Journal of Machine Learning Research 23 (2022) 1-41 Submitted 2/21; Revised 10/21; Published 2/22

Total Stability of SVMs and Localized SVMs

Hannes Köhler* HANNES.KOEHLER@UNI-BAYREUTH.DE
Andreas Christmann ANDREAS.CHRISTMANN@UNI-BAYREUTH.DE
Department of Mathematics
University of Bayreuth
95440 Bayreuth, Germany

Editor: Gabor Lugosi

Abstract

Regularized kernel-based methods such as support vector machines (SVMs) typically depend on the underlying probability measure P (respectively an empirical measure D_n in applications) as well as on the regularization parameter λ and the kernel k. Whereas classical statistical robustness only considers the effect of small perturbations in P, the present paper investigates the influence of simultaneous slight variations in the whole triple (P, λ, k), respectively $(\mathrm{D}_n, \lambda_n, k)$, on the resulting predictor. Existing results from the literature are considerably generalized and improved. In order to also make them applicable to big data, where regular SVMs suffer from their super-linear computational requirements, we show how our results can be transferred to the context of localized learning. Here, the effect of slight variations in the applied regionalization, which might for example stem from changes in P respectively D_n, is considered as well.

Keywords: statistical robustness, stability, localized learning, kernel methods, big data

1. Introduction

Let $\mathcal{X} \times \mathcal{Y}$ be a set and let P be the distribution of a pair of random variables (X, Y) with values in $\mathcal{X} \times \mathcal{Y}$, where X is the input variable and Y is the real-valued output variable. The goal of statistical machine learning is to find a function $f : \mathcal{X} \to \mathcal{Y}$ which relates X to Y, that is, which can be used to predict the unknown output variable based on a given input variable, with (almost) no prior knowledge about P. One way to approach such prediction problems, both for regression and classification purposes, is employing *support vector machines* (SVMs) which perform regularized empirical risk minimization on special Hilbert spaces of functions, so-called *reproducing kernel Hilbert spaces* (RKHSs), and which have been the focus of extensive theoretical investigations (cf. Vapnik, 1995, 1998; Schölkopf and Smola, 2002; Cucker and Zhou, 2007; Steinwart and Christmann, 2008, among others).

To be more specific, an SVM is based on a *loss function* $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ which quantifies the quality of a prediction $f(x)$ by $L(x, y, f(x))$ if the observed output variable belonging to x is y. The loss function specifies the exact goal of the prediction. For example, typical loss functions for classification tasks include the *hinge loss*, the *least squares loss* and the *logistic loss*. For regression tasks, the *least squares loss* is often used to estimate the conditional mean function, whereas the *pinball loss* is suited to quantile regression.

*. Corresponding author

Based on the loss function of choice, one can define the *L-risk* (or just *risk*) $\mathcal{R}_{L,\mathrm{P}}$ as the expectation of said loss function, that is,

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

Since the loss measures the quality of a specific prediction $f(x)$, the risk quantifies the quality of the whole predictor f and we aim at finding a predictor whose risk is small. However, because the true underlying distribution P is unknown in machine learning problems, it is impossible to minimize $\mathcal{R}_{L,\mathrm{P}}$ directly. Instead, one uses an available data set consisting of observations from P to calculate the *empirical risk* as an approximation of the theoretical risk. Since minimizing the empirical risk almost certainly leads to some extent of overfitting, a regularization term has to be added, which also makes the resulting minimization problem well-posed in Hadamard's sense (cf. Hable and Christmann, 2011). An SVM $f_{L,\mathrm{P},\lambda,k}$ is then defined as the solution of the minimization problem

$$f_{L,\mathrm{P},\lambda,k} := \arg\inf_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \left|\left|f\right|\right|_H^2 . \tag{1}$$

Here, $\lambda > 0$ is a regularization parameter which controls the amount of regularization and H is the RKHS of a measurable *kernel on* $\mathcal{X}$, that is, a symmetric and positive semidefinite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (cf. Aronszajn, 1950; Schölkopf and Smola, 2002; Berlinet and Thomas-Agnan, 2004; Cucker and Zhou, 2007). We will often be interested in *bounded kernels* for which we define $\left|\left|k\right|\right|_\infty := \sup_{x \in \mathcal{X}} \sqrt{k(x,x)}$. Furthermore, we need the so-called *canonical feature map* $\Phi : X \to H$ defined by $\Phi(x) := k(\cdot, x)$ for $x \in \mathcal{X}$. This canonical feature map satisfies the *reproducing property*

$$\langle f, \Phi(x) \rangle_H = f(x) \qquad \forall\, x \in \mathcal{X}, f \in H\,, \tag{2}$$

from which one can easily deduce

$$\langle \Phi(x_1), \Phi(x_2) \rangle_H = k(x_1, x_2) \qquad \forall\, x_1, x_2 \in \mathcal{X}\,, \tag{3}$$

cf. Schölkopf and Smola (2002, Definition 2.9). Lastly, we will usually assume $\mathcal{X}$ to be a complete and separable metric space equipped with the Borel σ-algebra $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{Y}$ to be a closed subset of $\mathbb{R}$ equipped with $\mathcal{B}_{\mathcal{Y}}$, and therefore also assume that $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ with $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ denoting the set of all Borel probability measures on $\mathcal{X} \times \mathcal{Y}$. For brevity of notation, we will from now on often omit explicitly stating the σ-algebras and instead always assume sets to be equipped with their respective Borel σ-algebra when not explicitly stated otherwise.

It can be shown that suitable choices of kernel, loss function and regularization parameter (with the last one depending on the size of the given data set) lead to desirable properties of SVMs under rather mild assumptions. These include existence, uniqueness and universal consistency as well as specific learning rates, with the last one typically requiring some more conditions on P than the former properties for which (almost) no such conditions are needed. In addition to the books mentioned at the beginning of this article, which include extensive introductions to SVMs as well as many results on the aforementioned properties, some more specific results on learning rates can, for example, be found in Caponnetto and De Vito (2007); Smale and Zhou (2007); Xiang and Zhou (2009); Steinwart et al. (2009);

Eberts and Steinwart (2011, 2013); Farooq and Steinwart (2019). We refer to Christmann and Hable (2012); Christmann and Zhou (2016a) for results on SVMs for additive models and to Christmann and Zhou (2016b); Gensler and Christmann (2020) for results on kernel-based pairwise learning.

In this article, we will mainly concern ourselves with the stability of SVMs, that is, how much influence simultaneous slight changes in the probability measure P (or in the data set in the empirical case), the regularization parameter λ and the kernel k have on the resulting SVM. Because of this special interest in the effect of varying (P, λ, k), we will usually just write $f_{\mathrm{P},\lambda,k}$ instead of $f_{L,\mathrm{P},\lambda,k}$ to shorten the notation whenever L is clear from the context or does not need to be specified. Since L has to be chosen by the user depending on the problem at hand (for example, least squares loss for regression or pinball loss for quantile regression), deviations stemming from changes in the loss function are indeed desired and we are not interested in bounding them—in contrast to deviations produced by slight changes in λ or k which can, for example, stem from slight changes in the underlying data since λ and the hyperparameter(s) of k are often chosen in a data-dependent way.

There are several already existing results on different notions of classical statistical robustness of SVMs, that is, on the influence of small changes in P (or in the data set) on the predictor, cf. Bousquet and Elisseeff (2002); Christmann and Steinwart (2004, 2007); Christmann and Van Messem (2008); Hable and Christmann (2011) among others. We will, however, base many of our considerations on Christmann et al. (2018), where the authors investigated the effect of simultaneous slight changes in the whole triple (P, λ, k) instead of only in P and obtained results like

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty = \mathcal{O}\left(||\mathrm{P}_1 - \mathrm{P}_2||_{tv}\right) + \mathcal{O}\left(|\lambda_1 - \lambda_2|\right) + \mathcal{O}\left(||k_1 - k_2||_\infty\right) \qquad (4)$$

with known constants, and with $||\nu||_{tv}$ denoting the norm of total variation of a signed measure ν.

We will first modify one such result (Theorem 2.7 from Christmann et al., 2018) slightly in Section 2 in order to generalize it and make it applicable to a larger class of loss functions and to arbitrary positive regularization parameters. We will additionally derive an analogous statement concerning the $L_p(\mathrm{P}_i^X)$-norm, $i = 1, 2$, $p \in [1, \infty)$, of $f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}$ instead of its supremum norm, with Q^X denoting the marginal distribution on $\mathcal{X}$ associated with Q, for all probability measures Q on $\mathcal{X} \times \mathcal{Y}$. Afterwards, we will investigate *localized SVMs* in Section 3.

The principal idea behind localized SVMs is to not calculate one SVM on the whole input space $\mathcal{X}$ but instead split $\mathcal{X}$ into different (not necessarily disjoint) subsets, calculate SVMs on these subsets and then join them together (combined with some weight functions in the case of overlapping subsets) in order to obtain a global predictor. There are three main advantages to this approach: Firstly, the calculation of SVMs is known to have super-linear (in the number of training samples) computational requirements (in time as well as in storage space), cf. Platt (1998); Joachims (1998) among others. For large data sets, it is therefore faster and more computationally feasible to calculate several small SVMs instead of a single large one. Secondly, localization allows the algorithm to deal with different structures in different regions of the input space in different ways. For global learning approaches, it can be difficult to accurately predict a function whose complexity and volatility vary among different areas of the input space because the complexity of a predictor is usually

controlled globally by some hyperparameters. For the very same reason, large differences between the conditional distributions $\mathrm{P}(Y|X = x)$ in different areas of $\mathcal{X}$ can cause similar difficulties. A good regionalization can separate such different areas and can therefore lead to better predictions based on a given training data set. Thirdly, the use of a bounded and continuous kernel k, which is popular in practice (with, for example, the Gaussian RBF kernel satisfying both properties) and yields some useful theoretical properties, leads to all functions from the RKHS H, and thus also the SVM, being continuous and bounded as well (cf. Steinwart and Christmann, 2008, Lemma 4.28). Hence, it can be difficult for such SVMs to accurately model discontinuities in the true function, and large oscillations and overshooting can occur near these discontinuities, similar to the well-known Gibbs phenomenon occurring for Fourier series (cf. Hewitt and Hewitt, 1979, and the references cited therein). By dividing the input space into separate regions at these discontinuities, a good regionalization can eliminate this hindrance.

Hence, localized approaches—not only for SVMs but also for other machine learning methods which often face similar challenges—have been of interest for many years. For this reason, we will in the following give a short overview of literature on this topic—with a similar overview also being found in Meister and Steinwart (2016). Early theoretical investigations can be found in Bottou and Vapnik (1992); Vapnik and Bottou (1993), and nowadays various approaches for splitting the data into local subsets before applying SVMs exist. For example, Bennett and Blue (1998); Wu et al. (1999); Tibshirani and Hastie (2007) as well as Chang et al. (2010) all employ decision trees. The methods proposed in these articles however differ in how the tree is generated: In the former three, an SVM is also used for each decision in the tree, whereas Chang et al. (2010) split the data in an axis-parallel way and only apply SVMs to the final regions. That is, the goal of the former articles is only to improve the accuracy of the predictor whereas the latter one is also concerned with reducing training time. Another popular approach is to combine SVMs with k-nearest neighbor (kNN) methods. This has, for example, been done by Zhang et al. (2006); Hable (2013) and by Blanzieri and Bryl (2007); Blanzieri and Melgani (2008); Segata and Blanzieri (2010), with the former two measuring distances (for selecting the k nearest neighbors) in the input space $\mathcal{X}$ and the rest measuring them in the feature space H. Segata and Blanzieri (2010) constitutes a special case among these, since the authors modified the procedure somewhat in order to speed up the process of classifying new data points: kNN methods usually suffer from a computationally intensive and slow prediction phase due to their construction. That is, they have to calculate a new SVM on the k-neighborhood of each test point whose output they have to predict. Even though these SVMs are not based on the whole training set and thus relatively faster to train, this still significantly slows down the prediction step if many predictions have to be made. To circumvent this problem, Segata and Blanzieri (2010) proposed to train an SVM on the k-neighborhood of each point from the training set during the training phase and then use the SVM belonging to the closest training point when having to predict the output belonging to a test point. Slightly different approaches have been proposed by Cheng et al. (2007, 2010); Gu and Han (2013), where the training data is being split into clusters by some variants of k-means and then an SVM is trained on each cluster. Similarly, Rida et al. (1999) combined SVMs with density-based clustering for the case of having a multimodal input. Zakai and Ritov (2009) showed that any consistent learning method has to be localizable because it has to behave

in a local manner in order to be consistent. However, they also only investigated a special localization technique where only training samples within a ball of some fixed radius around a test point are considered when the associated output has to be predicted.

Additionally, there are several articles that provide theoretical results about localized SVMs (whereas many of the aforementioned articles focused on experimental analyses, notable exceptions being Zakai and Ritov, 2009; Gu and Han, 2013; Hable, 2013, among others) and that do often not demand special localization techniques but instead only require the resulting regionalization to satisfy some (rather mild) conditions. These include Meister and Steinwart (2016); Thomann et al. (2017); Blaschzyk (2020), where the Gaussian RBF kernel in combination with the hinge or the least squares loss function is used to derive learning rates for such localized SVMs under some assumptions regarding the underlying distribution P, which are always needed in order to obtain learning rates because of the no-free-lunch theorem, cf. Devroye (1982); Devroye et al. (1996, Section 7.2). Lastly, Dumpert and Christmann (2018); Dumpert (2020) allow even more general regionalizations as well as more general kernels and loss functions, based on which they show localized SVMs' consistency as well as their statistical robustness with respect to the maxbias and the influence function without any restrictive assumptions about P.

Other approaches for reducing the computational requirements of SVMs (or similar methods) include, for example, distributed learning (see Christmann et al., 2007; Zhang et al., 2015; Lin et al., 2017; Guo et al., 2017a; Lin et al., 2020, among others) and online learning (see Ying and Zhou, 2006; Smale and Yao, 2006; Guo et al., 2017b, among others).

We will, however, focus on a localized approach and proceed similarly to Dumpert and Christmann (2018); Dumpert (2020) in Section 3—that is, impose only very mild assumptions on regionalization, kernel, loss function and the underlying distribution—and then transfer the stability results from Section 2 to our localized SVMs. Notably, we will not make any assumptions regarding the heaviness of the tails of the conditional distributions $\mathrm{P}(Y|X = x)$, $x \in \mathcal{X}$, and especially not require $\mathcal{Y}$ to be bounded (which is, for example, needed in the aforementioned results on learning rates). Lastly, we will look at the effect of not only probability measure, regularization parameters and kernels but also the regionalization varying. That is, we will investigate the stability of localized SVMs with respect to slight changes in the whole quadruple consisting of probability measure, regularization parameters, kernels and regionalization.

As this is a theoretical investigation, we will focus on such theoretical results, and numerical experiments will be published elsewhere.

2. Total Stability of SVMs

In this section, we will show stability of SVMs with respect to slight changes in the triple (P, λ, k) consisting of probability measure, regularization parameter and kernel. Our notion of stability will be similar to that of (4), with the slight difference that we additionally need to consider $\sqrt{||k_1 - k_2||_\infty}$ as an exchange for the result considerably generalizing the referenced theorem by Christmann et al. (2018), that is, the result being applicable to arbitrary positive regularization parameters and a larger class of loss functions. Thus, it

will be of the type

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty = \mathcal{O}\left(||\mathrm{P}_1 - \mathrm{P}_2||_{tv}\right) + \mathcal{O}\left(|\lambda_1 - \lambda_2|\right) + \mathcal{O}\left(||k_1 - k_2||_\infty\right) + \mathcal{O}\left(\sqrt{||k_1 - k_2||_\infty}\right). \tag{5}$$

Afterwards, we will derive a similar stability result which bounds the $L_p(\mathrm{P}_i^X)$-norm, $i = 1, 2$, of $f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}$. This will be of the type

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)} = \mathcal{O}\left(||\mathrm{P}_1 - \mathrm{P}_2||_{tv}\right) + \mathcal{O}\left(|\lambda_1 - \lambda_2|\right) + \mathcal{O}\left(||k_1 - k_2||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)}\right) + \mathcal{O}\left(\sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)}}\right) \tag{6}$$

and will require the same mild conditions as the one concerning the supremum norm.

As mentioned in the introduction, $||\nu||_{tv}$ denotes the norm of total variation of a signed measure ν on a measurable space $(\Omega, \mathcal{A})$ (in our case usually $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X}\times\mathcal{Y}})$), that is

$$||\nu||_{tv} := |\nu|(\Omega) := \sup\left\{ \sum_{i=1}^{n} |\nu(A_i)| \,\middle|\, A_1, \ldots, A_n \text{ is a measurable partition of } \Omega \right\}.$$

This is equivalent to

$$||\nu||_{tv} = 2 \cdot \sup_{A \in \mathcal{A}} |\mathrm{P}_1(A) - \mathrm{P}_2(A)|$$

in the case $\nu = \mathrm{P}_1 - \mathrm{P}_2$ occurring in our results, which can be seen from the fact that $(\mathrm{P}_1 - \mathrm{P}_2)(\Omega) = 0$.

In the following, we will see two examples of when two distributions are similar with regards to the norm of total variation and we therefore obtain a rather small bound on the difference between $f_{\mathrm{P}_1,\lambda_1,k_1}$ and $f_{\mathrm{P}_2,\lambda_2,k_2}$ by applying our stability results:

Example 1 *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 $D_1 := ((x_1, y_1), \ldots, (x_n, y_n))$.*

If $D_2 := ((\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_n, \tilde{y}_n))$ 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) $((x_1, y_1), \ldots, (x_{n-\ell}, y_{n-\ell})) = ((\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_{n-\ell}, \tilde{y}_{n-\ell}))$, then

$$||\mathrm{D}_1 - \mathrm{D}_2||_{tv} \leq \frac{2\ell}{n}.$$

Similarly, if D_2 was obtained by adding $m \in \mathbb{N}$ new data points to D_1, such that now $D_2 := ((x_1, y_1), \ldots, (x_n, y_n), (x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m}))$, then

$$||\mathrm{D}_1 - \mathrm{D}_2||_{tv} \leq \frac{2m}{n+m}.$$

Example 2 *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* $(\mathcal{X} \times \mathcal{Y}, \mathcal{A})$. *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*

$$||\mathrm{P}_n - \mathrm{P}||_{tv} = 2 \cdot \sup_{A \in \mathcal{A}} |\mathrm{P}_1(A) - \mathrm{P}_2(A)| \leq 2 \cdot \int_{\mathcal{X} \times \mathcal{Y}} |f_n - f| \, d\mu \to 0 \,, \qquad n \to \infty \,.$$

On the other hand, the following example shows a case in which the norm of total variation is not close to zero:

Example 3 *Suppose that* P *is a continuous distribution with Lebesgue density and that* D_n *is the empirical distribution belonging to a data set* $D_n := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ *drawn from* P. *Then, we have*

$$||\mathrm{P} - \mathrm{D}_n||_{tv} = 2 \cdot |\mathrm{P}(\mathrm{supp}(\mathrm{D}_n)) - \mathrm{D}_n(\mathrm{supp}(\mathrm{D}_n))| = 2 \cdot |0 - 1| = 2$$

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,* $||\mathrm{P} - \mathrm{D}_n||_{tv}$ *does not converge to 0 as* $n \to \infty$ *in such cases.*

Remark 1 *We saw in Example 3 that our stability results, which take the form* (5) *respectively* (6), *will not yield meaningful results for comparing an empirical SVM with a theoretical SVM that is based on a continuous distribution* P *with Lebesgue density. To tackle such comparisons, one would require different quantities in the bounds than the norm of total variation, for example the bounded Lipschitz metric, but as far as we know, no such result exists. Note that this problem does not occur if* P *is a discrete distribution, which is the case in many important applications of machine learning, with text mining as a leading example.*

However, the principal goal of this paper is not such a comparison between an empirical and a theoretical SVM, and the associated investigation of consistency, anyway. Instead, we aim at deriving results regarding the stability of empirical SVMs with respect to slight changes in the data set (for example stemming from chance variation in sampling or from observing additional data) respectively the stability of theoretical SVMs with respect to slight changes in the underlying distribution. For these situations, we saw in Example 1 and Example 2 that the norm of total variation is indeed suited.

As for the other quantities occurring in (5) and (6), note that $k_1 - k_2$ generally is no kernel and therefore $||k_1 - k_2||_\infty$ in (5) denotes the general supremum norm of a function instead of the special definition of $||\cdot||_\infty$ for kernels stated before (which coincides with the square root of the general definition when applied to kernels, cf. Cucker and Zhou, 2007, p. 22).

In the following, we will give two examples in order to illustrate the behavior of this supremum norm $||k_1 - k_2||_\infty$. First of all, Example 4 compares two Gaussian kernels with different bandwidths and examines how this difference influences $||k_1 - k_2||_\infty$. As the kernel's hyperparameter(s) (in this case the bandwidth) are usually chosen in a data-dependent way, often using grid search and cross-validation, such a difference can in practice for example arise from two practitioners using slightly different grids and cross-validation schemes, as well as from them having slightly different data at hand for performing the cross-validation.

γ_1/γ_2	1	1.01	1.05	1.1	1.5	2
$\|\|k_{\gamma_1} - k_{\gamma_2}\|\|_\infty$	0.000	0.007	0.036	0.070	0.290	0.472
$g(k_{\gamma_1}, k_{\gamma_2})$	0.000	0.089	0.207	0.300	0.684	0.924

Table 1: Ratio between the bandwidths γ_1 and γ_2 of two Gaussian kernels, as well as the resulting value of $||k_{\gamma_1} - k_{\gamma_2}||_\infty$ and the according value of $g(k_{\gamma_1}, k_{\gamma_2})$ introduced in equation (7) in Example 4.

Example 4 *Let $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$ and let k_γ be the Gaussian kernel with bandwidth $\gamma > 0$, which is defined by*

$$k_\gamma(x, x') := \exp\left(-\frac{||x - x'||_2^2}{\gamma^2}\right) \qquad \forall\, x, x' \in \mathcal{X}\,.$$

It is easy to see that 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 1. Additionally, that table also includes the according values of

$$g(k_{\gamma_1}, k_{\gamma_2}) := \frac{1}{2} \cdot ||k_{\gamma_1} - k_{\gamma_2}||_\infty + \max\{||k_{\gamma_1}||_\infty, ||k_{\gamma_2}||_\infty\} \cdot \sqrt{||k_{\gamma_1} - k_{\gamma_2}||_\infty}\,, \tag{7}$$

which is the term associated with the last two summands of (5) that will (in combination with a factor depending on loss function and regularization parameters) be used in the stability result Theorem 2 in order to bound the difference between the two resulting SVMs.

Similarly to Example 4, one might also be interested in the effect on the SVM of not slightly changing the Gaussian kernel's bandwidth, but of instead for example 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 5 *Let $\mathcal{X} \subseteq \mathbb{R}^5$ (other dimensions can be analyzed analogously and yield similar results). Let k_γ be the Gaussian kernel with bandwidth $\gamma > 0$, cf. Example 4, and let k_W be the normalised Wendland kernel defined by $k_W(x, x') := \psi_{6,3}(||x - x'||_2)$ with $\psi_{6,3}$ as in Chernih et al. (2014, Theorem 3.3, with $\alpha := \gamma^{-2}$); 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 from Wendland (2005). This results in $||k_\gamma - k_W||_\infty \approx 0.0037$, and thus $g(k_\gamma, k_W) \approx 0.0631$ with g as in equation (7) from Example 4, being quite small and hence the corresponding SVMs closely resembling each other because of their stability with respect to changes in the kernel which will be shown in our results.*

We now return to the already mentioned reduced conditions on the loss function L compared to the referenced theorem by Christmann et al. (2018). We more specifically only require L to be convex as well as Lipschitz continuous. Here, convexity refers to convexity in the last argument, that is, we call a loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ *convex* if

$L(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ is convex for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Furthermore, we call L *Lipschitz continuous* if there exists a constant $c \geq 0$ such that

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

We denote the smallest such constant by $|L|_1$ and call it *Lipschitz constant* of L.

Additionally, we require the kernels used in the definition of the SVMs to be bounded, which is, for example, satisfied by the popular Gaussian kernel. We will always denote the RKHS and the canonical feature map associated with a kernel by providing them with the same indices or other additional notation, for example, $\tilde{H}_1$ and $\tilde{\Phi}_1$ denote RKHS and canonical feature map belonging to $\tilde{k}_1$.

With these conditions on the loss function L and the kernels k_1 and k_2, it would now be possible to state a stability result of the type (5). This would however require us to impose an additional condition on the probability measures P_1 and P_2 in order to guarantee both of the SVMs which are to be compared to uniquely exist. More specifically, it would be necessary that the RKHS H_i contains at least one function f with finite risk with respect to P_i (cf. Steinwart and Christmann, 2008, Lemma 5.1 and Theorem 5.2). One way to ensure this is to impose the moment condition $\mathbb{E}_{\mathrm{P}_i}[|Y|] < \infty$ on P_i (cf. Christmann et al., 2009). Alas, this moment condition excludes heavy-tailed distributions such as the Cauchy distribution. In order to circumvent this problem, we will use so-called *shifted loss functions*. These special losses have been applied in robust statistics for a long time (cf. Huber, 1967) and have been introduced to SVMs by Christmann et al. (2009). The concept of shifted loss functions is simple enough: As the name suggests, they just shift a loss function by some fixed amount. More specifically, given a loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$, the *shifted loss function* $L^\star$ is defined by

$$\begin{aligned} L^\star : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} &\to \mathbb{R}\,, \\ (x, y, t) &\mapsto L(x, y, t) - L(x, y, 0)\,. \end{aligned}$$

Risks and SVMs can be defined in the same way as for normal loss functions, that is,

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

and

$$f_{L^\star,\mathrm{P},\lambda,k} := \arg\inf_{f \in H} \mathcal{R}_{L^\star,\mathrm{P}}(f) + \lambda\,||f||_H^2\,. \tag{8}$$

We will again just write $f_{\mathrm{P},\lambda,k}$ instead of $f_{L^\star,\mathrm{P},\lambda,k}$ whenever $L^\star$ is clear from the context or does not need to be specified. It is easy to see that $L^\star$ is convex respectively Lipschitz continuous if and only if L is convex respectively Lipschitz continuous and that they both have the same Lipschitz constant, that is, $|L^\star|_1 = |L|_1$ (cf. Christmann et al., 2009, Proposition 2). Hence, the properties of L and of $L^\star$ can be used interchangeably.

Christmann et al. (2009) additionally showed that $f_{L^\star,\mathrm{P},\lambda,k} = f_{L,\mathrm{P},\lambda,k}$ holds true whenever $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, i.e., that using shifted loss functions leads to the same results as using normal loss functions and is therefore justified, and that the use of $L^\star$ eliminates the need for the moment condition whenever L is Lipschitz continuous and the kernel is bounded.

Since we required these two properties anyway, using $L^\star$ instead of L rids us of the moment condition without imposing any additional conditions. When writing $f_{\mathrm{P},\lambda,k}$, we will usually refer to $f_{L^\star,\mathrm{P},\lambda,k}$ instead of $f_{L,\mathrm{P},\lambda,k}$ because of these advantages of $L^\star$. However, since the two functions coincide whenever both exist, we could obviously also use $f_{L,\mathrm{P},\lambda,k}$ in these cases.

We will now state our first main result about the stability of SVMs. As mentioned in the introduction, this is a generalization of a result (Theorem 2.7) by Christmann et al. (2018): First of all, we eliminated an additional condition on L that was required by Christmann et al. (2018). Previously, L did not only need to be convex and Lipschitz continuous but also differentiable. Since many 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. Christmann et al., 2009, Theorem 8). In order to circumvent this 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 $||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty$, as in (4), they bound $||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{H_1}$. Since

$$||f||_\infty \leq ||k||_\infty \cdot ||f||_H \tag{9}$$

for every RKHS H and $f \in H$ (cf. Cucker and Zhou, 2007, Theorem 2.9; Steinwart and Christmann, 2008, Lemma 4.23) and 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, we neither need λ_1 and λ_2 to be greater than some positive constant nor H_1 and H_2 to be nested:

Theorem 2 *Let $\mathcal{X}$ be a complete and separable metric space and $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathrm{P}_1, \mathrm{P}_2 \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be probability measures, $\lambda_1, \lambda_2 > 0$ and k_1, k_2 be measurable and bounded kernels on $\mathcal{X}$ with separable RKHSs H_1, H_2. Denote $\kappa := \max\{||k_1||_\infty, ||k_2||_\infty\}$ and $\tau := \min\{\lambda_1, \lambda_2\}$. Let L be a convex and Lipschitz continuous loss function. Then,*

$$||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty \leq \frac{|L|_1}{\tau} \cdot \left(\kappa^2 \cdot ||\mathrm{P}_1 - \mathrm{P}_2||_{tv} + \frac{\kappa^2}{\tau} \cdot |\lambda_1 - \lambda_2| \right.$$
$$\left. + \frac{1}{2} \cdot ||k_1 - k_2||_\infty + \kappa \cdot \sqrt{||k_1 - k_2||_\infty} \right).$$

Also recall Examples 1 to 5 and Remark 1, where we looked at how the different quantities on the right hand side of the bound behave in different situations.

Remark 3 *The condition of H_1 and H_2 being separable can be difficult to check. Because of the separability of $\mathcal{X}$, this however holds true whenever k_1 and k_2 are continuous (cf. Berlinet and Thomas-Agnan, 2004, Corollary 4; Steinwart and Christmann, 2008, Lemma 4.33), and it suffices to verify this continuity instead (which is satisfied by most of the typically used kernels and is easy to check).*

Now, the subsequent Theorem 4 states a result which is very similar to that from Theorem 2 but with respect to the $L_p(\mathrm{P}_i^X)$-norm:

Theorem 4 *Let $\mathcal{X}$ be a complete and separable metric space and $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathrm{P}_1, \mathrm{P}_2 \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be probability measures, $\lambda_1, \lambda_2 > 0$ and k_1, k_2 be measurable and bounded kernels on $\mathcal{X}$ with separable RKHSs H_1, H_2. Denote $\kappa := \max\{||k_1||_\infty, ||k_2||_\infty\}$ and $\tau := \min\{\lambda_1, \lambda_2\}$. Let L be a convex and Lipschitz continuous loss function. Then, for all $p \in [1, \infty)$ and all $i \in \{1, 2\}$,*

$$
\begin{aligned}
&||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)} \\
&\leq \frac{|L|_1}{\tau} \cdot \Bigg(\kappa^2 \cdot ||\mathrm{P}_1 - \mathrm{P}_2||_{tv} + \frac{\kappa^2}{\tau} \cdot |\lambda_1 - \lambda_2| \\
&\qquad + \frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_i^X \otimes \mathrm{P}_i^X)}} \Bigg) .
\end{aligned}
$$

This result will become particularly useful in section Section 3.2 where we will investigate the stability of *localized SVMs* by examining the difference between two localized SVMs that are based on two different regionalizations of the input space $\mathcal{X}$. These different regionalizations will lead to the two localized SVMs possibly vastly differing at some points where the regionalizations do not coincide and thus no interesting bound on the supremum norm of their difference being possible. On the other hand, if the regionalizations do not differ too much, we will still be able to derive meaningful bounds on the $L_1(\mathrm{P}_i^X)$-norm of the difference between the two localized SVMs based on Theorem 4.

3. Total Stability of Localized SVMs

We now want to take look at localized SVMs and show that they inherit the stability properties from Theorems 2 and 4 under certain conditions on the regionalization method. We will take a similar approach to Dumpert and Christmann (2018); Dumpert (2020), that is, divide the input space $\mathcal{X}$ into several, possibly overlapping, regions with relatively mild assumptions about the specific regionalization. On these subspaces, we will define local SVMs which we will then combine in order to obtain a global predictor that we will call *localized SVM*.

3.1 Same Regionalization for Both Localized SVMs

First, we assume that both of the global predictors we want to compare in order to assess the stability of this localized approach are based on the same regionalization. Our stability result will not be based on any specific regionalization method but rather can be applied to any regionalization satisfying some very mild conditions. We will denote this regionalization by $\boldsymbol{\mathcal{X}_B} := \{\mathcal{X}_1, \ldots, \mathcal{X}_B\}$ for some sets $\mathcal{X}_1, \ldots, \mathcal{X}_B$ such that the following holds true:

(R1) $\mathcal{X}_1, \ldots, \mathcal{X}_B \subseteq \mathcal{X}$ measurable and $\mathcal{X} = \bigcup_{b=1}^B \mathcal{X}_b$.

Additionally, $\boldsymbol{\mathcal{X}_B}$ needs to satisfy the following condition for whichever probability measure P we will base a localized SVM on:

(R2) $\mathrm{P}(\mathcal{X}_b \times \mathcal{Y}) > 0$ for all $b \in \{1, \ldots, B\}$.

Note that **(R1)** tells us that the regions need not necessarily be pairwise disjoint but can instead also overlap. Condition **(R2)** of no region having probability mass zero is trivially needed because the local SVM on the respective region would not be defined elsewise.

Since we want to investigate stability similarly to Theorems 2 and 4, we again assume to have two possibly different Borel probability measures P_1 and P_2 on $\mathcal{X} \times \mathcal{Y}$ which the predictors we want to compare are based on. For obtaining local SVMs on $\mathcal{X}_1, \ldots, \mathcal{X}_B$ for our localized approach, we now first need to introduce the associated local probability measures on these regions (respectively on $\mathcal{X}_b \times \mathcal{Y}$, $b = 1, \ldots, B$) by restricting P_1 and P_2: On an arbitrary $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ satisfying $\mathrm{P}_i(\tilde{\mathcal{X}} \times \mathcal{Y}) > 0$, we define the local probability measure based on P_i by

$$\mathrm{P}_{i,\tilde{\mathcal{X}}} := \frac{1}{\mathrm{P}_i(\tilde{\mathcal{X}} \times \mathcal{Y})} \cdot (\mathrm{P}_i)\big|_{\tilde{\mathcal{X}} \times \mathcal{Y}}\,, \qquad i \in \{1, 2\}. \tag{10}$$

If the regionalization satisfies **(R2)** for P_1 and P_2, these local probability measures are obviously well-defined on any $\mathcal{X}_b$, $b \in \{1, \ldots, B\}$. Since we will mainly need the local probability measures on these regions, we also write $\mathrm{P}_{i,b} := \mathrm{P}_{i,\mathcal{X}_b}$ to shorten the notation, $i \in \{1, 2\}$, $b \in \{1, \ldots, B\}$.

As mentioned in Section 1, one big advantage of such localized approaches lies in their increased flexibility with regards to learning a function whose complexity and volatility vary across the input space since areas with differing complexities of the function can be separated into different regions. Hence, it should obviously be possible to choose different regularization parameters and kernels in the different regions because they to some extent control the complexity of the resulting SVM. We therefore have vectors of regularization parameters $\boldsymbol{\lambda_i} := (\lambda_{i,1}, \ldots, \lambda_{i,B})$, with $\lambda_{i,b} > 0$ for all $b \in \{1, \ldots, B\}$, and vectors of kernels $\boldsymbol{k_i} := (k_{i,1}, \ldots, k_{i,B})$, for $i = 1, 2$. Based on these and a shifted loss function $L^\star$, we obtain from (8) SVMs

$$f_{\mathrm{P}_{i,b}, \lambda_{i,b}, k_{i,b}} : \mathcal{X}_b \to \mathbb{R}\,, \qquad i \in \{1, 2\},\, b \in \{1, \ldots, B\}$$

which we call *local SVMs* on $\mathcal{X}_b$.

For combining these local SVMs and thus obtaining a global predictor on $\mathcal{X}$, we first need to extend them in a way such that they are defined on all of $\mathcal{X}$. That is, for a function $g : \tilde{\mathcal{X}} \to \mathbb{R}$ on $\tilde{\mathcal{X}} \subseteq \mathcal{X}$, we define the zero-extension $\hat{g} : \mathcal{X} \to \mathbb{R}$ by

$$\hat{g}(x) := \begin{cases} g(x) & , \text{if } x \in \tilde{\mathcal{X}}\,, \\ 0 & , \text{else}\,. \end{cases}$$

Later on we will also need zero-extensions of kernels and probability measures which we indicate by using the $\hat{(\cdot)}$-notation as well: If $k : \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \to \mathbb{R}$ is a kernel on $\tilde{\mathcal{X}} \subseteq \mathcal{X}$, we define $\hat{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$\hat{k}(x_1, x_2) := \begin{cases} k(x_1, x_2) & , \text{if } x_1, x_2 \in \tilde{\mathcal{X}}\,, \\ 0 & , \text{else}\,. \end{cases}$$

If Q is a Borel probability measure on $\tilde{\mathcal{X}} \times \mathcal{Y}$ with $\tilde{\mathcal{X}} \subseteq \mathcal{X}$, we define $\hat{\mathrm{Q}} : \mathcal{B}_{\mathcal{X}\times\mathcal{Y}} \to [0,1]$ by

$$\hat{\mathrm{Q}}(A) := \mathrm{Q}\left(A \cap \left(\tilde{\mathcal{X}} \times \mathcal{Y}\right)\right) \qquad \forall\, A \in \mathcal{B}_{\mathcal{X}\times\mathcal{Y}}\,.$$

Before finally defining our global predictors, we lastly also need weight functions w_b, $b \in \{1,\ldots,B\}$, which pointwisely control the influence of the different local SVMs in areas where two or more regions overlap. We require these weight functions to satisfy the same three conditions as in Dumpert and Christmann (2018); Dumpert (2020):

(W1) $w_b : \mathcal{X} \to [0,1]$ measurable for all $b \in \{1,\ldots,B\}$.

(W2) $\sum_{b=1}^{B} w_b(x) = 1$ for all $x \in \mathcal{X}$.

(W3) $w_b(x) = 0$ for all $x \notin \mathcal{X}_b$ and all $b \in \{1,\ldots,B\}$.

With this, global predictors $f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1}}$ and $f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2}}$, which we will call *localized SVMs* even though they are not necessarily SVMs themselves, can be defined by

$$f_{\mathrm{P}_i,\boldsymbol{\lambda_i},\boldsymbol{k_i}} : \mathcal{X} \to \mathbb{R}\,,\; x \mapsto \sum_{b=1}^{B} w_b(x) \cdot \hat{f}_{\mathrm{P}_{i,b},\lambda_{i,b},k_{i,b}}(x) \tag{11}$$

for $i = 1, 2$, where we again omitted the shifted loss function $L^\star$ from the index to shorten the notation.

The succeeding theorem states that Theorem 2 can be transferred to the situation at hand, i.e., that such localized SVMs inherit a similar stability property from regular SVMs:

Theorem 5 *Let $\mathcal{X}$ be a complete and separable metric space and $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathrm{P}_1, \mathrm{P}_2 \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ be probability measures. Let $\boldsymbol{\mathcal{X}_B} := \{\mathcal{X}_1,\ldots,\mathcal{X}_B\}$ be a regionalization of $\mathcal{X}$ such that $\boldsymbol{\mathcal{X}_B}$ satisfies **(R1)** and, for P_1 as well as for P_2, **(R2)**. For all $i \in \{1,2\}$ and $b \in \{1,\ldots,B\}$, let $\lambda_{i,b} > 0$ and let $k_{i,b}$ be a bounded and measurable kernel on $\mathcal{X}_b$ with separable RKHS $H_{i,b}$. Denote $\kappa_b := \max\{||k_{1,b}||_\infty, ||k_{2,b}||_\infty\}$ and $\tau_b := \min\{\lambda_{1,b}, \lambda_{2,b}\}$ for all $b \in \{1,\ldots,B\}$. Let L be a convex and Lipschitz continuous loss function. Let $f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1}}$ and $f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2}}$ be defined as in (11) with the weight functions $w_1,\ldots,w_B$ satisfying **(W1)**, **(W2)** and **(W3)**. Then,*

$$\begin{aligned}
&||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2}}||_\infty \\
&\leq |L|_1 \cdot \max_{b\in\{1,\ldots,B\}} \frac{1}{\tau_b} \cdot \Bigg(\kappa_b^2 \cdot ||\mathrm{P}_{1,b} - \mathrm{P}_{2,b}||_{tv} + \frac{\kappa_b^2}{\tau_b} \cdot |\lambda_{1,b} - \lambda_{2,b}| \\
&\qquad\qquad + \frac{1}{2} \cdot ||k_{1,b} - k_{2,b}||_\infty + \kappa_b \cdot \sqrt{||k_{1,b} - k_{2,b}||_\infty}\Bigg)\,.
\end{aligned}$$

Similarly, we can also transfer Theorem 4 in order to bound the $L_p(\mathrm{P}_i^X)$-norm of the difference:

Theorem 6 *Let $\mathcal{X}$ be a complete and separable metric space and $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathrm{P}_1, \mathrm{P}_2 \in \mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ be probability measures. Let $\boldsymbol{\mathcal{X}_B} := \{\mathcal{X}_1,\ldots,\mathcal{X}_B\}$ be a regionalization of $\mathcal{X}$ such that $\boldsymbol{\mathcal{X}_B}$ satisfies **(R1)** and, for P_1 as well as for P_2, **(R2)**. For all $i \in \{1,2\}$*

*and $b \in \{1, \ldots, B\}$, let $\lambda_{i,b} > 0$ and let $k_{i,b}$ be a bounded and measurable kernel on $\mathcal{X}_b$ with separable RKHS $H_{i,b}$. Denote $\kappa_b := \max\{||k_{1,b}||_\infty, ||k_{2,b}||_\infty\}$ and $\tau_b := \min\{\lambda_{1,b}, \lambda_{2,b}\}$ for all $b \in \{1, \ldots, B\}$. Let L be a convex and Lipschitz continuous loss function. Let $f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1}}$ and $f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2}}$ be defined as in (11) with the weight functions $w_1, \ldots, w_B$ satisfying **(W1)**, **(W2)** and **(W3)**. Then, for all $p \in [1, \infty)$ and all $i \in \{1, 2\}$,*

$$
\begin{aligned}
&||f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2}}||_{L_p(\mathrm{P}_i^X)} \\
&\leq |L|_1 \cdot \sum_{b=1}^{B} \left(\mathrm{P}_i^X(\mathcal{X}_b)\right)^{1/p} \cdot \left(\frac{\kappa_b^2}{\tau_b} \cdot ||\mathrm{P}_{1,b} - \mathrm{P}_{2,b}||_{tv} + \frac{\kappa_b^2}{\tau_b^2} \cdot |\lambda_{1,b} - \lambda_{2,b}| \right. \\
&\qquad\qquad + \frac{1}{2\tau_b} \cdot ||k_{1,b} - k_{2,b}||_{L_p(\mathrm{P}_{i,b}^X \otimes \mathrm{P}_{i,b}^X)} \\
&\qquad\qquad \left. + \frac{\kappa_b}{\tau_b} \cdot \sqrt{||k_{1,b} - k_{2,b}||_{L_p(\mathrm{P}_{i,b}^X \otimes \mathrm{P}_{i,b}^X)}} \right).
\end{aligned}
$$

Remark 7 *Theorem 5 does actually not need **(W3)** and in Theorem 6 we can even waive **(W2)** as well as **(W3)**. We still included them in the assumptions of the two theorems since we think that weight functions should usually satisfy these conditions.*

As can be seen from the two preceding theorems, localizing the SVMs does not ruin their stability with respect to probability measure, regularization parameters and kernel: We can still bound the difference between two localized SVMs—with respect to its supremum or an L_p-norm—by a term that converges to zero whenever the norm of total variation between the two probability measures, the difference between the two regularization parameters and the supremum norm respectively the L_p-norm of the difference between the two kernels all converge to zero on all regions of the regionalization that is used.

3.2 Different Regionalizations for the Localized SVMs

In the previous section, we investigated stability of localized SVMs with respect to changes in the triple $(\mathrm{P}, \boldsymbol{\lambda}, \boldsymbol{k})$ but not in the regionalization. However, when there are changes in the distribution P used for calculating the localized SVM (for example, because of changes in the training data set in practice), it may very well happen that this also affects the regionalization if it is not predetermined but also based on a learning method (for example, decision trees, cf. Bennett and Blue, 1998; Wu et al., 1999; Tibshirani and Hastie, 2007; Chang et al., 2010, among others). Hence, we will take a closer look at the effect of slight changes in the regionalization (in addition to those in $(\mathrm{P}, \boldsymbol{\lambda}, \boldsymbol{k})$) on the resulting localized SVM in this section.

First of all, it has to be mentioned that we will sadly not be able to derive a meaningful result regarding the supremum norm of the difference of two such localized SVMs (like Theorem 5 in the case of coinciding regionalizations), which can readily be seen from the simple example visualized in Figure 1. There, 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

$$
X \sim \mathcal{U}(-1, 1)\,, \qquad Y|X \sim \mathrm{sign}(X) + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}(0, 0.5)\,,
$$

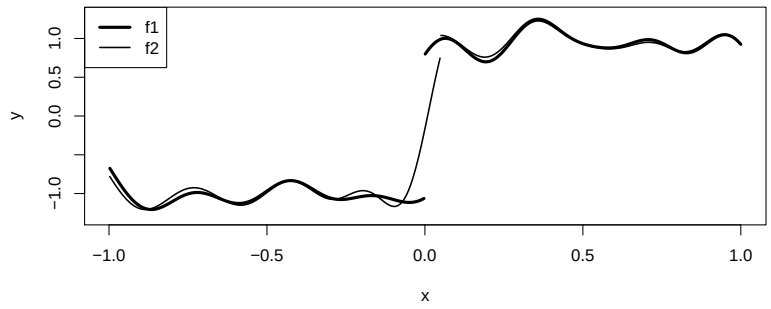

Figure 1: Comparison of two localized SVMs based on the same distribution, regularization parameters and kernels, but on slightly different regionalizations.

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 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 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 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 Theorem 6. This will indeed be the case, but before stating the corresponding theorem, we first need to modify some of the notation introduced in Section 3.1 such that it fits this new situation:

First of all, we have two different regionalizations $\boldsymbol{\mathcal{X}}_{\boldsymbol{A_1}}^{\boldsymbol{(1)}} := \{\mathcal{X}_1^{(1)}, \ldots, \mathcal{X}_{A_1}^{(1)}\}$ and $\boldsymbol{\mathcal{X}}_{\boldsymbol{A_2}}^{\boldsymbol{(2)}} := \{\mathcal{X}_1^{(2)}, \ldots, \mathcal{X}_{A_2}^{(2)}\}$ now. Contrary to Section 3.1, these regionalizations are now required to actually be partitions of $\mathcal{X}$, that is, to satisfy, for $i = 1, 2$, the following modified version of condition **(R1)**:

(R1') $\mathcal{X}_1^{(i)}, \ldots, \mathcal{X}_{A_i}^{(i)} \subseteq \mathcal{X}$ measurable and $\mathcal{X} = \bigcup_{a=1}^{A_i} \mathcal{X}_a^{(i)}$.

Furthermore, we also need to alter **(R2)** slightly since our proofs in this section will use auxiliary SVMs defined on all possible intersections of sets from $\boldsymbol{\mathcal{X}}_{\boldsymbol{A_1}}^{\boldsymbol{(1)}}$ with sets from $\boldsymbol{\mathcal{X}}_{\boldsymbol{A_2}}^{\boldsymbol{(2)}}$, which is why all these intersections need to have positive probability with respect to any probability measure P which a localized SVM will be based on:

(R2') $\mathrm{P}(\mathcal{X}_b^* \times \mathcal{Y}) > 0$ for all $\mathcal{X}_b^* \in \boldsymbol{\mathcal{X}}_{\boldsymbol{B}}^{\boldsymbol{*}}$, where

$$\begin{aligned}\boldsymbol{\mathcal{X}}_{\boldsymbol{B}}^{\boldsymbol{*}} &:= \{\mathcal{X}_1^*, \ldots, \mathcal{X}_B^*\} \\ &:= \left\{ \mathcal{X}^* \subseteq \mathcal{X} \;\middle|\; \exists\, \mathcal{X}_{a_1}^{(1)} \in \boldsymbol{\mathcal{X}}_{\boldsymbol{A_1}}^{\boldsymbol{(1)}}, \mathcal{X}_{a_2}^{(2)} \in \boldsymbol{\mathcal{X}}_{\boldsymbol{A_2}}^{\boldsymbol{(2)}} : \mathcal{X}^* = \mathcal{X}_{a_1}^{(1)} \cap \mathcal{X}_{a_2}^{(2)} \right\} \setminus \{\emptyset\}.\end{aligned}$$

Based on $\boldsymbol{\mathcal{X}_B^*}$ from **(R2')**, we denote $J_{i,a} := \{b \in \{1,\dots,B\} \,|\, \mathcal{X}_b^* \subseteq \mathcal{X}_a^{(i)}\} \neq \emptyset$ for $i = 1,2$ and $a = 1,\dots,A_i$. Additionally, for $i = 1,2$ and $b = 1,\dots,B$, we denote by $a(i,b)$ that index $a \in \{1,\dots,A_i\}$ such that $\mathcal{X}_b^* \subseteq \mathcal{X}_a^{(i)}$, which is well-defined because of **(R2')** (existence of such an index) and **(R1')** (uniqueness of that index).

Local probability measures can be defined as before, cf. (10), and we will again shorten the notation for the ones we will mainly use: $\mathrm{P}_{i,a} := \mathrm{P}_{i,\mathcal{X}_a^{(i)}}$, $i = 1,2$, $a = 1,\dots,A_i$. We then obtain local SVMs

$$f_{\mathrm{P}_{i,a},\lambda_{i,a},k_{i,a}} : \mathcal{X}_a^{(i)} \to \mathbb{R}\,, \qquad i \in \{1,2\},\, a \in \{1,\dots,A_i\}$$

based on regularization parameters $\lambda_{i,a} > 0$ and kernels $k_{i,a}$.

For combining these local SVMs in order to obtain the global predictors, no weight functions are needed this time, since the regions from the regionalizations are not allowed to overlap, cf. **(R1')**. Thus, the localized SVMs are now readily defined as

$$f_{\mathrm{P}_i,\boldsymbol{\lambda_i},\boldsymbol{k_i},\boldsymbol{\mathcal{X}_{A_i}^{(i)}}} : \mathcal{X} \to \mathbb{R}\,,\, x \mapsto \sum_{a=1}^{A_i} \hat{f}_{\mathrm{P}_{i,a},\lambda_{i,a},k_{i,a}}(x) \tag{12}$$

for $i = 1,2$, where $\boldsymbol{\lambda_i} := (\lambda_{i,1},\dots,\lambda_{i,A_i})$ and $\boldsymbol{k_i} := (k_{i,1},\dots,k_{i,A_i})$, again omitting the shifted loss function $L^\star$ from the index to shorten the notation.

As before, we will investigate stability by bounding the norm of the difference of two estimators, in this case of two localized SVMs, based on how much the underlying probability measures and the vectors of regularization parameters and kernels differ. Additionally, the regionalizations are now added as a fourth possible difference. In order to base our new bound on the difference between the two underlying regionalizations as well, this difference first has to be quantified somehow. We will base this quantification on the intersections between regions from the different regionalizations, that is, on the sets $\mathcal{X}_1^*,\dots,\mathcal{X}_B^*$ from $\boldsymbol{\mathcal{X}_B^*}$ introduced in **(R2')**. For each such intersection $\mathcal{X}_b^*$, we look at two properties which both in some sense characterize the difference between the two regionalizations: Firstly, it has to be considered how much the two intersecting regions producing $\mathcal{X}_b^*$ differ in size—relatively to these regions' own size and with respect to some probability measure Q on $\mathcal{X}$ satisfying $\mathrm{Q}(\mathcal{X}_a^{(i)}) > 0$ for all $i \in \{1,2\}$ and $a \in \{1,\dots,A_i\}$. That is, we have to include the term

$$\frac{\left|\mathrm{Q}(\mathcal{X}_{a(1,b)}^{(1)}) - \mathrm{Q}(\mathcal{X}_{a(2,b)}^{(2)})\right|}{\max\left\{\mathrm{Q}(\mathcal{X}_{a(1,b)}^{(1)}), \mathrm{Q}(\mathcal{X}_{a(2,b)}^{(2)})\right\}} \tag{13}$$

for each such intersection, that is, for each $b \in \{1,\dots,B\}$. This difference in size has to be accounted for since a large difference could possibly lead to the local SVM on 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 regions' intersection $\mathcal{X}_b^*$. Secondly, we also have to consider how closely each region from $\boldsymbol{\mathcal{X}_{A_1}^{(1)}}$ as well as from $\boldsymbol{\mathcal{X}_{A_2}^{(2)}}$ coincides with one of the intersections $\mathcal{X}_1^*,\dots,\mathcal{X}_B^*$, that is, how well each such region can be identified with a region from the other regionalization, again with respect to some probability measure Q on $\mathcal{X}$. In order to control this property, we will include the term

$$\mathrm{Q}_{\mathcal{X}_{a(i,b)}^{(i)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{Q}_{\mathcal{X}_{a(i,b)}^{(i)}}(\mathcal{X}_b^*)\right) \tag{14}$$

for each $b \in \{1, \ldots, B\}$ and $i \in \{1, 2\}$. If a region $\mathcal{X}_a^{(i)}$ closely coincides with $\mathcal{X}_{b_0}^*$ for some $b_0 \in J_{i,a}$, then $\mathrm{Q}_{\mathcal{X}_a^{(i)}}(\mathcal{X}_b^*)$ will be either close to 1 or close to 0, and (14) will hence be small, for each $b \in J_{i,a}$. If this is the case for all $i \in \{1, 2\}$ and $a \in \{1, \ldots, A_i\}$, each region can be identified well with a region from the other regionalization and the two regionalizations are therefore similar to each other in the sense of this second criterion.

We now combine these two criteria in order to obtain a quantity to measure the difference between $\boldsymbol{\mathcal{X}}_{\boldsymbol{A_1}}^{\boldsymbol{(1)}}$ and $\boldsymbol{\mathcal{X}}_{\boldsymbol{A_2}}^{\boldsymbol{(2)}}$. Since we analyzed the two criteria intersection-wise, we also define the quantification of this difference intersection-wise, that is,

$$\begin{aligned} \mathrm{d}_{\mathrm{Q},b}(\boldsymbol{\mathcal{X}}_{\boldsymbol{A_1}}^{\boldsymbol{(1)}}, \boldsymbol{\mathcal{X}}_{\boldsymbol{A_2}}^{\boldsymbol{(2)}}) := & \frac{\left|\mathrm{Q}(\mathcal{X}_{a(1,b)}^{(1)}) - \mathrm{Q}(\mathcal{X}_{a(2,b)}^{(2)})\right|}{\max\left\{\mathrm{Q}(\mathcal{X}_{a(1,b)}^{(1)}), \mathrm{Q}(\mathcal{X}_{a(2,b)}^{(2)})\right\}} \\ & + \sum_{i=1}^{2} \left(\frac{1}{2} \cdot \mathrm{Q}_{\mathcal{X}_{a(i,b)}^{(i)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{Q}_{\mathcal{X}_{a(i,b)}^{(i)}}(\mathcal{X}_b^*)\right) \right. \\ & \qquad \left. + \sqrt{\mathrm{Q}_{\mathcal{X}_{a(i,b)}^{(i)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{Q}_{\mathcal{X}_{a(i,b)}^{(i)}}(\mathcal{X}_b^*)\right)} \right), \end{aligned} \tag{15}$$

for all $b \in \{1, \ldots, B\}$ and for Q being a probability measure on $\mathcal{X}$ satisfying $\mathrm{Q}(\mathcal{X}_a^{(i)}) > 0$ for all $i \in \{1, 2\}$ and $a \in \{1, \ldots, A_i\}$. We additionally included the square root of the criterion from (14) since it will arise in the proof of the subsequent Theorem 8 that this square root is also relevant to the difference we want to investigate.

Finally, we need to introduce some new notation which will arise in the succeeding theorem because of the already mentioned auxiliary SVMs on the sets $\mathcal{X}_1^*, \ldots, \mathcal{X}_B^*$ from **(R2')** that are needed for proving the theorem: We will denote the auxiliary distributions and kernels on the sets $\mathcal{X}_1^*, \ldots, \mathcal{X}_B^*$ by using the $(\cdot)^*$-notation. That is, $\mathrm{P}_{i,b}^* := \mathrm{P}_{i,\mathcal{X}_b^*}$ and $k_{i,b}^* := k_{i,a(i,b)}\big|_{\mathcal{X}_b^* \times \mathcal{X}_b^*}$ for $i = 1, 2$ and $b = 1, \ldots, B$. By Berlinet and Thomas-Agnan (2004, Theorem 6), $k_{i,b}^*$ is actually a kernel (on $\mathcal{X}_b^*$) again.

With this, we can now state our stability result. As seen before, we will not be able to derive meaningful results regarding the supremum norm of the difference of two localized SVMs based on different regionalizations. However, it is indeed possible to obtain a meaningful bound on the $L_1(\mathrm{P}_i^X)$-norm of this difference:

Theorem 8 *Let $\mathcal{X}$ be a complete and separable metric space and $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathrm{P}_1, \mathrm{P}_2 \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be probability measures. Let $\boldsymbol{\mathcal{X}}_{\boldsymbol{A_i}}^{\boldsymbol{(i)}} := \{\mathcal{X}_1^{(i)}, \ldots, \mathcal{X}_{A_i}^{(i)}\}$, $i = 1, 2$, be regionalizations of $\mathcal{X}$ such that $\boldsymbol{\mathcal{X}}_{\boldsymbol{A_1}}^{\boldsymbol{(1)}}$ and $\boldsymbol{\mathcal{X}}_{\boldsymbol{A_2}}^{\boldsymbol{(2)}}$ both satisfy **(R1')** and that they together satisfy, for P_1 as well as for P_2, **(R2')**. For all $i \in \{1, 2\}$ and $a \in \{1, \ldots, A_i\}$, let $\lambda_{i,a} > 0$ and let $k_{i,a}$ be a bounded and measurable kernel on $\mathcal{X}_a^{(i)}$ with separable RKHS $H_{i,a}$. Denote $\kappa_b := \max\{\left|\left|k_{1,a(1,b)}\right|\right|_\infty, \left|\left|k_{2,a(2,b)}\right|\right|_\infty\}$, $\tau_b := \min\{\lambda_{1,a(1,b)}, \lambda_{2,a(2,b)}\}$ and $\rho_{1,b} := \max\left\{\mathrm{P}_1^X(\mathcal{X}_{a(1,b)}^{(1)}), \mathrm{P}_1^X(\mathcal{X}_{a(2,b)}^{(2)})\right\}$ for all $b \in \{1, \ldots, B\}$. Let L be a convex and Lipschitz continuous loss function. Let $f_{\mathrm{P}_1, \boldsymbol{\lambda_1}, \boldsymbol{k_1}, \boldsymbol{\mathcal{X}}_{\boldsymbol{A_1}}^{\boldsymbol{(1)}}}$ and $f_{\mathrm{P}_2, \boldsymbol{\lambda_2}, \boldsymbol{k_2}, \boldsymbol{\mathcal{X}}_{\boldsymbol{A_2}}^{\boldsymbol{(2)}}}$ be defined as in (12).*

Then,

$$
\begin{aligned}
&\left|\left| f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}^{(1)}_{\boldsymbol{A_1}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}^{(2)}_{\boldsymbol{A_2}}} \right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{a=1}^{A_2} \mathrm{P}_1^X(\mathcal{X}_a^{(2)}) \cdot \frac{||k_{2,a}||_\infty^2}{\lambda_{2,a}} \cdot \left|\left| \mathrm{P}_{1,\mathcal{X}_a^{(2)}} - \mathrm{P}_{2,\mathcal{X}_a^{(2)}} \right|\right|_{tv} \\
&\quad + |L|_1 \cdot \sum_{b=1}^{B} \Bigg(\rho_{1,b} \cdot \frac{\kappa_b^2}{\tau_b^2} \cdot \left|\lambda_{1,a(1,b)} - \lambda_{2,a(2,b)}\right| \\
&\qquad\qquad + \mathrm{P}_1^X(\mathcal{X}_b^*) \cdot \Bigg(\frac{1}{2\tau_b} \cdot \left|\left| k_{1,b}^* - k_{2,b}^* \right|\right|_{L_1((\mathrm{P}_{1,b}^*)^X \otimes (\mathrm{P}_{1,b}^*)^X)} \\
&\qquad\qquad\qquad + \frac{\kappa_b}{\tau_b} \cdot \sqrt{\left|\left| k_{1,b}^* - k_{2,b}^* \right|\right|_{L_1((\mathrm{P}_{1,b}^*)^X \otimes (\mathrm{P}_{1,b}^*)^X)}} \Bigg) \\
&\qquad\qquad + \rho_{1,b} \cdot \frac{\kappa_b^2}{\tau_b} \cdot \mathrm{d}_{\mathrm{P}_1^X,b}(\boldsymbol{\mathcal{X}}^{(1)}_{\boldsymbol{A_1}}, \boldsymbol{\mathcal{X}}^{(2)}_{\boldsymbol{A_2}}) \Bigg) .
\end{aligned}
$$

Note that the denominator occurring in $\mathrm{d}_{\mathrm{P}_1^X,b}(\boldsymbol{\mathcal{X}}^{(1)}_{\boldsymbol{A_1}}, \boldsymbol{\mathcal{X}}^{(2)}_{\boldsymbol{A_2}})$ is greater than zero for all $b \in \{1, \ldots, B\}$ in the situation of this theorem because of $\boldsymbol{\mathcal{X}}^{(1)}_{\boldsymbol{A_1}}$ and $\boldsymbol{\mathcal{X}}^{(2)}_{\boldsymbol{A_2}}$ being assumed to satisfy **(R2')** for P_1, and the bound from the theorem is therefore well-defined.

By interchanging the roles of $f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}^{(1)}_{\boldsymbol{A_1}}}$ and $f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}^{(2)}_{\boldsymbol{A_2}}}$, it is furthermore obvious that Theorem 8 also holds true with respect to the $L_1(\mathrm{P}_2^X)$-norm if the indices on the right hand side are adjusted accordingly. For the sake of notational clarity, we did not explicitly include this in the theorem.

Even though allowing for differing regionalizations makes this result on total stability look more complicated than those from Section 3.1 on first glance, the statement basically stays the same—with the main difference being that we can only bound the L_1-norm in a meaningful way, but not other L_p-norms or the supremum norm (for other L_p-norms, it would be possible to derive a similar result, but in this case, the factor in front of the difference between the regularization parameters could increase with increasing similarity of the two regionalizations and it would therefore not necessarily be possible to interpret the result as yielding total stability, i.e., particularly stability with respect to simultaneous slight changes in regularization parameters and regionalization). Here, we additionally need to consider the difference between the two regionalizations, but otherwise have the same statement as before: The L_1-norm of the difference between the two localized SVMs converges to zero if, on all regions respectively intersections of regions, the norm of total variation of the difference between the two probability measures, the difference between the two regularization parameters, the L_1-norm of the difference between the two kernels, and now additionally the difference between the two regionalizations, as measured by $\mathrm{d}_{\mathrm{P}_i^X,b}$, all converge to zero as well.

4. Discussion

This paper is composed of two main parts. In the first one, stability of SVMs with respect to slight changes in the full triple (P, λ, k), consisting of a probability measure P, a regularization parameter λ and a kernel k, was investigated. This part is related to Christmann et al. (2018), where the difference $||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty$ between two such SVMs based on slightly differing triples $(\mathrm{P}_i, \lambda_i, k_i)$, $i = 1, 2$, had already been bounded in a very similar way. We succeeded in considerably generalizing the referenced result by Christmann et al. (2018), such that we now know that the investigated notion of stability holds true for any SVM that uses a convex and Lipschitz continuous loss function. We also derived an analogous stability result regarding $||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)}$, $i = 1, 2$, before turning our attention to the second part of this paper.

Here, we investigated localized SVMs which, amongst other advantages, thrive on their reduced computational requirements compared to calculating a global SVM. They share this advantage with other methods mentioned in the introduction, like for example distributed learning. Distributed learning is similar to localized approaches in that both divide the training data into several subsets and then produce a global predictor by combining the predictors obtained on the subsets. However, whereas the subsets constitute subregions of the input space in localized approaches, they are usually generated by drawing simple random samples (without replacement) from the original data set and thus typically cover almost the entire input space in distributed learning (that is, each of the predictors on the subsets is defined on all of $\mathcal{X}$ and they are then typically combined by means of some weighted average). This leads to distributed learning reducing the computation time even further on the one hand (since the effort of regionalizing can be omitted) but not sharing the additional advantages of localized approaches (the ability to treat different structures in different regions of the input space in different ways and the ability to better model discontinuities in the true function) on the other hand, which is one reason why we think that localized learning can be interesting.

We managed to transfer our stability results to localized SVMs, adding to the list of properties such localized SVMs inherit from the local SVMs they are based on. This further substantiates the theoretical justification for localized SVMs to be used in order to accurately predict a function whose complexity varies across the input space or whenever a large data set drastically increases the computation time of a global SVM. It has even been possible to show stability with respect to the $L_1(\mathrm{P}_i^X)$-norm, $i = 1, 2$, if not only the triple $(\mathrm{P}, \boldsymbol{\lambda}, \boldsymbol{k})$ but also the regionalization slightly changes. Since variations in the underlying probability measure (respectively data set)—which are also of interest in considerations regarding classical statistical robustness, where only stability with respect to the probability measure is regarded—may very well lead to changes in the regionalization, it is especially reassuring to see that this does not ruin the localized SVMs' stability (as long as a statistically robust method is used for constructing the regionalization, such that small changes in P only lead to small changes in the regionalization).

Based on this influence of the probability measure on the regionalization, it might be interesting to take another look at the already existing results about localized SVMs' consistency, learning rates and classical statistical robustness and examine whether they still hold true if this influence is factored in, respectively what assumptions about the dependence

between probability measure and regionalization are necessary in order for the results to still hold true.

Appendix A. Auxiliary Results Regarding the Stability of SVMs

In order to prove Theorem 2, we will first apply the triangle inequality in order to decompose the difference which we have to bound:

$$\begin{aligned}||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty \leq\ & ||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_1,k_1}||_\infty + ||f_{\mathrm{P}_2,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_1}||_\infty \\ & + ||f_{\mathrm{P}_2,\lambda_2,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty \,. \end{aligned} \tag{16}$$

Of course, the order of this decomposition can also be varied. We will take this into account when actually proving Theorem 2 in Appendix B.1, but for now we will just investigate the three summands on the right hand side of (16) separately. The $L_p(\mathrm{P}_i^X)$-norm of $f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}$ can obviously be decomposed in the same way as

$$\begin{aligned}||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)} \leq\ & ||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_1,k_1}||_{L_p(\mathrm{P}_i^X)} + ||f_{\mathrm{P}_2,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_1}||_{L_p(\mathrm{P}_i^X)} \\ & + ||f_{\mathrm{P}_2,\lambda_2,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)} \end{aligned} \tag{17}$$

and thus, Theorem 4 can also be proven by examining the three summands separately.

Before doing this, first for the supremum norm and then for the $L_p(\mathrm{P}_i^X)$-norm, we have to state two auxiliary results needed for conducting the proofs. Firstly, we recall a representer theorem for SVMs (Christmann et al., 2009, Theorem 7) and prior to that the definition of a *subdifferential* (cf. Phelps, 1993; Christmann et al., 2009), which is referenced in the representer theorem:

Definition 9 *Let E be a Banach space and let $f : E \to \mathbb{R} \cup \{\infty\}$ be a convex function, and $w \in E$ with $f(w) < \infty$. Then, the subdifferential of f at w is defined by*

$$\partial f(w) := \{w' \in E' : \langle w', v - w\rangle \leq f(v) - f(w) \text{ for all } v \in E\}\,.$$

Remark 10 *For a convex loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ we denote by $\partial L(x, y, t_0)$ the subdifferential with respect to the third argument, that is, the subdifferential of the convex function defined by $t \mapsto L(x, y, t)$ at the point $t_0 \in \mathbb{R}$. We say that a function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is from the subdifferential of L with respect to a function $f : \mathcal{X} \to \mathbb{R}$ if $g(x, y) \in \partial L(x, y, f(x))$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We use an analogous notation for shifted loss functions $L^\star$.*

Theorem 11 (Christmann et al., 2009). *Let $\mathcal{X}$ be a complete and separable metric space and $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be a probability measure. Let L be a convex and Lipschitz continuous loss function, k be a bounded and measurable kernel on $\mathcal{X}$ with separable RKHS H. Then, for all $\lambda > 0$, there exists an $h \in L_\infty(\mathrm{P})$ such that*

$$\begin{aligned} h(x, y) &\in \partial L^\star(x, y, f_{L^\star,\mathrm{P},\lambda,k}(x)) \qquad \forall\, (x, y) \in \mathcal{X} \times \mathcal{Y} \\ f_{L^\star,\mathrm{P},\lambda,k} &= -\frac{1}{2\lambda}\mathbb{E}_\mathrm{P}\left[h\Phi\right] \\ ||h||_\infty &\leq |L|_1 \\ \left|\left|f_{L^\star,\mathrm{P},\lambda,k} - f_{L^\star,\bar{\mathrm{P}},\lambda,k}\right|\right|_H &\leq \frac{1}{\lambda}\left|\left|\mathbb{E}_\mathrm{P}\left[h\Phi\right] - \mathbb{E}_{\bar{\mathrm{P}}}\left[h\Phi\right]\right|\right|_H \end{aligned}$$

for all distributions $\bar{\mathrm{P}}$ *on* $\mathcal{X} \times \mathcal{Y}$.

Since the feature map Φ is H-valued, we need to consider H-valued *Bochner integrals* when examining the expectations from Theorem 11 and similar integrals. For a detailed introduction to Bochner integrals, see Diestel and Uhl (1977); Diestel (1984); Denkowski et al. (2003). We will additionally need the two succeeding inequalities (18) and (19) in order to bound the norm of a Bochner integral. Even though we suppose that these two inequalities are already established, we did not find them in the literature, which is why we prove them here:

Lemma 12 *Let* Q *be a probability measure on some measurable space* $(\Omega, \mathcal{A})$ *and let* k *be a bounded kernel on* Ω *with RKHS* H*. Let* $g : \Omega \to H$ *be a* Q*-Bochner integrable function. Then,*

$$\left|\left|\int_\Omega g(x)\, d\mathrm{Q}(x)\right|\right|_\infty \leq \int_\Omega ||g(x)||_\infty \, d\mathrm{Q}(x) \tag{18}$$

and, for all $p \in [1, \infty)$*,*

$$\left|\left|\int_\Omega g(x)\, d\mathrm{Q}(x)\right|\right|_{L_p(\mathrm{Q})} \leq \int_\Omega ||g(x)||_{L_p(\mathrm{Q})} \, d\mathrm{Q}(x)\,. \tag{19}$$

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 $s_n : \Omega \to H, \omega \mapsto \sum_{j=1}^{m_n} b_j^{(n)} \mathbb{1}_{A_j^{(n)}}(\omega)$, 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

$$\lim_{n\to\infty} \int_\Omega ||g(\omega) - s_n(\omega)||_H \, d\mathrm{Q}(\omega) = 0\,. \tag{20}$$

Then, the same definition tells us that

$$\int_\Omega g(\omega)\, d\mathrm{Q}(\omega) := \lim_{n\to\infty} \int_\Omega s_n(\omega)\, d\mathrm{Q}(\omega)\,,$$

where

$$\int_\Omega s_n(\omega)\, d\mathrm{Q}(\omega) := \sum_{j=1}^{m_n} b_j^{(n)} \mathrm{Q}\left(A_j^{(n)}\right)$$

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 $||\cdot||_\infty$ or $||\cdot||_{L_p(\mathrm{Q})}$. Then,

$$\begin{aligned}
&\left|\left|\int g(\omega)\, d\mathrm{Q}(\omega)\right|\right|_\bullet = \left|\left|\lim_{n\to\infty}\left(\sum_{j=1}^{m_n} b_j^{(n)} \mathrm{Q}\left(A_j^{(n)}\right)\right)\right|\right|_\bullet = \lim_{n\to\infty}\left|\left|\sum_{j=1}^{m_n} b_j^{(n)} \mathrm{Q}\left(A_j^{(n)}\right)\right|\right|_\bullet \\
&\leq \lim_{n\to\infty}\left(\sum_{j=1}^{m_n}\left|\left|b_j^{(n)}\right|\right|_\bullet \mathrm{Q}\left(A_j^{(n)}\right)\right) = \lim_{n\to\infty}\left(\sum_{j=1}^{m_n}\int \left|\left|b_j^{(n)}\right|\right|_\bullet \mathbb{1}_{A_j^{(n)}}(\omega)\, d\mathrm{Q}(\omega)\right) \\
&= \lim_{n\to\infty}\left(\int \sum_{j=1}^{m_n}\left|\left|b_j^{(n)}\right|\right|_\bullet \mathbb{1}_{A_j^{(n)}}(\omega)\, d\mathrm{Q}(\omega)\right) = \lim_{n\to\infty}\left(\int \left|\left|\sum_{j=1}^{m_n} b_j^{(n)} \mathbb{1}_{A_j^{(n)}}(\omega)\right|\right|_\bullet d\mathrm{Q}(\omega)\right) \\
&= \lim_{n\to\infty}\left(\int ||s_n(\omega)||_\bullet \, d\mathrm{Q}(\omega)\right) = \int ||g(\omega)||_\bullet \, d\mathrm{Q}(\omega)\,, \qquad (21)
\end{aligned}$$

where we applied the continuity of $||\cdot||_\bullet$ as a function on H in the second step and the pairwise disjointness of $A_1^{(n)}, \ldots, A_{m_n}^{(n)}$ in the second to last row, with the continuity of $||\cdot||_\bullet$ holding true because of $||h||_{L_p(\mathrm{Q})} \leq ||h||_\infty \leq ||k||_\infty ||h||_H$ and thus

$$||h||_\bullet \leq ||k||_\infty ||h||_H \qquad (22)$$

for all $h \in H$, cf. (9). Additionally, the equality in the last step of (21) holds true because of

$$\begin{aligned}
&\left|\int ||g(\omega)||_\bullet \, d\mathrm{Q}(\omega) - \lim_{n\to\infty}\left(\int ||s_n(\omega)||_\bullet \, d\mathrm{Q}(\omega)\right)\right| \\
&\leq \lim_{n\to\infty}\left(\int \left|\, ||g(\omega)||_\bullet - ||s_n(\omega)||_\bullet \,\right| d\mathrm{Q}(\omega)\right) \\
&\leq \lim_{n\to\infty}\left(\int ||g(\omega) - s_n(\omega)||_\bullet \, d\mathrm{Q}(\omega)\right) \\
&\leq ||k||_\infty \cdot \lim_{n\to\infty}\left(\int ||g(\omega) - s_n(\omega)||_H \, d\mathrm{Q}(\omega)\right) = 0\,. \qquad (23)
\end{aligned}$$

Here, we employed the finiteness of the two summands on the left hand side in the first step, and the reverse triangle inequality, (22) and (20) in the remaining steps. In the first step, the finiteness of the first summand follows directly from (22) and Theorem 3.10.9 from Denkowski et al. (2003), and the finiteness of the second one can be shown by again using (22) and then slightly adapting the mentioned theorem's proof:

$$\begin{aligned}
&\lim_{n\to\infty}\left(\int ||s_n(\omega)||_H \, d\mathrm{Q}(\omega)\right) \leq \lim_{n\to\infty}\left(\int ||s_n(\omega) - g(\omega)||_H \, d\mathrm{Q}(\omega) + \int ||g(\omega)||_H \, d\mathrm{Q}(\omega)\right) \\
&= \lim_{n\to\infty}\left(\int ||s_n(\omega) - g(\omega)||_H \, d\mathrm{Q}(\omega)\right) + \int ||g(\omega)||_H \, d\mathrm{Q}(\omega) = \int ||g(\omega)||_H \, d\mathrm{Q}(\omega) < \infty
\end{aligned}$$

with the first inequality holding true because of the second integral on its right hand side being finite (Denkowski et al., 2003, Theorem 3.10.9) and the first one being finite for n sufficiently large, cf. (20). The same equation (20) additionally tells us that

$\lim_{n\to\infty}\left(\int ||s_n(\omega)-g(\omega)||_H \, d\mathrm{Q}(\omega)\right)$ exists and the linearity of the limit can therefore be applied in the second step. Finally, (20) and the mentioned Theorem 3.10.9 yield the last two steps. ■

We will now turn our attention to the three summands on the right hand side of (16). That is, in the subsequent three lemmas, we will examine the effect of only one element of the triple (P,λ,k) varying at a time, with the proofs of Lemmas 13 and 14 being closely connected to their counterparts by Christmann et al. (2018) but generalizing them to the case of non-differentiable losses.

Lemma 13 *Let $\mathcal{X}$ be a complete and separable metric space and $\mathcal{Y}\subseteq\mathbb{R}$ be closed. Let $\mathrm{P}_1,\mathrm{P}_2\in\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ be probability measures, $\lambda>0$ and k be a bounded and measurable kernel on $\mathcal{X}$ with separable RKHS H. Let L be a convex and Lipschitz continuous loss function. Then,*

$$||f_{\mathrm{P}_1,\lambda,k}-f_{\mathrm{P}_2,\lambda,k}||_\infty \leq \frac{||k||_\infty^2\,|L|_1}{\lambda}\cdot||\mathrm{P}_1-\mathrm{P}_2||_{tv}\ .$$

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 (9). By Theorem 11, there exists a function h from the subdifferential of $L^\star$ with respect to $f_{\mathrm{P}_1,\lambda,k}$ such that (using the properties of Bochner integrals, cf. Christmann et al., 2018, Lemma 6.1, as well as (3))

$$\begin{aligned}
||f_{\mathrm{P}_1,\lambda,k}-f_{\mathrm{P}_2,\lambda,k}||_H &\leq \frac{1}{\lambda}\cdot\left|\left|\int h(x,y)\Phi(x)\,d\mathrm{P}_1(x,y)-\int h(x,y)\Phi(x)\,d\mathrm{P}_2(x,y)\right|\right|_H\\
&\leq \frac{1}{\lambda}\cdot\int||h(x,y)\Phi(x)||_H\;d|\mathrm{P}_1-\mathrm{P}_2|(x,y)\\
&\leq \frac{1}{\lambda}\cdot\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}}|h(x,y)|\cdot\sup_{x\in\mathcal{X}}||\Phi(x)||_H\cdot\int 1\,d|\mathrm{P}_1-\mathrm{P}_2|(x,y)\\
&= \frac{1}{\lambda}\cdot\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}}|h(x,y)|\cdot\sup_{x\in\mathcal{X}}\sqrt{k(x,x)}\cdot||\mathrm{P}_1-\mathrm{P}_2||_{tv}\\
&\leq \frac{1}{\lambda}\cdot|L|_1\cdot||k||_\infty\cdot||\mathrm{P}_1-\mathrm{P}_2||_{tv}\ ,
\end{aligned}$$

from which the assertion follows. ■

Lemma 14 *Let $\mathcal{X}$ be a complete and separable metric space and $\mathcal{Y}\subseteq\mathbb{R}$ be closed. Let $\mathrm{P}\in\mathcal{M}_1(\mathcal{X}\times\mathcal{Y})$ be a probability measure, $\lambda_1,\lambda_2>0$ and k be a bounded and measurable kernel on $\mathcal{X}$ with separable RKHS H. Let L be a convex and Lipschitz continuous loss function. Then,*

$$||f_{\mathrm{P},\lambda_1,k}-f_{\mathrm{P},\lambda_2,k}||_\infty \leq \frac{||k||_\infty^2\,|L|_1}{\min\{\lambda_1,\lambda_2\}^2}\cdot|\lambda_1-\lambda_2|\,.$$

Proof To shorten the notation, we define $f_i := f_{\mathrm{P},\lambda_i,k}$, $i = 1, 2$, in this proof. By (9) we know that

$$||f_1 - f_2||_\infty \leq ||k||_\infty \cdot ||f_1 - f_2||_H \ .$$

Assume now without loss of generality that $||f_1 - f_2||_H > 0$ since the case $||f_1 - f_2||_H = 0$ is trivial.

Theorem 11 yields functions h_1 and h_2 from the subdifferential of $L^\star$ (with respect to f_1 respectively f_2) such that

$$f_1 - f_2 = -\frac{1}{2\lambda_1} \cdot \int h_1(x,y)\Phi(x)\, d\mathrm{P}(x,y) + \frac{1}{2\lambda_2} \cdot \int h_2(x,y)\Phi(x)\, d\mathrm{P}(x,y)\,.$$

From this we obtain, by applying the reproducing property (2) in the last step,

$$\begin{aligned}
||f_1 - f_2||_H^2 &= \langle f_1 - f_2, f_1 - f_2\rangle_H \\
&= \left\langle \frac{1}{2\lambda_2} \cdot \int h_2(x,y)\Phi(x)\, d\mathrm{P}(x,y), f_1 - f_2 \right\rangle_H \\
&\quad - \left\langle \frac{1}{2\lambda_1} \cdot \int h_1(x,y)\Phi(x)\, d\mathrm{P}(x,y), f_1 - f_2 \right\rangle_H \\
&= \frac{1}{2\lambda_2} \cdot \int h_2(x,y)(f_1(x) - f_2(x))\, d\mathrm{P}(x,y) \\
&\quad - \frac{1}{2\lambda_1} \cdot \int h_1(x,y)(f_1(x) - f_2(x))\, d\mathrm{P}(x,y)\,. \qquad (24)
\end{aligned}$$

Because L (and thus also $L^\star$) 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

$$h_i(x,y) \cdot (t - f_i(x)) \leq L^\star(x,y,t) - L^\star(x,y,f_i(x)) \qquad \forall\, t \in \mathbb{R}, \qquad i = 1, 2\,,$$

more specifically

$$h_1(x,y) \cdot (f_2(x) - f_1(x)) \leq L^\star(x,y,f_2(x)) - L^\star(x,y,f_1(x))$$

and

$$h_2(x,y) \cdot (f_1(x) - f_2(x)) \leq L^\star(x,y,f_1(x)) - L^\star(x,y,f_2(x))\,.$$

Plugging these two inequalities into (24) yields

$$\begin{aligned}
||f_1 - f_2||_H^2 &\leq \left(\frac{1}{2\lambda_2} - \frac{1}{2\lambda_1}\right) \cdot \int L^\star(x,y,f_1(x)) - L^\star(x,y,f_2(x))\, d\mathrm{P}(x,y) \\
&= \left(\frac{1}{2\lambda_2} - \frac{1}{2\lambda_1}\right) \cdot \left(\mathbb{E}_\mathrm{P}\left[L^\star(X,Y,f_1(X))\right] - \mathbb{E}_\mathrm{P}\left[L^\star(X,Y,f_2(X))\right]\right) \qquad (25)
\end{aligned}$$

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 $\mathbb{E}_P\left[L^\star(X,Y,f_1(X))\right] - \mathbb{E}_P\left[L^\star(X,Y,f_2(X))\right]$ 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

$$\mathbb{E}_P\left[L^\star(X,Y,f_1(X))\right] + \lambda_1 \left|\left|f_1\right|\right|_H^2 \leq \mathbb{E}_P\left[L^\star(X,Y,f_2(X))\right] + \lambda_1 \left|\left|f_2\right|\right|_H^2 .$$

From this, it follows that

$$\begin{aligned} 0 &< \mathbb{E}_P\left[L^\star(X,Y,f_1(X))\right] - \mathbb{E}_P\left[L^\star(X,Y,f_2(X))\right] \leq \lambda_1 \cdot \left(\left|\left|f_2\right|\right|_H^2 - \left|\left|f_1\right|\right|_H^2\right) \\ &= \lambda_1 \cdot \left(\left|\left|f_1\right|\right|_H + \left|\left|f_2\right|\right|_H\right) \cdot \left(\left|\left|f_2\right|\right|_H - \left|\left|f_1\right|\right|_H\right) \leq \lambda_1 \cdot \left(\left|\left|f_1\right|\right|_H + \left|\left|f_2\right|\right|_H\right) \cdot \left|\left|f_1 - f_2\right|\right|_H \end{aligned}$$

with the last inequality holding true because of $\lambda_1(\left|\left|f_1\right|\right|_H + \left|\left|f_2\right|\right|_H) \geq 0$ and the reverse triangle inequality. Plugging this into (25) and dividing by $\left|\left|f_1 - f_2\right|\right|_H$, we obtain

$$\left|\left|f_1 - f_2\right|\right|_H \leq \frac{1}{2} \cdot \left(\frac{\max\{\lambda_1,\lambda_2\}}{\min\{\lambda_1,\lambda_2\}} - 1\right) \cdot \left(\left|\left|f_1\right|\right|_H + \left|\left|f_2\right|\right|_H\right) . \tag{26}$$

The case $\lambda_2 > \lambda_1$ yields the same inequality.

By additionally applying that $\left|\left|f_i\right|\right|_H \leq \frac{1}{\lambda_i} |L|_1 \left|\left|k\right|\right|_\infty$ (cf. Christmann et al., 2009, proof of Proposition 3), $i = 1, 2$, we now obtain

$$\begin{aligned} \left|\left|f_1 - f_2\right|\right|_H &\leq \frac{|L|_1 \left|\left|k\right|\right|_\infty}{2} \cdot \left(\frac{\max\{\lambda_1,\lambda_2\}}{\min\{\lambda_1,\lambda_2\}} - 1\right) \cdot \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right) \\ &\leq \frac{|L|_1 \left|\left|k\right|\right|_\infty}{2\min\{\lambda_1,\lambda_2\}} \cdot \big(\max\{\lambda_1,\lambda_2\} - \min\{\lambda_1,\lambda_2\}\big) \cdot \frac{2}{\min\{\lambda_1,\lambda_2\}} \\ &= \frac{|L|_1 \left|\left|k\right|\right|_\infty}{\min\{\lambda_1,\lambda_2\}^2} \cdot |\lambda_1 - \lambda_2| \end{aligned}$$

which yields the assertion. ■

Lemma 15 *Let $\mathcal{X}$ be a complete and separable metric space and $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be a probability measure, $\lambda > 0$ and k_1, k_2 be bounded and measurable kernels on $\mathcal{X}$ with separable RKHSs H_1, H_2. Denote $\kappa := \max\{\left|\left|k_1\right|\right|_\infty, \left|\left|k_2\right|\right|_\infty\}$. Let L be a convex and Lipschitz continuous loss function. Then,*

$$\left|\left|f_{P,\lambda,k_1} - f_{P,\lambda,k_2}\right|\right|_\infty \leq \frac{|L|_1}{\lambda} \cdot \left(\frac{1}{2} \cdot \left|\left|k_1 - k_2\right|\right|_\infty + \kappa \cdot \sqrt{\left|\left|k_1 - k_2\right|\right|_\infty}\right) .$$

In order to prove Lemma 15, we first need a short auxiliary statement which is probably well-known, but which we were unable to find a reference for. For reasons of better readability we therefore prove the following auxiliary lemma:

Lemma 16 *Let $\mathcal{X} \neq \emptyset$ and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with RKHS H. Let $\alpha > 0$ and define the kernel $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by $\tilde{k} := \alpha k$. Then, $\tilde{H} := H$ equipped with the norm $\left|\left|\cdot\right|\right|_{\tilde{H}} := \frac{1}{\sqrt{\alpha}} \left|\left|\cdot\right|\right|_H$ is the RKHS of $\tilde{k}$.*

Proof Steinwart and Christmann (2008, Lemma 4.5) yields that $\tilde{k}$ is actually a kernel. Hence, the assertion follows directly from Steinwart and Christmann (2008, Theorem 4.21): It can easily be seen for the pre-Hilbert space from equation (4.12) in that theorem and follows by completion for the whole Hilbert space. ■

Proof of Lemma 15 To shorten the notation, we define $f_i := f_{\mathrm{P},\lambda,k_i}$, $i = 1, 2$, in this proof.

Define $\tilde{k}_i := \frac{k_i}{2}$ for $i = 1, 2$. From Lemma 16 we know that $\tilde{H}_i = H_i$ (equipped with the norm $||\cdot||_{\tilde{H}_i} = \sqrt{2}\,||\cdot||_{H_i}$) is the RKHS of $\tilde{k}_i$. Thus, we obviously have $f_i \in \tilde{H}_i$ for $i = 1, 2$.

In the next step, we define a new space which contains f_1 as well as f_2 by

$$\tilde{H} := \tilde{H}_1 \oplus \tilde{H}_2 := \left\{ g : \mathcal{X} \to \mathbb{R} \;\middle|\; g = g_1 + g_2, g_1 \in \tilde{H}_1, g_2 \in \tilde{H}_2 \right\}.$$

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 will 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 $f_{L^\star,\mathrm{P},\lambda,\tilde{k}} =: \tilde{f}$ (cf. Theorem 11). The triangle inequality then yields

$$||f_1 - f_2||_\infty \leq \left|\left|f_1 - \tilde{f}\right|\right|_\infty + \left|\left|f_2 - \tilde{f}\right|\right|_\infty . \tag{27}$$

By applying Theorem 11, we can expand both of the differences on the right hand side as

$$\begin{aligned} f_i - \tilde{f} = & -\frac{1}{2\lambda} \cdot \int h_i(x,y)\Phi_i(x)\, d\mathrm{P}(x,y) \\ & + \frac{1}{2\lambda} \cdot \int \tilde{h}(x,y)\tilde{\Phi}(x)\, d\mathrm{P}(x,y) \\ = & \frac{1}{2\lambda} \cdot \int h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x)\right) d\mathrm{P}(x,y) \\ & + \frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y)\right) \tilde{\Phi}(x)\, d\mathrm{P}(x,y) \end{aligned} \tag{28}$$

with h_i and $\tilde{h}$ from the subdifferential of $L^\star$ (with respect to f_i respectively $\tilde{f}$). Thus, (9) yields for $i = 1, 2$

$$\begin{aligned}
\left|\left|f_i - \tilde{f}\right|\right|_\infty &\leq \left|\left|\frac{1}{2\lambda} \cdot \int h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x)\right) \, d\mathrm{P}(x,y)\right|\right|_\infty \\
&\quad + \left|\left|\frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y)\right) \tilde{\Phi}(x) \, d\mathrm{P}(x,y)\right|\right|_\infty \\
&\leq \left|\left|\frac{1}{2\lambda} \cdot \int h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x)\right) \, d\mathrm{P}(x,y)\right|\right|_\infty \\
&\quad + \left|\left|\tilde{k}\right|\right|_\infty \cdot \left|\left|\frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y)\right) \tilde{\Phi}(x) \, d\mathrm{P}(x,y)\right|\right|_{\tilde{H}} . \qquad (29)
\end{aligned}$$

Now, we can easily bound the first summand on the right hand side of (29) by

$$\begin{aligned}
\left|\left|\frac{1}{2\lambda} \cdot \int h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x)\right) \, d\mathrm{P}(x,y)\right|\right|_\infty &\leq \frac{1}{2\lambda} \cdot ||h_i||_\infty \cdot \sup_{x \in \mathcal{X}} \left|\left|\tilde{\Phi}(x) - \Phi_i(x)\right|\right|_\infty \\
&\leq \frac{|L|_1}{2\lambda} \cdot \left|\left|\tilde{k} - k_i\right|\right|_\infty , \qquad (30)
\end{aligned}$$

where we applied Lemma 12 in the first step and obtained the bound for h_i from Theorem 11.

As for the square of the $\tilde{H}$-norm in the second summand on the right hand side of (29), applying (28) yields

$$\begin{aligned}
&\left|\left|\frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y)\right) \tilde{\Phi}(x) \, d\mathrm{P}(x,y)\right|\right|_{\tilde{H}}^2 \\
&= \left\langle \frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y)\right) \tilde{\Phi}(x) \, d\mathrm{P}(x,y) \, , f_i - \tilde{f} \right\rangle_{\tilde{H}} \\
&\quad - \left\langle \frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y)\right) \tilde{\Phi}(x) \, d\mathrm{P}(x,y) \, , \right. \\
&\qquad \left. \frac{1}{2\lambda} \cdot \int h_i(x',y') \left(\tilde{\Phi}(x') - \Phi_i(x')\right) \, d\mathrm{P}(x',y') \right\rangle_{\tilde{H}} , \qquad (31)
\end{aligned}$$

where we can apply the reproducing property (2) to the first of these two inner products in order to obtain

$$\begin{aligned}
&\left\langle \frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y)\right) \tilde{\Phi}(x) \, d\mathrm{P}(x,y), f_i - \tilde{f} \right\rangle_{\tilde{H}} \\
&= \frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y)\right) \left(f_i(x) - \tilde{f}(x)\right) \, d\mathrm{P}(x,y) \leq 0 \, .
\end{aligned}$$

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 result into (31) results in

$$\begin{aligned}
&\left|\left|\frac{1}{2\lambda}\cdot\int\left(\tilde{h}(x,y)-h_i(x,y)\right)\tilde{\Phi}(x)\,d\mathrm{P}(x,y)\right|\right|_{\tilde{H}}^2\\
&\leq\left|\left\langle\frac{1}{2\lambda}\cdot\int\left(\tilde{h}(x,y)-h_i(x,y)\right)\tilde{\Phi}(x)\,d\mathrm{P}(x,y)\,,\right.\right.\\
&\qquad\left.\left.\frac{1}{2\lambda}\cdot\int h_i(x',y')\left(\tilde{\Phi}(x')-\Phi_i(x')\right)\,d\mathrm{P}(x',y')\right\rangle_{\tilde{H}}\right|\\
&=\frac{1}{4\lambda^2}\cdot\left|\int\int\left(\tilde{h}(x,y)-h_i(x,y)\right)h_i(x',y')\left(\tilde{k}(x,x')-k_i(x,x')\right)\,d\mathrm{P}(x',y')\,d\mathrm{P}(x,y)\right|\\
&\leq\frac{1}{4\lambda^2}\cdot\left|\left|\tilde{h}-h_i\right|\right|_\infty\cdot||h_i||_\infty\cdot\left|\left|\tilde{k}-k_i\right|\right|_\infty\\
&\leq\frac{|L|_1^2}{2\lambda^2}\cdot\left|\left|\tilde{k}-k_i\right|\right|_\infty\,,
\end{aligned}\tag{32}$$

where we again applied the reproducing property in the second step and Theorem 11 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 (27) with (29), (30) and (32):

$$\begin{aligned}
||f_1-f_2||_\infty&\leq\left|\left|f_1-\tilde{f}\right|\right|_\infty+\left|\left|f_2-\tilde{f}\right|\right|_\infty\\
&\leq\sum_{i=1}^{2}\left(\frac{|L|_1}{2\lambda}\cdot\left|\left|\tilde{k}-k_i\right|\right|_\infty+\left|\left|\tilde{k}\right|\right|_\infty\cdot\frac{|L|_1}{\sqrt{2}\lambda}\cdot\sqrt{\left|\left|\tilde{k}-k_i\right|\right|_\infty}\right)\\
&\leq\frac{|L|_1}{\lambda}\cdot\left(\frac{1}{2}\cdot||k_1-k_2||_\infty+\kappa\cdot\sqrt{||k_1-k_2||_\infty}\right)\,.
\end{aligned}$$

■

We can now progress to the analogous decomposition of $||f_{\mathrm{P}_1,\lambda_1,k_1}-f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_i^X)}$. However, we only need to prove an analogous result to Lemma 15 but not to Lemmas 13 and 14, since our analysis of the first two summands on the right hand side of (17) in the proof of Theorem 4 will be based directly on Lemma 13 respectively Lemma 14 and the fact that

$$||g||_{L_p(\mathrm{P}^X)}\leq||g||_\infty\tag{33}$$

for all bounded functions g and all $p\in[1,\infty)$.

Lemma 17 *Let $\mathcal{X}$ be a complete and separable metric space and $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ be a probability measure, $\lambda > 0$ and k_1, k_2 be bounded and measurable kernels on $\mathcal{X}$ with separable RKHSs H_1, H_2. Denote $\kappa := \max\{||k_1||_\infty, ||k_2||_\infty\}$. Let L be a convex and Lipschitz continuous loss function and 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 Lemma 15 with $||\cdot||_\infty$ being replaced by $||\cdot||_{L_p(\mathrm{P}^X)}$, for which reason we will only highlight the differences here.

First of all, because of (33), we obtain analogously to (29)

$$\begin{aligned}
\left|\left| f_i - \tilde{f} \right|\right|_{L_p(\mathrm{P}^X)} \leq & \left|\left| \frac{1}{2\lambda} \cdot \int h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x) \right) d\mathrm{P}(x,y) \right|\right|_{L_p(\mathrm{P}^X)} \\
& + \left|\left| \frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y) \right) \tilde{\Phi}(x) \, d\mathrm{P}(x,y) \right|\right|_\infty \\
\leq & \left|\left| \frac{1}{2\lambda} \cdot \int h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x) \right) d\mathrm{P}(x,y) \right|\right|_{L_p(\mathrm{P}^X)} \\
& + \left|\left| \tilde{k} \right|\right|_\infty \cdot \left|\left| \frac{1}{2\lambda} \cdot \int \left(\tilde{h}(x,y) - h_i(x,y) \right) \tilde{\Phi}(x) \, d\mathrm{P}(x,y) \right|\right|_{\tilde{H}} .
\end{aligned}$$

Then, the first summand on the right hand side can be bounded in an analogous way to (30):

$$\begin{aligned}
& \left|\left| \frac{1}{2\lambda} \cdot \int_{\mathcal{X}\times\mathcal{Y}} h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x) \right) d\mathrm{P}(x,y) \right|\right|_{L_p(\mathrm{P}^X)} \\
& \leq \frac{1}{2\lambda} \int_{\mathcal{X}\times\mathcal{Y}} \left|\left| h_i(x,y) \left(\tilde{\Phi}(x) - \Phi_i(x) \right) \right|\right|_{L_p(\mathrm{P}^X)} d\mathrm{P}(x,y) \\
& \leq \frac{1}{2\lambda} \cdot ||h_i||_\infty \cdot \int_{\mathcal{X}} \left|\left| \tilde{\Phi}(x) - \Phi_i(x) \right|\right|_{L_p(\mathrm{P}^X)} d\mathrm{P}^X(x) \\
& = \frac{1}{2\lambda} \cdot ||h_i||_\infty \cdot \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \left| \tilde{k}(x,x') - k_i(x,x') \right|^p d\mathrm{P}^X(x) \right)^{1/p} d\mathrm{P}^X(x) \\
& \leq \frac{|L|_1}{2\lambda} \cdot \left|\left| \tilde{k} - k_i \right|\right|_{L_p(\mathrm{P}^X \otimes \mathrm{P}^X)} ,
\end{aligned}$$

where we applied Lemma 12 in the first step, and Theorem 11 (for obtaining the bound on h_i) as well as Hölder's inequality in the last step. Finally, we can tighten the bound from

the last steps of (32) in the following way:

$$\begin{aligned}
&\frac{1}{4\lambda^2} \cdot \left| \int \int \left(\tilde{h}(x,y) - h_i(x,y) \right) h_i(x',y') \left(\tilde{k}(x,x') - k_i(x,x') \right) \, d\mathrm{P}(x',y') \, d\mathrm{P}(x,y) \right| \\
&\leq \frac{|L|_1^2}{2\lambda^2} \cdot \int \int \left| \tilde{k}(x,x') - k_i(x,x') \right| \, d\mathrm{P}(x',y') \, d\mathrm{P}(x,y) \\
&= \frac{|L|_1^2}{2\lambda^2} \cdot \left|\left| \tilde{k} - k_i \right|\right|_{L_1(\mathrm{P}^X \otimes \mathrm{P}^X)} \\
&\leq \frac{|L|_1^2}{2\lambda^2} \cdot \left|\left| \tilde{k} - k_i \right|\right|_{L_p(\mathrm{P}^X \otimes \mathrm{P}^X)} .
\end{aligned}$$

The assertion then follows in the same way as in the proof of Lemma 15. ■

Appendix B. Proofs

In this appendix, we will first prove the results from Section 2 in Appendix B.1 and then the ones from Section 3 in Appendix B.2.

B.1 Proofs for Section 2

Proof of Theorem 2 Applying Lemmas 13 to 15 to the decomposition (16) of the investigated norm $||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty$ yields

$$\begin{aligned}
||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_\infty \leq{}& \frac{||k_1||_\infty^2 \, |L|_1}{\lambda_1} \cdot ||\mathrm{P}_1 - \mathrm{P}_2||_{tv} + \frac{||k_1||_\infty^2 \, |L|_1}{\min\{\lambda_1,\lambda_2\}^2} \cdot |\lambda_1 - \lambda_2| \\
&+ \frac{|L|_1}{\lambda_2} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_\infty + \kappa \cdot \sqrt{||k_1 - k_2||_\infty} \right) .
\end{aligned}$$

Since the order of decomposition can of course be freely varied, 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. ■

Proof of Theorem 4 Applying Lemma 17 as well as Lemmas 13 and 14 in combination with (33) to the decomposition (17) 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 $i = 2$

$$\begin{aligned}
&||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_2^X)} \\
&\leq \frac{||k_1||_\infty^2 \, |L|_1}{\lambda_1} \cdot ||\mathrm{P}_1 - \mathrm{P}_2||_{tv} + \frac{||k_1||_\infty^2 \, |L|_1}{\min\{\lambda_1,\lambda_2\}^2} \cdot |\lambda_1 - \lambda_2| \\
&\quad + \frac{|L|_1}{\lambda_2} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_2^X \otimes \mathrm{P}_2^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_2^X \otimes \mathrm{P}_2^X)}} \right) \\
&\leq \frac{\kappa^2 |L|_1}{\tau} \cdot ||\mathrm{P}_1 - \mathrm{P}_2||_{tv} + \frac{\kappa^2 |L|_1}{\tau^2} \cdot |\lambda_1 - \lambda_2| \\
&\quad + \frac{|L|_1}{\tau} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_2^X \otimes \mathrm{P}_2^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_2^X \otimes \mathrm{P}_2^X)}} \right) .
\end{aligned}$$

Analogously, reversing the order of decomposition (such that Lemma 17 can be applied to a summand with probability measure P_1 in both SVMs) yields for $i = 1$

$$\begin{aligned}
&||f_{\mathrm{P}_1,\lambda_1,k_1} - f_{\mathrm{P}_2,\lambda_2,k_2}||_{L_p(\mathrm{P}_1^X)} \\
&\leq \frac{|L|_1}{\lambda_1} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_1^X \otimes \mathrm{P}_1^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_1^X \otimes \mathrm{P}_1^X)}} \right) \\
&\quad + \frac{||k_2||_\infty^2 \, |L|_1}{\min\{\lambda_1, \lambda_2\}^2} \cdot |\lambda_1 - \lambda_2| + \frac{||k_2||_\infty^2 \, |L|_1}{\lambda_2} \cdot ||\mathrm{P}_1 - \mathrm{P}_2||_{tv} \\
&\leq \frac{\kappa^2 |L|_1}{\tau} \cdot ||\mathrm{P}_1 - \mathrm{P}_2||_{tv} + \frac{\kappa^2 |L|_1}{\tau^2} \cdot |\lambda_1 - \lambda_2| \\
&\quad + \frac{|L|_1}{\tau} \cdot \left(\frac{1}{2} \cdot ||k_1 - k_2||_{L_p(\mathrm{P}_1^X \otimes \mathrm{P}_1^X)} + \kappa \cdot \sqrt{||k_1 - k_2||_{L_p(\mathrm{P}_1^X \otimes \mathrm{P}_1^X)}} \right) .
\end{aligned}$$

■

B.2 Proofs for Section 3

Proof of Theorem 5 To shorten the notation, we define $f_i := f_{\mathrm{P}_i, \boldsymbol{\lambda_i}, \boldsymbol{k_i}}$ and $f_{i,b} := f_{\mathrm{P}_{i,b}, \lambda_{i,b}, k_{i,b}}$, $i = 1, 2$, $b = 1, \ldots, B$, in this proof. By the definition of f_1 and f_2 we know that

$$\begin{aligned}
||f_1 - f_2||_\infty &\leq \sup_{x \in \mathcal{X}} \sum_{b=1}^{B} w_b(x) \cdot \left| \hat{f}_{1,b}(x) - \hat{f}_{2,b}(x) \right| \\
&\leq \sup_{x \in \mathcal{X}} \max_{b \in \{1,\ldots,B\}} \left| \hat{f}_{1,b}(x) - \hat{f}_{2,b}(x) \right| \\
&= \max_{b \in \{1,\ldots,B\}} \left| \left| \hat{f}_{1,b} - \hat{f}_{2,b} \right| \right|_\infty , \qquad (34)
\end{aligned}$$

where we applied **(W1)** and **(W2)** in the second step. Since the functions $\hat{f}_{i,b}$ have not been defined as SVMs but instead as zero-extensions of SVMs $f_{\mathrm{P}_{i,b}, \lambda_{i,b}, k_{i,b}}$ on $\mathcal{X}_b$, we cannot apply Theorem 2 to the right hand side of (34) yet. However, these functions can actually be seen as SVMs on $\mathcal{X}$ themselves, $\hat{f}_{i,b} = f_{\hat{\mathrm{P}}_{i,b}, \lambda_{i,b}, \hat{k}_{i,b}}$:

According to Meister and Steinwart (2016, Lemma 2), we have $\hat{H}_{i,b} = \{ \hat{g} \mid g \in H_{i,b} \}$ and $||\hat{g}||_{\hat{H}_{i,b}} = ||g||_{H_{i,b}}$ for all $g \in H_{i,b}$. Since additionally $\mathcal{R}_{L^\star, \hat{\mathrm{P}}_{i,b}}(\hat{g}) = \mathcal{R}_{L^\star, \mathrm{P}_{i,b}}(g)$ for all $g \in H_{i,b}$ (because the whole probability mass of $\hat{\mathrm{P}}_{i,b}$ is on $\mathcal{X}_b$ where $\hat{g}$ and g coincide), (8) yields $f_{\hat{\mathrm{P}}_{i,b}, \lambda_{i,b}, \hat{k}_{i,b}} = \hat{f}_{\mathrm{P}_{i,b}, \lambda_{i,b}, k_{i,b}} \, (= \hat{f}_{i,b})$.

Thus, we can apply Theorem 2 to the right hand side of (34) since the functions $\hat{f}_{i,b}$ are actually SVMs on the complete space $\mathcal{X}$ (whereas the functions $f_{i,b}$ are SVMs on the not necessarily complete spaces $\mathcal{X}_b$ for which reason the theorem can not be applied to $||f_{1,b} - f_{2,b}||_\infty$ even though this term is obviously equivalent to $||\hat{f}_{1,b} - \hat{f}_{2,b}||_\infty$). By doing this, the first assertion follows, but with every $\mathrm{P}_{i,b}$ replaced by $\hat{\mathrm{P}}_{i,b}$ and $k_{i,b}$ by $\hat{k}_{i,b}$. Because of them just being zero-extensions of $\mathrm{P}_{i,b}$ and $k_{i,b}$ respectively however, this does not influence the respective norms. ■

Proof of Theorem 6 To shorten the notation, we define $f_i := f_{\mathrm{P}_i,\boldsymbol{\lambda_i},\boldsymbol{k_i}}$ and $f_{i,b} := f_{\mathrm{P}_{i,b},\lambda_{i,b},k_{i,b}}$, $i = 1, 2$, $b = 1, \ldots, B$, in this proof. By the definition of f_1 and f_2 we know that

$$\begin{aligned}
||f_1 - f_2||_{L_p(\mathrm{P}_i^X)} &\leq \sum_{b=1}^{B} \left|\left| w_b \cdot \left(\hat{f}_{1,b} - \hat{f}_{2,b} \right) \right|\right|_{L_p(\mathrm{P}_i^X)} \\
&\leq \sum_{b=1}^{B} \left(\int_{\mathcal{X}} \left| \hat{f}_{1,b}(x) - \hat{f}_{2,b}(x) \right|^p \, d\mathrm{P}_i^X(x) \right)^{1/p} \\
&= \sum_{b=1}^{B} \left(\mathrm{P}_i^X(\mathcal{X}_b) \cdot \int_{\mathcal{X}_b} \left| \hat{f}_{1,b}(x) - \hat{f}_{2,b}(x) \right|^p \, d\mathrm{P}_{i,b}^X(x) \right)^{1/p} \\
&= \sum_{b=1}^{B} \left(\mathrm{P}_i^X(\mathcal{X}_b) \right)^{1/p} \cdot \left(\int_{\mathcal{X}} \left| \hat{f}_{1,b}(x) - \hat{f}_{2,b}(x) \right|^p \, d\hat{\mathrm{P}}_{i,b}^X(x) \right)^{1/p} \\
&= \sum_{b=1}^{B} \left(\mathrm{P}_i^X(\mathcal{X}_b) \right)^{1/p} \cdot \left|\left| \hat{f}_{1,b} - \hat{f}_{2,b} \right|\right|_{L_p(\hat{\mathrm{P}}_{i,b}^X)} . \qquad (35)
\end{aligned}$$

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 (10) 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 $\hat{f}_{1,b}$ and $\hat{f}_{2,b}$ are SVMs on $\mathcal{X}$ themselves, $\hat{f}_{i,b} = f_{\hat{\mathrm{P}}_{i,b},\lambda_{i,b},\hat{k}_{i,b}}$ (cf. proof of Theorem 5), we can now apply Theorem 4 to the norms on the right hand side of (35). This yields the assertion (as in the proof of Theorem 5 with $\hat{\mathrm{P}}_{i,b}$ and $\hat{k}_{i,b}$ instead ob $\mathrm{P}_{i,b}$ and $k_{i,b}$ which does not change the respective norms). ■

Proof of Theorem 8 In addition to the auxiliary distributions and kernels introduced prior to Theorem 8, we also need auxiliary regularization parameters in this proof. We denote these parameters by $\lambda_{i,j,b}^* := (\mathrm{P}^X_{j,\mathcal{X}_{a(i,b)}^{(i)}}(\mathcal{X}_b^*))^{-1}\lambda_{i,a(i,b)}$ for $i, j = 1, 2$ and $b = 1, \ldots, B$.

By applying the triangle inequality we can now expand the norm we have to investigate as

$$\begin{aligned}
\left|\left| f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}_{\boldsymbol{A_1}}^{(\mathbf{1})}} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}_{\boldsymbol{A_2}}^{(\mathbf{2})}} \right|\right|_{L_1(\mathrm{P}_1^X)} \leq\ & \left|\left| f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}_{\boldsymbol{A_1}}^{(\mathbf{1})}} - f_{\mathrm{P}_1,\boldsymbol{\lambda}_{\mathbf{1,1}}^*,\boldsymbol{k}_{\mathbf{1}}^*,\boldsymbol{\mathcal{X}}_{\boldsymbol{B}}^*} \right|\right|_{L_1(\mathrm{P}_1^X)} \\
&+ \left|\left| f_{\mathrm{P}_1,\boldsymbol{\lambda}_{\mathbf{1,1}}^*,\boldsymbol{k}_{\mathbf{1}}^*,\boldsymbol{\mathcal{X}}_{\boldsymbol{B}}^*} - f_{\mathrm{P}_1,\boldsymbol{\lambda}_{\mathbf{2,1}}^*,\boldsymbol{k}_{\mathbf{2}}^*,\boldsymbol{\mathcal{X}}_{\boldsymbol{B}}^*} \right|\right|_{L_1(\mathrm{P}_1^X)} \\
&+ \left|\left| f_{\mathrm{P}_1,\boldsymbol{\lambda}_{\mathbf{2,1}}^*,\boldsymbol{k}_{\mathbf{2}}^*,\boldsymbol{\mathcal{X}}_{\boldsymbol{B}}^*} - f_{\mathrm{P}_2,\boldsymbol{\lambda_2},\boldsymbol{k_2},\boldsymbol{\mathcal{X}}_{\boldsymbol{A_2}}^{(\mathbf{2})}} \right|\right|_{L_1(\mathrm{P}_1^X)} \qquad (36)
\end{aligned}$$

with $\boldsymbol{\lambda}_{\boldsymbol{i,j}}^* := (\lambda_{i,j,1}^*, \ldots, \lambda_{i,j,B}^*)$ and $\boldsymbol{k}_{\boldsymbol{i}}^* := (k_{i,1}^*, \ldots, k_{i,B}^*)$ for $i, j = 1, 2$, and the newly introduced localized SVMs being defined as in (12). We will now examine the three norms from the right hand side of (36) separately:

(i) For a function $g : \mathcal{X}_b^* \to \mathbb{R}$, denote by $\tilde{g}$ its zero-extension to $\mathcal{X}_{a(1,b)}^{(1)}$ (respectively to $\mathcal{X}_{a(1,b)}^{(1)} \times \mathcal{X}_{a(1,b)}^{(1)}$ 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 $a \in \{1,\dots,A_1\}$ new local SVMs $f_{\mathrm{P}_{1,a},\lambda_{1,a},k^\circ_{1,a}}$ which by (8) are defined as

$$f_{\mathrm{P}_{1,a},\lambda_{1,a},k^\circ_{1,a}} = \arg\inf_{f\in H^\circ_{1,a}} \mathcal{R}_{L^\star,\mathrm{P}_{1,a}}(f) + \lambda_{1,a}\,||f||^2_{H^\circ_{1,a}}\,. \tag{37}$$

Now, combining Theorem 5 from Berlinet and Thomas-Agnan (2004) and Lemma 2 from Meister and Steinwart (2016) yields that

$$H^\circ_{1,a} = \left\{ f : \mathcal{X}^{(1)}_a \to \mathbb{R} \;\middle|\; f = \sum_{b\in J_{1,a}} \tilde{f}_b,\ f_b \in H^*_{1,b} \text{ for } b = 1,\dots,B \right\},$$

with the decomposition of each such $f \in H^\circ_{1,a}$ being unique because of the sets $\mathcal{X}^*_1,\dots,\mathcal{X}^*_B$, the domains of the functions f_b, being pairwise disjoint, cf. **(R1')** and **(R2')**. Thus, the mentioned results also yield $||f||^2_{H^\circ_{1,a}} = \sum_{b\in J_{1,a}} ||f_b||^2_{H^*_{1,b}}$ for all $f \in H^\circ_{1,a}$. Additionally, again because of the domains of the functions f_b being pairwise disjoint, we are able to also expand the risk from (37) similarly to the preceding expansion of the $H^\circ_{1,a}$-norm:

$$\begin{aligned}
\mathcal{R}_{L^\star,\mathrm{P}_{1,a}}(f) &= \int_{\mathcal{X}^{(1)}_a} L^\star(x,y,f(x))\,d\mathrm{P}_{1,a}(x,y) \\
&= \sum_{b\in J_{1,a}} \int_{\mathcal{X}^*_b} L^\star(x,y,f_b(x))\,d\mathrm{P}_{1,a}(x,y) \\
&= \sum_{b\in J_{1,a}} \mathrm{P}^X_{1,a}(\mathcal{X}^*_b)\cdot \int_{\mathcal{X}^*_b} L^\star(x,y,f_b(x))\,d\mathrm{P}^*_{1,b}(x,y) \\
&= \sum_{b\in J_{1,a}} \mathrm{P}^X_{1,a}(\mathcal{X}^*_b)\cdot \mathcal{R}_{L^\star,\mathrm{P}^*_{1,b}}(f_b)\,,
\end{aligned}$$

where we applied (10) in the third step.

By plugging this into (37), we obtain

$$\begin{aligned}
f_{\mathrm{P}_{1,a},\lambda_{1,a},k^\circ_{1,a}} &= \arg\inf_{f\in H^\circ_{1,a}} \sum_{b\in J_{1,a}} \left(\mathrm{P}^X_{1,a}(\mathcal{X}^*_b)\cdot \mathcal{R}_{L^\star,\mathrm{P}^*_{1,b}}(f_b) + \lambda_{1,a}\,||f_b||^2_{H^*_{1,b}} \right) \\
&= \sum_{b\in J_{1,a}} \widetilde{\arg\inf_{f_b\in H^*_{1,b}}} \left(\mathrm{P}^X_{1,a}(\mathcal{X}^*_b)\cdot \mathcal{R}_{L^\star,\mathrm{P}^*_{1,b}}(f_b) + \lambda_{1,a}\,||f_b||^2_{H^*_{1,b}} \right) \\
&= \sum_{b\in J_{1,a}} \widetilde{\arg\inf_{f_b\in H^*_{1,b}}} \left(\mathcal{R}_{L^\star,\mathrm{P}^*_{1,b}}(f_b) + \frac{\lambda_{1,a}}{\mathrm{P}^X_{1,a}(\mathcal{X}^*_b)}\,||f_b||^2_{H^*_{1,b}} \right) \\
&= \sum_{b\in J_{1,a}} \tilde{f}_{\mathrm{P}^*_{1,b},\lambda^*_{1,1,b},k^*_{1,b}}
\end{aligned}$$

and thus

$$f_{\mathrm{P}_1,\boldsymbol{\lambda^*_{1,1}},\boldsymbol{k^*_1},\boldsymbol{\mathcal{X}^*_B}} = \sum_{b=1}^{B} \hat{f}_{\mathrm{P}^*_{1,b},\lambda^*_{1,1,b},k^*_{1,b}} = \sum_{a=1}^{A_1} \sum_{b\in J_{1,a}} \hat{f}_{\mathrm{P}^*_{1,b},\lambda^*_{1,1,b},k^*_{1,b}} = \sum_{a=1}^{A_1} \hat{f}_{\mathrm{P}_{1,a},\lambda_{1,a},k^\circ_{1,a}}\,.$$

We can therefore also interpret the first difference on the right hand side of (36) as the difference between two localized SVMs that are based on the same regionalization $\boldsymbol{\mathcal{X}}^{(\mathbf{1})}_{\boldsymbol{A_1}}$ (and on the same probability measure and vector of regularization parameters). An application of Theorem 6 hence yields

$$\begin{aligned}
&\left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}^{(\mathbf{1})}_{\boldsymbol{A_1}}} - f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{1},\mathbf{1}},\boldsymbol{k}^*_{\mathbf{1}},\boldsymbol{\mathcal{X}}^*_{\boldsymbol{B}}}\right|\right|_{L_1(\mathrm{P}_1^X)} = \left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}^{(\mathbf{1})}_{\boldsymbol{A_1}}} - \sum_{a=1}^{A_1} \hat{f}_{\mathrm{P}_{1,a},\lambda_{1,a},k^\circ_{1,a}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{a=1}^{A_1} \mathrm{P}_1^X(\mathcal{X}_a^{(1)}) \cdot \left(\frac{1}{2\lambda_{1,a}} \cdot \left|\left|k_{1,a} - k^\circ_{1,a}\right|\right|_{L_1(\mathrm{P}^X_{1,a}\otimes\mathrm{P}^X_{1,a})} \right. \\
&\qquad\qquad \left. + \frac{\max\left\{||k_{1,a}||_\infty, \left|\left|k^\circ_{1,a}\right|\right|_\infty\right\}}{\lambda_{1,a}} \cdot \sqrt{\left|\left|k_{1,a} - k^\circ_{1,a}\right|\right|_{L_1(\mathrm{P}^X_{1,a}\otimes\mathrm{P}^X_{1,a})}} \right).
\end{aligned}$$

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

$$\begin{aligned}
\left|\left|k_{1,a} - k^\circ_{1,a}\right|\right|_{L_1(\mathrm{P}^X_{1,a}\otimes\mathrm{P}^X_{1,a})} &= \int_{\mathcal{X}_a^{(1)}} \int_{\mathcal{X}_a^{(1)}} \left|k_{1,a}(x,x') - k^\circ_{1,a}(x,x')\right| \, d\mathrm{P}^X_{1,a}(x') d\mathrm{P}^X_{1,a}(x) \\
&= \sum_{b\in J_{1,a}} \int_{\mathcal{X}_b^*} \int_{\mathcal{X}_a^{(1)}\setminus\mathcal{X}_b^*} \left|k_{1,a}(x,x')\right| \, d\mathrm{P}^X_{1,a}(x') d\mathrm{P}^X_{1,a}(x) \\
&\leq ||k_{1,a}||^2_\infty \cdot \sum_{b\in J_{1,a}} \mathrm{P}^X_{1,a}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,a}(\mathcal{X}_b^*)\right)
\end{aligned}$$

which finally results in

$$\begin{aligned}
&\left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda_1},\boldsymbol{k_1},\boldsymbol{\mathcal{X}}^{(\mathbf{1})}_{\boldsymbol{A_1}}} - f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{1},\mathbf{1}},\boldsymbol{k}^*_{\mathbf{1}},\boldsymbol{\mathcal{X}}^*_{\boldsymbol{B}}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{a=1}^{A_1} \mathrm{P}_1^X(\mathcal{X}_a^{(1)}) \cdot \left(\frac{||k_{1,a}||^2_\infty}{2\lambda_{1,a}} \cdot \sum_{b\in J_{1,a}} \mathrm{P}^X_{1,a}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,a}(\mathcal{X}_b^*)\right) \right. \\
&\qquad\qquad \left. + \frac{||k_{1,a}||^2_\infty}{\lambda_{1,a}} \cdot \sqrt{\sum_{b\in J_{1,a}} \mathrm{P}^X_{1,a}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,a}(\mathcal{X}_b^*)\right)} \right).
\end{aligned}$$

(ii) The second norm on the right hand side of (36) already consists of the difference of two localized SVMs that are based on the same regionalization $\boldsymbol{\mathcal{X}}^*_{\boldsymbol{B}}$ (and the same probability measure), without us needing to make any changes before. We can therefore

directly apply Theorem 6 and obtain

$$\begin{aligned}
&\left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{1},\mathbf{1}},\boldsymbol{k}^*_{\mathbf{1}},\boldsymbol{\mathcal{X}}^*_{\boldsymbol{B}}} - f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{2},\mathbf{1}},\boldsymbol{k}^*_{\mathbf{2}},\boldsymbol{\mathcal{X}}^*_{\boldsymbol{B}}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{b=1}^{B} \mathrm{P}_1^X(\mathcal{X}^*_b) \cdot \left(\frac{(\kappa^*_b)^2}{(\tau^*_{1,b})^2} \cdot \left|\lambda^*_{1,1,b} - \lambda^*_{2,1,b}\right| \right. \\
&\qquad\qquad + \frac{1}{2\tau^*_{1,b}} \cdot \left|\left|k^*_{1,b} - k^*_{2,b}\right|\right|_{L_1((\mathrm{P}^*_{1,b})^X \otimes (\mathrm{P}^*_{1,b})^X)} \\
&\qquad\qquad \left. + \frac{\kappa^*_b}{\tau^*_{1,b}} \cdot \sqrt{\left|\left|k^*_{1,b} - k^*_{2,b}\right|\right|_{L_1((\mathrm{P}^*_{1,b})^X \otimes (\mathrm{P}^*_{1,b})^X)}} \right)
\end{aligned} \tag{38}$$

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, (38) 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 $\mathrm{P}^X_{1,\mathcal{X}^{(i)}_{a(i,b)}}$ from (10), yields

$$\begin{aligned}
\left|\lambda^*_{1,1,b} - \lambda^*_{2,1,b}\right| &= \frac{1}{\mathrm{P}_1^X(\mathcal{X}^*_b)} \cdot \left|\lambda_{1,a(1,b)} \cdot \mathrm{P}_1^X(\mathcal{X}^{(1)}_{a(1,b)}) - \lambda_{2,a(2,b)} \cdot \mathrm{P}_1^X(\mathcal{X}^{(2)}_{a(2,b)})\right| \\
&\leq \frac{1}{\mathrm{P}_1^X(\mathcal{X}^*_b)} \cdot \left(\lambda_{1,a(1,b)} \cdot \left|\mathrm{P}_1^X(\mathcal{X}^{(1)}_{a(1,b)}) - \mathrm{P}_1^X(\mathcal{X}^{(2)}_{a(2,b)})\right| \right. \\
&\qquad\qquad \left. + \mathrm{P}_1^X(\mathcal{X}^{(2)}_{a(2,b)}) \cdot \left|\lambda_{1,a(1,b)} - \lambda_{2,a(2,b)}\right| \right)
\end{aligned}$$

as well as analogously

$$\begin{aligned}
\left|\lambda^*_{1,1,b} - \lambda^*_{2,1,b}\right| \leq \frac{1}{\mathrm{P}_1^X(\mathcal{X}^*_b)} \cdot &\left(\lambda_{2,a(2,b)} \cdot \left|\mathrm{P}_1^X(\mathcal{X}^{(1)}_{a(1,b)}) - \mathrm{P}_1^X(\mathcal{X}^{(2)}_{a(2,b)})\right| \right. \\
&\left. + \mathrm{P}_1^X(\mathcal{X}^{(1)}_{a(1,b)}) \cdot \left|\lambda_{1,a(1,b)} - \lambda_{2,a(2,b)}\right| \right),
\end{aligned}$$

and hence

$$\begin{aligned}
&\left|\lambda^*_{1,1,b} - \lambda^*_{2,1,b}\right| \\
&\leq \frac{1}{\mathrm{P}_1^X(\mathcal{X}^*_b)} \cdot \left(\tau_b \cdot \left|\mathrm{P}_1^X(\mathcal{X}^{(1)}_{a(1,b)}) - \mathrm{P}_1^X(\mathcal{X}^{(2)}_{a(2,b)})\right| + \rho_{1,b} \cdot \left|\lambda_{1,a(1,b)} - \lambda_{2,a(2,b)}\right| \right).
\end{aligned}$$

Plugging this into (38) finally yields

$$\begin{aligned}
&\left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{1},\mathbf{1}},\boldsymbol{k}^*_{\mathbf{1}},\boldsymbol{\mathcal{X}}^*_{\boldsymbol{B}}} - f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{2},\mathbf{1}},\boldsymbol{k}^*_{\mathbf{2}},\boldsymbol{\mathcal{X}}^*_{\boldsymbol{B}}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{b=1}^{B} \Bigg(\rho_{1,b} \cdot \frac{\kappa_b^2}{\tau_b^2} \cdot \left|\lambda_{1,a(1,b)} - \lambda_{2,a(2,b)}\right| \\
&\qquad + \mathrm{P}_1^X(\mathcal{X}_b^*) \cdot \Bigg(\frac{1}{2\tau_b} \cdot \left|\left|k^*_{1,b} - k^*_{2,b}\right|\right|_{L_1((\mathrm{P}^*_{1,b})^X \otimes (\mathrm{P}^*_{1,b})^X)} \\
&\qquad\qquad + \frac{\kappa_b}{\tau_b} \cdot \sqrt{\left|\left|k^*_{1,b} - k^*_{2,b}\right|\right|_{L_1((\mathrm{P}^*_{1,b})^X \otimes (\mathrm{P}^*_{1,b})^X)}} \Bigg) \\
&\qquad + \frac{\kappa_b^2}{\tau_b} \cdot \left|\mathrm{P}_1^X(\mathcal{X}^{(1)}_{a(1,b)}) - \mathrm{P}_1^X(\mathcal{X}^{(2)}_{a(2,b)})\right| \Bigg) .
\end{aligned}$$

(iii) The third norm on the right hand side of (36) 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

$$f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{2},\mathbf{1}},\boldsymbol{k}^*_{\mathbf{2}},\boldsymbol{\mathcal{X}}^*_{\boldsymbol{B}}} = \sum_{b=1}^{B} \hat{f}_{\mathrm{P}^*_{1,b},\lambda^*_{2,1,b},k^*_{2,b}} = \sum_{a=1}^{A_2} \sum_{b \in J_{2,a}} \hat{f}_{\mathrm{P}^*_{1,b},\lambda^*_{2,1,b},k^*_{2,b}} = \sum_{a=1}^{A_2} \hat{f}_{\mathrm{P}_{1,\mathcal{X}_a^{(2)}},\lambda_{2,a},k^\circ_{2,a}} ,$$

where $k^\circ_{2,a} := \sum_{b \in J_{2,a}} \tilde{k}^*_{2,b}$ for $a = 1, \ldots, A_2$. We can thus also interpret the third difference on the right hand side of (36) as the difference between two localized SVMs that are based on the same regionalization $\boldsymbol{\mathcal{X}}^{(\mathbf{2})}_{\boldsymbol{A_2}}$ (and on the same vector of regularization parameters) and apply Theorem 6:

$$\begin{aligned}
&\left|\left|f_{\mathrm{P}_1,\boldsymbol{\lambda}^*_{\mathbf{2},\mathbf{1}},\boldsymbol{k}^*_{\mathbf{2}},\boldsymbol{\mathcal{X}}^*_{\boldsymbol{B}}} - f_{\mathrm{P}_2,\boldsymbol{\lambda}_{\mathbf{2}},\boldsymbol{k}_{\mathbf{2}},\boldsymbol{\mathcal{X}}^{(\mathbf{2})}_{\boldsymbol{A_2}}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
&= \left|\left|\sum_{a=1}^{A_2} \hat{f}_{\mathrm{P}_{1,\mathcal{X}_a^{(2)}},\lambda_{2,a},k^\circ_{2,a}} - f_{\mathrm{P}_2,\boldsymbol{\lambda}_{\mathbf{2}},\boldsymbol{k}_{\mathbf{2}},\boldsymbol{\mathcal{X}}^{(\mathbf{2})}_{\boldsymbol{A_2}}}\right|\right|_{L_1(\mathrm{P}_1^X)} \\
&\leq |L|_1 \cdot \sum_{a=1}^{A_2} \mathrm{P}_1^X(\mathcal{X}_a^{(2)}) \cdot \Bigg(\frac{||k_{2,a}||_\infty^2}{\lambda_{2,a}} \cdot \left|\left|\mathrm{P}_{1,\mathcal{X}_a^{(2)}} - \mathrm{P}_{2,\mathcal{X}_a^{(2)}}\right|\right|_{tv} \\
&\qquad + \frac{||k_{2,a}||_\infty^2}{2\lambda_{2,a}} \cdot \sum_{b \in J_{2,a}} \mathrm{P}^X_{1,\mathcal{X}_a^{(2)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_a^{(2)}}(\mathcal{X}_b^*)\right) \\
&\qquad + \frac{||k_{2,a}||_\infty^2}{\lambda_{2,a}} \cdot \sqrt{\sum_{b \in J_{2,a}} \mathrm{P}^X_{1,\mathcal{X}_a^{(2)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_a^{(2)}}(\mathcal{X}_b^*)\right)} \Bigg) ,
\end{aligned}$$

where we employed that $\max\left\{||k_{2,a}||_\infty, \left|\left|k^\circ_{2,a}\right|\right|_\infty\right\} = ||k_{2,a}||_\infty$ and

$$\left|\left|k_{2,a} - k^\circ_{2,a}\right|\right|_{L_1(\mathrm{P}^X_{1,\mathcal{X}_a^{(2)}} \otimes \mathrm{P}^X_{1,\mathcal{X}_a^{(2)}})} \leq ||k_{2,a}||_\infty^2 \cdot \sum_{b \in J_{2,a}} \mathrm{P}^X_{1,\mathcal{X}_a^{(2)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_a^{(2)}}(\mathcal{X}_b^*)\right) ,$$

which follows in the same way as the analogous statements in (i).

Plugging these three bounds into (36) and additionally observing

$$\begin{aligned}
&\sum_{a=1}^{A_i} \mathrm{P}_1^X(\mathcal{X}_a^{(i)}) \cdot \left(\frac{||k_{i,a}||_\infty^2}{2\lambda_{i,a}} \cdot \sum_{b \in J_{i,a}} \mathrm{P}^X_{1,\mathcal{X}_a^{(i)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_a^{(i)}}(\mathcal{X}_b^*)\right) \right. \\
&\qquad \left. + \frac{||k_{i,a}||_\infty^2}{\lambda_{i,a}} \cdot \sqrt{\sum_{b \in J_{i,a}} \mathrm{P}^X_{1,\mathcal{X}_a^{(i)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_a^{(i)}}(\mathcal{X}_b^*)\right)} \right) \\
&\leq \sum_{a=1}^{A_i} \sum_{b \in J_{i,a}} \mathrm{P}_1^X(\mathcal{X}_a^{(i)}) \cdot \frac{||k_{i,a}||_\infty^2}{\lambda_{i,a}} \cdot \left(\frac{1}{2} \cdot \mathrm{P}^X_{1,\mathcal{X}_a^{(i)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_a^{(i)}}(\mathcal{X}_b^*)\right) \right. \\
&\qquad \left. + \sqrt{\mathrm{P}^X_{1,\mathcal{X}_a^{(i)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}_a^{(i)}}(\mathcal{X}_b^*)\right)} \right) \\
&\leq \sum_{b=1}^{B} \rho_{1,b} \cdot \frac{\kappa_b^2}{\tau_b} \cdot \left(\frac{\mathrm{P}^X_{1,\mathcal{X}^{(i)}_{a(i,b)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}^{(i)}_{a(i,b)}}(\mathcal{X}_b^*)\right)}{2} \right. \\
&\qquad \left. + \sqrt{\mathrm{P}^X_{1,\mathcal{X}^{(i)}_{a(i,b)}}(\mathcal{X}_b^*) \cdot \left(1 - \mathrm{P}^X_{1,\mathcal{X}^{(i)}_{a(i,b)}}(\mathcal{X}_b^*)\right)} \right),
\end{aligned}$$

$i = 1, 2$, yields the assertion. ■

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

K. P. Bennett and J. A. Blue. A support vector machine approach to decision trees. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence*, volume 3, pages 2396–2401, 1998.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science+Business Media, New York, 2004.

P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1995.

E. Blanzieri and A. Bryl. Instance-based spam filtering using SVM nearest neighbor classifier. In *Proceedings of FLAIRS Conference*, pages 441–442, 2007.

E. Blanzieri and F. Melgani. Nearest neighbor classification of remote sensing images with the maximal margin principle. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6):1804–1811, 2008.

I. Blaschzyk. *Improved Classification Rates for Localized Algorithms under Margin Conditions*. Springer, Wiesbaden, 2020.

L. Bottou and V. Vapnik. Local learning algorithms. *Neural Computation*, 4(6):888–900, 1992.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

F. Chang, C.-Y. Guo, X.-R. Lin, and C.-J. Lu. Tree decomposition for large-scale SVM problems. *Journal of Machine Learning Research*, 11:2935–2972, 2010.

H. Cheng, P.-N. Tan, and R. Jin. Localized support vector machine and its efficient algorithm. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 461–466, 2007.

H. Cheng, P.-N. Tan, and R. Jin. Efficient algorithm for localized support vector machine. *IEEE Transactions on Knowledge and Data Engineering*, 22(4):537–549, 2010.

A. Chernih, I. H. Sloan, and R. S. Womersley. Wendland functions with increasing smoothness converge to a Gaussian. *Advances in Computational Mathematics*, 40:185–200, 2014.

A. Christmann and R. Hable. Consistency of support vector machines using additive kernels for additive models. *Computational Statistics & Data Analysis*, 56(4):854–873, 2012.

A. Christmann and I. Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 5:1007–1034, 2004.

A. Christmann and I. Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

A. Christmann and A. Van Messem. Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research*, 9:915–936, 2008.

A. Christmann and D.-X. Zhou. Learning rates for the risk of kernel-based quantile regression estimators in additive models. *Analysis and Applications*, 14(3):449–477, 2016a.

A. Christmann and D.-X. Zhou. On the robustness of regularized pairwise learning methods based on kernels. *Journal of Complexity*, 37:1–33, 2016b.

A. Christmann, I. Steinwart, and M. Hubert. Robust learning from bites for data mining. *Computational Statistics & Data Analysis*, 52(1):347–361, 2007.

A. Christmann, A. Van Messem, and I. Steinwart. On consistency and robustness properties of support vector machines for heavy-tailed distributions. *Statistics and Its Interface*, 2 (3):311–327, 2009.

A. Christmann, D. Xiang, and D.-X. Zhou. Total stability of kernel methods. *Neurocomputing*, 289:101–118, 2018.

F. Cucker and D.-X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2007.

Z. Denkowski, S. Migórski, and N. S. Papageorgiou. *An Introduction to Nonlinear Analysis: Theory*. Springer Science+Business Media, New York, 2003.

L. Devroye. 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, 1982.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Stochastic Modelling and Applied Probability. Springer, New York, 1996.

J. Diestel. *Sequences and Series in Banach Spaces*. Graduate Texts in Mathematics. Springer, New York, 1984.

J. Diestel and J. J. Uhl. *Vector Measures*. American Mathematical Society, Providence, Rhode Island, 1977.

F. Dumpert. Quantitative robustness of localized support vector machines. *Communications on Pure & Applied Analysis*, 19(8):3947–3956, 2020.

F. Dumpert and A. Christmann. Universal consistency and robustness of localized support vector machines. *Neurocomputing*, 315:96–106, 2018.

M. Eberts and I. Steinwart. Optimal learning rates for least squares SVMs using Gaussian kernels. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1539–1547, 2011.

M. Eberts and I. Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electronic Journal of Statistics*, 7:1–42, 2013.

M. Farooq and I. Steinwart. Learning rates for kernel-based expectile regression. *Machine Learning*, 108:203–227, 2019.

P. Gensler and A. Christmann. On the robustness of kernel-based pairwise learning. *arXiv preprint arXiv:2010.15527*, 2020. Accepted.

Q. Gu and J. Han. Clustered support vector machines. In *Artificial Intelligence and Statistics*, pages 307–315, 2013.

Z.-C. Guo, S.-B. Lin, and D.-X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017a.

Z.-C. Guo, Y. Ying, and D.-X. Zhou. Online regularized learning with pairwise loss functions. *Advances in Computational Mathematics*, 43(1):127–150, 2017b.

R. Hable. Universal consistency of localized versions of regularized kernel methods. *Journal of Machine Learning Research*, 14:153–186, 2013.

R. Hable and A. Christmann. On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, 102:993–1007, 2011.

E. Hewitt and R. E. Hewitt. The Gibbs-Wilbraham phenomenon: An episode in fourier analysis. *Archive for History of Exact Sciences*, 21(2):129–160, 1979.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In N. M. LeCam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 221–233, Berkeley, 1967.

T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. J. Smola, editors, *Kernel Methods: Support Vector Learning*. MIT Press, 1998.

S.-B. Lin, X. Guo, and D.-X. Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18:3202–3232, 2017.

S.-B. Lin, D. Wang, and D.-X. Zhou. Distributed kernel ridge regression with communications. *Journal of Machine Learning Research*, 21:1–38, 2020.

M. Meister and I. Steinwart. Optimal learning rates for localized SVMs. *Journal of Machine Learning Research*, 17:1–44, 2016.

R. R. Phelps. *Convex Functions, Monotone Operators and Differentiability*. Number 1364 in Lecture Notes in Mathematics. Springer, Berlin, 1993.

J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. J. Smola, editors, *Kernel Methods: Support Vector Learning*. MIT Press, 1998.

A. Rida, A. Labbi, and C. Pellegrini. Local experts combination through density decomposition. In *International Workshop on AI and Statistics, Uncertainty '99*. Morgan Kaufmann, 1999.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts, 2002.

N. Segata and E. Blanzieri. Fast and scalable local kernel machines. *Journal of Machine Learning Research*, 11:1883–1926, 2010.

S. Smale and Y. Yao. Online learning algorithms. *Foundations of Computational Mathematics*, 6(2):145–170, 2006.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.

P. Thomann, I. Blaschzyk, M. Meister, and I. Steinwart. Spatial decompositions for large scale SVMs. In *Artificial Intelligence and Statistics*, pages 1329–1337, 2017.

R. Tibshirani and T. Hastie. Margin trees for high-dimensional classification. *Journal of Machine Learning Research*, 8:637–652, 2007.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

V. N. Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications and Control. Wiley, New York, 1998.

V. N. Vapnik and L. Bottou. Local algorithms for pattern recognition and dependencies estimation. *Neural Computation*, 5(6):893–909, 1993.

H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2005.

D. Wu, K. P. Bennett, N. Cristianini, and J. Shawe-Taylor. Large margin trees for induction and transduction. In *Proceedings of the 17th International Conference on Machine Learning*, pages 474–483, 1999.

D.-H. Xiang and D.-X. Zhou. Classification with Gaussians and convex loss. *Journal of Machine Learning Research*, 10:1447–1468, 2009.

Y. Ying and D.-X. Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788, 2006.

A. Zakai and Y. Ritov. Consistency and localizability. *Journal of Machine Learning Research*, 10:827–856, 2009.

H. Zhang, C. Berg, A, M. Maire, and J. Malik. 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, 2006.

Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015.