectors of the same length, for which reason it is still valid to use these choices.

4.2.3 Differences between Kernels

Since the kernels are functions, it suggests itself to use an analogous norm for d_3 as the one on the left hand side in Definition 4.1.1, namely

$$||\cdot||_\bullet := \begin{cases} ||\cdot||_\infty & \text{for sup-stability,} \\ ||\cdot||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)} & \text{for } L_p\text{-stability.} \end{cases}$$

To be more specific, the results need a slightly adapted version of $||\cdot||_\bullet$, namely

$$d_3(k_1, k_2) = ||k_1 - k_2||_\bullet + \sqrt{||k_1 - k_2||_\bullet}\,. \tag{4.6}$$

Remark 4.2.12. Even though $\sqrt{||\cdot||_\bullet}$ is not a norm, it is easy to see that it still induces a metric. Hence, d_3 is also a metric since it is the sum of two metrics.[25]

Note that $k_1 - k_2$ is generally not a kernel. Therefore, if one chooses $||\cdot||_\bullet$ in (4.6) as the supremum norm, $||k_1 - k_2||_\infty$ denotes the general supremum norm of a function instead of the special definition of $||\cdot||_\infty$ for kernels stated in (2.2).

In the following, two examples are given in order to illustrate the behavior of this supremum norm $||k_1 - k_2||_\infty$ as well as the whole distance d_3 from (4.6).[26] First of all, Example 4.2.13 compares two Gaussian kernels with different bandwidths and examines how this difference influences $||k_1 - k_2||_\infty$. Such a difference can for example arise from two practitioners using slightly different grids for their cross-validation schemes or from them having slightly different data at hand cf. Section 4.1.

Example 4.2.13. Let $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$ and let k_γ be the Gaussian RBF kernel with bandwidth $\gamma > 0$ on $\mathcal{X}$, cf. Example 2.1.12. For $\gamma_1, \gamma_2 > 0$ and $x, x' \in \mathcal{X}$, we then have

$$\begin{aligned} |k_{\gamma_1}(x,x') - k_{\gamma_2}(x,x')| &= \left| \exp\left(-\frac{||x-x'||_2^2}{\gamma_1^2}\right) - \exp\left(-\frac{||x-x'||_2^2}{\gamma_2^2}\right) \right| \\ &= \left| \exp\left(-\frac{\left|\left|\frac{x}{\gamma_2} - \frac{x'}{\gamma_2}\right|\right|_2^2}{\gamma_1^2/\gamma_2^2}\right) - \exp\left(-\frac{\left|\left|\frac{x}{\gamma_2} - \frac{x'}{\gamma_2}\right|\right|_2^2}{1}\right) \right| \\ &= \left| k_{\gamma_1/\gamma_2}\left(\frac{x}{\gamma_2}, \frac{x'}{\gamma_2}\right) - k_1\left(\frac{x}{\gamma_2}, \frac{x'}{\gamma_2}\right) \right| \end{aligned}$$

and analogously

$$\left| k_{\gamma_1/\gamma_2}(x,x') - k_1(x,x') \right| = |k_{\gamma_1}(\gamma_2 x, \gamma_2 x') - k_{\gamma_2}(\gamma_2 x, \gamma_2 x')|\ .$$

[25] If $||\cdot||_\bullet = ||\cdot||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)}$, then $||\cdot||_\bullet$ of course strictly speaking only defines a seminorm, for which reason the distance d_3 between two distinct kernels can be zero. As usual, one formally needs to switch to equivalence classes of kernels in order for $||\cdot||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)}$ to actually define a norm, which we however omit for ease of notation.

[26] These two Examples 4.2.13 and 4.2.14 are slightly adapted versions of Examples 4 and 5 from the peer-reviewed paper Köhler and Christmann (2022).

γ_1/γ_2	1	1.01	1.05	1.1	1.5	2
$\lvert\lvert k_{\gamma_1} - k_{\gamma_2}\rvert\rvert_\infty$	0.000	0.007	0.036	0.070	0.290	0.472
$d_3(k_{\gamma_1}, k_{\gamma_2})$	0.000	0.093	0.225	0.335	0.829	1.160

Table 4.2.1: Ratio between the bandwidths γ_1 and γ_2 of two Gaussian RBF kernels, as well as the resulting value of $||k_{\gamma_1} - k_{\gamma_2}||_\infty$ and the corresponding value of $d_3(k_{\gamma_1}, k_{\gamma_2})$ for d_3 as defined in (4.6)

As $x, x' \in \mathcal{X}$ implies that $x/\gamma_2, x'/\gamma_2, \gamma_2 x, \gamma_2 x' \in \mathcal{X}$ as well (because $\mathcal{X} = \mathbb{R}^d$), we hence obtain

$$||k_{\gamma_1} - k_{\gamma_2}||_\infty = \left|\left|k_{\gamma_1/\gamma_2} - k_1\right|\right|_\infty .$$

Therefore, changing the bandwidth from γ_1 to γ_2 results in a value of $||k_{\gamma_1} - k_{\gamma_2}||_\infty$ which depends only on the ratio between γ_1 and γ_2. We computed $||k_{\gamma_1} - k_{\gamma_2}||_\infty$ for some such ratios γ_1/γ_2 and collected the results in Table 4.2.1. Additionally, that table also includes the corresponding values of d_3 as defined in (4.6).

Similarly to Example 4.2.13, one might also be interested in the effect on the SVM of not slightly changing the bandwidth of the Gaussian kernel, but instead for example be interested in the effect of switching to a suiting Wendland kernel (cf. Wendland, 2005, Definition 9.11 and the subsequent results), which possesses the numerical advantage of having a compact support:

Example 4.2.14. Let $\mathcal{X} = \mathbb{R}^5$ (other dimensions can be analyzed analogously and yield similar results). Let k_γ be the Gaussian kernel with bandwidth $\gamma > 0$ on $\mathcal{X}$ and let k_W be the normalized Wendland kernel on $\mathcal{X}$ defined by $k_\mathrm{W}(x, x') := \psi_{6,3}(||x - x'||_2)$ with $\psi_{6,3}$ as in Chernih et al. (2014, Theorem 3.3, with $\alpha := \gamma^{-2}$).[27] This results in $||k_\gamma - k_\mathrm{W}||_\infty \approx 0.0037$, and thus $d_3(k_\gamma, k_\mathrm{W}) \approx 0.0650$ with d_3 as in (4.6), being quite small and hence the corresponding SVMs closely resembling each other because of their stability with respect to changes in the kernel which is shown in Proposition 4.3.14.

For vectors $\boldsymbol{k_1}$ and $\boldsymbol{k_2}$ of kernels underlying localized SVMs that are both based on the same regionalization of size $A \in \mathbb{N}$ and that are hence of the same length, the results will use

$$d_3(\boldsymbol{k_1}, \boldsymbol{k_2}) = \max_{a \in \{1,\dots,A\}} \left(||k_{1,a} - k_{2,a}||_\infty + \sqrt{||k_{1,a} - k_{2,a}||_\infty} \right)$$

for sup-stability and

$$d_3(\boldsymbol{k_1}, \boldsymbol{k_2}) = \sum_{a=1}^{A} \left(||k_{1,a} - k_{2,a}||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)} + \sqrt{||k_{1,a} - k_{2,a}||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)}} \right)$$

for L_p-stability in Definition 4.1.3. Again, if the regionalizations differ, Section 4.4.2 yields that this can be reduced to the case of still comparing vectors of the same length, for which reason it is still valid to use one of the above two choices.

[27] Note that the notation used in that paper differs from the one used by Wendland (2005), such that the function $\phi_{6,3}$ used in the definition of $\psi_{6,3}$ in the mentioned theorem corresponds to $\phi_{5,3}$ in the notation of Wendland (2005).

4.2.4 Differences between Regionalizations

Assume throughout this section that the two regionalizations $\boldsymbol{\mathcal{X}_1} = \{\mathcal{X}_{1,1}, \dots, \mathcal{X}_{1,A_1}\}$ and $\boldsymbol{\mathcal{X}_2} = \{\mathcal{X}_{2,1}, \dots, \mathcal{X}_{2,A_2}\}$ that are compared are both partitioning regionalizations of $\mathcal{X}$, and that Q is a probability measure on $\mathcal{X}$ (in the results, this will be P_i^X, $i = 1, 2$). Them being partitioning regionalizations will also be required in Section 4.4.2 on total stability of localized SVMs, which is the only section in which the difference between two regionalizations comes into play. It is not straightforward how to quantify this difference. However, in Section 4.4.2, two different quantities connected to the regionalizations arise that are relevant for bounding the difference between two localized SVMs and which both in some sense characterize the difference between two regionalizations. Both of these quantities can be defined intersection-wise, that is, for each $b \in \{1, \dots, B\}$ and corresponding region of the combined regionalization $\boldsymbol{\mathcal{X}^*_{1,2}} = \{\mathcal{X}^*_1, \dots, \mathcal{X}^*_B\}$ separately.

First off,

$$\left|\mathrm{Q}(\mathcal{X}_{1,a(1,b)}) - \mathrm{Q}(\mathcal{X}_{2,a(2,b)})\right|$$

shows to play a role in the bound. This difference in size (according to the probability measure Q) has to be accounted for since a large difference could possibly lead to the local SVM in the smaller of the two regions being fitted much closer to its underlying data than its counterpart and the two local SVMs therefore greatly differing on the intersection $\mathcal{X}^*_b$.

Secondly,

$$\mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}^*_b) \cdot \left(1 - \mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}^*_b)\right)$$

plays a role as well for $i = 1, 2$. If each region from $\boldsymbol{\mathcal{X}_1}$ closely coincides with a region from $\boldsymbol{\mathcal{X}_2}$, then the probability $\mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}^*_b)$ appearing in this term will be either close to 0 or close to 1 for all $b \in \{1, \dots, B\}$ and $i \in \{1, 2\}$ and the regionalizations are therefore similar to each other in the sense of this second criterion.

In the bound in Section 4.4.2, these two quantities need to be combined in the following way:

$$\begin{aligned}\xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) \coloneqq{}& \left|\mathrm{Q}(\mathcal{X}_{1,a(1,b)}) - \mathrm{Q}(\mathcal{X}_{2,a(2,b)})\right| \\ &+ \max\left\{\mathrm{Q}(\mathcal{X}_{1,a(1,b)}), \mathrm{Q}(\mathcal{X}_{2,a(2,b)})\right\} \\ &\cdot \sum_{i=1}^{2} \Bigg(\frac{1}{2} \cdot \mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}^*_b) \cdot \left(1 - \mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}^*_b)\right) \\ &\qquad + \sqrt{\mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}^*_b) \cdot \left(1 - \mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}^*_b)\right)} \Bigg), \end{aligned} \tag{4.7}$$

for all $b \in \{1, \dots, B\}$.

Summing this over all intersections from $\boldsymbol{\mathcal{X}^*_{1,2}}$, one can choose

$$d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) = \sum_{b=1}^{B} \xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2})$$

in Definition 4.1.3.

In the following, we give a few examples on the behavior of $d_{4,\mathrm{Q}}$ in different situations.

Example 4.2.15. Let $\boldsymbol{\mathcal{X}_1} = \{\mathcal{X}_{1,1}, \dots, \mathcal{X}_{1,A_1}\}$ be a partitioning regionalization of size $A_1 \in \mathbb{N}$. Let $\boldsymbol{\mathcal{X}_2} = \{\mathcal{X}_{2,1}, \dots, \mathcal{X}_{2,A_1+1}\}$ be a partitioning regionalization of size $A_1 + 1$ that satisfies $\mathcal{X}_{2,a} = \mathcal{X}_{1,a}$ for all $a \in \{1, \dots, A_1 - 1\}$, such that the only difference between the two regionalizations is that $\mathcal{X}_{1,A_1}$ was split in two, which results in $\boldsymbol{\mathcal{X}^*_{1,2}} = \boldsymbol{\mathcal{X}_2}$. Denoting $q \coloneqq \mathrm{Q}(\mathcal{X}_{1,A_1})$, we have $\mathrm{Q}(\mathcal{X}_{2,A_1}) = \theta q$ and $\mathrm{Q}(\mathcal{X}_{2,A_1+1}) = (1-\theta)q$ for some $\theta \in (0,1)$.[28] This yields

$$\begin{aligned}
\xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) &= 0 \qquad \forall\, b \in \{1, \dots, A_1 - 1\}\,, \\
\xi_{\mathrm{Q},A_1}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) &= |q - \theta q| \\
&\quad + \max\{q, \theta q\} \cdot \left(\frac{1}{2}\theta(1-\theta) + \sqrt{\theta(1-\theta)} + \frac{1}{2} \cdot 1 \cdot 0 + \sqrt{1 \cdot 0}\right) \\
&= q \cdot \left((1-\theta) + \frac{\theta(1-\theta)}{2} + \sqrt{\theta(1-\theta)}\right)\,, \\
\xi_{\mathrm{Q},A_1+1}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) &= |q - (1-\theta)q| \\
&\quad + \max\{q, (1-\theta)q\} \left(\frac{1}{2}(1-\theta)\theta + \sqrt{(1-\theta)\theta} + \frac{1}{2} \cdot 1 \cdot 0 + \sqrt{1 \cdot 0}\right) \\
&= q \cdot \left(\theta + \frac{\theta(1-\theta)}{2} + \sqrt{\theta(1-\theta)}\right)\,,
\end{aligned}$$

and hence

$$d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) = q \cdot \left(1 + \theta(1-\theta) + 2\sqrt{\theta(1-\theta)}\right)\,.$$

Example 4.2.16. Let $\boldsymbol{\mathcal{X}_1} = \{\mathcal{X}_{1,1}, \dots, \mathcal{X}_{1,A_1}\}$ be a partitioning regionalization of size $A_1 \in \mathbb{N}$. Let $\boldsymbol{\mathcal{X}_2} = \{\mathcal{X}_{2,1}, \dots, \mathcal{X}_{2,A_1}\}$ be a second partitioning regionalization of size A_1 that satisfies $\mathcal{X}_{2,a} = \mathcal{X}_{1,a}$ for all $a \in \{1, \dots, A_1 - 2\}$, and $\mathcal{X}_{1,A_1-1} \subset \mathcal{X}_{2,A_1-1}$ (and hence $\mathcal{X}_{1,A_1} \supset \mathcal{X}_{2,A_1}$), such that the only difference between the two regionalizations is that a part of $\mathcal{X}_{1,A_1}$ was moved to $\mathcal{X}_{2,A_1-1}$, which results in

$$\boldsymbol{\mathcal{X}^*_{1,2}} = \{\mathcal{X}_{1,1}, \dots, \mathcal{X}_{1,A_1-2}, \underbrace{\mathcal{X}_{1,A_1-1}}_{=\mathcal{X}^*_{A_1-1}}, \underbrace{\mathcal{X}_{1,A_1} \cap \mathcal{X}_{2,A_1-1}}_{=\mathcal{X}^*_{A_1}}, \underbrace{\mathcal{X}_{2,A_1}}_{=\mathcal{X}^*_{A_1+1}}\}\,.$$

Assume that $\mathrm{Q}(\mathcal{X}_{1,A_1-1}) = \mathrm{Q}(\mathcal{X}_{1,A_1}) \eqqcolon q$ to slightly simplify the notation and calculations. We hence have $\mathrm{Q}(\mathcal{X}_{2,A_1-1}) = (1+\theta)q$ and $\mathrm{Q}(\mathcal{X}_{2,A_1}) = (1-\theta)q$ for some $\theta \in (0,1)$.[29] This

[28] We assume that $\mathrm{Q}(\mathcal{X}_{2,A_1}), \mathrm{Q}(\mathcal{X}_{2,A_1+1}) > 0$ because we would trivially obtain $d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) = 0$ otherwise.

[29] We assume that $\mathrm{Q}(\mathcal{X}_{2,A_1-1}), \mathrm{Q}(\mathcal{X}_{2,A_1}) > 0$ because we would trivially obtain $d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) = 0$ otherwise.

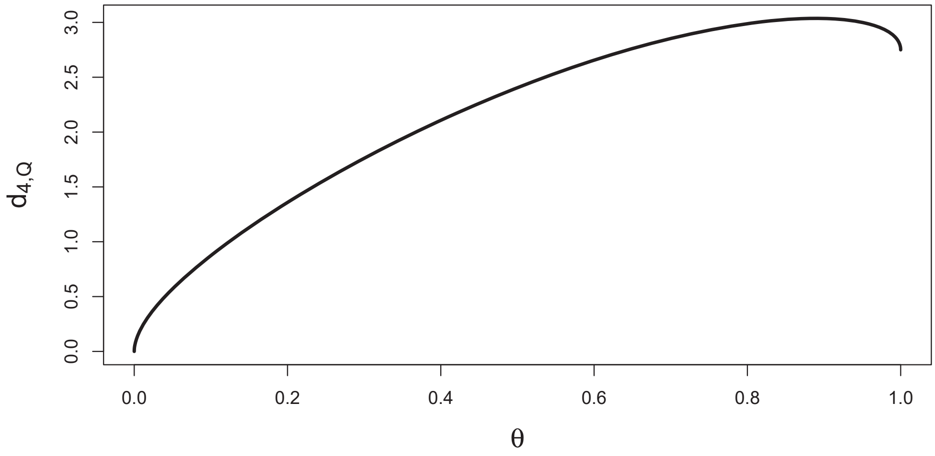

Figure 4.2.1: Behavior of $d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2})$ as a function of θ in the situation of Example 4.2.16 for $q = \frac{1}{2}$. Please note that $d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2})$ is decreasing for values of θ close to 1.

yields

$$\begin{aligned}
\xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) &= 0 \qquad \forall\, b \in \{1, \dots, A_1 - 2\}\,, \\
\xi_{\mathrm{Q},A_1-1}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) &= q \cdot \left(\theta + \frac{\theta}{2(1+\theta)} + \sqrt{\theta}\right)\,, \\
\xi_{\mathrm{Q},A_1}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) &= q \cdot \left(\theta + \frac{\theta(1-\theta)(1+\theta)}{2} + (1+\theta)\sqrt{\theta(1-\theta)} + \frac{\theta}{2(1+\theta)} + \sqrt{\theta}\right)\,, \\
\xi_{\mathrm{Q},A_1+1}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) &= q \cdot \left(\theta + \frac{\theta(1-\theta)}{2} + \sqrt{\theta(1-\theta)}\right)\,,
\end{aligned}$$

and hence

$$d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) = q \cdot \left(3\theta + \frac{\theta}{1+\theta} + 2\sqrt{\theta} + \frac{\theta(1-\theta)(2+\theta)}{2} + (2+\theta)\sqrt{\theta(1-\theta)}\right)\,,$$

whose behavior can be seen in Figure 4.2.1 for the case $q = \frac{1}{2}$, i.e. for the case of $\boldsymbol{\mathcal{X}_1}$ and $\boldsymbol{\mathcal{X}_2}$ both consisting of only two regions (for different q, the principal behavior of $d_{4,\mathrm{Q}}$ depending on θ does of course not change, but all values get scaled according to q).

Example 4.2.17. Let $\boldsymbol{\mathcal{X}_1} = \{\mathcal{X}_{1,1}, \dots, \mathcal{X}_{1,A_1}\}$ be a partitioning regionalization of size $A_1 \in \mathbb{N}$ satisfying $\mathrm{Q}(\mathcal{X}_{1,a}) = 1/A_1$ for all $a \in \{1, \dots, A_1\}$. Let $\boldsymbol{\mathcal{X}_2} = \{\mathcal{X}_{2,1}, \dots, \mathcal{X}_{2,A_1}\}$ be a second partitioning regionalization of size A_1 that satisfies

$$\begin{aligned}
\mathrm{Q}(\mathcal{X}_{1,a} \cap \mathcal{X}_{2,a}) &= \frac{1-\theta}{A_1} && \forall\, a \in \{1, \dots, A_1\}\,, \\
\mathrm{Q}(\mathcal{X}_{1,a} \cap \mathcal{X}_{2,a+1}) &= \frac{\theta}{A_1} && \forall\, a \in \{1, \dots, A_1 - 1\}\,, \\
\mathrm{Q}(\mathcal{X}_{1,A_1} \cap \mathcal{X}_{2,1}) &= \frac{\theta}{A_1}
\end{aligned}$$

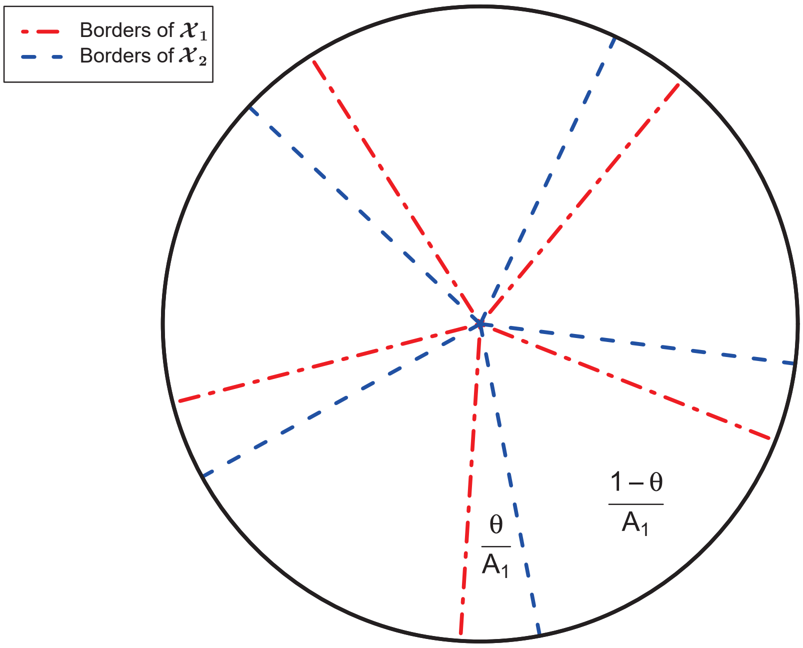

Figure 4.2.2: Visualization of situation from Example 4.2.17 for a two-dimensional $\mathcal{X}$ that is assumed to be equipped with a uniform distribution Q.

(and hence also $\mathrm{Q}(\mathcal{X}_{2,a}) = 1/A_1$ for all $a \in \{1, \dots, A_1\}$) for some $\theta \in (0, 1)$, which can be interpreted as the borders between the regions all being moved by a share of θ in the probability mass of the regions, see Figure 4.2.2 for a visualization for a two-dimensional $\mathcal{X}$. This results in

$$\boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}} = \{\underbrace{\mathcal{X}_{1,1} \cap \mathcal{X}_{2,1}}_{=\mathcal{X}^*_1}, \mathcal{X}_{1,1} \cap \mathcal{X}_{2,2}, \mathcal{X}_{1,2} \cap \mathcal{X}_{2,2}, \dots, \mathcal{X}_{1,A_1} \cap \mathcal{X}_{2,A_1}, \underbrace{\mathcal{X}_{1,A_1} \cap \mathcal{X}_{2,1}}_{=\mathcal{X}^*_{2A_1}}\}$$

and yields

$$\xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}}_{\mathbf{1}}, \boldsymbol{\mathcal{X}}_{\mathbf{2}}) = \frac{1}{A_1} \cdot \left(\theta(1-\theta) + 2\sqrt{\theta(1-\theta)}\right) \qquad \forall\, b \in \{1, \dots, 2A_1\}$$

and hence

$$d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}}_{\mathbf{1}}, \boldsymbol{\mathcal{X}}_{\mathbf{2}}) = 2\theta(1-\theta) + 4\sqrt{\theta(1-\theta)}$$

independently of A_1.

Lastly note that, even though the preceding examples showed that $d_{4,\mathrm{Q}}$ seems to be sensible for quantifying the difference between two regionalizations because it takes small values when one would intuitively describe the regionalizations as being similar and vice versa, $d_{4,\mathrm{Q}}$ does in general *not* define a metric as can be seen in the following.

Example 4.2.18. Let $\mathcal{X} = \mathbb{R}$ and $\mathrm{Q} = \mathcal{U}(0,4)$ be the uniform distribution on $(0,4)$. Let

$$\boldsymbol{\mathcal{X}_1} = \{\mathcal{X}\}\,, \quad \boldsymbol{\mathcal{X}_2} = \{(-\infty,2),[2,\infty)\}\,, \quad \boldsymbol{\mathcal{X}_3} = \{(-\infty,2),[2,3),[3,\infty)\}\,.$$

From Example 4.2.15, we immediately obtain

$$d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1},\boldsymbol{\mathcal{X}_2}) = 1\cdot\left(1+\frac{1}{2}\cdot\frac{1}{2}+2\sqrt{\frac{1}{2}\cdot\frac{1}{2}}\right) = \frac{9}{4}\,,$$
$$d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_2},\boldsymbol{\mathcal{X}_3}) = \frac{1}{2}\cdot\left(1+\frac{1}{2}\cdot\frac{1}{2}+2\sqrt{\frac{1}{2}\cdot\frac{1}{2}}\right) = \frac{9}{8}\,.$$

$d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1},\boldsymbol{\mathcal{X}_3})$ can be obtained in a similar way: The structure of $\boldsymbol{\mathcal{X}_1}$ and $\boldsymbol{\mathcal{X}_3}$ immediately yields $\boldsymbol{\mathcal{X}^*_{1,3}} = \boldsymbol{\mathcal{X}_3}$ and with that

$$\begin{aligned}\xi_{\mathrm{Q},1}(\boldsymbol{\mathcal{X}_1},\boldsymbol{\mathcal{X}_3}) &= \left|1-\frac{1}{2}\right| + \max\left\{1,\frac{1}{2}\right\}\cdot\left(\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}+\sqrt{\frac{1}{2}\cdot\frac{1}{2}}+\frac{1}{2}\cdot 1\cdot 0+\sqrt{1\cdot 0}\right)\\ &= \frac{9}{8}\,,\\ \xi_{\mathrm{Q},2}(\boldsymbol{\mathcal{X}_1},\boldsymbol{\mathcal{X}_3}) &= \left|1-\frac{1}{4}\right| + \max\left\{1,\frac{1}{4}\right\}\cdot\left(\frac{1}{2}\cdot\frac{1}{4}\cdot\frac{3}{4}+\sqrt{\frac{1}{4}\cdot\frac{3}{4}}+\frac{1}{2}\cdot 1\cdot 0+\sqrt{1\cdot 0}\right)\\ &= \frac{27+8\sqrt{3}}{32} = \xi_{\mathrm{Q},3}(\boldsymbol{\mathcal{X}_1},\boldsymbol{\mathcal{X}_3})\,,\end{aligned}$$

which altogether yields

$$d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1},\boldsymbol{\mathcal{X}_3}) = \frac{90+16\sqrt{3}}{32} \approx 3.68 > 3.375 = d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_1},\boldsymbol{\mathcal{X}_2}) + d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}_2},\boldsymbol{\mathcal{X}_3})\,.$$

Hence, $d_{4,\mathrm{Q}}$ does not satisfy the triangle inequality and does therefore not define a metric.

Remark 4.2.19. Apart from the triangle inequality, $d_{4,\mathrm{Q}}$ satisfies all properties of a metric if we define it on equivalence classes of regionalizations instead of on regionalizations themselves, where two regionalizations $\boldsymbol{\mathcal{X}_1} = \{\mathcal{X}_{1,1},\ldots,\mathcal{X}_{1,A_1}\}$ and $\boldsymbol{\mathcal{X}_2} = \{\mathcal{X}_{2,1},\ldots,\mathcal{X}_{2,A_2}\}$ are called equivalent, $\boldsymbol{\mathcal{X}_1} \equiv \boldsymbol{\mathcal{X}_2}$, if

$$\begin{aligned}&\forall\, a_1 \in \{1,\ldots,A_1\}\,\forall\, a_2 \in \{1,\ldots,A_2\} \text{ with } \mathcal{X}_{1,a_1}\cap\mathcal{X}_{2,a_2} \neq \emptyset :\\ &\qquad \mathrm{Q}(\mathcal{X}_{1,a_1}) = \mathrm{Q}(\mathcal{X}_{2,a_2}) \quad\text{and}\quad \mathrm{Q}(\mathcal{X}_{1,a_1}\cap\mathcal{X}_{2,a_2}) \in \{0,\mathrm{Q}(\mathcal{X}_{1,a_1})\}\;. \end{aligned}\tag{4.8}$$

This does indeed define an equivalence relation: Reflexivity and symmetry are trivial, so only transitivity remains to show. For this, assume that $\boldsymbol{\mathcal{X}_1} \equiv \boldsymbol{\mathcal{X}_2}$ and $\boldsymbol{\mathcal{X}_2} \equiv \boldsymbol{\mathcal{X}_3}$ and let $a_1 \in \{1,\ldots,A_1\}$ and $a_3 \in \{1,\ldots,A_3\}$ be such that $\mathcal{X}_{1,a_1}\cap\mathcal{X}_{3,a_3} \neq \emptyset$. Then, $\boldsymbol{\mathcal{X}_2}$ being a regionalization implies that there exists $a_2 \in \{1,\ldots,A_2\}$ such that

$$(\mathcal{X}_{1,a_1}\cap\mathcal{X}_{3,a_3})\cap\mathcal{X}_{2,a_2} \neq \emptyset\,,$$

and hence, because of $\boldsymbol{\mathcal{X}}_{\mathbf{1}} \equiv \boldsymbol{\mathcal{X}}_{\mathbf{2}}$ and $\boldsymbol{\mathcal{X}}_{\mathbf{2}} \equiv \boldsymbol{\mathcal{X}}_{\mathbf{3}}$,

$$\mathrm{Q}(\mathcal{X}_{1,a_1}) = \mathrm{Q}(\mathcal{X}_{2,a_2}) = \mathrm{Q}(\mathcal{X}_{3,a_3})\,,$$

which yields the first property from (4.8). For proving the second property, assume without loss of generality that $\mathrm{Q}(\mathcal{X}_{1,a_1} \cap \mathcal{X}_{3,a_3}) > 0$, and choose

$$\tilde{a}_2 \coloneqq \arg \max_{a_2 \in \{1,\ldots,A_2\}} \mathrm{Q}\Big((\mathcal{X}_{1,a_1} \cap \mathcal{X}_{3,a_3}) \cap \mathcal{X}_{2,a_2}\Big)\,.$$

This implies, for $i = 1, 3$,

$$\mathrm{Q}(\mathcal{X}_{i,a_i} \cap \mathcal{X}_{2,\tilde{a}_2}) \geq \mathrm{Q}\Big((\mathcal{X}_{1,a_1} \cap \mathcal{X}_{3,a_3}) \cap \mathcal{X}_{2,\tilde{a}_2}\Big) > 0$$

and hence, because of $\boldsymbol{\mathcal{X}}_{\mathbf{1}} \equiv \boldsymbol{\mathcal{X}}_{\mathbf{2}}$ and $\boldsymbol{\mathcal{X}}_{\mathbf{2}} \equiv \boldsymbol{\mathcal{X}}_{\mathbf{3}}$,

$$\begin{aligned}
&\mathrm{Q}(\mathcal{X}_{i,a_i} \cap \mathcal{X}_{2,\tilde{a}_2}) = \mathrm{Q}(\mathcal{X}_{i,a_i}) = \mathrm{Q}(\mathcal{X}_{2,\tilde{a}_2})\,,\\
&\mathrm{Q}(\mathcal{X}_{i,a_i} \cap \complement\mathcal{X}_{2,\tilde{a}_2}) = \mathrm{Q}(\mathcal{X}_{i,a_i}) - \mathrm{Q}(\mathcal{X}_{i,a_i} \cap \mathcal{X}_{2,\tilde{a}_2}) = 0\,,\\
&\mathrm{Q}(\complement\mathcal{X}_{i,a_i} \cap \mathcal{X}_{2,\tilde{a}_2}) = \mathrm{Q}(\mathcal{X}_{2,\tilde{a}_2}) - \mathrm{Q}(\mathcal{X}_{i,a_i} \cap \mathcal{X}_{2,\tilde{a}_2}) = 0\,.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{Q}(\mathcal{X}_{1,a_1} \cap \mathcal{X}_{3,a_3}) &= \mathrm{Q}\Big((\mathcal{X}_{1,a_1} \cap \mathcal{X}_{3,a_3}) \cap \mathcal{X}_{2,\tilde{a}_2}\Big) + \underbrace{\mathrm{Q}\Big((\mathcal{X}_{1,a_1} \cap \mathcal{X}_{3,a_3}) \cap \complement\mathcal{X}_{2,\tilde{a}_2}\Big)}_{\leq \mathrm{Q}(\mathcal{X}_{1,a_1} \cap \complement\mathcal{X}_{2,\tilde{a}_2}) = 0}\\
&= \mathrm{Q}(\mathcal{X}_{1,a_1} \cap \mathcal{X}_{2,\tilde{a}_2}) - \underbrace{\mathrm{Q}(\mathcal{X}_{1,a_1} \cap \mathcal{X}_{2,\tilde{a}_2} \cap \complement\mathcal{X}_{3,a_3})}_{\leq \mathrm{Q}(\mathcal{X}_{2,\tilde{a}_2} \cap \complement\mathcal{X}_{3,a_3}) = 0}\\
&= \mathrm{Q}(\mathcal{X}_{1,a_1})\,,
\end{aligned}$$

which finally yields $\boldsymbol{\mathcal{X}}_{\mathbf{1}} \equiv \boldsymbol{\mathcal{X}}_{\mathbf{3}}$.

Now, if $\boldsymbol{\mathcal{X}}_{\mathbf{1}}$ and $\boldsymbol{\mathcal{X}}_{\mathbf{2}}$ are element of the same equivalence class $[\boldsymbol{\mathcal{X}}_{\mathbf{1}}]$, then we have by definition that $\mathrm{Q}(\mathcal{X}_{1,a(1,b)}) = \mathrm{Q}(\mathcal{X}_{2,a(2,b)})$ and $\mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}^*_b) \in \{0,1\}$ for $i = 1, 2$ and $b = 1, \ldots, B$, and hence $\xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}}_{\mathbf{1}}, \boldsymbol{\mathcal{X}}_{\mathbf{2}}) = 0$ for $b = 1, \ldots, B$, where B denotes the number of regions in the combined regionalization $\boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}} = \{\mathcal{X}^*_1, \ldots, \mathcal{X}^*_B\}$. Thus,

$$d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}}_{\mathbf{1}}, \boldsymbol{\mathcal{X}}_{\mathbf{2}}) = \sum_{b=1}^{B} \xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}}_{\mathbf{1}}, \boldsymbol{\mathcal{X}}_{\mathbf{2}}) = 0$$

in this case. If on the other hand $[\boldsymbol{\mathcal{X}}_{\mathbf{1}}] \neq [\boldsymbol{\mathcal{X}}_{\mathbf{2}}]$, there exists at least one $b_0 \in \{1, \ldots, B\}$ such that $\mathrm{Q}(\mathcal{X}_{1,a(1,b_0)}) \neq \mathrm{Q}(\mathcal{X}_{2,a(2,b_0)})$ or $\mathrm{Q}_{\mathcal{X}_{i,a(i,b_0)}}(\mathcal{X}^*_{b_0}) \notin \{0,1\}$ and hence—as this trivially implies that $\max\{\mathrm{Q}(\mathcal{X}_{1,a(1,b_0)}), \mathrm{Q}(\mathcal{X}_{2,a(2,b_0)})\} > 0$ holds true—also $\xi_{\mathrm{Q},b_0}(\boldsymbol{\mathcal{X}}_{\mathbf{1}}, \boldsymbol{\mathcal{X}}_{\mathbf{2}}) > 0$. As $\xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}}_{\mathbf{1}}, \boldsymbol{\mathcal{X}}_{\mathbf{2}}) \geq 0$ for all $b \in \{1, \ldots, B\}$, we have

$$d_{4,\mathrm{Q}}(\boldsymbol{\mathcal{X}}_{\mathbf{1}}, \boldsymbol{\mathcal{X}}_{\mathbf{2}}) = \sum_{b=1}^{B} \xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}}_{\mathbf{1}}, \boldsymbol{\mathcal{X}}_{\mathbf{2}}) \geq \xi_{\mathrm{Q},b_0}(\boldsymbol{\mathcal{X}}_{\mathbf{1}}, \boldsymbol{\mathcal{X}}_{\mathbf{2}}) > 0$$

in this case. As symmetry is trivially satisfied by $d_{4,\mathrm{Q}}$, it does indeed satisfy all properties of a metric on these equivalence classes of regionalizations except for the triangle inequality.

4.3 Total Stability of Support Vector Machines

In this section, total stability of SVMs, as it was defined in Definition 4.1.1, is derived based on the distance measures from Section 4.2. As intermediate steps, Sections 4.3.1 to 4.3.3 consider stability with respect to only one component of the triple (P, λ, k) changing at a time, before these results are finally combined in Section 4.3.4 to actually derive total stability.

In Sections 4.3.1 and 4.3.2, only sup-stability but not L_p-stability is investigated. As

$$||g||_{L_p(\mathrm{Q})} \leq ||g||_\infty \tag{4.9}$$

for all probability measures Q on $\mathcal{X}$, $p \in [1, \infty)$ and measurable functions g, the corresponding results can still be used to derive total L_p-stability in Section 4.3.4.

All main results from this section impose the following assumptions on loss function, probability measures, regularization parameters and kernels:

Assumption 4.3.1. Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex, Lipschitz continuous loss function and let $L^\star$ be its shifted version. Let $\mathrm{P}, \mathrm{P}_1, \mathrm{P}_2$ be probability measures on $\mathcal{X} \times \mathcal{Y}$, $\lambda, \lambda_1, \lambda_2 > 0$, and k, k_1, k_2 be bounded and measurable kernels on $\mathcal{X}$ with separable RKHSs H, H_1, H_2.

Note that the requested separability of the RKHSs is always satisfied for continuous kernels, cf. Lemma 2.1.10(iii). The results using the Wasserstein distance to measure the difference between two probability measures additionally require the following:

Assumption 4.3.2. Let $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$. Let L be distance-based with representing function ψ such that ψ is differentiable and ψ as well as its derivative ψ' are both Lipschitz continuous. Let $\tilde{\mathcal{X}} \supseteq \mathcal{X}$ with $\tilde{\mathcal{X}} \subseteq \mathbb{R}^d$ be open and assume that $k = \tilde{k}|_{\mathcal{X}\times\mathcal{X}}$ for a kernel $\tilde{k}$ on $\tilde{\mathcal{X}}$ that can be written as $\tilde{k}(x, x') = \varphi(||x - x'||_2^2)$ for all $x, x' \in \tilde{\mathcal{X}}$, where $\varphi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a twice continuously differentiable function satisfying $\varphi(0) - \varphi(r) \leq \frac{c_k^2}{2} \cdot r$ for all $r \geq 0$ for some $c_k \geq 0$. Analogously for k_1, k_2 with open sets $\tilde{\mathcal{X}_1}, \tilde{\mathcal{X}_2} \supseteq \mathcal{X}$, functions φ_1, φ_2 and constants $c_{k_1}, c_{k_2} \geq 0$.

Note that these additional assumptions on the kernels are satisfied by popular kernels such as the Gaussian RBF kernels, cf. Corollary 4.3.7.

Several parts of this section coincide with parts of the peer-reviewed paper Köhler and Christmann (2022) that was published in *Journal of Machine Learning Research*: Proposition 4.3.3 as well as the results from Section 4.3.3 appear in Appendix A of that paper. Proposition 4.3.10 is similar to a further result from the same Appendix A, but improves the derived bound by a factor of 2, which makes the bound asymptotically sharp. In Section 4.3.4, the parts using the total variation distance have already been published similarly in Section 2 of that paper, but the results have been improved based of the improvement achieved in Proposition 4.3.10. The results using the Wasserstein distance have not been published before.

4.3.1 Stability Regarding Changes in the Probability Measure

First, stability with respect to the total variation distance is examined. The proof of the subsequent proposition closely coincides with that given in the proof of Theorem 2.7 of Christmann et al. (2018), but slightly generalizes it to not necessarily differentiable loss functions. Notably, the proposition shows that $||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty$ grows at most linearly in the difference between P_1 and P_2 as measured by the total variation distance.

Proposition 4.3.3. *Let Assumption 4.3.1 be satisfied. Then,*

$$||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty \leq \frac{||k||_\infty^2 \, |L|_1}{\lambda} \cdot d_{\mathrm{tv}}(\mathrm{P}_1, \mathrm{P}_2) \, .$$

Proof. First of all,

$$||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty \leq ||k||_\infty \cdot ||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_H$$

by Lemma 2.1.10(i). By Christmann et al. (2009, Theorem 7), there exists a function h from the subdifferential of $L^\star$ with respect to $f_{\mathrm{P}_1,\lambda,k}$ such that

$$\begin{aligned}
||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_H &\leq \frac{1}{\lambda} \cdot \left|\left| \int h(x,y)\Phi(x) \, \mathrm{dP}_1(x,y) - \int h(x,y)\Phi(x) \, \mathrm{dP}_2(x,y) \right|\right|_H \\
&\leq \frac{1}{\lambda} \cdot \int ||h(x,y)\Phi(x)||_H \, \mathrm{d}|\mathrm{P}_1 - \mathrm{P}_2|(x,y) \\
&\leq \frac{1}{\lambda} \cdot \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |h(x,y)| \cdot \sup_{x\in\mathcal{X}} ||\Phi(x)||_H \cdot \int 1 \, \mathrm{d}|\mathrm{P}_1 - \mathrm{P}_2|(x,y) \\
&= \frac{1}{\lambda} \cdot \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |h(x,y)| \cdot \sup_{x\in\mathcal{X}} \sqrt{k(x,x)} \cdot d_{\mathrm{tv}}(\mathrm{P}_1, \mathrm{P}_2) \\
&\leq \frac{1}{\lambda} \cdot |L|_1 \cdot ||k||_\infty \cdot d_{\mathrm{tv}}(\mathrm{P}_1, \mathrm{P}_2) \, ,
\end{aligned}$$

where we used Christmann et al. (2018, Lemma 6.1) in the second and the reproducing property in the fourth step. This yields the assertion. □

It is not clear whether the bound from Proposition 4.3.3 is sharp or not. For minimal examples such as the subsequent one, it is off by a factor of 2. We look at such simple cases in most of the examples in this chapter—mostly at distributions whose support consists only of a small amount of points—because this simplicity makes it feasible to actually analytically derive SVMs and therefore assess the quality of the different bounds derived in this chapter.

Example 4.3.4. Let $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$, $\mathcal{Y} = \mathbb{R}$, $L^\star = L^*_{0.5\text{-pin}}$ be the shifted 0.5-pinball loss, $\lambda > 0$, and k be a Gaussian RBF kernel.[30] Let further $\mathrm{P}_1 = \delta_{(x_1,y_1)}$ and $\mathrm{P}_2 = \delta_{(x_2,y_2)}$ be Dirac distributions in some $(x_1, y_1), (x_2, y_2) \in \mathcal{X} \times \mathcal{Y}$ satisfying $y_1 > 0 > y_2$.

[30] Other loss functions and kernels satisfying certain conditions can also be used without changing much in the example.

For $i = 1, 2$, using the representation of SVMs from Christmann et al. (2009, Theorem 7), one obtains

$$f_{\mathrm{P}_i,\lambda,k} = -\frac{1}{2\lambda} \cdot \int h_i(x,y)\Phi(x)\,\mathrm{dP}_i(x,y) = -\frac{1}{2\lambda} \cdot h_i(x_i,y_i)\Phi(x_i)\,,$$

for some $h_i \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ from the subdifferential of $L^\star$ with respect to $f_{\mathrm{P}_i,\lambda,k}$ (see Definition 2.1.14). By the definition of the pinball loss, we obtain

$$h_i(x_i,y_i) = \begin{cases} -\frac{1}{2} & \text{, if } y_i > f_{\mathrm{P}_i,\lambda,k}(x_i)\,, \\ c \in \left[-\frac{1}{2}, +\frac{1}{2}\right] & \text{, if } y_i = f_{\mathrm{P}_i,\lambda,k}(x_i)\,, \\ +\frac{1}{2} & \text{, if } y_i < f_{\mathrm{P}_i,\lambda,k}(x_i)\,, \end{cases}$$

that is, $|h_i(x_i,y_i)| \leq \frac{1}{2}$ and hence

$$|f_{\mathrm{P}_i,\lambda,k}(x_i)| \leq \frac{1}{4\lambda} \cdot k(x_i,x_i) = \frac{1}{4\lambda}$$

because $k(x_i,x_i) = 1$ for the Gaussian RBF kernel.

Now assume that λ is chosen in such a way that $y_1 > (4\lambda)^{-1}$ and $y_2 < -(4\lambda)^{-1}$. The former condition implies that $y_1 > f_{\mathrm{P}_1,\lambda,k}(x_1)$ and hence $h_1(x_1,y_1) = -\frac{1}{2}$, which yields that

$$f_{\mathrm{P}_1,\lambda,k} = \frac{1}{4\lambda} \cdot \Phi(x_1)\,.$$

Analogously, the latter condition yields that

$$f_{\mathrm{P}_2,\lambda,k} = -\frac{1}{4\lambda} \cdot \Phi(x_2)\,.$$

Hence, we obtain

$$||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty = \frac{1}{4\lambda} \cdot ||\Phi(x_1) + \Phi(x_2)||_\infty$$

for these values of λ. If additionally P_1 and P_2 are such that $x_1 = x_2$, this reduces to

$$||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty = \frac{1}{4\lambda} \cdot ||2\Phi(x_1)||_\infty = \frac{1}{2\lambda}$$

by the definition of Gaussian RBF kernels.

On the other hand, the bound from Proposition 4.3.3 yields

$$||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty \leq \frac{||k||_\infty^2\,|L_{\tau\text{-pin}}|_1}{\lambda} \cdot d_{\mathrm{tv}}(\mathrm{P}_1,\mathrm{P}_2) = \frac{1}{2\lambda} \cdot d_{\mathrm{tv}}(\mathrm{P}_1,\mathrm{P}_2) = \frac{1}{\lambda}\,.$$

Thus, the bound is greater than the actual supremum norm by a factor of 2 in this situation.

Because stability with respect to the Wasserstein distance would also yield meaningful bounds in situations in which stability with respect to the total variation distance does not bound $||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty$ in a meaningful way, cf. Section 4.2.1, we now aim at deriving such a stability result as well. To our knowledge, the only existing result on stability with respect to the Wasserstein distance is Theorem 7 from Eckstein et al. (2023). Comparing the subsequent result to the referenced Theorem 7 however yields several differences regarding the assumptions and the situations for which the results are applicable. One major advantage of the subsequent result is not requiring $\mathcal{Y}$ to be bounded. Additionally, it yields a bound not only on an L_2- but on the stronger ∞-norm, but in exchange has to include λ^{-2} instead of only λ^{-1} (cf. Eckstein et al., 2023, Remark 8) in the bound. Finally, the two results are valid for completely distinct sets of loss functions and thus applicable in different learning scenarios. Whereas the referenced Theorem 7 is applicable for the popular least squares loss (which is possible because of the authors assuming $\mathcal{Y}$ to be bounded) but no other loss functions, the subsequent result can be used for a whole class of loss functions, however excluding the least squares loss because of requiring the loss to be Lipschitz continuous (as already stated in Assumption 4.3.1).

As did Proposition 4.3.3, the subsequent proposition also shows that $||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty$ grows at most linearly in the difference between P_1 and P_2, which is however measured by the Wasserstein distance now.

Proposition 4.3.5. *Let Assumptions 4.3.1 and 4.3.2 be satisfied. Then,*

$$\begin{aligned}&||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty \\ &\le \frac{||k||_\infty}{\lambda} \cdot \max\left\{ |\psi|_1 c_k + (-2\varphi'(0))^{1/2}\, \frac{|\psi|_1 |\psi'|_1\, ||k||_\infty^2}{\lambda} \,,\ |\psi'|_1\, ||k||_\infty \right\} \cdot d_{\mathrm{W}}(\mathrm{P}_1, \mathrm{P}_2)\,.\end{aligned}$$

It is not clear yet whether it is possible to further weaken the conditions on L from Assumptions 4.3.1 and 4.3.2 and still obtain an analogous stability result using the Wasserstein distance.

In order to prove Proposition 4.3.5, the following lemma bounding the slope of an SVM is needed, which follows as a special case from Steinwart and Christmann (2008, Corollary 4.36):

Lemma 4.3.6. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$. Let k be a kernel on $\mathcal{X}$ with RKHS H. Let $\tilde{\mathcal{X}} \supseteq \mathcal{X}$ with $\tilde{\mathcal{X}} \subseteq \mathbb{R}^d$ be open and assume that $k = \tilde{k}|_{\mathcal{X}\times\mathcal{X}}$ for a kernel $\tilde{k}$ on $\tilde{\mathcal{X}}$ that can be written as $\tilde{k}(x,x') = \varphi(||x-x'||_2^2)$ for all $x, x' \in \tilde{\mathcal{X}}$, where $\varphi \colon \mathbb{R}_{\ge 0} \to \mathbb{R}$ is a twice continuously differentiable function. Then, every $f \in H$ satisfies*

$$\frac{|f(x) - f(x')|}{||x-x'||_1} \le (-2\varphi'(0))^{1/2} \cdot ||f||_H \qquad \forall\, x, x' \in \mathcal{X}\,, x \ne x'\,.$$

Proof. For $x, x' \in \mathbb{R}^d$, denote in this proof by x_i, x_i', $i = 1, \dots, d$, the components of x, x'. Further denoting the argument of φ by r, one obtains for all $i \in \{1, \dots, d\}$ and all $x, x' \in \tilde{\mathcal{X}}$

$$\frac{\partial \tilde{k}(x,x')}{\partial x_i} = \frac{\partial \varphi(||x-x'||_2^2)}{\partial r} \cdot \frac{\partial\, ||x-x'||_2^2}{\partial x_i} = \frac{\partial \varphi(||x-x'||_2^2)}{\partial r} \cdot 2(x_i - x_i')\,,$$

and hence

$$\begin{aligned}\frac{\partial^2 \tilde{k}(x,x')}{\partial x_i' \partial x_i} &= \frac{\partial^2 \varphi(||x-x'||_2^2)}{\partial x_i' \partial r} \cdot 2(x_i - x_i') + \frac{\partial \varphi(||x-x'||_2^2)}{\partial r} \cdot (-2) \\ &= -4 \cdot \frac{\partial^2 \varphi(||x-x'||_2^2)}{\partial r^2} \cdot (x_i - x_i')^2 - 2 \cdot \frac{\partial \varphi(||x-x'||_2^2)}{\partial r} \,.\end{aligned}$$

With $\tilde{H}$ denoting the RKHS of $\tilde{k}$, Steinwart and Christmann (2008, Corollary 4.36) yields that all $\tilde{f} \in \tilde{H}$ satisfy

$$\begin{aligned}\left|\frac{\partial \tilde{f}(x)}{\partial x_i}\right| &\leq \left|\left|\tilde{f}\right|\right|_{\tilde{H}} \cdot \left(-4 \cdot \frac{\partial^2 \varphi(||x-x||_2^2)}{\partial r^2} \cdot (x_i - x_i)^2 - 2 \cdot \frac{\partial \varphi(||x-x||_2^2)}{\partial r}\right)^{1/2} \\ &= \left|\left|\tilde{f}\right|\right|_{\tilde{H}} \cdot \left(-2 \cdot \frac{\partial \varphi(0)}{\partial r}\right)^{1/2}\end{aligned}$$

for all $i \in \{1, \ldots, d\}$ and all $x \in \tilde{\mathcal{X}}$. With us denoting $\varphi' := \frac{\partial \varphi}{\partial r}$, we hence obtain

$$\frac{|\tilde{f}(x) - \tilde{f}(x')|}{||x-x'||_1} \leq (-2\varphi'(0))^{1/2} \cdot \left|\left|\tilde{f}\right|\right|_{\tilde{H}}$$

for all $\tilde{f} \in \tilde{H}$ and all $x, x' \in \tilde{\mathcal{X}}$. Finally, Berlinet and Thomas-Agnan (2004, Theorem 6) yields that H consists exactly of the restrictions to $\mathcal{X}$ of the elements of $\tilde{H}$ and that

$$\begin{aligned}(-2\varphi'(0))^{1/2} \cdot ||f||_H = (-2\varphi'(0))^{1/2} \cdot \min_{\substack{\tilde{f} \in \tilde{H}: \\ \tilde{f}|_{\mathcal{X}} = f}} \left|\left|\tilde{f}\right|\right|_{\tilde{H}} &\geq \min_{\substack{\tilde{f} \in \tilde{H}: \\ \tilde{f}|_{\mathcal{X}} = f}} \frac{|\tilde{f}(x) - \tilde{f}(x')|}{||x-x'||_1} \\ &= \frac{|f(x) - f(x')|}{||x-x'||_1}\end{aligned}$$

for all $f \in H$ and all $x, x' \in \mathcal{X}$. □

Proof of Proposition 4.3.5. First note that $\mathcal{Y} \subseteq \mathbb{R}$ being closed by Assumption 4.0.1 directly implies that $\mathcal{Y}$ is separable and complete, see also Bauer (2001, p. 157). Hence, $\mathcal{X} \times \mathcal{Y}$ is separable and complete as well and the definition of the Wasserstein distance is applicable.

By Lemma 2.1.10(i),

$$||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty \leq ||k||_\infty \cdot ||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_H \,.$$

Now, since L is differentiable (which is equivalent to ψ being differentiable), Christmann et al. (2009, Theorem 7) yields that for the function $h \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ defined by $h(x,y) = (L^\star)'(y, f_{\mathrm{P}_1,\lambda,k}(x)) = L'(y, f_{\mathrm{P}_1,\lambda,k}(x)) = -\psi'(y - f_{\mathrm{P}_1,\lambda,k}(x))$,[31] we have

$$||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_H \leq \frac{1}{\lambda} \cdot \left|\left|\int h(x,y)\Phi(x)\, \mathrm{dP}_1(x,y) - \int h(x,y)\Phi(x)\, \mathrm{dP}_2(x,y)\right|\right|_H \,.$$

[31] As usual, $(L^\star)'$ and L' denote the derivatives with respect to the last argument of the (shifted) loss function.

Therefore, for any probability measure Q on $(\mathcal{X}\times\mathcal{Y})\times(\mathcal{X}\times\mathcal{Y})$ with marginal distributions P_1 and P_2,[32] we have

$$\begin{aligned}||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_H &\leq \frac{1}{\lambda}\cdot\left|\left|\int \Big(h(x,y)\Phi(x) - h(x',y')\Phi(x')\Big)\,\mathrm{dQ}((x,y),(x',y'))\right|\right|_H \\ &\leq \frac{1}{\lambda}\cdot\int ||h(x,y)\Phi(x) - h(x',y')\Phi(x')||_H \,\mathrm{dQ}((x,y),(x',y')) \\ &\leq \frac{1}{\lambda}\cdot\int ||h(x,y)\Phi(x) - h(x,y)\Phi(x')||_H \,\mathrm{dQ}((x,y),(x',y')) \\ &\quad + \frac{1}{\lambda}\cdot\int ||h(x,y)\Phi(x') - h(x',y')\Phi(x')||_H \,\mathrm{dQ}((x,y),(x',y'))\,,\end{aligned} \tag{4.10}$$

where Diestel and Uhl (1977, Theorem II.2.4) was applied in the second step.

The integrands of the two summands on the right hand side of (4.10) can now be examined separately, starting with the first one:

$$\begin{aligned}||h(x,y)\Phi(x) - h(x,y)\Phi(x')||_H &= |h(x,y)|\cdot\Big\langle \Phi(x)-\Phi(x')\,,\,\Phi(x)-\Phi(x')\Big\rangle_H^{1/2} \\ &= |h(x,y)|\cdot\Big(k(x,x)+k(x',x')-2k(x,x')\Big)^{1/2} \\ &= |h(x,y)|\cdot\sqrt{2}\cdot\Big(\varphi(0)-\varphi(||x-x'||_2^2)\Big)^{1/2} \\ &\leq |\psi|_1\cdot c_k\cdot||x-x'||_2\,,\end{aligned} \tag{4.11}$$

where the reproducing property was applied in the second and $||h||_\infty \leq |L|_1 = |\psi|_1$ (cf. Christmann et al., 2009, Theorem 7) in the last step.

Now, we can take a look at the the integrand of the second summand: By Lemma 4.3.6,

$$|f(x)-f(x')| \leq (-2\varphi'(0))^{1/2}\cdot||f||_H\cdot||x-x'||_1$$

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

Combining this with

$$||\Phi(x')||_H = \langle\Phi(x')\,,\,\Phi(x')\rangle_H^{1/2} = (k(x',x'))^{1/2} = (\varphi(0))^{1/2} = ||k||_\infty$$

by the reproducing property and because of (2.2), and with ψ' being Lipschitz continuous, yields

$$\begin{aligned}&||h(x,y)\Phi(x') - h(x',y')\Phi(x')||_H \\ &= ||\Phi(x')||_H\cdot\Big|h(x,y)-h(x',y')\Big| \\ &= ||k||_\infty\cdot\Big|\psi'(y-f_{\mathrm{P}_1,\lambda,k}(x)) - \psi'(y'-f_{\mathrm{P}_1,\lambda,k}(x'))\Big| \\ &\leq ||k||_\infty\cdot|\psi'|_1\cdot\Big|(y-f_{\mathrm{P}_1,\lambda,k}(x)) - (y'-f_{\mathrm{P}_1,\lambda,k}(x'))\Big| \\ &\leq ||k||_\infty\cdot|\psi'|_1\cdot\Big(|y-y'| + |f_{\mathrm{P}_1,\lambda,k}(x) - f_{\mathrm{P}_1,\lambda,k}(x')|\Big) \\ &\leq ||k||_\infty\cdot|\psi'|_1\cdot\Big(|y-y'| + (-2\varphi'(0))^{1/2}\,||f_{\mathrm{P}_1,\lambda,k}||_H\,||x-x'||_1\Big)\,.\end{aligned} \tag{4.12}$$

[32] As Q can be chosen as the product measure of P_1 and P_2, at least one such Q always exists.

Since additionally $||x-x'||_2 \leq ||x-x'||_1$ as well as $||f_{\mathrm{P}_1,\lambda,k}||_H \leq \lambda^{-1}|L|_1\,||k||_\infty$ (cf. Christmann et al., 2009, equations (16) and (17)) and $|L|_1 = |\psi|_1$ (cf. Lemma 2.1.16(iii)), plugging (4.11) and (4.12) into (4.10) yields

$$\begin{aligned}
&||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_H \\
&\leq \frac{1}{\lambda} \cdot \int |\psi|_1 \cdot c_k \cdot ||x-x'||_2 \,\mathrm{dQ}((x,y),(x',y')) \\
&\quad + \frac{1}{\lambda} \cdot \int (||k||_\infty \cdot |\psi'|_1 \cdot \left(|y-y'| + (-2\varphi'(0))^{1/2} \frac{|\psi|_1\,||k||_\infty}{\lambda}\,||x-x'||_1 \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathrm{dQ}((x,y),(x',y')) \\
&\leq \frac{1}{\lambda} \cdot \max\left\{ |\psi|_1 c_k + (-2\varphi'(0))^{1/2} \frac{|\psi|_1 |\psi'|_1\,||k||_\infty^2}{\lambda} \,,\, |\psi'|_1\,||k||_\infty \right\} \\
&\quad \cdot \int \left(|y-y'| + ||x-x'||_1 \right) \mathrm{dQ}((x,y),(x',y'))\,.
\end{aligned}$$

Because $|y-y'| + ||x-x'||_1 = ||(x,y)-(x',y')||_1$ and Q was allowed to be an arbitrary probability measure with marginal distributions P_1 and P_2, this completes the proof. □

Note that the assumptions imposed on the kernel in Proposition 4.3.5 are satisfied by popular kernels such as for example the Gaussian RBF kernels, cf. Example 2.1.12, which gets captured by the subsequent corollary.

Corollary 4.3.7. *Let Assumptions 4.3.1 and 4.3.2 be satisfied.*[33] *Let $\gamma \in (0,\infty)$ and k_γ be the Gaussian RBF kernel on $\mathcal{X}$ with bandwidth γ and RKHS H_γ. Then,*

$$\left|\left| f_{\mathrm{P}_1,\lambda,k_\gamma} - f_{\mathrm{P}_2,\lambda,k_\gamma} \right|\right|_\infty \leq \frac{1}{\lambda} \cdot \max\left\{ \frac{\sqrt{2}\cdot|\psi|_1}{\gamma}\left(1 + \frac{|\psi'|_1}{\lambda}\right) \,,\, |\psi'|_1 \right\} \cdot d_{\mathrm{W}}(\mathrm{P}_1,\mathrm{P}_2)\,.$$

Proof. k_γ is measurable and bounded by 1. Furthermore, $k_\gamma(x,x') = \varphi_\gamma(||x-x'||_2^2)$ for all $x,x' \in \mathcal{X}$ if one defines

$$\varphi_\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R},\, r \mapsto \exp\left(-\frac{r}{\gamma^2}\right)\,.$$

φ_γ is twice continuously differentiable and, because $\exp(t) \geq 1+t$ for all $t \in \mathbb{R}$, it satisfies

$$\varphi_\gamma(0) - \varphi_\gamma(r) = 1 - \exp\left(-\frac{r}{\gamma^2}\right) \leq \frac{c_{k_\gamma}^2}{2} \cdot r \qquad \forall\, r \geq 0$$

for $c_{k_\gamma} := \frac{\sqrt{2}}{\gamma}$.

Hence, Proposition 4.3.5 can be applied and yields the assertion because $||k_\gamma||_\infty = 1$ and

$$\varphi_\gamma'(0) = \exp\left(-\frac{0}{\gamma^2}\right) \cdot \left(-\frac{1}{\gamma^2}\right) = -\frac{1}{\gamma^2}\,. \qquad □$$

[33] Those parts of the assumptions that concern the kernels k, k_1, k_2 are of course of no relevance for this corollary as it already specifies the use of special kernels, namely Gaussian RBF kernels. Instead, the proof of this corollary shows that the Gaussian RBF kernels indeed possess the properties required from k, k_1, k_2, such that they can be plugged in for k in Proposition 4.3.5.

Indeed, Proposition 4.3.5 using the Wasserstein distance can yield considerably better bounds than Proposition 4.3.3 using the total variation distance. To observe this, we look at a similar situation as that in Example 4.3.4. However, we do not use the pinball loss now because this loss function does not satisfy Assumption 4.3.2 which is required by Proposition 4.3.5.

Example 4.3.8. Let $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$, $\mathcal{Y} = \mathbb{R}$, $\lambda > 0$, and k be the Gaussian RBF kernel of bandwidth $\gamma > 0$. Let $\alpha > 0$, L be the α-Huber loss defined by

$$L(x, y, t) := \begin{cases} \frac{(y-t)^2}{2} & , \text{ if } |y-t| \leq \alpha, \\ \alpha|y-t| - \frac{\alpha^2}{2} & , \text{ if } |y-t| > \alpha\,, \end{cases}$$

and $L^\star$ be its shifted version.[34] Let further $\mathrm{P}_1 = \delta_{(x_1,y_1)}$ and $\mathrm{P}_2 = \delta_{(x_2,y_2)}$ be Dirac distributions in some $(x_1, y_1), (x_2, y_2) \in \mathcal{X} \times \mathcal{Y}$ satisfying $y_1, y_2 > 0$.

As in Example 4.3.4, we obtain for $i = 1, 2$

$$f_{\mathrm{P}_i,\lambda,k} = -\frac{1}{2\lambda} \cdot h_i(x_i, y_i)\Phi(x_i)\,,$$

for some $h_i \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ from the subdifferential of $L^\star$ with respect to $f_{\mathrm{P}_i,\lambda,k}$. By the definition of the Huber loss, we obtain

$$h_i(x_i, y_i) = \begin{cases} -\alpha & , \text{ if } y_i > f_{\mathrm{P}_i,\lambda,k}(x_i) + \alpha\,, \\ -(y - f_{\mathrm{P}_i,\lambda,k}(x_i)) & , \text{ if } y_i \in [f_{\mathrm{P}_i,\lambda,k}(x_i) - \alpha, f_{\mathrm{P}_i,\lambda,k}(x_i) + \alpha]\,, \\ \alpha & , \text{ if } y_i < f_{\mathrm{P}_i,\lambda,k}(x_i) - \alpha\,, \end{cases}$$

that is, $|h_i(x_i, y_i)| \leq \alpha$ and hence

$$|f_{\mathrm{P}_i,\lambda,k}(x_i)| \leq \frac{\alpha}{2\lambda} \cdot k(x_i, x_i) = \frac{\alpha}{2\lambda}\,.$$

Now assume that λ is chosen in such a way that $y_1, y_2 > \alpha \cdot (2\lambda)^{-1} + \alpha$. This implies, for $i = 1, 2$, that $y_i > f_{\mathrm{P}_i,\lambda,k}(x_i) + \alpha$ and hence $h_i(x_i, y_i) = -\alpha$, which yields that

$$f_{\mathrm{P}_i,\lambda,k} = \frac{\alpha}{2\lambda} \cdot \Phi(x_i)\,.$$

Hence, we obtain

$$||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty = \frac{\alpha}{2\lambda} \cdot ||\Phi(x_1) - \Phi(x_2)||_\infty$$

for these values of λ.

For the bounds from Proposition 4.3.3 and Proposition 4.3.5 (or more specifically the special case from Corollary 4.3.7 as we are using Gaussian RBF kernels), note that one

[34] Other loss functions and kernels satisfying certain conditions can also be used without changing much in the example.

$\lvert x_1^{(1)} - x_2^{(1)}\rvert$	0.01	0.1	0.5	1	2	5	10
$\text{bound}_{\text{tv}} / \lvert\lvert f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}\rvert\rvert_\infty$	466.34	46.71	9.72	5.48	4.07	4.00	4.00
$\text{bound}_{\text{W}} / \lvert\lvert f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}\rvert\rvert_\infty$	6.59	6.61	6.87	7.74	11.51	28.28	56.59

Table 4.3.1: Ratio between the bound using the total variation distance (bound_{tv}) as well as the bound using the Wasserstein distance (bound_{W}) and the actual supremum norm $||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty$ for $\alpha = \lambda = \gamma = 1$ and different distances $|x_1^{(1)} - x_2^{(1)}|$ in the scenario described in Example 4.3.8.

can easily derive $|L|_1 = |\psi|_1 = \alpha$ and $|\psi'|_1 = 1$, where ψ denotes the representing function of L. The bound from Proposition 4.3.3 using the total variation distance therefore yields

$$||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty \leq \frac{||k||_\infty^2 \, |L|_1}{\lambda} \cdot d_{\text{tv}}(\mathrm{P}_1, \mathrm{P}_2) = \frac{\alpha}{\lambda} \cdot d_{\text{tv}}(\mathrm{P}_1, \mathrm{P}_2) = \frac{2\alpha}{\lambda} \, .$$

The bound from Corollary 4.3.7 using the Wasserstein distance can be simplified to

$$\begin{aligned} ||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty &\leq \frac{1}{\lambda} \cdot \max\left\{ \frac{\sqrt{2} \cdot |\psi|_1}{\gamma} \left(1 + \frac{|\psi'|_1}{\lambda}\right) \, , \, |\psi'|_1 \right\} \cdot d_{\mathrm{W}}(\mathrm{P}_1, \mathrm{P}_2) \\ &= \frac{1}{\lambda} \cdot \max\left\{ \frac{\sqrt{2} \cdot \alpha}{\gamma} \left(1 + \frac{1}{\lambda}\right) \, , \, 1 \right\} \cdot ||(x_1, y_1) - (x_2, y_2)||_1 \end{aligned}$$

As the latter bound additionally considers the distance between (x_1, y_1) and (x_2, y_2), it seems likely that it might be the superior one if this distance is small and the inferior one if it is large—as long as the bandwidth γ, which also only appears in the latter bound, is fixed. Table 4.3.1 affirms this by collecting how both the bound using the total variation distance and the one using the Wasserstein distance compare to the actual supremum norm $||f_{\mathrm{P}_1,\lambda,k} - f_{\mathrm{P}_2,\lambda,k}||_\infty$ for different distances between (x_1, y_1) and (x_2, y_2). To be more specific, we assumed that this distance comes entirely from the first component $x_i^{(1)}$ of x_i, i.e. that y_1 and y_2 coincide and that x_1 and x_2 coincide in all components but the first one.

Lastly, we present another lemma, which is similar to Lemma 4.3.6, also bounding the slope of an SVM, but which is only applicable for the Gaussian RBF kernel and empirical SVMs. In exchange for this reduced generality, the subsequent lemma strengthens Lemma 4.3.6 by considering $||x - x'||_2$ instead of $||x - x'||_1$ and deriving the exact same bound apart from that (because $\varphi'(0) = -\gamma^{-2}$ for the Gaussian kernel, cf. proof of Corollary 4.3.7). Even though this strengthened bound is not used in the results on (total) stability (because the Wasserstein distance needs the 1-norm instead of the 2-norm of the difference anyway), we still explicitly state this lemma since it might also be interesting in its own right because of the popularity of the Gaussian RBF kernel.

Lemma 4.3.9. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$. Let $L \colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex loss function and let $L^\star$ be its shifted version. Let $\gamma \in (0, \infty)$ and k_γ be the Gaussian RBF kernel on $\mathcal{X}$ with bandwidth γ and RKHS H_γ. Let $D_n := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ and $\lambda > 0$. Then,*

$$\frac{\left|f_{\mathrm{D}_n,\lambda,k_\gamma}(x) - f_{\mathrm{D}_n,\lambda,k_\gamma}(x')\right|}{||x - x'||_2} \leq \frac{\sqrt{2}}{\gamma} \cdot \left|\left|f_{\mathrm{D}_n,\lambda,k_\gamma}\right|\right|_{H_\gamma} \qquad \forall\, x, x' \in \mathcal{X} \, , x \neq x' \, . \tag{4.13}$$

Proof. The assertion gets proven by showing that the absolute value of the directional derivative of $f_{\mathrm{D}_n,\lambda,k_\gamma}$ in direction of any unit vector (with respect to $||\cdot||_2$) can not be greater than the right hand side of (4.13). For the purpose of this proof, elements of a vector $x \in \mathcal{X} \subseteq \mathbb{R}^d$ are denoted by $x^{(\ell)}$, $\ell = 1, \ldots, d$.

By Steinwart and Christmann (2008, Theorem 5.5), there exist $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that $f_{\mathrm{D}_n,\lambda,k_\gamma}$ can be written as

$$f_{\mathrm{D}_n,\lambda,k_\gamma}(x) = \sum_{i=1}^{n} \alpha_i k_\gamma(x, x_i) \qquad \forall\, x \in \mathcal{X}\,. \tag{4.14}$$

Now, let $x = (x^{(1)}, \ldots, x^{(d)})^T \in \mathcal{X}$ be arbitrary but fixed. We only examine the derivative of $f_{\mathrm{D}_n,\lambda,k_\gamma}$ in direction of the unit vector $(d^{-1/2}, \ldots, d^{-1/2})^T$ because each derivative in the direction of a different unit vector can be viewed as a derivative in the direction $(d^{-1/2}, \ldots, d^{-1/2})^T$ by just rotating x as well as $x_1, \ldots, x_n$ around the origin accordingly (because of k_γ being rotationally invariant). With a slight abuse of notation, we denote this directional derivative by $f'_{\mathrm{D}_n,\lambda,k_\gamma}$ in the following. For investigating $f'_{\mathrm{D}_n,\lambda,k_\gamma}$, we first need the partial derivatives of $f_{\mathrm{D}_n,\lambda,k_\gamma}$ with respect to each of the components $x^{(1)}, \ldots, x^{(d)}$. With $z_i \coloneqq x - x_i$ for $i = 1, \ldots, n$, we obtain

$$\begin{aligned} \frac{\partial f_{\mathrm{D}_n,\lambda,k_\gamma}(x)}{\partial x^{(\ell)}} &= -\frac{2}{\gamma^2} \cdot \sum_{i=1}^{n} \alpha_i \exp\left(-\frac{||x - x_i||_2^2}{\gamma^2}\right) \left(x^{(\ell)} - x_i^{(\ell)}\right) \\ &= -\frac{2}{\gamma^2} \cdot \sum_{i=1}^{n} \alpha_i \exp\left(-\frac{||z_i||_2^2}{\gamma^2}\right) z_i^{(\ell)} \end{aligned}$$

for $\ell = 1, \ldots, d$. Observing that

$$f'_{\mathrm{D}_n,\lambda,k_\gamma}(x) = \frac{1}{\sqrt{d}} \cdot \sum_{\ell=1}^{d} \frac{\partial f_{\mathrm{D}_n,\lambda,k_\gamma}(x)}{\partial x^{(\ell)}}$$

and applying the Cauchy-Schwarz inequality then yields

$$\begin{aligned} \left(f'_{\mathrm{D}_n,\lambda,k_\gamma}(x)\right)^2 &= \frac{1}{d} \cdot \left(\sum_{\ell=1}^{d} \frac{\partial f_{\mathrm{D}_n,\lambda,k_\gamma}(x)}{\partial x^{(\ell)}}\right)^2 \\ &\leq \sum_{\ell=1}^{d} \left(\frac{\partial f_{\mathrm{D}_n,\lambda,k_\gamma}(x)}{\partial x^{(\ell)}}\right)^2 \\ &= \frac{4}{\gamma^4} \cdot \sum_{\ell=1}^{d} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \exp\left(-\frac{||z_i||_2^2 + ||z_j||_2^2}{\gamma^2}\right) z_i^{(\ell)} z_j^{(\ell)} \\ &= \frac{4}{\gamma^4} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \exp\left(-\frac{||z_i||_2^2 + ||z_j||_2^2}{\gamma^2}\right) \langle z_i, z_j \rangle \end{aligned} \tag{4.15}$$

Because of the representation from (4.14), the reproducing property additionally yields that

$$\begin{aligned}\left|\left|f_{\mathrm{D}_n,\lambda,k_\gamma}\right|\right|_{H_\gamma}^2 &= \left\langle \sum_{i=1}^n \alpha_i k_\gamma(\cdot,x_i)\,,\, \sum_{j=1}^n \alpha_j k_\gamma(\cdot,x_j)\right\rangle_{H_\gamma}\\ &= \sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j \exp\left(-\frac{||x_i-x_j||_2^2}{\gamma^2}\right)\\ &= \sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j \exp\left(-\frac{||z_i-z_j||_2^2}{\gamma^2}\right) \qquad (4.16)\end{aligned}$$

because $x_i - x_j = z_j - z_i$ for all $i, j \in \{1, \dots, n\}$.

To prove the assertion, it hence suffices to show that the right hand side of (4.15) can be bounded by $2\gamma^{-2}$ times the right hand side of (4.16). That is, we have to prove that

$$\frac{2}{\gamma^2}\cdot\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j \left(\exp\left(-\frac{||z_i-z_j||_2^2}{\gamma^2}\right) - \frac{2}{\gamma^2}\exp\left(-\frac{||z_i||_2^2+||z_j||_2^2}{\gamma^2}\right)\langle z_i, z_j\rangle\right) \geq 0 \quad (4.17)$$

for all $\alpha_1, \dots, \alpha_n, z_1, \dots, z_n \in \mathbb{R}$. As this is equivalent to

$$k_1\colon \mathbb{R}^d\times\mathbb{R}^d\,,\ (z,z')\mapsto \exp\left(-\frac{||z-z'||_2^2}{\gamma^2}\right) - \frac{2}{\gamma^2}\exp\left(-\frac{||z||_2^2+||z'||_2^2}{\gamma^2}\right)\langle z, z'\rangle$$

being positive definite and as k_1 is obviously symmetric, (4.17) holding true is equivalent to k_1 being a kernel on $\mathbb{R}^d$, which is what gets proven in the following.

First of all, define

$$\begin{aligned}&k_2\colon \mathbb{R}^d\times\mathbb{R}^d\,,\ (z,z')\mapsto \exp\left(\frac{2\langle z,z'\rangle}{\gamma^2}\right) - \frac{2}{\gamma^2}\langle z,z'\rangle\,,\\ &k_3\colon \mathbb{R}^d\times\mathbb{R}^d\,,\ (z,z')\mapsto \exp\left(\langle z,z'\rangle\right) - \langle z,z'\rangle\\ &k_4\colon \mathbb{R}^d\times\mathbb{R}^d\,,\ (z,z')\mapsto \langle z,z'\rangle\,.\end{aligned}$$

As k_4 is known to define a kernel on $\mathbb{R}^d$ (cf. Berlinet and Thomas-Agnan, 2004, Lemma 1), it follows analogously to Cristianini and Shawe-Taylor (2000, Corollary 3.13), by using the polynomial coefficients $\alpha_1 := \frac{1}{1!} - 1 = 0$ and $\alpha_m := \frac{1}{m!} > 0$ for $m \neq 1$, that

$$\sum_{m=0}^\infty \alpha_m k_4^m(z,z') = \exp\left(k_4(z,z')\right) - k_4(z,z') = k_3(z,z')$$

also defines a kernel on $\mathbb{R}^d$. Because

$$k_2(z,z') = k_3(\psi(z),\psi(z')) \qquad \forall\, z,z' \in \mathbb{R}^d$$

for $\psi\colon \mathbb{R}^d \to \mathbb{R}^d$, $z \mapsto \frac{\sqrt{2}}{\gamma}z$, Cristianini and Shawe-Taylor (2000, Proposition 3.12) yields that k_2 is a kernel on $\mathbb{R}^d$ as well. By definition, there therefore exist a Hilbert space H and a feature map $\Phi_2\colon \mathbb{R}^d \to H$ such that

$$k_2(z,z') = \langle \Phi_2(z)\,,\,\Phi_2(z')\rangle_H \qquad \forall\, z,z' \in \mathbb{R}^d\,.$$

Defining

$$\Phi_1 \colon \mathbb{R}^d \to H\,,\ z \mapsto \exp\left(-\frac{||z||_2^2}{\gamma^2}\right) \cdot \Phi_2(z)$$

finally yields

$$\begin{aligned} k_1(z,z') &= \exp\left(-\frac{||z||_2^2 + ||z'||_2^2}{\gamma^2}\right) \cdot k_2(z,z') \\ &= \exp\left(-\frac{||z||_2^2 + ||z'||_2^2}{\gamma^2}\right) \cdot \langle \Phi_2(z), \Phi_2(z')\rangle_H \\ &= \langle \Phi_1(z), \Phi_1(z')\rangle_H \qquad \forall\, z, z' \in \mathbb{R}^d\,. \end{aligned}$$

Hence, H and Φ_1 are feature space respectively feature map of k_1. Thus, k_1 is a kernel on $\mathbb{R}^d$, which completes the proof. □

4.3.2 Stability Regarding Changes in the Regularization Parameter

Similarly to Proposition 4.3.3 on stability regarding changes in the probability measure with respect to the total variation distance, the proof of the subsequent result on stability regarding changes in the regularization parameter also closely coincides with that of a result from Christmann et al. (2018), in this case Theorem 2.6, but slightly generalizes it to not necessarily differentiable losses, while at the same time improving the derived bound by a factor of 2. It also constitutes an improvement by a factor of 2 compared with Köhler and Christmann (2022, Lemma 14). The proposition shows that $||f_{\mathrm{P},\lambda_1,k} - f_{\mathrm{P},\lambda_2,k}||_\infty$ grows at most linearly in $|\lambda_1 - \lambda_2|$ as long as $\min\{\lambda_1, \lambda_2\}$ does not decrease.

Proposition 4.3.10. *Let Assumption 4.3.1 be satisfied. Then,*

$$||f_{\mathrm{P},\lambda_1,k} - f_{\mathrm{P},\lambda_2,k}||_\infty \leq \frac{||k||_\infty^2\,|L|_1}{2\min\{\lambda_1,\lambda_2\}^2} \cdot |\lambda_1 - \lambda_2|\,.$$

In order to achieve the aspired improvement by a factor of 2 in the bound, the following lemma is needed:

Lemma 4.3.11. *Let Assumption 4.3.1 be satisfied. Then,*

$$||f_{\mathrm{P},\lambda,k}||_H \leq \frac{||k||_\infty\,|L|_1}{2\lambda}\,.$$

Proof. Assume without loss of generality that $||f_{\mathrm{P},\lambda,k}||_H > 0$ since the case $||f_{\mathrm{P},\lambda,k}||_H = 0$ is trivial.

Christmann et al. (2009, Theorem 7) yields the existence of a function $h \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfying $||h||_\infty \leq |L|_1$ as well as

$$f_{\mathrm{P},\lambda,k} = -\frac{1}{2\lambda}\int h(x,y)\Phi(x)\,\mathrm{dP}(x,y)\,.$$

By using this function h and applying the reproducing property as well as in the last step Lemma 2.1.10(i), we obtain

$$\begin{aligned}
||f_{\mathrm{P},\lambda,k}||_H^2 &= \Big\langle f_{\mathrm{P},\lambda,k}\,,\, -\frac{1}{2\lambda}\int h(x,y)\Phi(x)\,\mathrm{dP}(x,y)\Big\rangle_H \\
&= -\frac{1}{2\lambda}\int h(x,y)\cdot f_{\mathrm{P},\lambda,k}(x)\,\mathrm{dP}(x,y) \\
&\le \frac{1}{2\lambda}\int |h(x,y)|\cdot|f_{\mathrm{P},\lambda,k}(x)|\,\mathrm{dP}(x,y) \\
&\le \frac{|L|_1}{2\lambda}\,||f_{\mathrm{P},\lambda,k}||_\infty \\
&\le \frac{||k||_\infty\,|L|_1}{2\lambda}\,||f_{\mathrm{P},\lambda,k}||_H\,.
\end{aligned}$$

Dividing by $||f_{\mathrm{P},\lambda,k}||_H$ yields the assertion. □

Proof of Proposition 4.3.10. To shorten the notation, define $f_i := f_{\mathrm{P},\lambda_i,k}$, $i=1,2$, in this proof. By Lemma 2.1.10(i),

$$||f_1 - f_2||_\infty \le ||k||_\infty \cdot ||f_1 - f_2||_H\,.$$

Assume now without loss of generality that $||f_1 - f_2||_H > 0$ since the case $||f_1 - f_2||_H = 0$ is trivial.

Christmann et al. (2009, Theorem 7) yields functions h_1 and h_2 from the subdifferential of $L^\star$ (with respect to f_1 respectively f_2) such that

$$f_1 - f_2 = -\frac{1}{2\lambda_1}\cdot\int h_1(x,y)\Phi(x)\,\mathrm{dP}(x,y) + \frac{1}{2\lambda_2}\cdot\int h_2(x,y)\Phi(x)\,\mathrm{dP}(x,y)\,.$$

From this we obtain, by applying the reproducing property in the last step,

$$\begin{aligned}
||f_1 - f_2||_H^2 &= \langle f_1 - f_2, f_1 - f_2\rangle_H \\
&= \Big\langle \frac{1}{2\lambda_2}\cdot\int h_2(x,y)\Phi(x)\,\mathrm{dP}(x,y), f_1 - f_2\Big\rangle_H \\
&\quad - \Big\langle \frac{1}{2\lambda_1}\cdot\int h_1(x,y)\Phi(x)\,\mathrm{dP}(x,y), f_1 - f_2\Big\rangle_H \\
&= \frac{1}{2\lambda_2}\cdot\int h_2(x,y)(f_1(x) - f_2(x))\,\mathrm{dP}(x,y) \\
&\quad - \frac{1}{2\lambda_1}\cdot\int h_1(x,y)(f_1(x) - f_2(x))\,\mathrm{dP}(x,y)\,. \qquad (4.18)
\end{aligned}$$

Because L (and thus also $L^\star$, cf. Lemma 2.1.30) is convex and $h_i(x,y) \in \partial L^\star(x,y,f_i(x))$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$ and for $i=1,2$, we know that

$$h_i(x,y)\cdot(t - f_i(x)) \le L^\star(x,y,t) - L^\star(x,y,f_i(x)) \qquad \forall\, t\in\mathbb{R}, \qquad i=1,2\,,$$

more specifically

$$h_1(x,y)\cdot(f_2(x) - f_1(x)) \le L^\star(x,y,f_2(x)) - L^\star(x,y,f_1(x))$$

and

$$h_2(x,y)\cdot(f_1(x)-f_2(x)) \le L^\star(x,y,f_1(x)) - L^\star(x,y,f_2(x))\,.$$

Plugging these two inequalities into (4.18) yields

$$\begin{aligned}||f_1-f_2||_H^2 &\le \left(\frac{1}{2\lambda_2}-\frac{1}{2\lambda_1}\right)\cdot\int L^\star(x,y,f_1(x)) - L^\star(x,y,f_2(x))\,\mathrm{dP}(x,y)\\ &= \left(\frac{1}{2\lambda_2}-\frac{1}{2\lambda_1}\right)\cdot\Big(\mathbb{E}_\mathrm{P}\left[L^\star(X,Y,f_1(X))\right]-\mathbb{E}_\mathrm{P}\left[L^\star(X,Y,f_2(X))\right]\Big) \qquad (4.19)\end{aligned}$$

Now, $||f_1-f_2||_H^2$ being positive implies that the right hand side of this inequality has to be positive as well. That is, both factors need to have the same sign. First assume $\lambda_1>\lambda_2$:

In this case $\frac{1}{2\lambda_2}-\frac{1}{2\lambda_1}>0$ and thus $\mathbb{E}_\mathrm{P}\left[L^\star(X,Y,f_1(X))\right]-\mathbb{E}_\mathrm{P}\left[L^\star(X,Y,f_2(X))\right]$ has to be positive as well. Because of the definition of f_1 as the minimizer of the regularized risk with regularization parameter λ_1, we know that

$$\mathbb{E}_\mathrm{P}\left[L^\star(X,Y,f_1(X))\right]+\lambda_1\,||f_1||_H^2 \le \mathbb{E}_\mathrm{P}\left[L^\star(X,Y,f_2(X))\right]+\lambda_1\,||f_2||_H^2\ .$$

From this, it follows that

$$\begin{aligned}&0<\mathbb{E}_\mathrm{P}\left[L^\star(X,Y,f_1(X))\right]-\mathbb{E}_\mathrm{P}\left[L^\star(X,Y,f_2(X))\right]\le\lambda_1\cdot\Big(||f_2||_H^2-||f_1||_H^2\Big)\\ &\quad=\lambda_1\cdot(||f_1||_H+||f_2||_H)\cdot(||f_2||_H-||f_1||_H)\le\lambda_1\cdot(||f_1||_H+||f_2||_H)\cdot||f_1-f_2||_H\end{aligned}$$

with the last inequality holding true because of $\lambda_1(||f_1||_H+||f_2||_H)\ge 0$ and the reverse triangle inequality. Plugging this into (4.19) and dividing by $||f_1-f_2||_H$, we obtain

$$||f_1-f_2||_H \le \frac{1}{2}\cdot\left(\frac{\max\{\lambda_1,\lambda_2\}}{\min\{\lambda_1,\lambda_2\}}-1\right)\cdot(||f_1||_H+||f_2||_H)\ . \qquad (4.20)$$

The case $\lambda_2>\lambda_1$ yields the same inequality.

By additionally applying Lemma 4.3.11, we now obtain

$$\begin{aligned}||f_1-f_2||_H &\le \frac{|L|_1\,||k||_\infty}{2}\cdot\left(\frac{\max\{\lambda_1,\lambda_2\}}{\min\{\lambda_1,\lambda_2\}}-1\right)\cdot\left(\frac{1}{2\lambda_1}+\frac{1}{2\lambda_2}\right)\\ &\le \frac{|L|_1\,||k||_\infty}{2\min\{\lambda_1,\lambda_2\}}\cdot\Big(\max\{\lambda_1,\lambda_2\}-\min\{\lambda_1,\lambda_2\}\Big)\cdot\frac{2}{2\min\{\lambda_1,\lambda_2\}}\\ &= \frac{|L|_1\,||k||_\infty}{2\min\{\lambda_1,\lambda_2\}^2}\cdot|\lambda_1-\lambda_2|\end{aligned}$$

which yields the assertion. □

The subsequent minimal example shows that the bound from Proposition 4.3.10 is asymptotically sharp:

Example 4.3.12. Let $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$, $\mathcal{Y} = \mathbb{R}$, $L^\star = L^*_{0.5\text{-pin}}$ be the shifted 0.5-pinball loss, k be a Gaussian RBF kernel, and $\mathrm{P} = \delta_{(x_0,y_0)}$ be the Dirac distribution in some $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, where we assume $y_0 > 0$ for simplicity.[35] As in Example 4.3.4, we obtain

$$f_{\mathrm{P},\lambda_i,k} = \frac{1}{4\lambda_i} \cdot \Phi(x_0)$$

for $i = 1, 2$ if $\lambda_i > (4y_0)^{-1}$. For $\lambda_2 > \lambda_1 > (4y_0)^{-1}$, we therefore obtain

$$||f_{\mathrm{P},\lambda_1,k} - f_{\mathrm{P},\lambda_2,k}||_\infty = \left(\frac{1}{4\lambda_1} - \frac{1}{4\lambda_2}\right) \cdot ||\Phi(x_0)||_\infty = \frac{1}{4\lambda_1\lambda_2} \cdot (\lambda_2 - \lambda_1)\,.$$

On the other hand, the bound from Proposition 4.3.10 yields

$$||f_{\mathrm{P},\lambda_1,k} - f_{\mathrm{P},\lambda_2,k}||_\infty \leq \frac{||k||_\infty^2 \, |L_{\tau\text{-pin}}|_1}{2\lambda_1^2} \cdot (\lambda_2 - \lambda_1) = \frac{1}{4\lambda_1^2} \cdot (\lambda_2 - \lambda_1)\,.$$

Thus, the bound is greater than the actual supremum norm only by a factor of λ_2/λ_1, which converges to 1 if we replace λ_2 by a sequence $(\lambda_{2,n})_{n\in\mathbb{N}}$ satisfying $\lambda_{2,n} \searrow \lambda_1$ as $n \to \infty$.

4.3.3 Stability Regarding Changes in the Kernel

The stability results from this section are the ones leading to the main differences to those derived by Christmann et al. (2018). In addition to not requiring the underlying loss function to be differentiable, the results from this section also eliminate additional assumptions regarding the regularization parameter respectively the kernels from Christmann et al. (2018, Lemmas 6.4 and 6.5). These generalizations are discussed in more detail in Section 4.3.4, where the results from Sections 4.3.1 to 4.3.3 are combined in order to derive total stability.

We start by giving a result on sup-stability, before then turning the attention to L_p-stability. Contrary to the results from the previous sections considering changes in P and λ, the subsequent proposition does not yield a bound growing strictly linearly in the difference between k_1 and k_2 but instead also includes the square root of this difference, which dominates the behavior of the bound for small differences.

Remark 4.3.13. It would also be possible to state an analogous result that only uses the linear part in the bound $||k_1 - k_2||_\infty$, cf. Christmann et al. (2018, Lemma 6.4). This would however require the additional assumptions regarding loss function and regularization parameters that were mentioned before and that are laid out in more detail at the beginning of Section 4.3.4.

Proposition 4.3.14. *Let Assumption 4.3.1 be satisfied and let* $\kappa := \max\{||k_1||_\infty, ||k_2||_\infty\}$. *Then,*

$$||f_{\mathrm{P},\lambda,k_1} - f_{\mathrm{P},\lambda,k_2}||_\infty \leq \frac{|L|_1}{\lambda} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_\infty + \kappa \cdot \sqrt{||k_1 - k_2||_\infty}\right)\,.$$

[35] Other loss functions and kernels satisfying certain conditions can also be used without changing much in the example.

In order to prove Proposition 4.3.14 we need an auxiliary statement giving two inequalities which are well-known for Lebesgue integrals, but which are needed for RKHS-valued Bochner integrals in the proof. Even though we suppose that this statement is also already established for Bochner integrals, we did not find a reference in the literature, for which reason we prove it here. See Diestel and Uhl (1977), Diestel (1984), Denkowski et al. (2003) for a detailed introduction to Bochner integrals.

Lemma 4.3.15. *Let* Q *be a probability measure on some measurable space* $(\Omega, \mathcal{A})$ *and let* k *be a bounded and measurable kernel on* Ω *with RKHS* H. *Let* $g\colon \Omega \to H$ *be a* Q*-Bochner integrable function. Then,*

$$\left|\left|\int_\Omega g(x)\,\mathrm{dQ}(x)\right|\right|_\infty \leq \int_\Omega ||g(x)||_\infty\,\mathrm{dQ}(x) \tag{4.21}$$

and, for all $p \in [1, \infty)$,

$$\left|\left|\int_\Omega g(x)\,\mathrm{dQ}(x)\right|\right|_{L_p(\mathrm{Q})} \leq \int_\Omega ||g(x)||_{L_p(\mathrm{Q})}\,\mathrm{dQ}(x)\,. \tag{4.22}$$

Proof. By Denkowski et al. (2003, Definition 3.10.7), g being Q-Bochner integrable means that there exists a sequence $(s_n)_{n\in\mathbb{N}}$ of so-called simple functions $s_n\colon \Omega \to H, \omega \mapsto \sum_{j=1}^{m_n} b_j^{(n)} \mathbb{1}_{A_j^{(n)}}(\omega)$, with $b_j^{(n)} \in H$, $A_j^{(n)} \in \mathcal{A}$ and $\mathbb{1}_{A_j^{(n)}}$ denoting the indicator function on $A_j^{(n)}$ for all $n \in \mathbb{N}$ and $j \in \{1, \ldots, m_n\}$, such that

$$\lim_{n\to\infty} \int_\Omega ||g(\omega) - s_n(\omega)||_H\,\mathrm{dQ}(\omega) = 0\,. \tag{4.23}$$

Then, the same definition tells us that

$$\int_\Omega g(\omega)\,\mathrm{dQ}(\omega) := \lim_{n\to\infty} \int_\Omega s_n(\omega)\,\mathrm{dQ}(\omega)\,,$$

where

$$\int_\Omega s_n(\omega)\,\mathrm{dQ}(\omega) := \sum_{j=1}^{m_n} b_j^{(n)} \mathrm{Q}\left(A_j^{(n)}\right)$$

for all $n \in \mathbb{N}$. Additionally, we know from Diestel (1984, Chapter IV) that we can without loss of generality assume $A_1^{(n)}, \ldots, A_{m_n}^{(n)}$ to be pairwise disjoint for all $n \in \mathbb{N}$.

Let now $||\cdot||_\bullet$ denote either of $||\cdot||_\infty$ and $||\cdot||_{L_p(\mathrm{Q})}$. Then,

$$\begin{aligned}
&\left|\left|\int g(\omega)\,\mathrm{dQ}(\omega)\right|\right|_\bullet = \left|\left|\lim_{n\to\infty}\left(\sum_{j=1}^{m_n} b_j^{(n)} \mathrm{Q}\left(A_j^{(n)}\right)\right)\right|\right|_\bullet = \lim_{n\to\infty}\left|\left|\sum_{j=1}^{m_n} b_j^{(n)} \mathrm{Q}\left(A_j^{(n)}\right)\right|\right|_\bullet \\
&\leq \lim_{n\to\infty}\left(\sum_{j=1}^{m_n} \left|\left|b_j^{(n)}\right|\right|_\bullet \mathrm{Q}\left(A_j^{(n)}\right)\right) = \lim_{n\to\infty}\left(\sum_{j=1}^{m_n} \int \left|\left|b_j^{(n)}\right|\right|_\bullet \mathbb{1}_{A_j^{(n)}}(\omega)\,\mathrm{dQ}(\omega)\right) \\
&= \lim_{n\to\infty}\left(\int \sum_{j=1}^{m_n} \left|\left|b_j^{(n)}\right|\right|_\bullet \mathbb{1}_{A_j^{(n)}}(\omega)\,\mathrm{dQ}(\omega)\right) = \lim_{n\to\infty}\left(\int \left|\left|\sum_{j=1}^{m_n} b_j^{(n)} \mathbb{1}_{A_j^{(n)}}(\omega)\right|\right|_\bullet \mathrm{dQ}(\omega)\right) \\
&= \lim_{n\to\infty}\left(\int ||s_n(\omega)||_\bullet\,\mathrm{dQ}(\omega)\right) = \int ||g(\omega)||_\bullet\,\mathrm{dQ}(\omega)\,,
\end{aligned} \tag{4.24}$$

where we applied the continuity of $||\cdot||_\bullet$ as a function on H in the second step and the pairwise disjointness of $A_1^{(n)}, \dots, A_{m_n}^{(n)}$ in the second to last row, with the continuity of $||\cdot||_\bullet$ holding true because $||h||_{L_p(\mathrm{Q})} \leq ||h||_\infty \leq ||k||_\infty \, ||h||_H$ and thus

$$||h||_\bullet \leq ||k||_\infty \, ||h||_H \tag{4.25}$$

for all $h \in H$, cf. Lemma 2.1.10(i). Additionally, the equality in the last step of (4.24) holds true because

$$\begin{aligned}
&\left| \int ||g(\omega)||_\bullet \, \mathrm{dQ}(\omega) - \lim_{n\to\infty} \Big(\int ||s_n(\omega)||_\bullet \, \mathrm{dQ}(\omega) \Big) \right| \\
&\leq \lim_{n\to\infty} \Big(\int \Big| \, ||g(\omega)||_\bullet - ||s_n(\omega)||_\bullet \Big| \, \mathrm{dQ}(\omega) \Big) \\
&\leq \lim_{n\to\infty} \Big(\int ||g(\omega) - s_n(\omega)||_\bullet \, \mathrm{dQ}(\omega) \Big) \\
&\leq ||k||_\infty \cdot \lim_{n\to\infty} \Big(\int ||g(\omega) - s_n(\omega)||_H \, \mathrm{dQ}(\omega) \Big) = 0 \, .
\end{aligned} \tag{4.26}$$

Here, we employed the finiteness of the two summands on the left hand side in the first step, and the reverse triangle inequality, (4.25) and (4.23) in the remaining steps. In the first step, the finiteness of the first summand follows directly from (4.25) and Denkowski et al. (2003, Theorem 3.10.9), and the finiteness of the second one can be shown by again using (4.25) and then slightly adapting the proof of the mentioned theorem:

$$\begin{aligned}
&\lim_{n\to\infty} \Big(\int ||s_n(\omega)||_H \, \mathrm{dQ}(\omega) \Big) \\
&\leq \lim_{n\to\infty} \Big(\int ||s_n(\omega) - g(\omega)||_H \, \mathrm{dQ}(\omega) + \int ||g(\omega)||_H \, \mathrm{dQ}(\omega) \Big) \\
&= \lim_{n\to\infty} \Big(\int ||s_n(\omega) - g(\omega)||_H \, \mathrm{dQ}(\omega) \Big) + \int ||g(\omega)||_H \, \mathrm{dQ}(\omega) \\
&= \int ||g(\omega)||_H \, \mathrm{dQ}(\omega) < \infty
\end{aligned}$$

with the first inequality holding true because of the second integral on its right hand side being finite (cf. Denkowski et al., 2003, Theorem 3.10.9) and the first one being finite for n sufficiently large, cf. (4.23). The same equation (4.23) additionally yields that $\lim_{n\to\infty} (\int ||s_n(\omega) - g(\omega)||_H \, \mathrm{dQ}(\omega))$ exists and the linearity of the limit can therefore be applied in the second step. Finally, (4.23) and the mentioned Theorem 3.10.9 yield the last two steps. □

Proof of Proposition 4.3.14. To shorten the notation, define $f_i := f_{\mathrm{P},\lambda,k_i}$, $i = 1, 2$, in this proof.

Define $\tilde{k}_i := \frac{k_i}{2}$ for $i = 1, 2$. By Lemma 2.1.11, $\tilde{H}_i = H_i$ (equipped with the norm $||\cdot||_{\tilde{H}_i} = \sqrt{2} \, ||\cdot||_{H_i}$) is the RKHS of $\tilde{k}_i$. Thus, $f_i \in \tilde{H}_i$ for $i = 1, 2$.

In the next step, define a new space which contains f_1 as well as f_2 by

$$\tilde{H} := \tilde{H}_1 \oplus \tilde{H}_2 := \left\{ g \colon \mathcal{X} \to \mathbb{R} \;\Big|\; g = g_1 + g_2, g_1 \in \tilde{H}_1, g_2 \in \tilde{H}_2 \right\} .$$

Berlinet and Thomas-Agnan (2004, Theorem 5) tells us that $\tilde{H}$ equipped with the norm

$$||g||_{\tilde{H}}^2 := \min_{g_1 \in \tilde{H}_1,\, g_2 \in \tilde{H}_2 \,:\, g_1+g_2=g} \left(||g_1||_{\tilde{H}_1}^2 + ||g_2||_{\tilde{H}_2}^2 \right) \qquad \forall\, g \in \tilde{H}$$

is the RKHS of the reproducing kernel $\tilde{k} := \tilde{k}_1 + \tilde{k}_2 = (k_1+k_2)/2$. Since obviously $f_1, f_2 \in \tilde{H}$, we now use this new RKHS as an aid for investigating the difference between f_1 and f_2.

First of all, because $\tilde{k}$ is measurable and bounded by $||\tilde{k}||_\infty \leq \frac{1}{2} \left(||k_1||_\infty + ||k_2||_\infty \right) < \infty$ and $\tilde{H}$ is obviously separable, there exists a unique SVM $f_{\mathrm{P},\lambda,\tilde{k}} =: \tilde{f}$ (Christmann et al., 2009, Theorem 7). The triangle inequality then yields

$$||f_1 - f_2||_\infty \leq ||f_1 - \tilde{f}||_\infty + ||f_2 - \tilde{f}||_\infty \,. \tag{4.27}$$

By applying Christmann et al. (2009, Theorem 7), both of the differences on the right hand side can be expanded as

$$\begin{aligned} f_i - \tilde{f} = & -\frac{1}{2\lambda} \cdot \int h_i(x,y) \Phi_i(x) \, \mathrm{dP}(x,y) \\ & + \frac{1}{2\lambda} \cdot \int \tilde{h}(x,y) \tilde{\Phi}(x) \, \mathrm{dP}(x,y) \\ = & \frac{1}{2\lambda} \cdot \int h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x) \right) \, \mathrm{dP}(x,y) \\ & + \frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y) \right) \tilde{\Phi}(x) \, \mathrm{dP}(x,y) \end{aligned} \tag{4.28}$$

with h_i and $\tilde{h}$ from the subdifferential of $L^\star$ (with respect to f_i respectively $\tilde{f}$). Thus, Lemma 2.1.10(i) yields for $i = 1, 2$

$$\begin{aligned} \left|\left| f_i - \tilde{f} \right|\right|_\infty \leq & \left|\left| \frac{1}{2\lambda} \cdot \int h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x) \right) \, \mathrm{dP}(x,y) \right|\right|_\infty \\ & + \left|\left| \frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y) \right) \tilde{\Phi}(x) \, \mathrm{dP}(x,y) \right|\right|_\infty \\ \leq & \left|\left| \frac{1}{2\lambda} \cdot \int h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x) \right) \, \mathrm{dP}(x,y) \right|\right|_\infty \\ & + ||\tilde{k}||_\infty \cdot \left|\left| \frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y) \right) \tilde{\Phi}(x) \, \mathrm{dP}(x,y) \right|\right|_{\tilde{H}} \,. \end{aligned} \tag{4.29}$$

Now, the first summand on the right hand side of (4.29) can easily be bounded by

$$\begin{aligned} \left|\left| \frac{1}{2\lambda} \cdot \int h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x) \right) \, \mathrm{dP}(x,y) \right|\right|_\infty & \leq \frac{1}{2\lambda} \cdot ||h_i||_\infty \cdot \sup_{x \in \mathcal{X}} \left|\left| \tilde{\Phi}(x) - \Phi_i(x) \right|\right|_\infty \\ & \leq \frac{|L|_1}{2\lambda} \cdot \left|\left| \tilde{k} - k_i \right|\right|_\infty \,, \end{aligned} \tag{4.30}$$

where we applied Lemma 4.3.15 in the first step and obtained the bound for h_i from Christmann et al. (2009, Theorem 7).

As for the square of the $\tilde{H}$-norm in the second summand on the right hand side of (4.29), applying (4.28) yields

$$\begin{aligned}
&\left|\left|\frac{1}{2\lambda}\cdot\int\Big(\tilde{h}(x,y)-h_i(x,y)\Big)\,\tilde{\Phi}(x)\,\mathrm{dP}(x,y)\right|\right|_{\tilde{H}}^2\\
&=\left\langle\frac{1}{2\lambda}\cdot\int\Big(\tilde{h}(x,y)-h_i(x,y)\Big)\,\tilde{\Phi}(x)\,\mathrm{dP}(x,y)\,,f_i-\tilde{f}\right\rangle_{\tilde{H}}\\
&\quad-\left\langle\frac{1}{2\lambda}\cdot\int\Big(\tilde{h}(x,y)-h_i(x,y)\Big)\,\tilde{\Phi}(x)\,\mathrm{dP}(x,y)\,,\right.\\
&\qquad\quad\left.\frac{1}{2\lambda}\cdot\int h_i(x',y')\left(\tilde{\Phi}(x')-\Phi_i(x')\right)\,\mathrm{dP}(x',y')\right\rangle_{\tilde{H}}\,,
\end{aligned}\tag{4.31}$$

where the reproducing property can be applied to the first of these two inner products in order to obtain

$$\begin{aligned}
&\left\langle\frac{1}{2\lambda}\cdot\int\Big(\tilde{h}(x,y)-h_i(x,y)\Big)\,\tilde{\Phi}(x)\,\mathrm{dP}(x,y),f_i-\tilde{f}\right\rangle_{\tilde{H}}\\
&=\frac{1}{2\lambda}\cdot\int\Big(\tilde{h}(x,y)-h_i(x,y)\Big)\left(f_i(x)-\tilde{f}(x)\right)\,\mathrm{dP}(x,y)\leq 0\,.
\end{aligned}$$

This inequality holds true because $L^\star$ is convex which implies that for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$ we have $s_1\leq s_2$ for every $s_1\in\partial L^\star(x,y,t_1)$, $s_2\in\partial L^\star(x,y,t_2)$ with $t_1\leq t_2$. Now there are two cases: Either at least one of the two factors in the integrand is zero or the two factors have different signs. Therefore, the integrand, and hence also the whole integral, is non-positive.

Plugging this into (4.31) results in

$$\begin{aligned}
&\left|\left|\frac{1}{2\lambda}\cdot\int\Big(\tilde{h}(x,y)-h_i(x,y)\Big)\,\tilde{\Phi}(x)\,\mathrm{dP}(x,y)\right|\right|_{\tilde{H}}^2\\
&\leq\left|\left\langle\frac{1}{2\lambda}\cdot\int\Big(\tilde{h}(x,y)-h_i(x,y)\Big)\,\tilde{\Phi}(x)\,\mathrm{dP}(x,y)\,,\right.\right.\\
&\qquad\quad\left.\left.\frac{1}{2\lambda}\cdot\int h_i(x',y')\left(\tilde{\Phi}(x')-\Phi_i(x')\right)\,\mathrm{dP}(x',y')\right\rangle_{\tilde{H}}\right|\\
&=\frac{1}{4\lambda^2}\cdot\left|\int\int\Big(\tilde{h}(x,y)-h_i(x,y)\Big)\,h_i(x',y')\left(\tilde{k}(x,x')-k_i(x,x')\right)\,\mathrm{dP}(x',y')\,\mathrm{dP}(x,y)\right|\\
&\leq\frac{1}{4\lambda^2}\cdot\left|\left|\tilde{h}-h_i\right|\right|_\infty\cdot||h_i||_\infty\cdot\left|\left|\tilde{k}-k_i\right|\right|_\infty\\
&\leq\frac{|L|_1^2}{2\lambda^2}\cdot\left|\left|\tilde{k}-k_i\right|\right|_\infty\,,
\end{aligned}\tag{4.32}$$

where we again applied the reproducing property in the second step and Christmann et al. (2009, Theorem 7) for bounding $\tilde{h}-h_i$ and h_i in the last step.

By the definition of $\tilde{k}$, we further know that

$$\left|\left|\tilde{k}-k_i\right|\right|_\infty=\left|\left|\frac{k_1+k_2}{2}-k_i\right|\right|_\infty=\left|\left|\frac{k_1-k_2}{2}\right|\right|_\infty=\frac{||k_1-k_2||_\infty}{2}$$

for $i = 1, 2$, as well as

$$\left|\left|\tilde{k}\right|\right|_\infty = \left|\left|\frac{k_1 + k_2}{2}\right|\right|_\infty \leq \frac{||k_1||_\infty + ||k_2||_\infty}{2} \leq \max\left\{||k_1||_\infty, ||k_2||_\infty\right\} = \kappa\,.$$

Thus, we obtain the assertion by combining (4.27) with (4.29), (4.30) and (4.32):

$$\begin{aligned}||f_1 - f_2||_\infty &\leq \left|\left|f_1 - \tilde{f}\right|\right|_\infty + \left|\left|f_2 - \tilde{f}\right|\right|_\infty \\ &\leq \sum_{i=1}^{2}\left(\frac{|L|_1}{2\lambda}\cdot\left|\left|\tilde{k} - k_i\right|\right|_\infty + \left|\left|\tilde{k}\right|\right|_\infty\cdot\frac{|L|_1}{\sqrt{2}\lambda}\cdot\sqrt{\left|\left|\tilde{k} - k_i\right|\right|_\infty}\right) \\ &\leq \frac{|L|_1}{\lambda}\cdot\left(\frac{1}{2}\cdot||k_1 - k_2||_\infty + \kappa\cdot\sqrt{||k_1 - k_2||_\infty}\right)\,.\end{aligned}$$

□

For similar minimal examples as those investigated in Examples 4.3.4 and 4.3.12, the linear part of the bound from Proposition 4.3.14 actually suffices. We slightly adapt the mentioned examples by not only considering Gaussian RBF kernels because the kernel is the focus of this example:

Example 4.3.16. Let $\mathcal{Y} = \mathbb{R}$, $L^\star = L^*_{0.5\text{-pin}}$ be the shifted 0.5-pinball loss,[36] $\lambda > 0$, and $\mathrm{P} = \delta_{(x_0,y_0)}$ be the Dirac distribution in some $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, where we assume $y_0 > 0$ for simplicity. Let further k_1 and k_2 be measurable and bounded kernels with separable RKHSs (for example, Gaussian RBF kernels, see Example 2.1.12). Similarly to Examples 4.3.4 and 4.3.12, one then obtains

$$f_{\mathrm{P},\lambda,k_i} = \frac{1}{4\lambda}\cdot\Phi_i(x_0)$$

for $i = 1, 2$ if $k_i(x_0, x_0) < 4\lambda y_0$. Hence, if both $k_1(x_0, x_0) < 4\lambda y_0$ and $k_2(x_0, x_0) < 4\lambda y_0$ hold true, we have

$$||f_{\mathrm{P},\lambda,k_1} - f_{\mathrm{P},\lambda,k_2}||_\infty = \frac{1}{4\lambda}\cdot||\Phi_1(x_0) - \Phi_2(x_0)||_\infty \leq \frac{1}{4\lambda}\cdot||k_1 - k_2||_\infty\,,$$

which exactly coincides with the linear part of the bound from Proposition 4.3.14 because $|L_{\tau\text{-pin}}|_1 = 1/2$.

Remark 4.3.17. We also examined further examples that were slightly more complex than Example 4.3.16 but still simple enough that it was feasible to derive a closed formula for the SVMs. That is, we looked at distributions P whose support did consist of different small amounts of points instead of only a single point, and considered kernels such as Gaussian RBF kernels or constant kernels. For all the examples we considered, the linear part of the bound from Proposition 4.3.14 always sufficed, for which reason we suspect that it might actually be possible to eliminate the other part from the bound from Proposition 4.3.14.

Additionally, an analogous result to Proposition 4.3.14 which however considers L_p-stability instead of sup-stability also holds true:

[36] Other loss functions satisfying certain conditions can also be used without changing much in the example.

Proposition 4.3.18. *Let Assumption 4.3.1 be satisfied and let* $\kappa := \max\{||k_1||_\infty, ||k_2||_\infty\}$*. Let* $p \in [1,\infty)$*. Then,*

$$||f_{\mathrm{P},\lambda,k_1} - f_{\mathrm{P},\lambda,k_2}||_{L_p(\mathrm{P}^X)} \leq \frac{|L|_1}{\lambda} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}^X \otimes \mathrm{P}^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}^X \otimes \mathrm{P}^X)}}\right) .$$

Proof. The proof is almost identical to that of Proposition 4.3.14 with $||\cdot||_\infty$ being replaced by $||\cdot||_{L_p(\mathrm{P}^X)}$, for which reason only the differences are highlighted here.

First of all, because of (4.9), we obtain analogously to (4.29)

$$\begin{aligned}
\left|\left|f_i - \tilde{f}\right|\right|_{L_p(\mathrm{P}^X)} &\leq \left|\left|\frac{1}{2\lambda} \cdot \int h_i(x,y)\left(\tilde{\Phi}(x) - \Phi_i(x)\right) \mathrm{dP}(x,y)\right|\right|_{L_p(\mathrm{P}^X)} \\
&\quad + \left|\left|\frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y)\right) \tilde{\Phi}(x) \, \mathrm{dP}(x,y)\right|\right|_\infty \\
&\leq \left|\left|\frac{1}{2\lambda} \cdot \int h_i(x,y)\left(\tilde{\Phi}(x) - \Phi_i(x)\right) \mathrm{dP}(x,y)\right|\right|_{L_p(\mathrm{P}^X)} \\
&\quad + \left|\left|\tilde{k}\right|\right|_\infty \cdot \left|\left|\frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y)\right) \tilde{\Phi}(x) \, \mathrm{dP}(x,y)\right|\right|_{\tilde{H}} .
\end{aligned}$$

Then, the first summand on the right hand side can be bounded in an analogous way to (4.30):

$$\begin{aligned}
&\left|\left|\frac{1}{2\lambda} \cdot \int_{\mathcal{X}\times\mathcal{Y}} h_i(x,y)\left(\tilde{\Phi}(x) - \Phi_i(x)\right) \mathrm{dP}(x,y)\right|\right|_{L_p(\mathrm{P}^X)} \\
&\leq \frac{1}{2\lambda} \int_{\mathcal{X}\times\mathcal{Y}} \left|\left|h_i(x,y)\left(\tilde{\Phi}(x) - \Phi_i(x)\right)\right|\right|_{L_p(\mathrm{P}^X)} \mathrm{dP}(x,y) \\
&\leq \frac{1}{2\lambda} \cdot ||h_i||_\infty \cdot \int_{\mathcal{X}} \left|\left|\tilde{\Phi}(x) - \Phi_i(x)\right|\right|_{L_p(\mathrm{P}^X)} \mathrm{dP}^X(x) \\
&= \frac{1}{2\lambda} \cdot ||h_i||_\infty \cdot \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \left|\tilde{k}(x,x') - k_i(x,x')\right|^p \mathrm{dP}^X(x)\right)^{1/p} \mathrm{dP}^X(x) \\
&\leq \frac{|L|_1}{2\lambda} \cdot \left|\left|\tilde{k} - k_i\right|\right|_{L_p(\mathrm{P}^X \otimes \mathrm{P}^X)} ,
\end{aligned}$$

where we applied Lemma 4.3.15 in the first step, and Christmann et al. (2009, Theorem 7) (for obtaining the bound on h_i) as well as Hölder's inequality in the last step. Finally, it is possible to tighten the bound from the last steps of (4.32) in the following way:

$$\begin{aligned}
&\frac{1}{4\lambda^2} \cdot \left|\int\int \left(\tilde{h}(x,y) - h_i(x,y)\right) h_i(x',y') \left(\tilde{k}(x,x') - k_i(x,x')\right) \mathrm{dP}(x',y') \, \mathrm{dP}(x,y)\right| \\
&\leq \frac{|L|_1^2}{2\lambda^2} \cdot \int\int \left|\tilde{k}(x,x') - k_i(x,x')\right| \mathrm{dP}(x',y') \, \mathrm{dP}(x,y) \\
&= \frac{|L|_1^2}{2\lambda^2} \cdot \left|\left|\tilde{k} - k_i\right|\right|_{L_1(\mathrm{P}^X \otimes \mathrm{P}^X)} \\
&\leq \frac{|L|_1^2}{2\lambda^2} \cdot \left|\left|\tilde{k} - k_i\right|\right|_{L_p(\mathrm{P}^X \otimes \mathrm{P}^X)} .
\end{aligned}$$

The assertion then follows in the same way as in the proof of Proposition 4.3.14. □

4.3.4 Total Stability

In this section, the results from Sections 4.3.1 to 4.3.3 are combined into bounds on the difference between two SVMs differing in the whole triple (P, λ, k), that is, into results on total stability of SVMs. All results are based on decomposing the difference between the two SVMs as

$$\begin{aligned} ||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty \leq\ & ||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_1,k_1}||_\infty + ||f_{\mathrm{P}_2,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_1}||_\infty \\ & + ||f_{\mathrm{P}_2,\lambda_2,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty \end{aligned} \tag{4.33}$$

for total sup-stability and as

$$\begin{aligned} ||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)} \leq\ & ||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_1,k_1}||_{L_p(\mathrm{P}_i^X)} \\ & + ||f_{\mathrm{P}_2,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_1}||_{L_p(\mathrm{P}_i^X)} \\ & + ||f_{\mathrm{P}_2,\lambda_2,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)} \end{aligned} \tag{4.34}$$

for total L_p-stability. Of course, the order of decomposition can also be varied, for example changing λ instead of P in the first step, thus considering the difference between $f_{\mathrm{P}_1,\lambda_1,k_1}$ and $f_{\mathrm{P}_1,\lambda_2,k_1}$ in the first summand on the right hand side of the decompositions. This is also taken into account in the results.

The first theorem considers total sup-stability (with the difference between the probability measures being measured by the total variation distance) and, as already mentioned in Section 4.1, considerably generalizes Christmann et al. (2018, Theorem 2.7): First of all, an additional condition on L that was required by Christmann et al. (2018) gets eliminated. Previously, L did not only need to be convex and Lipschitz continuous but also differentiable. Since many popular loss functions are not differentiable (e.g., pinball loss, ε-insensitive loss, hinge loss), this change makes the result applicable to a considerably larger class of learning tasks. Secondly, in Christmann et al. (2018, Theorem 2.7) it was assumed that the regularization parameters λ_1 and λ_2 were greater than some specified positive constant, which is unsatisfactory because the regularization parameter used by an SVM has to converge to zero as the size of the training data set tends to infinity in order to achieve consistency, cf. Chapter 3.

In order to circumvent the latter problem, Christmann et al. (2018) additionally provided another result (Theorem 2.10), in which λ_1 and λ_2 are allowed to be arbitrarily close to zero and instead of $||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty$ they bound $||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{H_1}$. Because of Lemma 2.1.10(i) and because k_1 is assumed to be bounded, this also translates to a bound for $||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty$. Alas, this result obviously requires the RKHSs H_1 and H_2 be nested, $H_2 \subseteq H_1$, and additionally uses $||k_1 - k_2||_{H_1}$ instead of the more easily interpretable $||k_1 - k_2||_\infty$ in the bound.

In the subsequent theorem, it is neither required that λ_1 and λ_2 are greater than some positive constant nor that H_1 and H_2 are nested. Furthermore, the improvement in the bound coming from the difference between λ_1 and λ_2 derived in Section 4.3.2 transfers to this theorem as well.

Theorem 4.3.19. *Let Assumption 4.3.1 be satisfied. Denote* $\kappa := \max\{||k_1||_\infty, ||k_2||_\infty\}$ *and* $\tau := \min\{\lambda_1, \lambda_2\}$*. Then,*

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty \leq \frac{|L|_1}{\tau} \cdot \Bigg(\kappa^2 \cdot d_{\mathrm{tv}}(\mathrm{P}_1,\mathrm{P}_2) + \frac{\kappa^2}{2\tau} \cdot |\lambda_1 - \lambda_2| \\ + \frac{1}{2} \cdot ||k_1 - k_2||_\infty + \kappa \cdot \sqrt{||k_1 - k_2||_\infty}\Bigg) .$$

Proof. Applying Propositions 4.3.3, 4.3.10 and 4.3.14 to the decomposition (4.33) of the investigated norm $||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty$ yields

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty \leq \frac{||k_1||_\infty^2 |L|_1}{\lambda_1} \cdot d_{\mathrm{tv}}(\mathrm{P}_1,\mathrm{P}_2) + \frac{||k_1||_\infty^2 |L|_1}{2\min\{\lambda_1,\lambda_2\}^2} \cdot |\lambda_1 - \lambda_2| \\ + \frac{|L|_1}{\lambda_2} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_\infty + \kappa \cdot \sqrt{||k_1 - k_2||_\infty}\right) .$$

Since the order of decomposition can of course be varied freely, we also obtain analogous bounds with k_1 being replaced by k_2 (and vice versa) as well as λ_1 by λ_2 (and vice versa) in some of these summands. Since the right hand side of the assertion is greater or equal to the right hand sides of all of the bounds generated this way, the assertion directly follows. □

Recall the examples from Section 4.2 on how the different quantities on the right hand side of the bound behave in different situations.

Remark 4.3.20. To formally use the metric d_3 from (4.6) in Definition 4.1.1 of total stability of SVMs, the bound from Theorem 4.3.19 can of course be bounded further by applying

$$\frac{1}{2} \cdot ||k_1 - k_2||_\infty + \kappa \cdot \sqrt{||k_1 - k_2||_\infty} \leq \max\left\{\frac{1}{2}, \kappa\right\} \cdot \left(||k_1 - k_2||_\infty + \sqrt{||k_1 - k_2||_\infty}\right) \\ = \max\left\{\frac{1}{2}, \kappa\right\} \cdot d_3(k_1, k_2) .$$

This can be done similarly in the remaining results of this section.

Remark 4.3.21. It is also possible to state an analogous result to Theorem 4.3.19 that uses $||k_1 - k_2||_\infty$ instead of $||k_1 - k_2||_\infty + \sqrt{||k_1 - k_2||_\infty}$, in exchange for imposing the additional assumptions regarding loss function and regularization parameters that were explained before this theorem, cf. Christmann et al. (2018, Theorem 2.7).

In addition, it is also possible to analogously derive a new result on total sup-stability using the Wasserstein distance to measure the distance between the probability measures. This has the advantage of also yielding meaningful bounds in situations in which Theorem 4.3.19 does not yield such meaningful bounds, cf. Section 4.2.1.

Theorem 4.3.22. *Let Assumptions 4.3.1 and 4.3.2 be satisfied. Denote*

$$\kappa := \max\{||k_1||_\infty, ||k_2||_\infty\} , \\ \tau := \min\{\lambda_1, \lambda_2\} , \\ \eta := \max_{j=1,2}\left(||k_j||_\infty \cdot \max\left\{|\psi|_1 c_{k_j} + \left(-2\varphi_j'(0)\right)^{1/2} \frac{|\psi|_1 |\psi'|_1 ||k_j||_\infty^2}{\tau}, |\psi'|_1 ||k_j||_\infty\right\}\right) .$$

Then,

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty \leq \frac{1}{\tau} \cdot \Bigg(\eta \cdot d_{\mathrm{W}}(\mathrm{P}_1, \mathrm{P}_2) + \frac{|\psi|_1 \kappa^2}{2\tau} \cdot |\lambda_1 - \lambda_2| \\ + \frac{|\psi|_1}{2} \cdot ||k_1 - k_2||_\infty + |\psi|_1 \kappa \cdot \sqrt{||k_1 - k_2||_\infty} \Bigg) .$$

Proof. The proof works analogously to that of Theorem 4.3.19 by just applying Proposition 4.3.5 instead of Proposition 4.3.3. □

Again, recall the examples from Section 4.2 on how the different quantities on the right hand side of the bound behave in different situations. Additionally, applying Corollary 4.3.7 instead of Proposition 4.3.5 in the proof yields a more specialized result for the Gaussian RBF kernel, in which the properties of that kernel are plugged in.

Remark 4.3.23. Similarly to Remark 4.3.21, it is also possible to derive an analogous result to Theorem 4.3.22 that uses $||k_1 - k_2||_\infty$ instead of $||k_1 - k_2||_\infty + \sqrt{||k_1 - k_2||_\infty}$, in exchange for imposing the additional assumptions regarding regularization parameters that were explained at the beginning of this section by applying Christmann et al. (2018, Lemma 6.4) instead of Proposition 4.3.14 in the proof.

Now, the subsequent Theorem 4.3.24 states a result which is very similar to Theorem 4.3.19 but which shows total L_p-stability instead of total sup-stability:

Theorem 4.3.24. *Let Assumption 4.3.1 be satisfied. Denote* $\kappa := \max\{||k_1||_\infty, ||k_2||_\infty\}$ *and* $\tau := \min\{\lambda_1, \lambda_2\}$. *Then, for all* $p \in [1, \infty)$ *and* $i \in \{1, 2\}$,

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)} \\ \leq \frac{|L|_1}{\tau} \cdot \Bigg(\kappa^2 \cdot d_{\mathrm{tv}}(\mathrm{P}_1, \mathrm{P}_2) + \frac{\kappa^2}{2\tau} \cdot |\lambda_1 - \lambda_2| \\ + \frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)}} \Bigg) .$$

Proof. Applying Proposition 4.3.18 as well as Propositions 4.3.3 and 4.3.10 in combination with (4.9) to the decomposition (4.34) of $||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)}$ yields for $i = 2$

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_2^X)} \\ \leq \frac{||k_1||_\infty^2 |L|_1}{\lambda_1} \cdot d_{\mathrm{tv}}(\mathrm{P}_1, \mathrm{P}_2) + \frac{||k_1||_\infty^2 |L|_1}{2 \min\{\lambda_1, \lambda_2\}^2} \cdot |\lambda_1 - \lambda_2| \\ + \frac{|L|_1}{\lambda_2} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_2^X \otimes \mathrm{P}_2^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_2^X \otimes \mathrm{P}_2^X)}} \right) \\ \leq \frac{\kappa^2 |L|_1}{\tau} \cdot d_{\mathrm{tv}}(\mathrm{P}_1, \mathrm{P}_2) + \frac{\kappa^2 |L|_1}{2\tau^2} \cdot |\lambda_1 - \lambda_2| \\ + \frac{|L|_1}{\tau} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_2^X \otimes \mathrm{P}_2^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_2^X \otimes \mathrm{P}_2^X)}} \right) .$$

Analogously, reversing the order of decomposition (such that Proposition 4.3.18 can be applied to a summand with probability measure P_1 in both SVMs) yields for $i = 1$

$$\begin{aligned}
&||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_1^X)} \\
&\le \frac{|L|_1}{\lambda_1} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_1^X \otimes \mathrm{P}_1^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_1^X \otimes \mathrm{P}_1^X)}} \right) \\
&\quad + \frac{||k_2||_\infty^2 \, |L|_1}{2 \min\{\lambda_1, \lambda_2\}^2} \cdot |\lambda_1 - \lambda_2| + \frac{||k_2||_\infty^2 \, |L|_1}{\lambda_2} \cdot d_{\mathrm{tv}}(\mathrm{P}_1, \mathrm{P}_2) \\
&\le \frac{\kappa^2 |L|_1}{\tau} \cdot d_{\mathrm{tv}}(\mathrm{P}_1, \mathrm{P}_2) + \frac{\kappa^2 |L|_1}{2\tau^2} \cdot |\lambda_1 - \lambda_2| \\
&\quad + \frac{|L|_1}{\tau} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_1^X \otimes \mathrm{P}_1^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_1^X \otimes \mathrm{P}_1^X)}} \right) . \qquad \square
\end{aligned}$$

As it was the case for the result on total sup-stability, it is also possible to analogously derive a result on total L_p-stability that uses the Wasserstein distance instead of the total variation distance in the bound.

Theorem 4.3.25. *Let Assumptions 4.3.1 and 4.3.2 be satisfied. Denote*

$$\begin{aligned}
&\kappa := \max\left\{||k_1||_\infty, ||k_2||_\infty\right\} , \\
&\tau := \min\left\{\lambda_1, \lambda_2\right\} , \\
&\eta := \max_{j=1,2} \left(||k_j||_\infty \cdot \max\left\{ |\psi|_1 c_{k_j} + \left(-2\varphi_j'(0)\right)^{1/2} \frac{|\psi|_1 |\psi'|_1 \, ||k_j||_\infty^2}{\tau} , \, |\psi'|_1 \, ||k_j||_\infty \right\} \right) .
\end{aligned}$$

Then, for all $p \in [1, \infty)$ and $i \in \{1, 2\}$,

$$\begin{aligned}
&||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)} \\
&\le \frac{1}{\tau} \cdot \Bigg(\eta \cdot d_{\mathrm{W}}(\mathrm{P}_1, \mathrm{P}_2) + \frac{|\psi|_1 \kappa^2}{2\tau} \cdot |\lambda_1 - \lambda_2| \\
&\qquad + \frac{|\psi|_1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)} + |\psi|_1 \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)}} \Bigg) .
\end{aligned}$$

Proof. The proof works analogously to that of Theorem 4.3.24 by just applying Proposition 4.3.5 instead of Proposition 4.3.3. □

The results on total L_p-stability become particularly useful in Section 4.4.2 where the total stability of localized SVMs is investigated. In that section, it will be explained that no meaningful result on total sup-stability can be derived in this situation, but it will at least still be possible to derive results on total L_p-stability, which are based on Theorems 4.3.24 and 4.3.25.

4.4 Total Stability of Localized Support Vector Machines

The goal of this section lies in deriving total stability of localized SVMs. As already hinted at in Section 4.1, it is however only possible to derive meaningful results on total

L_p-stability and not on total sup-stability. Therefore, before investigating total stability in Section 4.4.2, regionalization-subtotal stability is examined in Section 4.4.1 because for this it is also possible to obtain meaningful results on sup-stability.

The parts of this section using the total variation distance already appeared in the peer-reviewed paper Köhler and Christmann (2022, Section 3), which was published in *Journal of Machine Learning Research*. The parts using the Wasserstein distance have not been published before.

4.4.1 Regionalization-Subtotal Stability

As this section considers regionalization-subtotal stability, it is about the comparison of two localized SVMs $f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}}$ and $f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}}$ that are based on the same regionalization $\boldsymbol{\mathcal{X}} \coloneqq \{\mathcal{X}_1,\ldots,\mathcal{X}_A\}$ of size $A \in \mathbb{N}$ (and on the same weight functions). As in the definition of regionalization-subtotal stability (cf. Definition 4.1.2), these localized SVMs are allowed to be based on different probability measures P_1 and P_2, on different vectors of regularization parameters $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ as well as on different vectors of kernels $\boldsymbol{k_1}$ and $\boldsymbol{k_2}$ on $\boldsymbol{\mathcal{X}}$. The results from this section need the following assumptions regarding the components influencing the localized SVMs:

Assumption 4.4.1.

- Let $L\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R}\to[0,\infty)$ be a convex, Lipschitz continuous loss function and let $L^\star$ be its shifted version.
- Let $\boldsymbol{\mathcal{X}} \coloneqq \{\mathcal{X}_1,\ldots,\mathcal{X}_A\}$ be a regionalization of $\mathcal{X}$ of size $A\in\mathbb{N}$ and let the weight functions w_a, $a=1,\ldots,A$, satisfy **(W1)**, **(W2)**, **(W3)**.
- For $i=1,2$, let P_i be a probability measure on $\mathcal{X}\times\mathcal{Y}$ that is positive on $\boldsymbol{\mathcal{X}}$.
- For $i=1,2$, let $\boldsymbol{\lambda_i} \coloneqq (\lambda_{i,1},\ldots,\lambda_{i,A})\in(0,\infty)^A$.
- For $i=1,2$, let $\boldsymbol{k_i} \coloneqq (k_{i,1},\ldots,k_{i,A})$ be a vector of bounded and measurable kernels on $\boldsymbol{\mathcal{X}}$ with separable RKHSs $H_{i,a}$, $a=1,\ldots,A$.

Note again that the separability of the RKHSs is always satisfied if the associated kernels are continuous, cf. Lemma 2.1.10(iii). For $i\in\{1,2\}$ and $a\in\{1,\ldots,A\}$, further introduce the shortening notation $\mathrm{P}_{i,a} \coloneqq \mathrm{P}_{i,\mathcal{X}_a}$ based on the local probability measures defined in (2.6).

For the results using the Wasserstein distance in the bound, the following is required as well:

Assumption 4.4.2.

- Let $\mathcal{X}\subseteq\mathbb{R}^d$ for some $d\in\mathbb{N}$.
- Let L be distance-based with representing function ψ such that ψ is differentiable and ψ as well as its derivative ψ' are both Lipschitz continuous.

- For $a = 1, \dots, A$, let $\mathcal{X}_a$ be complete.
- For $i = 1, 2$ and $a = 1, \dots, A$, let $\tilde{\mathcal{X}}_{i,a} \supseteq \mathcal{X}_a$ with $\tilde{\mathcal{X}}_{i,a} \subseteq \mathbb{R}^d$ be open and assume that $k_{i,a} = \tilde{k}_{i,a}|_{\mathcal{X}_a \times \mathcal{X}_a}$ for a kernel $\tilde{k}_{i,a}$ on $\tilde{\mathcal{X}}_{i,a}$ which can be written as $\tilde{k}_{i,a}(x, x') = \varphi_{i,a}(||x - x'||_2)$ for all $x, x' \in \tilde{\mathcal{X}}_{i,a}$, where $\varphi_{i,a} : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a twice continuously differentiable function satisfying $\varphi_{i,a}(0) - \varphi_{i,a}(r) \leq \frac{c_{k_{i,a}}^2}{2} \cdot r$ for all $r \geq 0$ for some $c_{k_{i,a}} \geq 0$.

Recall that popular kernels such as the Gaussian RBF kernels can be written is such a way as it is requested in Assumption 4.4.2, cf. Corollary 4.3.7.

The succeeding theorem states that Theorem 4.3.19 can be transferred to the situation at hand, i.e., that localized SVMs inherit regionalization-subtotal sup-stability from the total sup-stability of regular SVMs:

Theorem 4.4.3. *Let Assumption 4.4.1 be satisfied. Denote*

$$\kappa_a := \max\left\{||k_{1,a}||_\infty, ||k_{2,a}||_\infty\right\} ,$$
$$\tau_a := \min\{\lambda_{1,a}, \lambda_{2,a}\} ,$$

for all $a \in \{1, \dots, A\}$. Then,

$$\begin{aligned} &||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}}||_\infty \\ &\leq |L|_1 \cdot \max_{a \in \{1,\dots,A\}} \frac{1}{\tau_a} \cdot \left(\kappa_a^2 \cdot d_{\mathrm{tv}}(\mathrm{P}_{1,a}, \mathrm{P}_{2,a}) + \frac{\kappa_a^2}{2\tau_a} \cdot |\lambda_{1,a} - \lambda_{2,a}| \right. \\ &\qquad \left. + \frac{1}{2} \cdot ||k_{1,a} - k_{2,a}||_\infty + \kappa_a \cdot \sqrt{||k_{1,a} - k_{2,a}||_\infty} \right) . \end{aligned}$$

Proof. To shorten the notation, define $f_i := f_{\mathrm{P}_i,\boldsymbol{\lambda_i},\boldsymbol{k_i},\boldsymbol{\mathcal{X}}}$ and $f_{i,a} := f_{\mathrm{P}_{i,a},\lambda_{i,a},k_{i,a}}$, $i = 1, 2$, $a = 1, \dots, A$, in this proof. By the definition of f_1 and f_2,

$$\begin{aligned} ||f_1 - f_2||_\infty &\leq \sup_{x \in \mathcal{X}} \sum_{a=1}^{A} w_a(x) \cdot \left| \hat{f}_{1,a}(x) - \hat{f}_{2,a}(x) \right| \\ &\leq \sup_{x \in \mathcal{X}} \max_{a \in \{1,\dots,A\}} \left| \hat{f}_{1,a}(x) - \hat{f}_{2,a}(x) \right| \\ &= \max_{a \in \{1,\dots,A\}} \left|\left| \hat{f}_{1,a} - \hat{f}_{2,a} \right|\right|_\infty , \end{aligned} \tag{4.35}$$

where we applied **(W1)** and **(W2)** in the second step. Since the functions $\hat{f}_{i,a}$ have not been defined as SVMs but instead as zero-extensions of SVMs $f_{\mathrm{P}_{i,a},\lambda_{i,a},k_{i,a}}$ on $\mathcal{X}_a$, we cannot apply Theorem 4.3.19 to the right hand side of (4.35) yet. However, these functions can actually be seen as SVMs on $\mathcal{X}$ themselves, $\hat{f}_{i,a} = f_{\hat{\mathrm{P}}_{i,a},\lambda_{i,a},\hat{k}_{i,a}}$ (where $\hat{\mathrm{P}}_{i,a}$ and $\hat{k}_{i,a}$ denote the zero-extensions of $\mathrm{P}_{i,a}$ and $k_{i,a}$ respectively):

According to Meister and Steinwart (2016, Lemma 2), we have $\hat{H}_{i,a} = \{\hat{g} \mid g \in H_{i,a}\}$ and $||\hat{g}||_{\hat{H}_{i,a}} = ||g||_{H_{i,a}}$ for all $g \in H_{i,a}$ for $\hat{H}_{i,a}$ denoting the RKHS of $\hat{k}_{i,a}$. Since additionally

$\mathcal{R}_{L^\star,\hat{\mathrm{P}}_{i,a}}(\hat{g}) = \mathcal{R}_{L^\star,\mathrm{P}_{i,a}}(g)$ for all $g \in H_{i,a}$ (because the whole probability mass of $\hat{\mathrm{P}}_{i,a}$ is on $\mathcal{X}_a$ where $\hat{g}$ and g coincide), the definition of SVMs yields $f_{\hat{\mathrm{P}}_{i,a},\lambda_{i,a},\hat{k}_{i,a}} = \hat{f}_{\mathrm{P}_{i,a},\lambda_{i,a},k_{i,a}}$ $(= \hat{f}_{i,a})$.

Thus, Theorem 4.3.19 can be applied to the right hand side of (4.35) since the functions $\hat{f}_{i,a}$ are actually SVMs on the complete space $\mathcal{X}$ (whereas the functions $f_{i,a}$ are SVMs on the not necessarily complete spaces $\mathcal{X}_a$, for which reason the theorem can not be applied to $||f_{1,a} - f_{2,a}||_\infty$ even though this term is obviously equivalent to $||\hat{f}_{1,a} - \hat{f}_{2,a}||_\infty$). By doing this, the assertion follows, but with every $\mathrm{P}_{i,a}$ replaced by $\hat{\mathrm{P}}_{i,a}$ and $k_{i,a}$ by $\hat{k}_{i,a}$. Because of them just being zero-extensions of $\mathrm{P}_{i,a}$ and $k_{i,a}$ respectively, this does however not influence the respective distances. □

Remark 4.4.4. To formally obtain a bound exactly as proposed in Definition 4.1.2 and based on the distance measures from Section 4.2, it suffices to further bound the inequality from Theorem 4.4.3 by

$$\begin{aligned}
&||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}}||_\infty \\
&\leq |L|_1 \cdot \max_{a\in\{1,\dots,A\}} \frac{1}{\tau_a} \cdot \Bigg(\kappa_a^2 \cdot d_{\mathrm{tv}}(\mathrm{P}_{1,a},\mathrm{P}_{2,a}) + \frac{\kappa_a^2}{2\tau_a} \cdot |\lambda_{1,a} - \lambda_{2,a}| \\
&\qquad\qquad\qquad\qquad + \frac{1}{2} \cdot ||k_{1,a} - k_{2,a}||_\infty + \kappa_a \cdot \sqrt{||k_{1,a} - k_{2,a}||_\infty} \Bigg) \\
&\leq |L|_1 \cdot \Bigg(\max_{a\in\{1,\dots,A\}} \frac{\kappa_a^2}{\tau_a} \cdot \max_{a\in\{1,\dots,A\}} d_{\mathrm{tv}}(\mathrm{P}_{1,a},\mathrm{P}_{2,a}) + \max_{a\in\{1,\dots,A\}} \frac{\kappa_a^2}{2\tau_a^2} \cdot \max_{a\in\{1,\dots,A\}} |\lambda_{1,a} - \lambda_{2,a}| \\
&\qquad\quad + \max_{a\in\{1,\dots,A\}} \frac{\max\{\frac{1}{2},\kappa_a\}}{\tau_a} \cdot \max_{a\in\{1,\dots,A\}} \Big(||k_{1,a} - k_{2,a}||_\infty + \sqrt{||k_{1,a} - k_{2,a}||_\infty} \Big) \Bigg) \\
&\leq \left(\max_{i\in\{1,2\}} \max_{a\in\{1,\dots,A\}} \frac{|L|_1}{\lambda_{i,a}} \right) \cdot \left(\max_{i\in\{1,2\}} \max_{a\in\{1,\dots,A\}} ||k_{i,a}||_\infty^2 \right) \cdot d_1(\mathbf{P_{1,\boldsymbol{\mathcal{X}}}}, \mathbf{P_{2,\boldsymbol{\mathcal{X}}}}) \\
&\quad + \left(\max_{i\in\{1,2\}} \max_{a\in\{1,\dots,A\}} \frac{|L|_1}{2\lambda_{i,a}^2} \right) \cdot \left(\max_{i\in\{1,2\}} \max_{a\in\{1,\dots,A\}} ||k_{i,a}||_\infty^2 \right) \cdot d_2(\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}) \\
&\quad + \left(\max_{i\in\{1,2\}} \max_{a\in\{1,\dots,A\}} \frac{|L|_1}{\lambda_{i,a}} \right) \cdot \left(\max_{i\in\{1,2\}} \max_{a\in\{1,\dots,A\}} \max\left\{ \frac{1}{2}, ||k_{i,a}||_\infty \right\} \right) \cdot d_3(\boldsymbol{k_1}, \boldsymbol{k_2}),
\end{aligned}$$

denoting by $\mathbf{P_{1,\boldsymbol{\mathcal{X}}}}, \mathbf{P_{2,\boldsymbol{\mathcal{X}}}}$ vectors of local measures as defined in Section 2.2.2. For the remaining results from this section as well as for those from Section 4.4.2, one can proceed analogously to formally obtain a bound in the exact shape of that in Definition 4.1.2 respectively that in Definition 4.1.3.

As for non-localized SVMs, it is again possible to derive a result on regionalization-subtotal sup-stability using the Wasserstein distance in an analogous way.

Theorem 4.4.5. *Let Assumptions 4.4.1 and 4.4.2 be satisfied. Denote*

$$\begin{aligned}
&\kappa_a := \max\left\{||k_{1,a}||_\infty, ||k_{2,a}||_\infty\right\} , \\
&\tau_a := \min\left\{\lambda_{1,a}, \lambda_{2,a}\right\} , \\
&\eta_a := \max_{j=1,2}\Bigg(||k_{j,a}||_\infty \\
&\qquad\qquad \cdot \max\left\{ |\psi|_1 c_{k_{j,a}} + \left(-2\varphi'_{j,a}(0)\right)^{1/2} \frac{|\psi|_1|\psi'|_1\, ||k_{j,a}||_\infty^2}{\tau_a} ,\ |\psi'|_1\, ||k_{j,a}||_\infty \right\} \Bigg) ,
\end{aligned}$$

for all $a \in \{1, \dots, A\}$. Then,

$$\begin{aligned}
&||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}}||_\infty \\
&\leq \max_{a\in\{1,\dots,A\}} \frac{1}{\tau_a} \cdot \Bigg(\eta_a \cdot d_{\mathrm{W}}(\mathrm{P}_{1,a}, \mathrm{P}_{2,a}) + \frac{|\psi|_1 \kappa_a^2}{2\tau_a} \cdot |\lambda_{1,a} - \lambda_{2,a}| \\
&\qquad\qquad + \frac{|\psi|_1}{2} \cdot ||k_{1,a} - k_{2,a}||_\infty + |\psi|_1 \kappa_a \cdot \sqrt{||k_{1,a} - k_{2,a}||_\infty} \Bigg) .
\end{aligned}$$

Proof. We have

$$\begin{aligned}
||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}}||_\infty &\leq \max_{a\in\{1,\dots,A\}} \left|\left|\hat{f}_{\mathrm{P}_{1,a},\lambda_{1,a},k_{1,a}} - \hat{f}_{\mathrm{P}_{2,a},\lambda_{2,a},k_{2,a}}\right|\right|_\infty \\
&= \max_{a\in\{1,\dots,A\}} \left|\left|f_{\mathrm{P}_{1,a},\lambda_{1,a},k_{1,a}} - f_{\mathrm{P}_{2,a},\lambda_{2,a},k_{2,a}}\right|\right|_\infty ,
\end{aligned}$$

where the first step is equivalent to (4.35) and the second follows from $f_{\mathrm{P}_{1,a},\lambda_{1,a},k_{1,a}}$ and $f_{\mathrm{P}_{2,a},\lambda_{2,a},k_{2,a}}$ being defined on the same region $\mathcal{X}_a$. By Remark 3.4.2, each region $\mathcal{X}_a$, $a = 1, \dots, A$, is separable again. As the regions are additionally assumed to be complete, Theorem 4.3.25 can be applied to the norms on the right hand side, which yields the assertion. □

Similarly, Theorem 4.3.24 can be transferred as well in order to obtain results on regionalization-subtotal L_p-stability, first based on the total variation distance.

Theorem 4.4.6. *Let Assumption 4.4.1 be satisfied. Denote*

$$\begin{aligned}
&\kappa_a := \max\left\{||k_{1,a}||_\infty, ||k_{2,a}||_\infty\right\} , \\
&\tau_a := \min\left\{\lambda_{1,a}, \lambda_{2,a}\right\} ,
\end{aligned}$$

for all $a \in \{1, \dots, A\}$. Then, for all $p \in [1, \infty)$ and $i \in \{1, 2\}$,

$$\begin{aligned}
&||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}}||_{L_p(\mathrm{P}_i^X)} \\
&\leq |L|_1 \cdot \sum_{a=1}^{A} \frac{\left(\mathrm{P}_i^X(\mathcal{X}_a)\right)^{1/p}}{\tau_a} \cdot \Bigg(\kappa_a^2 \cdot d_{\mathrm{tv}}(\mathrm{P}_{1,a}, \mathrm{P}_{2,a}) + \frac{\kappa_a^2}{2\tau_a} \cdot |\lambda_{1,a} - \lambda_{2,a}| \\
&\qquad\qquad + \frac{1}{2} \cdot ||k_{1,a} - k_{2,a}||_{L_p(\mathrm{P}_{i,a}^X \otimes \mathrm{P}_{i,a}^X)} \\
&\qquad\qquad + \kappa_a \cdot \sqrt{||k_{1,a} - k_{2,a}||_{L_p(\mathrm{P}_{i,a}^X \otimes \mathrm{P}_{i,a}^X)}} \Bigg) .
\end{aligned}$$

Proof. To shorten the notation, define $f_i := f_{\mathrm{P}_i,\boldsymbol{\lambda_i},\boldsymbol{k_i},\boldsymbol{\mathcal{X}}}$ and $f_{i,a} := f_{\mathrm{P}_{i,a},\lambda_{i,a},k_{i,a}}$, $i = 1, 2$, $a = 1, \dots, A$, in this proof. By the definition of f_1 and f_2,

$$\begin{aligned}||f_1 - f_2||_{L_p(\mathrm{P}_i^X)} &\leq \sum_{a=1}^{A} \left|\left|w_a \cdot \left(\hat{f}_{1,a} - \hat{f}_{2,a}\right)\right|\right|_{L_p(\mathrm{P}_i^X)} \\ &\leq \sum_{a=1}^{A} \left(\int_{\mathcal{X}} \left|\hat{f}_{1,a}(x) - \hat{f}_{2,a}(x)\right|^p \,\mathrm{d}\mathrm{P}_i^X(x)\right)^{1/p} \\ &= \sum_{a=1}^{A} \left(\mathrm{P}_i^X(\mathcal{X}_a) \cdot \int_{\mathcal{X}_a} \left|\hat{f}_{1,a}(x) - \hat{f}_{2,a}(x)\right|^p \,\mathrm{d}\mathrm{P}_{i,a}^X(x)\right)^{1/p} \\ &= \sum_{a=1}^{A} \left(\mathrm{P}_i^X(\mathcal{X}_a)\right)^{1/p} \cdot \left(\int_{\mathcal{X}} \left|\hat{f}_{1,a}(x) - \hat{f}_{2,a}(x)\right|^p \,\mathrm{d}\hat{\mathrm{P}}_{i,a}^X(x)\right)^{1/p} \\ &= \sum_{a=1}^{A} \left(\mathrm{P}_i^X(\mathcal{X}_a)\right)^{1/p} \cdot \left|\left|\hat{f}_{1,a} - \hat{f}_{2,a}\right|\right|_{L_p(\hat{\mathrm{P}}_{i,a}^X)} . \end{aligned} \tag{4.36}$$

Here, we applied **(W1)** in the second, $\hat{f}_{1,b}$ and $\hat{f}_{2,b}$ being zero on $\mathcal{X} \setminus \mathcal{X}_b$ in combination with the definition of the local measures $\mathrm{P}_{i,a}$ (cf. (2.6)) in the third, and the definition of $\hat{\mathrm{P}}_{i,b}$ as zero-extension of $\mathrm{P}_{i,b}$ in the fourth step.

Noting that $\hat{f}_{1,a}$ and $\hat{f}_{2,a}$ are SVMs on $\mathcal{X}$ themselves, $\hat{f}_{i,a} = f_{\hat{\mathrm{P}}_{i,a},\lambda_{i,a},\hat{k}_{i,a}}$ (cf. proof of Theorem 4.4.3), Theorem 4.3.24 can now be applied to the norms on the right hand side of (4.36). This yields the assertion (as in the proof of Theorem 4.4.3 with $\hat{\mathrm{P}}_{i,a}$ and $\hat{k}_{i,a}$ instead ob $\mathrm{P}_{i,a}$ and $k_{i,a}$ which does not change the respective norms). □

Similarly to how it was the case for sup-stability, an analogous result using the Wasserstein distance also follows directly.

Theorem 4.4.7. *Let Assumptions 4.4.1 and 4.4.2 be satisfied. Denote*

$$\begin{aligned}\kappa_a &:= \max\left\{||k_{1,a}||_\infty, ||k_{2,a}||_\infty\right\} , \\ \tau_a &:= \min\left\{\lambda_{1,a}, \lambda_{2,a}\right\} , \\ \eta_a &:= \max_{j=1,2} \Bigg(||k_{j,a}||_\infty \\ &\qquad \cdot \max\left\{|\psi|_1 c_{k_{j,a}} + \left(-2\varphi'_{j,a}(0)\right)^{1/2} \frac{|\psi|_1 |\psi'|_1 \, ||k_{j,a}||_\infty^2}{\tau_a} , \, |\psi'|_1 \, ||k_{j,a}||_\infty\right\}\Bigg) ,\end{aligned}$$

for all $a \in \{1, \dots, A\}$*. Then, for all* $p \in [1, \infty)$ *and* $i \in \{1, 2\}$*,*

$$\begin{aligned}
&||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}}||_{L_p(\mathrm{P}_i^X)} \\
&\leq \sum_{a=1}^{A} \frac{\left(\mathrm{P}_i^X(\mathcal{X}_a)\right)^{1/p}}{\tau_a} \cdot \Bigg(\eta_a \cdot d_{\mathrm{W}}(\mathrm{P}_{1,a}, \mathrm{P}_{2,a}) + \frac{|\psi|_1 \kappa_a^2}{2\tau_a} \cdot |\lambda_{1,a} - \lambda_{2,a}| \\
&\qquad\qquad + \frac{|\psi|_1}{2} \cdot ||k_{1,a} - k_{2,a}||_{L_p(\mathrm{P}_{i,a}^X \otimes \mathrm{P}_{i,a}^X)} \\
&\qquad\qquad + |\psi|_1 \kappa_a \cdot \sqrt{||k_{1,a} - k_{2,a}||_{L_p(\mathrm{P}_{i,a}^X \otimes \mathrm{P}_{i,a}^X)}}\Bigg) .
\end{aligned}$$

Proof. One obtains

$$\begin{aligned}
&||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}}||_{L_p(\mathrm{P}_i^X)} \\
&\leq \sum_{a=1}^{A} \left(\mathrm{P}_i^X(\mathcal{X}_a)\right)^{1/p} \cdot \left|\left|\hat{f}_{\mathrm{P}_{1,a},\lambda_{1,a},k_{1,a}} - \hat{f}_{\mathrm{P}_{2,a},\lambda_{2,a},k_{2,a}}\right|\right|_{L_p(\hat{\mathrm{P}}_{i,a}^X)} \\
&= \sum_{a=1}^{A} \left(\mathrm{P}_i^X(\mathcal{X}_a)\right)^{1/p} \cdot \left|\left|f_{\mathrm{P}_{1,a},\lambda_{1,a},k_{1,a}} - f_{\mathrm{P}_{2,a},\lambda_{2,a},k_{2,a}}\right|\right|_{L_p(\mathrm{P}_{i,a}^X)} ,
\end{aligned}$$

where the first step is equivalent to (4.36) and the second follows from $f_{\mathrm{P}_{1,a},\lambda_{1,a},k_{1,a}}$ and $f_{\mathrm{P}_{2,a},\lambda_{2,a},k_{2,a}}$ being defined on the same region $\mathcal{X}_a$ as $\mathrm{P}_{i,a}^X$. By Remark 3.4.2, each region $\mathcal{X}_a$, $a = 1, \dots, A$, is separable again. As the regions are additionally assumed to be complete, Theorem 4.3.25 can be applied to the norms on the right hand side, which yields the assertion. □

To conclude, regionalization-subtotal stability of localized SVMs—be it sup- or L_p-stability and with respect to the total variation or the Wasserstein distance—follows from the corresponding total stability of the underlying non-localized SVMs seamlessly and it is possible to accordingly bound the difference between two such localized SVMs based on the differences between the underlying probability measures, regularization parameters and kernels.

4.4.2 Total Stability

As explained in Section 4.1, one would hope that localized SVMs are stable with respect to changes in the underlying regionalization as well because this regionalization is often also chosen in a data-dependent way (for example, using decision trees, cf. Bennett and Blue, 1998; Wu et al., 1999; Tibshirani and Hastie, 2007; Chang et al., 2010, among others). That is, the goal of this section now lies in deriving not only regionalization-subtotal stability (cf. Section 4.4.1) but even total stability results.

However, it can readily be seen from the simple example visualized in Figure 4.4.1 that we will not be able to derive meaningful results regarding total sup-stability. In that figure, two localized SVMs are being compared. Both of them are based on the same training data (that is, on the same empirical distribution) generated according to

$$X \sim \mathcal{U}(-1, 1) , \qquad Y|X \sim \mathrm{sign}(X) + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}(0, 0.5) ,$$

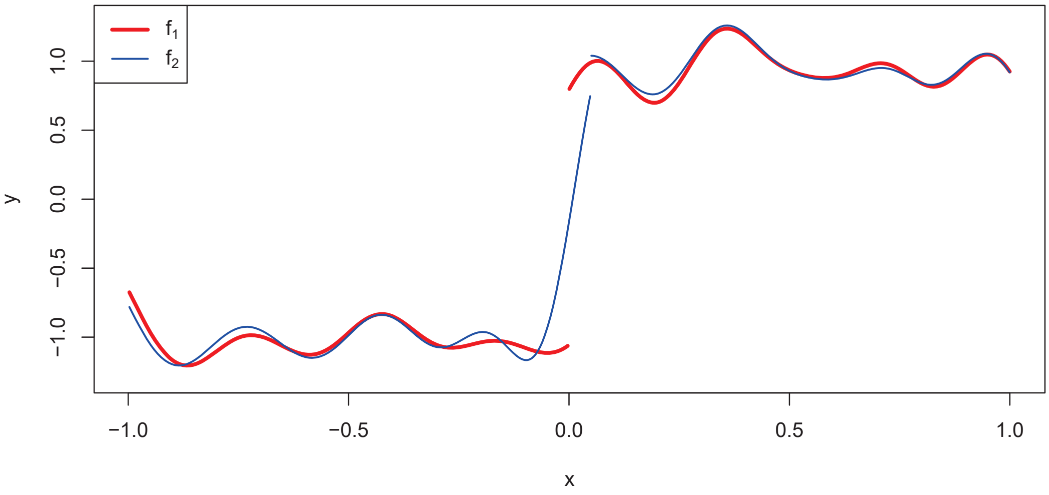

Figure 4.4.1: Comparison of two localized SVMs based on the same distribution, regularization parameters and kernels, but on slightly different regionalizations. [This figure was first published in Köhler and Christmann, 2022.]

with $\mathcal{U}(a,b)$ denoting the uniform distribution on (a,b) and $\mathcal{N}(\mu,\sigma^2)$ the normal distribution with mean μ and variance σ^2. Furthermore, both localized SVMs use the same regularization parameter and the same Gaussian RBF kernel on every region. They only differ in the underlying regionalization: The input space is split into two parts in both cases, but for f_1 the border between the two regions is at $x = 0$ (thus exactly capturing the pattern in the data) whereas it is moved slightly to the right, to $x = 0.05$, for f_2.

It can easily be seen from Figure 4.4.1 that this very minor change in the regionalization greatly impacts the maximum difference between f_1 and f_2 and it is thus obviously not possible to bound this maximum difference between two localized SVMs in any meaningful way. However, the same Figure 4.4.1 also suggests that it might still be possible to find such meaningful bounds on the $L_1(\mathrm{P}_i^X)$-norm of the difference (which is rather small in the example, approximately 0.06, compared to the supremum norm of about 0.95), similarly to Theorems 4.4.6 and 4.4.7. In the following, we simplify this example even further in order to also showcase the observed behavior analytically:

Example 4.4.8. Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $\mathrm{P}^X = \mathcal{U}(-1,1)$ and, for $x \in \mathcal{X}$,

$$\mathrm{P}(\cdot \mid X = x) = \begin{cases} \delta_0 & , \text{ if } x < 0\,, \\ \delta_1 & , \text{ if } x \geq 0 \end{cases}$$

with $\mathcal{U}(a,b)$ denoting the uniform distribution on (a,b) and δ_y denoting the Dirac distribution in $y \in \mathcal{Y}$. Let $\boldsymbol{\mathcal{X}_0} = \{(-\infty,0),[0,\infty)\}$ and $\boldsymbol{\mathcal{X}_n} = \{(-\infty,1/n),[1/n,\infty)\}$ for $n \in \mathbb{N}$. Let $L^\star = L^*_{0.5\text{-pin}}$ be the shifted 0.5-pinball loss and, for each $i \in \mathbb{N}_0$ and $a \in \{1,2\}$, let

$k_{i,a} \equiv 1$ be a constant kernel on $\mathcal{X}_{i,a}$ and $\lambda_{i,a} = 1$.[37]

For the local SVMs $f_{\mathrm{P}_{\mathcal{X}_{i,a}},\lambda_{i,a},k_{i,a}}$, $i \in \mathbb{N}_0$ and $a = 1, 2$, we obtain from Christmann et al. (2009, Theorem 7)

$$f_{\mathrm{P}_{\mathcal{X}_{i,a}},\lambda_{i,a},k_{i,a}} = -\frac{1}{2\lambda} \cdot \int h_{i,a}(x,y)\Phi_{i,a}(x) \,\mathrm{d}\mathrm{P}_{\mathcal{X}_{i,a}}(x,y) \tag{4.37}$$

for some $h_{i,a} \colon \mathcal{X}_{i,a} \times \mathcal{Y} \to \mathbb{R}$ from the subdifferential of $L^\star$ with respect to $f_{\mathrm{P}_{\mathcal{X}_{i,a}},\lambda_{i,a},k_{i,a}}$. Notably, as $(\Phi_{i,a}(x))(x') = k_{i,a}(x',x) = 1$ for all $x, x' \in \mathcal{X}_{i,a}$, (4.37) yields that

$$f_{\mathrm{P}_{\mathcal{X}_{i,a}},\lambda_{i,a},k_{i,a}}(x') = -\frac{1}{2} \cdot \int h_{i,a}(x,y) \,\mathrm{d}\mathrm{P}_{\mathcal{X}_{i,a}}(x,y) \qquad \forall\, x' \in \mathcal{X}_{i,a}\,, \tag{4.38}$$

i.e. that $f_{\mathrm{P}_{\mathcal{X}_{i,a}},\lambda_{i,a},k_{i,a}}$ is a constant function. By the definition of the 0.5-pinball loss, we further have

$$h_{i,a}(x,y) = \begin{cases} -\frac{1}{2} & \text{, if } y > f_{\mathrm{P}_{\mathcal{X}_{i,a}},\lambda_{i,a},k_{i,a}}(x)\,, \\ c \in \left[-\frac{1}{2}, +\frac{1}{2}\right] & \text{, if } y = f_{\mathrm{P}_{\mathcal{X}_{i,a}},\lambda_{i,a},k_{i,a}}(x)\,, \\ +\frac{1}{2} & \text{, if } y_i < f_{\mathrm{P}_{\mathcal{X}_{i,a}},\lambda_{i,a},k_{i,a}}(x)\,, \end{cases}$$

that is, $|h_i(x,y)| \leq \frac{1}{2}$ for all $(x,y) \in \mathcal{X}_{i,a} \times \mathcal{Y}$. Therefore,

$$|f_{\mathrm{P}_{\mathcal{X}_{i,a}},\lambda_{i,a},k_{i,a}}(x')| \leq \frac{1}{4} \qquad \forall\, x' \in \mathcal{X}_{i,a}\,.$$

For $i \in \mathbb{N}_0$, the definition of $\mathrm{P}(\cdot \,|\, X = x)$ and that of the regionalization $\boldsymbol{\mathcal{X}_i}$ then yields that $y = 1 > f_{\mathrm{P}_{\mathcal{X}_{i,a}},\lambda_{i,a},k_{i,a}}(x)$ and hence $h_{i,2}(x,y) = -\frac{1}{2}$ for $\mathrm{P}_{\mathcal{X}_{i,2}}$-almost all (x,y). Therefore, (4.38) yields that indeed

$$f_{\mathrm{P}_{\mathcal{X}_{i,2}},\lambda_{i,2},k_{i,2}}(x') = \frac{1}{4} \qquad \forall\, x' \in \mathcal{X}_{i,2}\,.$$

At the same time, we have, for all $i \in \mathbb{N}_0$,

$$f_{\mathrm{P}_{\mathcal{X}_{i,1}},\lambda_{i,1},k_{i,1}}(x') = 0 \qquad \forall x' \in \mathcal{X}_{i,1}\,,$$

because the function is constant by (4.38) and because all other possible constant functions would lead to contradictions: $f_{\mathrm{P}_{\mathcal{X}_{i,1}},\lambda_{i,1},k_{i,1}} < 0$ would by the definition of $\mathrm{P}(\cdot \,|\, X = x)$ imply that $h_{i,1}(x,y) = -\frac{1}{2}$ for $\mathrm{P}_{\mathcal{X}_{i,1}}$-almost all (x,y), which would by (4.38) yield the contradiction $f_{\mathrm{P}_{\mathcal{X}_{i,1}},\lambda_{i,1},k_{i,1}} \equiv \frac{1}{4} \geq 0$. Similarly, $f_{\mathrm{P}_{\mathcal{X}_{i,1}},\lambda_{i,1},k_{i,1}} > 0$ would imply that

$$\mathrm{P}_{\mathcal{X}_{i,1}}\left(h_{i,1} = \frac{1}{2}\right) \geq \mathrm{P}_{\mathcal{X}_{i,1}}\left(f_{\mathrm{P}_{\mathcal{X}_{i,1}},\lambda_{i,1},k_{i,1}}(x) > y\right) \geq \mathrm{P}_{\mathcal{X}_{i,1}}\left(\mathcal{X}_{i,1} \times \{0\}\right) \geq \frac{1}{2}$$

by the definition of $\mathrm{P}(\cdot \,|\, X = x)$ and $\boldsymbol{\mathcal{X}_i}$, which would by (4.38) yield the contradiction

$$f_{\mathrm{P}_{\mathcal{X}_{i,1}},\lambda_{i,1},k_{i,1}} \equiv -\frac{1}{2} \cdot \int h_{i,1}(x,y) \,\mathrm{d}\mathrm{P}_{\mathcal{X}_{i,1}}(x,y) \leq -\frac{1}{2} \cdot \left(\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \left(-\frac{1}{2}\right)\right) = 0\,.$$

[37] $k_{i,a}$ is indeed a kernel on $\mathcal{X}_{i,a}$ because choosing the feature space $H_{i,a} = \mathbb{R}$ and the feature map $\Phi_{i,a} \colon \mathcal{X}_{i,a} \to H_{i,a}\,,\ x \mapsto 1$ yields $k_{i,a}(x,x') = 1 = \langle \Phi_{i,a}(x), \Phi_{i,a}(x') \rangle_{H_{i,a}}$ for all $x, x' \in \mathcal{X}_{i,a}$.

Summing up, we obtain for the respective localized SVMs (using the simplified definition from Remark 2.2.8 which omits the weight functions because of the regionalizations being partitioning):

$$f_{\mathrm{P},\lambda,k,\boldsymbol{\mathcal{X}_0}}(x) = \begin{cases} 0 & , \text{if } x < 0\,, \\ \frac{1}{4} & , \text{if } x \geq 0\,, \end{cases}$$

$$f_{\mathrm{P},\lambda,k,\boldsymbol{\mathcal{X}_n}}(x) = \begin{cases} 0 & , \text{if } x < \frac{1}{n}\,, \\ \frac{1}{4} & , \text{if } x \geq \frac{1}{n}\,, \end{cases} \qquad \forall\, n \in \mathbb{N}\,.$$

This yields

$$||f_{\mathrm{P},\lambda,k,\boldsymbol{\mathcal{X}_0}} - f_{\mathrm{P},\lambda,k,\boldsymbol{\mathcal{X}_n}}||_\infty = \frac{1}{4}$$

for all $n \in \mathbb{N}$, even though both localized SVMs that are being compared are based on the same probability measure, regularization parameters and kernels, and the underlying regionalizations become arbitrarily similar as $n \to \infty$. On the other hand,

$$||f_{\mathrm{P},\lambda,k,\boldsymbol{\mathcal{X}_0}} - f_{\mathrm{P},\lambda,k,\boldsymbol{\mathcal{X}_n}}||_{L_1(\mathrm{P}^X)} = \frac{1}{8n}$$

does indeed converge to 0 as $n \to \infty$, i.e. as the regionalizations $\boldsymbol{\mathcal{X}_0}$ and $\boldsymbol{\mathcal{X}_n}$ become more similar.

As suggested by the preceding examples, it will indeed be possible to derive meaningful bounds on the $L_1(\mathrm{P}_i^X)$-norm of the difference of two localized SVMs, i.e. to derive results on the total L_1-stability of localized SVMs. Before stating the corresponding theorems, Assumption 4.4.1 from the preceding section first needs to be modified such that it fits the situation of this section. For this, recall Definition 2.2.4 of the combined regionalization $\boldsymbol{\mathcal{X}^*_{1,2}}$ of two regionalizations $\boldsymbol{\mathcal{X}_1}$ and $\boldsymbol{\mathcal{X}_2}$, as much of the analysis of this section is based on this combined regionalization and auxiliary local SVMs on its regions.

Assumption 4.4.9.

- Let $L\colon \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, Lipschitz continuous loss function and let $L^\star$ be its shifted version.
- For $i = 1,2$, let $\boldsymbol{\mathcal{X}_i} \coloneqq \{\mathcal{X}_{i,1}, \ldots, \mathcal{X}_{i,A_i}\}$ be a partitioning regionalization of $\mathcal{X}$ of size $A_i \in \mathbb{N}$.
- For $i = 1,2$, let P_i be a probability measure on $\mathcal{X} \times \mathcal{Y}$ that is positive on the combined regionalization $\boldsymbol{\mathcal{X}^*_{1,2}} \coloneqq \{\mathcal{X}^*_1, \ldots, \mathcal{X}^*_B\}$ of $\boldsymbol{\mathcal{X}_1}$ and $\boldsymbol{\mathcal{X}_2}$.
- For $i = 1,2$, let $\boldsymbol{\lambda_i} \coloneqq (\lambda_{i,1}, \ldots, \lambda_{i,A_i}) \in (0,\infty)^{A_i}$.
- For $i = 1,2$, let $\boldsymbol{k_i} \coloneqq (k_{i,1}, \ldots, k_{i,A_i})$ be a vector of bounded and measurable kernels on $\boldsymbol{\mathcal{X}_i}$ with separable RKHSs $H_{i,a}$, $a = 1, \ldots, A_i$.

The main additional assumption coming into play compared to the previous section is that the regionalizations need to be partitioning. Note that this comes with the slight notation-wise advantage of being able to just use the simplified definition of localized SVMs from Remark 2.2.8 and completely omit the weight functions and the associated assumptions.

Similarly, the additional assumptions needed in results using the Wasserstein distance in the bound need slight modifications as well:

Assumption 4.4.10.

- Let $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$.
- Let L be distance-based with representing function ψ such that ψ is differentiable and ψ as well as its derivative ψ' are both Lipschitz continuous.
- For $a = 1, \dots, A_2$, let $\mathcal{X}_{2,a}$ be complete.
- For $a = 1, \dots, A_2$, let $\tilde{\mathcal{X}}_{2,a} \supseteq \mathcal{X}_{2,a}$ with $\tilde{\mathcal{X}}_{2,a} \subseteq \mathbb{R}^d$ be open and assume that $k_{2,a} = \tilde{k}_{2,a}\big|_{\mathcal{X}_{2,a}\times\mathcal{X}_{2,a}}$ for a kernel $\tilde{k}_{2,a}$ on $\tilde{\mathcal{X}}_{2,a}$ which can be written as $\tilde{k}_{2,a}(x,x') = \varphi_{2,a}(||x-x'||_2)$ for all $x, x' \in \tilde{\mathcal{X}}_{2,a}$, where $\varphi_{2,a} : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a twice continuously differentiable function satisfying $\varphi_{2,a}(0) - \varphi_{2,a}(r) \leq \frac{c_{k_{2,a}}^2}{2} \cdot r$ for all $r \geq 0$ for some $c_{k_{2,a}} \geq 0$.

Before stating the results, some additional notation needs to be introduced. For $i = 1, 2$ and $a = 1, \dots, A_i$, the shortening notation $\mathrm{P}_{i,a} \coloneqq \mathrm{P}_{i,\mathcal{X}_{i,a}}$ (based on the local measures defined in (2.6)) gets used, and additionally denote

$$J_{i,a} \coloneqq \Big\{ b \in \{1, \dots, B\} \,\big|\, \mathcal{X}_b^* \subseteq \mathcal{X}_{i,a} \Big\} \neq \emptyset \,.$$

Further additional notation arises from the already mentioned auxiliary SVMs on the regions $\mathcal{X}_1^*, \dots, \mathcal{X}_B^*$ from $\boldsymbol{\mathcal{X}}_{\mathbf{1,2}}^*$ that are used in the results. For $i = 1, 2$ and $b = 1, \dots, B$, denote by $a(i,b)$ that index $a \in \{1, \dots, A_i\}$ such that $\mathcal{X}_b^* \subseteq \mathcal{X}_{i,a}$ (which is well-defined because of $\boldsymbol{\mathcal{X}_i}$ being partitioning) and by $\mathrm{P}_{i,b}^* \coloneqq \mathrm{P}_{i,\mathcal{X}_b^*}$ and $k_{i,b}^* \coloneqq k_{i,a(i,b)}\big|_{\mathcal{X}_b^* \times \mathcal{X}_b^*}$ auxiliary distributions and kernels on the sets $\mathcal{X}_1^*, \dots, \mathcal{X}_B^*$. By Berlinet and Thomas-Agnan (2004, Theorem 6), $k_{i,b}^*$ is actually a kernel (on $\mathcal{X}_b^*$) again.

Even though, as explained at the beginning of this section (cf. Figure 4.4.1 and Example 4.4.8), we cannot derive meaningful results on total sup-stability, it is now indeed possible to prove the subsequent result on total L_1-stability of localized SVMs.[38] For this, recall the definition of $\xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2})$, for a probability measure Q on $\mathcal{X}$, that was given in

[38] Theorem 4.4.11 is only stated with respect to the $L_1(\mathrm{P}_1^X)$- but not with respect to the $L_1(\mathrm{P}_2^X)$-norm. This was done only for the sake of notational clarity, and the theorem of course also holds true with respect to $L_1(\mathrm{P}_2^X)$-norm if the indices on the right hand side of the bound are adjusted accordingly, which is immediately obvious if one interchanges the roles of the two localized SVMs whose difference gets bounded.

(4.7):

$$
\begin{aligned}
\xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}_1},\boldsymbol{\mathcal{X}_2}) \coloneqq\ & \left|\mathrm{Q}(\mathcal{X}_{1,a(1,b)}) - \mathrm{Q}(\mathcal{X}_{2,a(2,b)})\right| \\
& + \max\left\{\mathrm{Q}(\mathcal{X}_{1,a(1,b)}), \mathrm{Q}(\mathcal{X}_{2,a(2,b)})\right\} \\
& \quad \cdot \sum_{i=1}^{2} \left(\frac{1}{2}\cdot \mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}_b^*)\cdot\left(1-\mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}_b^*)\right)\right. \\
& \qquad\qquad \left. + \sqrt{\mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}_b^*)\cdot\left(1-\mathrm{Q}_{\mathcal{X}_{i,a(i,b)}}(\mathcal{X}_b^*)\right)}\right) .
\end{aligned}
$$

Theorem 4.4.11. *Let Assumption 4.4.9 be satisfied. Denote*

$$
\begin{aligned}
&\kappa_b \coloneqq \max\left\{\left|\left|k_{1,a(1,b)}\right|\right|_\infty, \left|\left|k_{2,a(2,b)}\right|\right|_\infty\right\} , \\
&\tau_b \coloneqq \min\left\{\lambda_{1,a(1,b)}, \lambda_{2,a(2,b)}\right\} , \\
&\rho_{1,b} \coloneqq \max\left\{\mathrm{P}_1^X(\mathcal{X}_{1,a(1,b)}), \mathrm{P}_1^X(\mathcal{X}_{2,a(2,b)})\right\} ,
\end{aligned}
$$

for all $b \in \{1,\ldots,B\}$. *Then,*

$$
\begin{aligned}
&||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}}||_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{a=1}^{A_2} \mathrm{P}_1^X(\mathcal{X}_{2,a}) \cdot \frac{||k_{2,a}||_\infty^2}{\lambda_{2,a}} \cdot d_{\mathrm{tv}}(\mathrm{P}_{1,\mathcal{X}_{2,a}}, \mathrm{P}_{2,\mathcal{X}_{2,a}}) \\
&\quad + |L|_1 \cdot \sum_{b=1}^{B} \left(\rho_{1,b}\cdot\frac{\kappa_b^2}{2\tau_b^2}\cdot\left|\lambda_{1,a(1,b)} - \lambda_{2,a(2,b)}\right|\right. \\
&\qquad\qquad + \mathrm{P}_1^X(\mathcal{X}_b^*)\cdot\left(\frac{1}{2\tau_b}\cdot\left|\left|k_{1,b}^* - k_{2,b}^*\right|\right|_{L_1((\mathrm{P}_{1,b}^*)^X\otimes(\mathrm{P}_{1,b}^*)^X)}\right. \\
&\qquad\qquad\qquad \left. + \frac{\kappa_b}{\tau_b}\cdot\sqrt{\left|\left|k_{1,b}^* - k_{2,b}^*\right|\right|_{L_1((\mathrm{P}_{1,b}^*)^X\otimes(\mathrm{P}_{1,b}^*)^X)}}\right) \\
&\qquad\qquad \left. + \frac{\kappa_b^2}{\tau_b}\cdot\xi_{\mathrm{P}_1^X,b}(\boldsymbol{\mathcal{X}_1},\boldsymbol{\mathcal{X}_2})\right) .
\end{aligned}
$$

Proof. In addition to the auxiliary distributions and kernels introduced prior to the theorem, we also need auxiliary regularization parameters in this proof. Denote these parameters by $\lambda_{i,j,b}^* \coloneqq (\mathrm{P}^X_{j,\mathcal{X}_{i,a(i,b)}}(\mathcal{X}_b^*))^{-1}\lambda_{i,a(i,b)}$ for $i,j = 1,2$ and $b = 1,\ldots,B$.

By applying the triangle inequality, the left hand side of the assertion can be expanded as

$$
\begin{aligned}
||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}}||_{L_1(\mathrm{P}_1^X)} \leq\ & \left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_1,\boldsymbol{\lambda_{1,1}^*},\boldsymbol{k_1^*},\boldsymbol{\mathcal{X}_{1,2}^*}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
& + \left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda_{1,1}^*},\boldsymbol{k_1^*},\boldsymbol{\mathcal{X}_{1,2}^*}} - f_{\mathrm{P}_1,\boldsymbol{\lambda_{2,1}^*},\boldsymbol{k_2^*},\boldsymbol{\mathcal{X}_{1,2}^*}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
& + \left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda_{2,1}^*},\boldsymbol{k_2^*},\boldsymbol{\mathcal{X}_{1,2}^*}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}}\right|\right|_{L_1(\mathrm{P}_1^X)} \qquad (4.39)
\end{aligned}
$$

with $\boldsymbol{\lambda_{i,j}^*} \coloneqq (\lambda_{i,j,1}^*,\ldots,\lambda_{i,j,B}^*)$ and $\boldsymbol{k_i^*} \coloneqq (k_{i,1}^*,\ldots,k_{i,B}^*)$ for $i,j = 1,2$. We will now examine the three norms from the right hand side of (4.39) separately:

(i) For a function $g : \mathcal{X}_b^* \to \mathbb{R}$, denote by $\tilde{g}$ its zero-extension to $\mathcal{X}_{1,a(1,b)}$ (respectively to $\mathcal{X}_{1,a(1,b)} \times \mathcal{X}_{1,a(1,b)}$ if the function is instead defined on $\mathcal{X}_b^* \times \mathcal{X}_b^*$). Defining $k_{1,a}^\circ := \sum_{b \in J_{1,a}} \tilde{k}_{1,b}^*$ yields for all $a \in \{1, \ldots, A_1\}$ new local SVMs $f_{\mathrm{P}_{1,a}, \lambda_{1,a}, k_{1,a}^\circ}$ which are defined as

$$f_{\mathrm{P}_{1,a}, \lambda_{1,a}, k_{1,a}^\circ} = \arg \inf_{f \in H_{1,a}^\circ} \mathcal{R}_{L^\star, \mathrm{P}_{1,a}}(f) + \lambda_{1,a} \, ||f||_{H_{1,a}^\circ}^2 \, . \tag{4.40}$$

Now, combining Berlinet and Thomas-Agnan (2004, Theorem 5) and Meister and Steinwart (2016, Lemma 2) yields that $k_{1,a}^\circ$ is indeed a kernel on $\mathcal{X}_{1,a}$ and that its RKHS is given by

$$H_{1,a}^\circ = \left\{ f : \mathcal{X}_{1,a} \to \mathbb{R} \;\middle|\; f = \sum_{b \in J_{1,a}} \tilde{f}_b, \ f_b \in H_{1,b}^* \text{ for } b = 1, \ldots, B \right\} ,$$

with the decomposition of each such $f \in H_{1,a}^\circ$ being unique because of the sets $\mathcal{X}_1^*, \ldots, \mathcal{X}_B^*$, the domains of the functions f_b, being pairwise disjoint because $\boldsymbol{\mathcal{X}_1}$ and $\boldsymbol{\mathcal{X}_2}$ being partitioning. Thus, the mentioned results also yield

$$||f||_{H_{1,a}^\circ}^2 = \sum_{b \in J_{1,a}} ||f_b||_{H_{1,b}^*}^2$$

for all $f \in H_{1,a}^\circ$. Additionally, again because of the domains of the functions f_b being pairwise disjoint, it is possible to also expand the risk from (4.40) similarly to the preceding expansion of the $H_{1,a}^\circ$-norm:

$$\begin{aligned}
\mathcal{R}_{L^\star, \mathrm{P}_{1,a}}(f) &= \int_{\mathcal{X}_{1,a}} L^\star(x, y, f(x)) \, \mathrm{dP}_{1,a}(x, y) \\
&= \sum_{b \in J_{1,a}} \int_{\mathcal{X}_b^*} L^\star(x, y, f_b(x)) \, \mathrm{dP}_{1,a}(x, y) \\
&= \sum_{b \in J_{1,a}} \mathrm{P}_{1,a}^X(\mathcal{X}_b^*) \cdot \int_{\mathcal{X}_b^*} L^\star(x, y, f_b(x)) \, \mathrm{dP}_{1,b}^*(x, y) \\
&= \sum_{b \in J_{1,a}} \mathrm{P}_{1,a}^X(\mathcal{X}_b^*) \cdot \mathcal{R}_{L^\star, \mathrm{P}_{1,b}^*}(f_b) \, ,
\end{aligned}$$

where (2.6) was applied in the third step.

Plugging this into (4.40) yields

$$\begin{aligned}
f_{\mathrm{P}_{1,a}, \lambda_{1,a}, k_{1,a}^\circ} &= \arg \inf_{f \in H_{1,a}^\circ} \sum_{b \in J_{1,a}} \left(\mathrm{P}_{1,a}^X(\mathcal{X}_b^*) \cdot \mathcal{R}_{L^\star, \mathrm{P}_{1,b}^*}(f_b) + \lambda_{1,a} \, ||f_b||_{H_{1,b}^*}^2 \right) \\
&= \sum_{b \in J_{1,a}} \widetilde{\arg \inf_{f_b \in H_{1,b}^*}} \left(\mathrm{P}_{1,a}^X(\mathcal{X}_b^*) \cdot \mathcal{R}_{L^\star, \mathrm{P}_{1,b}^*}(f_b) + \lambda_{1,a} \, ||f_b||_{H_{1,b}^*}^2 \right) \\
&= \sum_{b \in J_{1,a}} \widetilde{\arg \inf_{f_b \in H_{1,b}^*}} \left(\mathcal{R}_{L^\star, \mathrm{P}_{1,b}^*}(f_b) + \frac{\lambda_{1,a}}{\mathrm{P}_{1,a}^X(\mathcal{X}_b^*)} \, ||f_b||_{H_{1,b}^*}^2 \right) \\
&= \sum_{b \in J_{1,a}} \tilde{f}_{\mathrm{P}_{1,b}^*, \lambda_{1,1,b}^*, k_{1,b}^*}
\end{aligned}$$

and thus

$$f_{\mathrm{P}_1,\boldsymbol{\lambda^*_{1,1}},\boldsymbol{k^*_1},\boldsymbol{\mathcal{X}^*_{1,2}}} = \sum_{b=1}^{B} \hat{f}_{\mathrm{P}^*_{1,b},\lambda^*_{1,1,b},k^*_{1,b}} = \sum_{a=1}^{A_1} \sum_{b\in J_{1,a}} \hat{f}_{\mathrm{P}^*_{1,b},\lambda^*_{1,1,b},k^*_{1,b}} = \sum_{a=1}^{A_1} \hat{f}_{\mathrm{P}_{1,a},\lambda_{1,a},k^\circ_{1,a}} \,.$$

The first difference on the right hand side of (4.39) can therefore also be interpreted as the difference between two localized SVMs that are based on the same regionalization $\boldsymbol{\mathcal{X}_1}$ (and on the same probability measure and vector of regularization parameters). An application of Theorem 4.4.6 hence yields

$$\begin{aligned}
&\left|\left| f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_1,\boldsymbol{\lambda^*_{1,1}},\boldsymbol{k^*_1},\boldsymbol{\mathcal{X}^*_{1,2}}} \right|\right|_{L_1(\mathrm{P}_1^X)} = \left|\left| f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - \sum_{a=1}^{A_1} \hat{f}_{\mathrm{P}_{1,a},\lambda_{1,a},k^\circ_{1,a}} \right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{a=1}^{A_1} \mathrm{P}_1^X(\mathcal{X}_{1,a}) \\
&\qquad\qquad \cdot \left(\frac{1}{2\lambda_{1,a}} \cdot \left|\left| k_{1,a} - k^\circ_{1,a} \right|\right|_{L_1(\mathrm{P}^X_{1,a}\otimes \mathrm{P}^X_{1,a})} \right. \\
&\qquad\qquad\quad \left. + \frac{\max\left\{ ||k_{1,a}||_\infty \,, \left|\left|k^\circ_{1,a}\right|\right|_\infty \right\}}{\lambda_{1,a}} \cdot \sqrt{\left|\left| k_{1,a} - k^\circ_{1,a} \right|\right|_{L_1(\mathrm{P}^X_{1,a}\otimes \mathrm{P}^X_{1,a})}} \right) .
\end{aligned}$$

Because

$$k^\circ_{1,a}(x,x') = \begin{cases} k_{1,a}(x,x') & \text{, if } \exists\, b \in J_{1,a} : x, x' \in \mathcal{X}^*_b \,, \\ 0 & \text{, else} \,, \end{cases}$$

we furthermore know that $\max\left\{ ||k_{1,a}||_\infty \,, \left|\left|k^\circ_{1,a}\right|\right|_\infty \right\} = ||k_{1,a}||_\infty$ and

$$\begin{aligned}
\left|\left| k_{1,a} - k^\circ_{1,a} \right|\right|_{L_1(\mathrm{P}^X_{1,a}\otimes \mathrm{P}^X_{1,a})} &= \int_{\mathcal{X}_{1,a}} \int_{\mathcal{X}_{1,a}} \left| k_{1,a}(x,x') - k^\circ_{1,a}(x,x') \right| \, d\mathrm{P}^X_{1,a}(x') \, d\mathrm{P}^X_{1,a}(x) \\
&= \sum_{b\in J_{1,a}} \int_{\mathcal{X}^*_b} \int_{\mathcal{X}_{1,a}\setminus \mathcal{X}^*_b} \left| k_{1,a}(x,x') \right| \, d\mathrm{P}^X_{1,a}(x') \, d\mathrm{P}^X_{1,a}(x) \\
&\leq ||k_{1,a}||^2_\infty \cdot \sum_{b\in J_{1,a}} \mathrm{P}^X_{1,a}(\mathcal{X}^*_b) \cdot \Big(1 - \mathrm{P}^X_{1,a}(\mathcal{X}^*_b) \Big)
\end{aligned}$$

which finally results in

$$\begin{aligned}
&\left|\left| f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_1,\boldsymbol{\lambda^*_{1,1}},\boldsymbol{k^*_1},\boldsymbol{\mathcal{X}^*_{1,2}}} \right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{a=1}^{A_1} \mathrm{P}_1^X(\mathcal{X}_{1,a}) \cdot \left(\frac{||k_{1,a}||^2_\infty}{2\lambda_{1,a}} \cdot \sum_{b\in J_{1,a}} \mathrm{P}^X_{1,a}(\mathcal{X}^*_b) \cdot \Big(1 - \mathrm{P}^X_{1,a}(\mathcal{X}^*_b) \Big) \right. \\
&\qquad\qquad\qquad \left. + \frac{||k_{1,a}||^2_\infty}{\lambda_{1,a}} \cdot \sqrt{\sum_{b\in J_{1,a}} \mathrm{P}^X_{1,a}(\mathcal{X}^*_b) \cdot \Big(1 - \mathrm{P}^X_{1,a}(\mathcal{X}^*_b) \Big)} \right) .
\end{aligned}$$

(ii) The second norm on the right hand side of (4.39) already consists of the difference of two localized SVMs that are based on the same regionalization $\boldsymbol{\mathcal{X}^*_{1,2}}$ (and the same probability measure), without us needing to make any changes beforehand. We can therefore directly apply Theorem 4.4.6 and obtain

$$\begin{aligned}
&\left|\left| f_{\mathrm{P}_1,\boldsymbol{\lambda^*_{1,1}},\boldsymbol{k^*_1},\boldsymbol{\mathcal{X}^*_{1,2}}} - f_{\mathrm{P}_1,\boldsymbol{\lambda^*_{2,1}},\boldsymbol{k^*_2},\boldsymbol{\mathcal{X}^*_{1,2}}} \right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{b=1}^{B} \mathrm{P}_1^X(\mathcal{X}^*_b) \cdot \left(\frac{(\kappa^*_b)^2}{2(\tau^*_{1,b})^2} \cdot \left| \lambda^*_{1,1,b} - \lambda^*_{2,1,b} \right| \right. \\
&\qquad\qquad + \frac{1}{2\tau^*_{1,b}} \cdot \left|\left| k^*_{1,b} - k^*_{2,b} \right|\right|_{L_1((\mathrm{P}^*_{1,b})^X \otimes (\mathrm{P}^*_{1,b})^X)} \\
&\qquad\qquad \left. + \frac{\kappa^*_b}{\tau^*_{1,b}} \cdot \sqrt{\left|\left| k^*_{1,b} - k^*_{2,b} \right|\right|_{L_1((\mathrm{P}^*_{1,b})^X \otimes (\mathrm{P}^*_{1,b})^X)}} \right)
\end{aligned} \tag{4.41}$$

with

$$\kappa^*_b := \max\left\{ \left|\left| k^*_{1,b} \right|\right|_\infty, \left|\left| k^*_{2,b} \right|\right|_\infty \right\} \leq \max\left\{ \left|\left| k_{1,a(1,b)} \right|\right|_\infty, \left|\left| k_{2,a(2,b)} \right|\right|_\infty \right\} = \kappa_b \,,$$

because $k^*_{i,b}$ and $k_{i,a(i,b)}$ coincide everywhere $k^*_{i,b}$ is defined, and

$$\tau^*_{1,b} := \min\left\{ \lambda^*_{1,1,b}, \lambda^*_{2,1,b} \right\} \geq \min\left\{ \lambda_{1,a(1,b)}, \lambda_{2,a(2,b)} \right\} = \tau_b$$

because of $\lambda^*_{i,1,b}$ being defined as $(\mathrm{P}^X_{1,\mathcal{X}_{i,a(i,b)}}(\mathcal{X}^*_b))^{-1}\lambda_{i,a(i,b)}$. Thus, (4.41) still holds true after replacing κ^*_b and $\tau^*_{1,b}$ by κ_b and τ_b. Additionally, applying the definition of $\lambda^*_{i,1,b}$ again as well as the definition of $\mathrm{P}^X_{1,\mathcal{X}_{i,a(i,b)}}$ yields

$$\begin{aligned}
\left| \lambda^*_{1,1,b} - \lambda^*_{2,1,b} \right| &= \frac{1}{\mathrm{P}_1^X(\mathcal{X}^*_b)} \cdot \left| \lambda_{1,a(1,b)} \cdot \mathrm{P}_1^X(\mathcal{X}_{1,a(1,b)}) - \lambda_{2,a(2,b)} \cdot \mathrm{P}_1^X(\mathcal{X}_{2,a(2,b)}) \right| \\
&\leq \frac{1}{\mathrm{P}_1^X(\mathcal{X}^*_b)} \cdot \Big(\lambda_{1,a(1,b)} \cdot \left| \mathrm{P}_1^X(\mathcal{X}_{1,a(1,b)}) - \mathrm{P}_1^X(\mathcal{X}_{2,a(2,b)}) \right| \\
&\qquad\qquad + \mathrm{P}_1^X(\mathcal{X}_{2,a(2,b)}) \cdot \left| \lambda_{1,a(1,b)} - \lambda_{2,a(2,b)} \right| \Big)
\end{aligned}$$

as well as analogously

$$\begin{aligned}
\left| \lambda^*_{1,1,b} - \lambda^*_{2,1,b} \right| &\leq \frac{1}{\mathrm{P}_1^X(\mathcal{X}^*_b)} \cdot \Big(\lambda_{2,a(2,b)} \cdot \left| \mathrm{P}_1^X(\mathcal{X}_{1,a(1,b)}) - \mathrm{P}_1^X(\mathcal{X}_{2,a(2,b)}) \right| \\
&\qquad\qquad + \mathrm{P}_1^X(\mathcal{X}_{1,a(1,b)}) \cdot \left| \lambda_{1,a(1,b)} - \lambda_{2,a(2,b)} \right| \Big) \,,
\end{aligned}$$

and hence

$$\begin{aligned}
&\left| \lambda^*_{1,1,b} - \lambda^*_{2,1,b} \right| \\
&\leq \frac{1}{\mathrm{P}_1^X(\mathcal{X}^*_b)} \cdot \Big(\tau_b \cdot \left| \mathrm{P}_1^X(\mathcal{X}_{1,a(1,b)}) - \mathrm{P}_1^X(\mathcal{X}_{2,a(2,b)}) \right| + \rho_{1,b} \cdot \left| \lambda_{1,a(1,b)} - \lambda_{2,a(2,b)} \right| \Big) \,.
\end{aligned}$$

Plugging this into (4.41) finally yields

$$\begin{aligned}
&\left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{1,1}},\boldsymbol{k}^*_{\mathbf{1}},\boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}}} - f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{2,1}},\boldsymbol{k}^*_{\mathbf{2}},\boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{b=1}^{B} \Bigg(\rho_{1,b} \cdot \frac{\kappa_b^2}{2\tau_b^2} \cdot \left|\lambda_{1,a(1,b)} - \lambda_{2,a(2,b)}\right| \\
&\qquad\qquad + \mathrm{P}_1^X(\mathcal{X}_b^*) \cdot \Bigg(\frac{1}{2\tau_b} \cdot \left|\left|k^*_{1,b} - k^*_{2,b}\right|\right|_{L_1((\mathrm{P}^*_{1,b})^X \otimes (\mathrm{P}^*_{1,b})^X)} \\
&\qquad\qquad\qquad + \frac{\kappa_b}{\tau_b} \cdot \sqrt{\left|\left|k^*_{1,b} - k^*_{2,b}\right|\right|_{L_1((\mathrm{P}^*_{1,b})^X \otimes (\mathrm{P}^*_{1,b})^X)}} \Bigg) \\
&\qquad\qquad + \frac{\kappa_b^2}{\tau_b} \cdot \left|\mathrm{P}_1^X(\mathcal{X}_{1,a(1,b)}) - \mathrm{P}_1^X(\mathcal{X}_{2,a(2,b)})\right| \Bigg) .
\end{aligned}$$

(iii) The third norm on the right hand side of (4.39) can be analyzed similarly to the first one. Let the $(\tilde{\cdot})$-notation now denote zero-extensions to $\mathcal{X}_{2,a(2,b)}$ instead of $\mathcal{X}_{1,a(1,b)}$. Analogously to (i), it can be shown that

$$f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{2,1}},\boldsymbol{k}^*_{\mathbf{2}},\boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}}} = \sum_{b=1}^{B} \hat{f}_{\mathrm{P}^*_{1,b},\lambda^*_{2,1,b},k^*_{2,b}} = \sum_{a=1}^{A_2} \sum_{b\in J_{2,a}} \hat{f}_{\mathrm{P}^*_{1,b},\lambda^*_{2,1,b},k^*_{2,b}} = \sum_{a=1}^{A_2} \hat{f}_{\mathrm{P}_{1,\mathcal{X}_{2,a}},\lambda_{2,a},k^\circ_{2,a}} ,$$

where $k^\circ_{2,a} := \sum_{b\in J_{2,a}} \tilde{k}^*_{2,b}$ for $a = 1,\dots,A_2$. We can thus also interpret the third difference on the right hand side of (4.39) as the difference between two localized SVMs that are based on the same regionalization $\boldsymbol{\mathcal{X}}_{\mathbf{2}}$ (and on the same vector of regularization parameters) and apply Theorem 4.4.6:

$$\begin{aligned}
&\left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{2,1}},\boldsymbol{k}^*_{\mathbf{2}},\boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda}_{\mathbf{2}},\boldsymbol{k}_{\mathbf{2}},\boldsymbol{\mathcal{X}}_{\mathbf{2}}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
&= \left|\left|\sum_{a=1}^{A_2} \hat{f}_{\mathrm{P}_{1,\mathcal{X}_{2,a}},\lambda_{2,a},k^\circ_{2,a}} - f_{\mathrm{P}_2,\boldsymbol{\lambda}_{\mathbf{2}},\boldsymbol{k}_{\mathbf{2}},\boldsymbol{\mathcal{X}}_{\mathbf{2}}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{a=1}^{A_2} \mathrm{P}_1^X(\mathcal{X}_{2,a}) \cdot \Bigg(\frac{||k_{2,a}||_\infty^2}{\lambda_{2,a}} \cdot d_{\mathrm{tv}}(\mathrm{P}_{1,\mathcal{X}_{2,a}}, \mathrm{P}_{2,\mathcal{X}_{2,a}}) \\
&\qquad\qquad + \frac{||k_{2,a}||_\infty^2}{2\lambda_{2,a}} \cdot \sum_{b\in J_{2,a}} \mathrm{P}^X_{1,\mathcal{X}_{2,a}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_{2,a}}(\mathcal{X}_b^*)\right) \\
&\qquad\qquad + \frac{||k_{2,a}||_\infty^2}{\lambda_{2,a}} \cdot \sqrt{\sum_{b\in J_{2,a}} \mathrm{P}^X_{1,\mathcal{X}_{2,a}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_{2,a}}(\mathcal{X}_b^*)\right)} \Bigg) ,
\end{aligned}$$

where we employed that $\max\{||k_{2,a}||_\infty\, , ||k^\circ_{2,a}||_\infty\} = ||k_{2,a}||_\infty$ and

$$\left|\left|k_{2,a} - k^\circ_{2,a}\right|\right|_{L_1(\mathrm{P}^X_{1,\mathcal{X}_{2,a}} \otimes \mathrm{P}^X_{1,\mathcal{X}_{2,a}})} \leq ||k_{2,a}||_\infty^2 \cdot \sum_{b\in J_{2,a}} \mathrm{P}^X_{1,\mathcal{X}_{2,a}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_{2,a}}(\mathcal{X}_b^*)\right) ,$$

which follows in the same way as the analogous statements in (i).

Plugging these three bounds into (4.39) and additionally observing

$$\begin{aligned}
&\sum_{a=1}^{A_i} \mathrm{P}_1^X(\mathcal{X}_{i,a}) \cdot \left(\frac{||k_{i,a}||_\infty^2}{2\lambda_{i,a}} \cdot \sum_{b\in J_{i,a}} \mathrm{P}^X_{1,\mathcal{X}_{i,a}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_{i,a}}(\mathcal{X}_b^*)\right) \right. \\
&\qquad\qquad\qquad \left. + \frac{||k_{i,a}||_\infty^2}{\lambda_{i,a}} \cdot \sqrt{\sum_{b\in J_{i,a}} \mathrm{P}^X_{1,\mathcal{X}_{i,a}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_{i,a}}(\mathcal{X}_b^*)\right)} \right) \\
&\leq \sum_{a=1}^{A_i} \sum_{b\in J_{i,a}} \mathrm{P}_1^X(\mathcal{X}_{i,a}) \cdot \frac{||k_{i,a}||_\infty^2}{\lambda_{i,a}} \cdot \left(\frac{1}{2} \cdot \mathrm{P}^X_{1,\mathcal{X}_{i,a}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_{i,a}}(\mathcal{X}_b^*)\right) \right. \\
&\qquad\qquad\qquad\qquad \left. + \sqrt{\mathrm{P}^X_{1,\mathcal{X}_{i,a}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_{i,a}}(\mathcal{X}_b^*)\right)} \right) \\
&\leq \sum_{b=1}^{B} \rho_{1,b} \cdot \frac{\kappa_b^2}{\tau_b} \cdot \left(\frac{\mathrm{P}^X_{1,\mathcal{X}_{i,a(i,b)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_{i,a(i,b)}}(\mathcal{X}_b^*)\right)}{2} \right. \\
&\qquad\qquad\qquad \left. + \sqrt{\mathrm{P}^X_{1,\mathcal{X}_{i,a(i,b)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_{i,a(i,b)}}(\mathcal{X}_b^*)\right)} \right) ,
\end{aligned}$$

$i = 1, 2$, yields the assertion. □

Even though allowing for differing regionalizations makes this result on total stability look more complicated than those on regionalization-subtotal stability from Section 4.4.1 at first glance, the statement indeed keeps the nice structure of bounding the difference between the two localized SVMs based on the difference between the vectors of local measures, vectors of regularization parameters, vectors of kernels, and now additionally the regionalizations. The main difference to Section 4.4.1 lies in the fact that we only derived such a result on L_1- but not on sup-consistency because of the difficulties explained in the context of Figure 4.4.1.

Looking at the proof of Theorem 4.4.11, an analogous result using the Wasserstein distance instead of the total variation distance can not be derived by just replacing all occurrences of Theorem 4.4.6 (result on regionalization-subtotal L_p-stability using the total variation distance) by Theorem 4.4.7 (analogous result using the Wasserstein distance) because the auxiliary kernels $k_{i,a}^\circ$ from the proof can not be written as $k_{i,a}^\circ(x, x') = \varphi_{i,a}^\circ(||x - x'||_2)$ for functions $\varphi_{i,a}^\circ$—even if one assumes that the kernels $k_{i,a}$ can be written in such a way—as they do not only depend on $||x - x'||_2$ but also on whether x and x' are contained in the same $\mathcal{X}_b^* \in \boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}}$. However, the situation at hand can easily be reduced to two parts which can be handled by Theorem 4.4.11 and Theorem 4.4.7 respectively, yielding the following:

Theorem 4.4.12. *Let Assumptions 4.4.9 and 4.4.10 be satisfied. Denote*

$$\begin{aligned}
\kappa_b &:= \max\left\{ \left|\left|k_{1,a(1,b)}\right|\right|_\infty , \left|\left|k_{2,a(2,b)}\right|\right|_\infty \right\} , \\
\tau_b &:= \min\left\{ \lambda_{1,a(1,b)}, \lambda_{2,a(2,b)} \right\} , \\
\rho_{1,b} &:= \max\left\{ \mathrm{P}_1^X(\mathcal{X}_{1,a(1,b)}), \mathrm{P}_1^X(\mathcal{X}_{2,a(2,b)}) \right\} ,
\end{aligned}$$

for all $b \in \{1, \dots, B\}$. Then,

$$
\begin{aligned}
&||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}}||_{L_1(\mathrm{P}_1^X)} \\
&\leq \sum_{a=1}^{A_2} \mathrm{P}_1^X(\mathcal{X}_{2,a}) \cdot \frac{||k_{2,a}||_\infty}{\lambda_{2,a}} \\
&\qquad \cdot \max\left\{ |\psi|_1 c_{k_{2,a}} + (-2\varphi_{2,a}'(0))^{1/2} \frac{|\psi|_1 |\psi'|_1 \, ||k_{2,a}||_\infty^2}{\lambda_{2,a}} \,,\; |\psi'|_1 \, ||k_{2,a}||_\infty \right\} \\
&\qquad \cdot d_{\mathrm{W}}(\mathrm{P}_{1,\mathcal{X}_{2,a}}, \mathrm{P}_{2,\mathcal{X}_{2,a}}) \\
&\quad + |L|_1 \cdot \sum_{b=1}^{B} \Bigg(\rho_{1,b} \cdot \frac{\kappa_b^2}{2\tau_b^2} \cdot \left| \lambda_{1,a(1,b)} - \lambda_{2,a(2,b)} \right| \\
&\qquad\qquad + \mathrm{P}_1^X(\mathcal{X}_b^*) \cdot \Bigg(\frac{1}{2\tau_b} \cdot \left|\left| k_{1,b}^* - k_{2,b}^* \right|\right|_{L_1((\mathrm{P}_{1,b}^*)^X \otimes (\mathrm{P}_{1,b}^*)^X)} \\
&\qquad\qquad\qquad + \frac{\kappa_b}{\tau_b} \cdot \sqrt{\left|\left| k_{1,b}^* - k_{2,b}^* \right|\right|_{L_1((\mathrm{P}_{1,b}^*)^X \otimes (\mathrm{P}_{1,b}^*)^X)}} \Bigg) \\
&\qquad\qquad + \frac{\kappa_b^2}{\tau_b} \cdot \xi_{\mathrm{P}_1^X,b}(\boldsymbol{\mathcal{X}_1}, \boldsymbol{\mathcal{X}_2}) \Bigg) .
\end{aligned}
$$

Proof. Bound $||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}}||_{L_1(\mathrm{P}_1^X)}$ by

$$
\begin{aligned}
&||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}}||_{L_1(\mathrm{P}_1^X)} \\
&\leq ||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}_1}} - f_{\mathrm{P}_1,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}}||_{L_1(\mathrm{P}_1^X)} + ||f_{\mathrm{P}_1,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}_2}}||_{L_1(\mathrm{P}_1^X)} .
\end{aligned}
$$

The assertion then follows directly from applying Theorem 4.4.11 to the first summand on the right hand side and Theorem 4.4.7 to the second one. □

Chapter 5

Conclusion and Outlook

The past few years have seen a rapid surge in popularity of machine learning and artificial intelligence. Machine learning approaches have been greatly successful at learning unknown relations between input and output variables in scenarios that are far too complex to model them by hand or by classic statistical methods. Some such approaches—like deep learning—take their popularity mainly from this empirical success and have their theoretical justification just slowly catching up. Support Vector Machines (SVMs) on the other hand are already investigated very well and are known to possess many desirable theoretical properties, even though there are still some open questions. In addition, even though SVMs also observe good performance if the training data set is not too large, they do—for large data sets—suffer from their computational requirements growing super-linearly in the size of the training data set (with respect to computation time as well as computer memory).

The contribution of this thesis is twofold: On the one hand, we further refined the list of theoretical properties of SVMs by deriving new properties as well as considerably generalizing some that were already known. On the other hand, we tackled the super-linear computational requirements of SVMs by also investigating analogous theoretical properties of *localized* SVMs, which are based on the idea of dividing the input space into different spatial regions and then training local SVMs on these regions instead of a single global SVM on the whole input space. In addition to reducing the computational costs, localized SVMs can also offer benefits over regular SVMs (as well as over other approaches that also reduce the computational costs of SVMs) when it comes to the quality of predictions—as we discussed in Section 2.2.1, with Example 2.2.1 showing this possible improvement regarding the quality of predictions for simulated data.

The examined theoretical properties can be split into two groups which correspond to Chapter 3 and Chapter 4 respectively, with Chapter 3 containing the results from the peer-reviewed papers Köhler (2024a,b) and Chapter 4 those from the peer-reviewed paper Köhler and Christmann (2022), but both chapters also adding previously unpublished results. To be more specific, Chapter 3 gives results on different types of consistency—namely risk consistency, L_p-consistency and H-consistency—, with Section 3.2 notably containing results that are valid not only for (localized) SVMs, but which instead state very general connections between different types of consistency that can also be applied to other learning methods. It was shown that risk consistency—for the risk employing a distance-based loss function of growth type p—and L_p-consistency are equivalent under relatively mild assumptions (Theorems 3.2.1 and 3.2.3) and that both are implied by H-consistency (Corollary 3.2.4). These mild assumptions however include the finiteness of the averaged p-th moment of the distribution P. As P is usually unknown in practice, this moment condition is disadvantageous and we hoped to be able to eliminate it by switching to shifted loss functions, which is an approach that can in many cases lead to the moment condition not being required anymore (at least for loss functions of growth type 1), see Section 2.1.4. Surprisingly and even though we succeeded at showing that the moment condition is indeed not required for L_1-consistency implying risk consistency (Theorem 3.2.13) as well as for the connection to H-consistency (Corollary 3.2.14) when using shifted loss functions, it was however possible to prove that this benefit of shifted loss functions does not take full effect for the other direction of the connection between L_p- and risk consistency: It is not possible to just omit the moment condition and still obtain L_p-consistency directly from risk consistency, even when using shifted loss functions (Propositions 3.2.6 and 3.2.8). It might however be possible to replace the moment condition by some different condition(s) that might be less restrictive. We succeeded in deriving such alternative and in some sense weaker (see Examples 3.2.11 and 3.2.12) conditions in the special case of the applied loss function being the pinball loss (Theorem 3.2.10). An interesting open question is how such alternative conditions could look like for other loss functions and especially whether it is possible to derive weaker conditions that can replace the moment condition not only for singular loss functions such as the pinball loss but instead for wider classes of loss functions.

The general connections between the different types of consistency were then used as an aid in deriving new consistency results for SVMs in Section 3.3 as well as for localized SVMs in Section 3.4. To our knowledge, such results for general loss functions only existed for risk consistency but not for L_p- and H-consistency before, which is in parts due to SVMs being defined as minimizers of regularized risk functionals and risk consistency therefore being a property that is more closely connected to this definition than the other two types of consistency are. We thus derived results on L_p- and H-consistency which are completely new in that generality (Sections 3.3.1 and 3.3.3 to 3.3.5 for regular SVMs and Section 3.4.2 for localized SVMs).[39]

Regarding risk consistency, we have to differentiate between regular and localized SVMs. For regular SVMs, there already existed very general results on risk consistency. The contribution of this thesis is to apply the discovered connection between L_p- and risk consistency to derive slightly modified conditions which also yield risk consistency (Corollary 3.3.5): On the one hand, we were able to slightly relax the conditions imposed on the sequence of regularization parameters λ_n, requiring $\lambda_n^{p^*} n \to \infty$ only for $p^* = \max\{p+1, p(p+1)/2\}$ instead of for $p^* = \max\{2p, p^2\}$ as it was required by Christmann and Steinwart (2007, Theorem 12), where p again denotes the growth type of the applied distance-based loss function. This does not change the condition if $p = 1$ but constitutes a relaxation whenever $p > 1$ (Example 3.3.1). On the other hand, we had to add the assumption of the Bayes function $f^*_{L,\mathrm{P}}$ P^X-a.s. uniquely existing as a tradeoff. For localized SVMs, existing re-

[39] Note that the results on H-consistency were only derived for regular but not for localized SVMs because there does not necessarily exist a self-evident RKHS H containing the respective localized SVMs.

sults such as Hable (2013, Theorem 1) and Dumpert and Christmann (2018, Theorem 3.1) indeed offer significantly less generality than our Theorem 3.4.14, for example requiring special methods for obtaining the regionalizations underlying the localized SVMs, the output space $\mathcal{Y}$ being bounded, the regionalization not changing as the size of the training data set increases or only considering Lipschitz continuous loss functions.[40]

In addition to consistency, we also examined the total stability of SVMs and localized SVMs, which was the topic of Chapter 4. That is, we looked at the effect that changes in the underlying probability measure P respectively data set as well as in the applied regularization parameter λ and kernel k—and in case of localized SVMs additionally in the regionalization $\boldsymbol{\mathcal{X}}$—have on the resulting (localized) SVM. More specifically, we derived bounds of the type

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{\bullet} \leq c_1 \cdot d_1(\mathrm{P}_1, \mathrm{P}_2) + c_2 \cdot d_2(\lambda_1, \lambda_2) + c_3 \cdot d_3(k_1, k_2)$$

in the case of global SVMs and analogous bounds that additionally consider the difference between the two underlying regionalizations in the case of localized SVMs. Note that the constants c_1, c_2, c_3 are known, see Sections 4.3.4 and 4.4. We derived such bounds for $||\cdot||_{\bullet}$ being either the supremum norm or a suitable L_p-norm and called the corresponding properties total sup-stability respectively total L_p-stability.[41] The motivation behind investigating total stability is the same as that behind investigating classic statistical robustness: In practice, there can always be slight errors in the data one has at hand—be it small measurement or rounding errors affecting all data points or more drastic errors affecting a smaller share of the data points, for example stemming from human errors in writing down the data—and one would hope that such slight errors do only lead to small deviations in the resulting SVM if it is compared to the one that would have been obtained by using the "correct" data. Total stability takes this approach one step further and additionally considers the effect of slight changes in regularization parameter, kernel and regionalization.

To derive the corresponding bounds, each influence was examined separately, starting with the probability measure. Here, we considered two different ways to measure the difference between P_1 and P_2, namely the total variation distance and the Wasserstein distance. The total variation distance had already been used by Christmann et al. (2018) whose bound we slightly generalized by also allowing for loss functions that are not differentiable, such as the pinball loss (Proposition 4.3.3). The Wasserstein distance on the other hand had already been used in a stability result by Eckstein et al. (2023), which we however also generalized in several aspects, see also the discussion preceding the result in Proposition 4.3.5. In Section 4.2.1, we discussed in detail how both the total variation and the Wasserstein distance behave in different possible scenarios. For example, we noted that the Wasserstein distance offers the advantage of also being able to provide meaningful results if two empirical distributions are being compared, where one is obtained from the other

[40] Even though our conditions can generally be seen as the less restrictive ones, this is no strict comparison and there also exist situations in which those by Hable (2013) or those by Dumpert and Christmann (2018) are satisfied even though ours are not, see also the discussion at the beginning of Section 3.4.3.

[41] In contrast to the results on L_p-consistency in Chapter 3, the loss function did not need to be distance-based of growth type p in these stability results.

by slightly shifting the whole data set (Example 4.2.8). Looking at minimal examples, we saw that the derived bound using the Wasserstein distance can indeed yield considerably better results than the one using the total variation distance in such a situation, but that this comparison flips if the shift between the two underlying data sets increases (Example 4.3.8). For this reason, each of the two bounds has its merit. Analogous minimal examples further showed that the bound using the total variation distance is sharp up to a factor of at most 2 (Example 4.3.4). It remains an open question whether there also exist examples that completely exhaust the bound, thus making it sharp, or whether it is possible to further improve the bound.

The bounds regarding the differences between regularization parameters and kernels are also similar in shape to bounds already derived by Christmann et al. (2018). They however also considerably generalize the bounds by Christmann et al. (2018), especially the one considering the difference between the kernels, which does not only eliminate the need for the loss function to be differentiable but also additional assumptions on the regularization parameter respectively kernels that were required by Christmann et al. (2018). In addition to the achieved generalization, Proposition 4.3.10 also improves the existing bounds regarding the difference between the regularization parameters by a factor of at least 2, thus making it asymptotically sharp (Example 4.3.12). The bound with respect to the difference between the kernels on the other hand pays for the mentioned considerable generalizations by adding a non-linear term to the bound

$$||f_{\mathrm{P},\lambda,k_1} - f_{\mathrm{P},\lambda,k_2}||_\infty \leq \frac{|L|_1}{\lambda} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_\infty + \max\{||k_1||_\infty , ||k_2||_\infty\} \cdot \sqrt{||k_1 - k_2||_\infty}\right) ,$$

which dominates the bound whenever the two kernels are reasonably close. For minimal examples such as the one considered in Example 4.3.16, the linear part of the bound actually suffices and is sharp. So far, we have, however, not been able to actually eliminate the non-linear part from the bound (without imposing additional assumptions like the ones used by Christmann et al., 2018), for which reason it remains an open question whether this is possible or whether there also exist examples for which the linear part does not suffice.

These different bounds were then combined in order to obtain results on total stability of SVMs (Section 4.3.4) as well as localized SVMs (Section 4.4.1). To be more specific, the bounds directly resulting from this actually only yielded regionalization-subtotal stability instead of total stability in the latter case, because our notion of total stability in the localized case additionally includes stability with respect to changes in the regionalization. For deriving such stability also with respect to changes in the regionalization, it was necessary to define a notion of difference between two regionalizations. This was less straightforward than for the difference between regularization parameters, kernels and—because there already existed many well-investigated distance measures for this, such as the total variation and the Wasserstein distance—also probability measures. In Section 4.2.4, we introduced such a possible notion of difference and gave examples on its behavior showing that it successfully captures what one would intuitively describe as regionalizations being similar or dissimilar, even though it does not define a metric because it does in general not satisfy the triangle inequality (Example 4.2.18). With this, it was possible to indeed derive results on total stability of localized SVMs (i.e. bounds for the difference between localized SVMs that also differ with respect to their regionalization) as well in Section 4.4.2, however only on L_1-stability but not on sup-stability as we were able to give examples showing that it is not possible to derive meaningful results on total sup-stability of localized SVMs (Example 4.4.8). For future work, it might be interesting to examine whether it is also possible to change the derived bound in such a way that it uses a notion of difference between regionalizations which actually defines a metric.

In summary, the focus of this thesis was to mathematically derive certain properties—consistency and total stability—of SVMs and localized SVMs. Whereas we still also included empirical investigations (based on simulated data) affirming the results on consistency (Examples 3.4.13 and 3.4.16), we did for this reason not perform any empirical studies on total stability. Hence, another extension of this thesis that might be of interest is to, for example, look at simulated or real-world data and compute (localized) SVMs based on different probability measures,[42] regularization parameters, kernels and regionalizations, and then to compare their actual difference with the bounds that were derived in Chapter 4 (which we did analytically for minimal examples in Section 4.3). It has to be noted that the bounds will oftentimes take values that greatly exceed the actual differences—even though the minimal examples from Section 4.3 showed that parts of the bounds are indeed sharp or at least almost sharp—, which is however a completely natural drawback of their *strong, non-probabilistic nature* and *generality* that cannot be circumvented without impeding at least one of these two desirable properties.

[42] This can for example be achieved by taking different parts of the available data into account for computing the different (localized) SVMs.

List of Symbols

Sets and Spaces

$\mathbb{N}$	set of natural numbers	
$\mathbb{N}_0$	$\mathbb{N} \cup \{0\}$	
$\mathbb{R}$	set of real numbers	
$\mathbb{C}$	set of complex numbers	
$\mathcal{X}$	input space (usually assumed to be a complete separable metric space)	
$\mathcal{Y}$	output space (assumed to be a closed subset of $\mathbb{R}$)	
$\boldsymbol{\mathcal{X}}$	regionalization of $\mathcal{X}$	33
$\boldsymbol{\mathcal{X}}^*_{\mathbf{1,2}}$	combined regionalization of $\boldsymbol{\mathcal{X}}_{\mathbf{1}}$ and $\boldsymbol{\mathcal{X}}_{\mathbf{2}}$	33
$\mathcal{B}_S$	Borel σ-algebra on S	
$\mathcal{W}_1(\Omega)$	Wasserstein space (of order 1) on Ω	94
$\mathcal{L}_0(\mathcal{X})$	space of measurable functions $f\colon (\mathcal{X}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$	
$L_p(\mu)$	L_p-space, space of (equivalence classes of) measurable functions whose p-th absolute power is μ-integrable (for a measure μ)	
$L_p(\mathcal{X})$	L_p-space with respect to the Lebesgue measure on $\mathcal{X}$	
$W^{m,q}(\mathcal{X})$	Sobolev space	51
$I_{\boldsymbol{\mathcal{X}},\mathrm{P}}$	indices of those regions in $\boldsymbol{\mathcal{X}}$ that probability measure P assigns positive probability to	67

Probability Measures and Related Concepts

$\mathcal{U}(a,b)$	uniform distribution on (a,b)	
δ_x	Dirac distribution in x	
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2	
P	probability measure on $\mathcal{X} \times \mathcal{Y}$	
P^X	marginal distribution of X on $\mathcal{X}$	
$\mathrm{P}(\cdot \,\vert\, X = x)$	conditional distribution of Y on $\mathcal{Y}$ given $X = x$	
$\mathrm{P}(\cdot \,\vert\, X)$	shortening notation for conditional distributions $\mathrm{P}(\cdot \,\vert\, X = x)$, $x \in \mathcal{X}$	
$\vert\mathrm{P}\vert_p$	average p-th moment of P	21
$\mathrm{P}_{\tilde{\mathcal{X}}}$	local measure on $\tilde{\mathcal{X}} \times \mathcal{Y}$, for $\tilde{\mathcal{X}} \subseteq \mathcal{X}$, based on P	33
$\mathbf{P}_{\boldsymbol{\mathcal{X}}}$	vector of local measures on $\boldsymbol{\mathcal{X}}$ based on P	34
D_n	data set consisting of $n \in \mathbb{N}$ observations $(x_1, y_1), \ldots, (x_n, y_n)$ sampled from P	

D_n	empirical distribution associated with D_n	9
(X, Y)	pair of random variables with distribution P	
$\mathbb{E}_{\mathrm{Q}}\left[Z\right]$	expectation of Z with respect to probability measure Q	

Loss Functions and Risks

L	loss function	8
$L_{\tau\text{-pin}}$	τ-pinball loss function with parameter $\tau \in (0,1)$	9
$L^\star$	shifted loss function	24
L'	derivative of L with respect to its last argument	17
∂L	subdifferential of L with respect to its last argument	18
$\mathcal{R}_{L,\mathrm{P}}$	L-risk	8
$\mathcal{R}_{L,\mathrm{D}_n}$	empirical L-risk	9
$\mathcal{R}^*_{L,\mathrm{P}}$	Bayes L-risk	9
$\mathcal{C}_{L,\mathrm{P}(\cdot\,\vert\,x),x}$	inner L-risk	23
$\mathcal{C}^*_{L,\mathrm{P}(\cdot\,\vert\,x),x}$	inner Bayes L-risk	23

Kernels, Reproducing Kernel Hilbert Spaces and Regularization Parameters

k	kernel	13
k_γ	Gaussian RBF kernel with parameter $\gamma > 0$	16
Φ	canonical feature map of k	14
H	reproducing kernel Hilbert space of k	14
λ	regularization parameter > 0	11
$\boldsymbol{k}$	vector of kernels on $\boldsymbol{\mathcal{X}}$	35
$\boldsymbol{\lambda}$	vector of regularization parameters	35

Functions

$\hat{f}$	zero-extension of f	34
$f\vert_S$	restriction of f to S	
f^+	positive part $\max\{f, 0\}$ of f	
f^-	negative part $\max\{-f, 0\}$ of f	
∂f	subdifferential of f	18
$f_{L,\mathrm{P},\lambda,k}$	support vector machine (also $f_{\mathrm{P},\lambda,k}$)	11
$f_{L,\mathrm{D}_n,\lambda,k}$	empirical support vector machine	11
$f_{L,\mathrm{P},\boldsymbol{\lambda},\boldsymbol{k},\boldsymbol{\mathcal{X}}}$	localized support vector machine (also $f_{\mathrm{P},\boldsymbol{\lambda},\boldsymbol{k},\boldsymbol{\mathcal{X}}}$)	35
$f^*_{L,\mathrm{P}}$	Bayes function	9
$f^*_{\tau,\mathrm{P}}$	conditional τ-quantile function	49
$\mathbb{1}_S$	indicator function, $\mathbb{1}_S(x) = 1$ if $x \in S$, $\mathbb{1}_S(x) = 0$ else	
w_a	weight function associated with region $\mathcal{X}_a$	34

Norms, Distances and Related Concepts

$\lvert \cdot \rvert_1$	Lipschitz constant	17
$\lvert\lvert\cdot\rvert\rvert_1$	1-norm of vectors	
$\lvert\lvert\cdot\rvert\rvert_\infty$	supremum norm	
$\lvert\lvert\cdot\rvert\rvert_{L_p(\mathrm{Q})}$	$L_p(\mathrm{Q})$-norm	
$\lvert\lvert\cdot\rvert\rvert_H$	norm in Hilbert space H	
$\langle \cdot, \cdot \rangle_H$	inner product in Hilbert space H	
d_{tv}	total variation distance	91
d_{W}	Wasserstein distance	93
$\xi_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}}_\mathbf{1}, \boldsymbol{\mathcal{X}}_\mathbf{2})$	distance between regionalizations $\boldsymbol{\mathcal{X}}_\mathbf{1}$ and $\boldsymbol{\mathcal{X}}_\mathbf{2}$ on b-th element of $\boldsymbol{\mathcal{X}}^*_\mathbf{1,2}$	100

Miscellaneous

$\mathcal{L}$	measurable learning method	37
v^T	transpose of vector v	

Abbreviations

a.s.	almost surely
i.i.d.	independent and identically distributed
RBF	radial basis function
RKHS	reproducing kernel Hilbert space
SVM	support vector machine

List of Figures

List of Tables

Bibliography

Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*. Pure and Applied Mathematics. Elsevier, Amsterdam.

Agarwal, S. and Niyogi, P. (2009). Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10:441–474.

Alpaydın, E. (2020). *Introduction to Machine Learning*. MIT Press, Cambridge, Massachusetts.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, doi:10.1090/S0002-9947-1950-0051437-7.

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In Shalev-Shwartz, S. and Steinwart, I., editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 185–209, Princeton. PMLR.

Bauer, H. (2001). *Measure and Integration Theory*. de Gruyter Studies in Mathematics. de Gruyter, Berlin, doi:10.1515/9783110866209.

Bellet, A. and Habrard, A. (2015). Robustness and generalization for metric learning. *Neurocomputing*, 151:259–267, doi:10.1016/j.neucom.2014.09.044.

Bennett, K. P. and Blue, J. A. (1998). A support vector machine approach to decision trees. In *The 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence*, volume 3, pages 2396–2401. IEEE, doi:10.1109/IJCNN.1998.687237.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science+Business Media, New York, doi:10.1007/978-1-4419-9096-9.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033.

Billingsley, P. (1995). *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Blanzieri, E. and Bryl, A. (2007). Instance-based spam filtering using SVM nearest neighbor classifier. In Wilson, D. and Sutcliffe, G., editors, *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference*, pages 441–442. AAAI Press.

Blanzieri, E. and Melgani, F. (2008). Nearest neighbor classification of remote sensing images with the maximal margin principle. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1804–1811, doi:10.1109/TGRS.2008.916090.

Blaschzyk, I. and Steinwart, I. (2022). Improved classification rates for localized SVMs. *Journal of Machine Learning Research*, 23:1–59.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. Association for Computing Machinery, doi:10.1145/130385.130401.

Bottou, L. and Vapnik, V. (1992). Local learning algorithms. *Neural Computation*, 4(6):888–900, doi:10.1162/neco.1992.4.6.888.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Cao, Q., Guo, Z.-C., and Ying, Y. (2016). Generalization bounds for metric and similarity learning. *Machine Learning*, 102:115–132, doi:10.1007/s10994-015-5499-7.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, doi:10.1007/s10208-006-0196-8.

Carratino, L., Vigogna, S., Calandriello, D., and Rosasco, L. (2021). ParK: Sound and efficient kernel ridge regression by feature space partitions. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6430–6441. Curran Associates, Inc.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chang, F., Guo, C.-Y., Lin, X.-R., and Lu, C.-J. (2010). Tree decomposition for large-scale SVM problems. *Journal of Machine Learning Research*, 11:2935–2972.

Cheng, H., Tan, P.-N., and Jin, R. (2010). Efficient algorithm for localized support vector machine. *IEEE Transactions on Knowledge and Data Engineering*, 22(4):537–549, doi:10.1109/TKDE.2009.116.

Cherkassky, V. and Mulier, F. (2007). *Learning from Data*. Wiley, Hoboken, New Jersey, doi:10.1002/9780470140529.

Chernih, A., Sloan, I. H., and Womersley, R. S. (2014). Wendland functions with increasing smoothness converge to a Gaussian. *Advances in Computational Mathematics*, 40:185–200, doi:10.1007/s10444-013-9304-5.

Christmann, A. and Steinwart, I. (2004). On robustness 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 risk minimization. *Bernoulli*, 13(3):799–819, doi:10.3150/07-BEJ5102.

Christmann, A., Steinwart, I., and Hubert, M. (2007). Robust learning from bites for data mining. *Computational Statistics & Data Analysis*, 52(1):347–361, doi:10.1016/j.csda.2006.12.009.

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(3):311–327, doi:10.4310/SII.2009.v2.n3.a5.

Christmann, A., Xiang, D.-H., and Zhou, D.-X. (2018). Total stability of kernel methods. *Neurocomputing*, 289:101–118, doi:10.1016/j.neucom.2018.02.009.

Christmann, A. and Zhou, D.-X. (2016). On the robustness of regularized pairwise learning methods based on kernels. *Journal of Complexity*, 37:1–33, doi:10.1016/j.jco.2016.07.001.

Clarke, B., Fokoué, E., and Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer Series in Statistics. Springer Science+Business Media, New York, doi:10.1007/978-0-387-98135-2.

Clémençon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, doi:10.1214/009052607000000910.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2010). Generalization bounds for learning kernels. In Fürnkranz, J. and Joachims, T., editors, *Proceedings of the 27th International Conference on Machine Learning*, pages 247–254, Haifa. Omnipress.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297, doi:10.1007/BF00994018.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, doi:10.1017/CBO9780511801389.

Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, doi:10.1090/S0273-0979-01-00923-5.

Cucker, F. and Zhou, D.-X. (2007). *Learning Theory: An Approximation Theory Viewpoint.* Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, doi:10.1017/CBO9780511618796.

Denkowski, Z., Migórski, S., and Papageorgiou, N. S. (2003). *An Introduction to Nonlinear Analysis: Theory.* Springer Science+Business Media, New York, doi:10.1007/978-1-4419-9158-4.

Dereich, S., Scheutzow, M., and Schottstedt, R. (2013). Constructive quantization: Approximation by empirical measures. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 49(4):1183–1203, doi:10.1214/12-AIHP489.

Devroye, L. (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(2):154–157, doi:10.1109/TPAMI.1982.4767222.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Stochastic Modelling and Applied Probability. Springer, New York, doi:10.1007/978-1-4612-0711-5.

Diestel, J. (1984). *Sequences and Series in Banach Spaces.* Graduate Texts in Mathematics. Springer, New York, doi:10.1007/978-1-4612-5200-9.

Diestel, J. and Uhl, J. J. (1977). *Vector Measures.* American Mathematical Society, Providence, Rhode Island, doi:10.1090/surv/015.

Dieuleveut, A. and Bach, F. (2016). Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, doi:10.1214/15-AOS1391.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., and Vapnik, V. (1996). Support vector regression machines. In Mozer, M., Jordan, M., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, pages 155–161. MIT Press.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification.* Wiley, New York.

Dudley, R. M. (2004). *Real Analysis and Probability.* Cambridge University Press, Cambridge, doi:10.1017/CBO9780511755347.

Dumpert, F. (2020). Quantitative robustness of localized support vector machines. *Communications on Pure & Applied Analysis*, 19(8):3947–3956, doi:10.3934/cpaa.2020174.

Dumpert, F. and Christmann, A. (2018). Universal consistency and robustness of localized support vector machines. *Neurocomputing*, 315:96–106, doi:10.1016/j.neucom.2018.06.061.

Dunford, N. and Schwartz, J. T. (1957). *Linear Operators, Part I: General Theory.* Pure and Applied Mathematics. A Series of Texts and Monographs. John Wiley & Sons, New York.

Eberts, M. and Steinwart, I. (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electronic Journal of Statistics*, 7:1–42, doi:10.1214/12-EJS760.

Eckstein, S., Iske, A., and Trabs, M. (2023). Dimensionality reduction and Wasserstein stability for kernel regression. *Journal of Machine Learning Research*, 24:1–35.

El Alaoui, A. and Mahoney, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28, pages 775–783. Curran Associates, Inc.

Farooq, M. and Steinwart, I. (2019). Learning rates for kernel-based expectile regression. *Machine Learning*, 108:203–227, doi:10.1007/s10994-018-5762-9.

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21:1–38.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, doi:10.1007/s00440-014-0583-7.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, doi:10.1214/aos/1176347963.

Gensler, P. and Christmann, A. (2022). On the robustness of kernel-based pairwise learning. In Steland, A. and Tsui, K.-L., editors, *Artificial Intelligence, Big Data and Data Science in Statistics*, pages 111–153. Springer, Cham, doi:10.1007/978-3-031-07155-3_5.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, doi:10.1111/j.1751-5823.2002.tb00178.x.

Gibbs, S. (2017). AlphaZero AI beats champion chess program after teaching itself in four hours. *The Guardian*, 7 December 2017. Available: `https://www.theguardian.com/technology/2017/dec/07/alphazero-google-deepmind-ai-beats-champion-program-teaching-itself-to-play-four-hours` [Last accessed: 20 September 2024].

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006). A kernel method for the two-sample-problem. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19, pages 513–520. MIT Press, doi:10.7551/mitpress/7503.003.0069.

Gu, Q. and Han, J. (2013). Clustered support vector machines. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 307–315, Scottsdale, Arizona. PMLR.

Guo, Z.-C., Lin, S.-B., and Zhou, D.-X. (2017). Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, doi:10.1088/1361-6420/aa72b2.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer, New York, doi:10.1007/b97848.

Hable, R. (2013). Universal consistency of localized versions of regularized kernel methods. *Journal of Machine Learning Research*, 14:153–186.

Hable, R. and Christmann, A. (2011). On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, 102:993–1007, doi:10.1016/j.jmva.2011.01.009.

Hable, R. and Christmann, A. (2014). Estimation of scale functions to model heteroscedasticity by regularised kernel-based quantile methods. *Journal of Nonparametric Statistics*, 26(2):219–239, doi:10.1080/10485252.2013.875547.

Hamm, T. and Steinwart, I. (2022). Intrinsic dimension adaptive partitioning for kernel methods. *SIAM Journal on Mathematics of Data Science*, 4(2):721–749, doi:10.1137/21M1435690.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, doi:10.1214/aoms/1177693054.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Hang, H. and Steinwart, I. (2017). A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2):708–743, doi:10.1214/16-AOS1465.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, New York, doi:10.1007/978-0-387-84858-7.

Hewitt, E. and Hewitt, R. E. (1979). The Gibbs-Wilbraham phenomenon: An episode in Fourier analysis. *Archive for History of Exact Sciences*, 21(2):129–160, doi:10.1007/BF00330404.

Hu, T. and Zhou, D.-X. (2021). Distributed regularized least squares with flexible Gaussian kernels. *Applied and Computational Harmonic Analysis*, 53:349–377, doi:10.1016/j.acha.2021.03.008.

Huang, S. and Wu, Q. (2021). Robust pairwise learning with Huber loss. *Journal of Complexity*, 66:101570, doi:10.1016/j.jco.2021.101570.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, doi:10.1214/aoms/1177703732.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Le Cam, L. M. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 221–233, Berkeley. University of California Press.

Huber, P. J. (1981). *Robust Statistics.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, doi:10.1002/0471725250.

Joachims, T. (1998). Making large-scale SVM learning practical. In Schölkopf, B., Burges, C., and Smola, A. J., editors, *Kernel Methods: Support Vector Learning.* MIT Press, Cambridge, Massachusetts.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-normalizing neural networks. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 971–980. Curran Associates, Inc.

Koenker, R. (2005). *Quantile Regression.* Econometric Society Monographs. Cambridge University Press, Cambridge, doi:10.1017/CBO9780511754098.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50, doi:10.2307/1913643.

Köhler, H. (2024a). Lp- and risk consistency of localized SVMs. *Neurocomputing*, 598(128060):1–13, doi:10.1016/j.neucom.2024.128060.

Köhler, H. (2024b). On the connection between Lp- and risk consistency and its implications on regularized kernel methods. *Journal of Machine Learning Research*, 25(213):1–33.

Köhler, H. and Christmann, A. (2022). Total stability of SVMs and localized SVMs. *Journal of Machine Learning Research*, 23(100):1–41.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L., and Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72.

Leisch, F. and Dimitriadou, E. (2021). *mlbench: Machine Learning Benchmark Problems.* R package version 2.1-3.1.

Lin, J. and Rosasco, L. (2017). Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18:1–47.

Lin, S.-B., Guo, X., and Zhou, D.-X. (2017). Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18:3202–3232.

Lin, S.-B., Wang, K., Wang, Y., and Zhou, D.-X. (2022). Universal consistency of deep convolutional neural networks. *IEEE Transactions on Information Theory*, 68(7):4610–4617, doi:10.1109/TIT.2022.3151753.

Liu, F., Huang, X., Chen, Y., and Suykens, J. A. K. (2022). Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148.

Liu, J. and Shi, L. (2024). Statistical optimality of divide and conquer kernel-based functional linear regression. *Journal of Machine Learning Research*, 25:1–56.

Liu, Y. and Liao, S. (2015). Eigenvalues ratio for kernel selection of kernel methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29(1), pages 2814–2820. doi:10.1609/aaai.v29i1.9554.

Liu, Y., Liao, S., Jiang, S., Ding, L., Lin, H., and Wang, W. (2020). Fast cross-validation for kernel-based algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1083–1096, doi:10.1109/TPAMI.2019.2892371.

Lohr, S. (2012). The Age of Big Data. *The New York Times*, 12 February 2012, section SR, page 1. Available: `https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html` [Last accessed: 20 September 2024].

Lv, S., Wang, J., Liu, J., and Liu, Y. (2021). Improved learning rates of a functional lasso-type SVM with sparse multi-kernel representation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21467–21479. Curran Associates, Inc.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, doi:10.1002/0470010940.

McFarland, M. (2016). What AlphaGo's sly move says about machine creativity. *The Washington Post*, 15 March 2016. Available: `https://www.washingtonpost.com/news/innovations/wp/2016/03/15/what-alphagos-sly-move-says-about-machine-creativity` [Last accessed: 20 September 2024].

Meanti, G., Carratino, L., Rosasco, L., and Rudi, A. (2020). Kernel methods through the roof: Handling billions of points efficiently. In Laronchelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14410–14422.

Mei, S., Misiakiewicz, T., and Montanari, A. (2022). Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, doi:10.1016/j.acha.2021.12.003.

Meister, M. and Steinwart, I. (2016). Optimal learning rates for localized SVMs. *Journal of Machine Learning Research*, 17:1–44.

Mendelson, S. and Neeman, J. (2010). Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, doi:10.1214/09-AOS728.

Micchelli, C. A. and Pontil, M. (2005). Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill Series in Computer Science. McGraw-Hill, Singapore.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, doi:10.1561/2200000060.

Mücke, N. (2019). Reducing training time by efficient localized kernel regression. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2603–2610. PMLR.

Mücke, N. and Blanchard, G. (2018). Parallelizing spectrally regularized kernel algorithms. *Journal of Machine Learning Research*, 19:1–29.

Mukherjee, S., Niyogi, P., Poggio, T., and Rifkin, R. (2006). Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, doi:10.1007/s10444-004-7634-z.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, doi:10.2307/1428011.

Ong, C. S., Smola, A. J., and Williamson, R. C. (2005). Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 6:1043–1071.

Panaretos, V. M. and Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space.* SpringerBriefs in Probability and Mathematical Statistics. Springer, Cham, doi:10.1007/978-3-030-38438-8.

Paoletti, M. E., Haut, J. M., Plaza, J., and Plaza, A. (2019). Deep learning classifiers for hyperspectral imaging: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158:279–317, doi:10.1016/j.isprsjprs.2019.09.006.

Platt, J. C. (1998). Fast training of support vector machines using sequential minimal optimization. In Schölkopf, B., Burges, C., and Smola, A. J., editors, *Advances in Kernel Methods: Support Vector Learning*, pages 185–208. MIT Press, Cambridge, Massachusetts, doi:10.7551/mitpress/1130.003.0016.

Poggio, T., Rifkin, R., Mukherjee, S., and Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, 428:419–422, doi:10.1038/nature02341.

R Core Team (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rachev, S. T. (1991). *Probability Metrics and the Stability of Stochastic Models.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Chichester.

Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems. Volume I: Theory.* Probability and its Applications. Springer, New York.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184. Curran Associates, Inc.

Rockafellar, R. T. (1972). *Convex Analysis.* Princeton University Press, Princeton, New Jersey, doi:10.1515/9781400873173.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, doi:10.1002/0471725382.

Rudi, A., Camoriano, R., and Rosasco, L. (2015). Less is more: Nyström computational regularization. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28, pages 1657–1665. Curran Associates, Inc.

Rudi, A., Carratino, L., and Rosasco, L. (2017). FALKON: An optimal large scale kernel method. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3888–3898. Curran Associates, Inc.

Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with random features. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3215–3225. Curran Associates, Inc.

Saitoh, S. and Sawano, Y. (2016). *Theory of Reproducing Kernels and Applications.* Developments in Mathematics. Springer, Singapore, doi:10.1007/978-981-10-0530-5.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, doi:10.1147/rd.33.0210.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, doi:10.1162/089976698300017467.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts, doi:10.7551/mitpress/4175.001.0001.

Segata, N. and Blanzieri, E. (2010). Fast and scalable local kernel machines. *Journal of Machine Learning Research*, 11:1883–1926.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, New York, doi:10.1017/CBO9781107298019.

Sheng, B., Liu, H., and Wang, H. (2020). Learning rates for the kernel regularized regression with a differentiable strongly convex loss. *Communications on Pure & Applied Analysis*, 19(8):3973–4005, doi:10.3934/cpaa.2020176.

Smale, S. and Yao, Y. (2006). Online learning algorithms. *Foundations of Computational Mathematics*, 6(2):145–170, doi:10.1007/s10208-004-0160-z.

Smale, S. and Zhou, D.-X. (2003). Estimating the approximation error in learning theory. *Analysis and Applications*, 1(1):17–41, doi:10.1142/S0219530503000089.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, doi:10.1023/B:STCO.0000035301.49549.88.

Sriperumbudur, B. (2016). On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, doi:10.3150/15-BEJ713.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561.

Sriperumbudur, B. K. and Szabó, Z. (2015). Optimal rates for random Fourier features. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28, pages 1144–1152. Curran Associates, Inc.

Steinwart, I. (2003). Sparseness of support vector machines—some asymptotically sharp bounds. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16, pages 1069–1076. MIT Press.

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, doi:10.1109/TIT.2004.839514.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines.* Information Science and Statistics. Springer, New York, doi:10.1007/978-0-387-77242-4.

Steinwart, I. and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, doi:10.3150/10-BEJ267.

Steinwart, I., Hush, D., and Scovel, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory.*

Steinwart, I. and Thomann, P. (2017). liquidSVM: A fast and versatile SVM package. *arXiv e-prints 1702.06899.* Software available at `http://pnp.mathematik.uni-stuttgart.de/isa/steinwart/software/liquidSVM.html`.

Sun, H. and Wu, Q. (2021). Optimal rates of distributed regression with imperfect kernels. *Journal of Machine Learning Research*, 22:1–34.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264.

Thomann, P., Blaschzyk, I., Meister, M., and Steinwart, I. (2017). Spatial decompositions for large scale SVMs. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 1329–1337. PMLR.

Tibshirani, R. and Hastie, T. (2007). Margin trees for high-dimensional classification. *Journal of Machine Learning Research*, 8:637–652.

Tierney, L. (1996). Introduction to general state-space Markov chain theory. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, pages 59–74. Chapman & Hall, Boca Raton, Florida.

Tong, H. and Ng, M. K. (2019). Calibration of ε-insensitive loss in support vector machines regression. *Journal of the Franklin Institute*, 356:2111–2129, doi:10.1016/j.jfranklin.2018.11.021.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer, New York, doi:10.1007/b13794.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer, New York, doi:10.1007/978-1-4757-2440-0.

Vapnik, V. N. (1998). *Statistical Learning Theory.* Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley & Sons, New York.

Vapnik, V. N. and Bottou, L. (1993). Local algorithms for pattern recognition and dependencies estimation. *Neural Computation*, 5(6):893–909, doi:10.1162/neco.1993.5.6.893.

Vapnik, V. N. and Chervonenkis, A. Y. (1964). On a class of perceptrons. *Automation and Remote Control*, 25:103–109.

Vapnik, V. N., Golowich, S., and Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. In Mozer, M., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, pages 281–287. MIT Press.

Vapnik, V. N. and Lerner, A. Y. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780.

Villani, C. (2009). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin.

von Luxburg, U. and Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In Gabbay, D. M., Hartmann, S., and Woods, J., editors, *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 651–706. Elsevier, Oxford, doi:10.1016/B978-0-444-52936-7.50016-1.

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, doi:10.3150/18-BEJ1065.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244.

Wendland, H. (2005). *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, doi:10.1017/CBO9780511617539.

Williams, C. and Seeger, M. (2000). Using the Nyström method to speed up kernel machines. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13, pages 682–688. MIT Press.

Wu, D., Bennett, K. P., Cristianini, N., and Shawe-Taylor, J. (1999). Large margin trees for induction and transduction. In Bratko, I. and Dzeroski, S., editors, *Proceedings of the 16th International Conference on Machine Learning*, pages 474–483. Morgan Kaufmann.

Wu, Q., Ying, Y., and Zhou, D.-X. (2007). Multi-kernel regularized classifiers. *Journal of Complexity*, 23:108–134, doi:10.1016/j.jco.2006.06.007.

Xiang, D.-H. (2013). Conditional quantiles with varying Gaussians. *Advances in Computational Mathematics*, 38(4):723–735, doi:10.1007/s10444-011-9257-5.

Xiang, D.-H., Hu, T., and Zhou, D.-X. (2012). Approximation analysis of learning algorithms for support vector regression and quantile regression. *Journal of Applied Mathematics*, 2012:1–17.

Xiang, D.-H. and Zhou, D.-X. (2009). Classification with Gaussians and convex loss. *Journal of Machine Learning Research*, 10:1447–1468.

Yang, T., Li, Y.-F., Mahdavi, M., Jin, R., and Zhou, Z.-H. (2012). Nyström method vs random Fourier features: A theoretical and empirical comparison. In Pereira, F., Burges, C. J., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25, pages 476–484. Curran Associates, Inc.

Ying, Y. and Campbell, C. (2009). Generalization bounds for learning the kernel. In *Proceedings of the 22nd Annual Conference on Learning Theory*.

Ying, Y. and Zhou, D.-X. (2006). Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788, doi:10.1109/TIT.2006.883632.

Ying, Y. and Zhou, D.-X. (2007). Learnability of Gaussians with flexible variances. *Journal of Machine Learning Research*, 8:249–276.

Ying, Y. and Zhou, D.-X. (2016). Online pairwise learning algorithms. *Neural Computation*, 28(4):743–777, doi:10.1162/NECO_a_00817.

Ying, Y. and Zhou, D.-X. (2017). Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42:224–244, doi:10.1016/j.acha.2015.08.007.

Zhang, H., Berg, A, C., Maire, M., and Malik, J. (2006). SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2126–2136. IEEE Computer Society, doi:10.1109/CVPR.2006.301.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, doi:10.1214/009053605000000255.

Zhang, Y., Duchi, J., and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340.

Zhao, Y., Fan, J., and Shi, L. (2017). Learning rates for regularized least squares ranking algorithm. *Analysis and Applications*, 15(6):815–836, doi:10.1142/S0219530517500063.

Zhou, Z.-H. (2021). *Machine Learning*. Springer, Singapore, doi:10.1007/978-981-15-1967-3. Translated by Shaowu Liu.

Zolotarev, V. M. (1976). Metric distances in spaces of random variables and their distributions. *Mathematics of the USSR-Sbornik*, 30(3):373–401, doi:10.1070/SM1976v030n03ABEH002280.

Publications

- Köhler, H. and Christmann, A. (2022). Total stability of SVMs and localized SVMs. *Journal of Machine Learning Research*, 23(100):1–41. Licensed under CC BY 4.0.
- Köhler, H. (2024a). Lp- and risk consistency of localized SVMs. *Neurocomputing*, 598(128060):1–13, doi:10.1016/j.neucom.2024.128060 Licensed under CC BY 4.0.
- Köhler, H. (2024b). On the connection between Lp- and risk consistency and its implications on regularized kernel methods. *Journal of Machine Learning Research*, 25(213):1–33. Licensed under CC BY 4.0.

The content of the first publication is included in Chapter 4, while the content of the second and the third publication is included in Chapter 3.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet habe.

Weiterhin erkläre ich, dass ich die Hilfe von gewerblichen Promotionsberatern bzw. -vermittlern oder ähnlichen Dienstleistern weder bisher in Anspruch genommen habe, noch künftig in Anspruch nehmen werde.

Zusätzlich erkläre ich hiermit, dass ich keinerlei frühere Promotionsversuche unternommen habe.

Bayreuth, den

Hannes Köhler