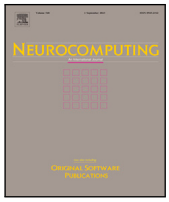
Neurocomputing 598 (2024) 128060

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Lp- and risk consistency of localized SVMs

Hannes Köhler

Department of Mathematics, University of Bayreuth, 95440 Bayreuth, Germany

ARTICLE INFO

Communicated by Y. Tian

Keywords:
Localized learning
Consistency
Kernel methods
Support vector machines
Big data

ABSTRACT

Kernel-based regularized risk minimizers, also called support vector machines (SVMs), are known to possess many desirable properties but suffer from their super-linear computational requirements when dealing with large data sets. This problem can be tackled by using localized SVMs instead, which also offer the additional advantage of being able to apply different hyperparameters to different regions of the input space. In this paper, localized SVMs are analyzed with regards to their consistency. It is proven that they inherit L_p- as well as risk consistency from global SVMs under very weak conditions. Though there already exist results on the latter of these two properties, this paper significantly generalizes them, notably also allowing the regions that underlie the localized SVMs to change as the size of the training data set increases, which is a situation also typically occurring in practice.

1. Introduction

Kernel-based regularized risk minimizers based on a general loss function, which are also known as (general) support vector machines (SVMs), play an important role in statistical machine learning, which is due to two main reasons: First, they are known to possess many desirable theoretical properties such as universal consistency, statistical robustness and stability, and good learning rates [1–5]. Secondly, they are the solutions of finite-dimensional convex programs [6] and empirically observe good performance [7,8] – at least if the data set is not too large. For large data sets, SVMs however suffer from their computational requirements growing at least quadratically in the number of training samples, with regards to both time and memory [9–11].

There exist different approaches to circumvent this problem, one of them being the use of localized SVMs, which implement the idea of not computing one SVM on the whole input space but instead dividing this input space into different (not necessarily disjoint) regions, computing SVMs on each of these regions, and then joining them together in order to obtain a global predictor. In addition to the computational advantage this approach offers, it can also yield improved predictions as it adds flexibility by allowing for differing underlying hyperparameters being chosen in the different regions. In Section 3.1, we discuss these advantages in more detail, as well as briefly mentioning some of the different approaches for circumventing the computational challenges.

Note that there exist many different approaches on how to divide the input space into regions and thus also on how to compute a localized SVM. The main goal of this paper is not to propose and analyze a new method of computing localized SVMs, but instead to derive new theoretical results which are as general as possible and therefore applicable to many of the different existing methods. As we therefore do not focus on any specific way of localizing the input space and as training and testing times of course depend on the method chosen for localization, the empirical analysis of these computation times is not the main focus of this paper. Some computation times for different amounts of regions are however still given in Example 4.6, and we also give a quick review on some existing analyses of computation times for localized SVMs based on different localization methods in Section 3.1.

To be more specific on the theoretical results derived in this paper, we prove that localized SVMs are risk consistent as well as L_p-consistent under certain mild conditions. Notably, we also allow for the regionalization, which underlies a localized SVM, to change as the size of the data set increases. This is natural to allow for because the regionalization is often not predefined but instead data-dependent in practice (cf. Section 3.1 and also Remark 3.1) and might for example become finer as the size of the data set increases.

Because of SVMs being defined as minimizers of some regularized risk function, risk consistency is the natural type of consistency to consider, and there already exist some results on risk consistency respectively learning rates (which imply risk consistency) of localized SVMs, see [12–15] among others. However, all of these in some aspects offer considerably less generality than the result we derive (see Section 4 for more details). On the other hand, L_p-consistency is of interest as it compares functions themselves instead of their risks, and, to our knowledge, there do not exist any results on L_p-consistency of localized SVMs so far.

E-mail address: hannes.koehler@uni-bayreuth.de.

https://doi.org/10.1016/j.neucom.2024.128060
Received 7 June 2023; Received in revised form 20 February 2024; Accepted 12 June 2024
Available online 18 June 2024

The results derived in this paper build on those from [16], where a general connection between risk consistency and L_p-consistency was established and this connection was used to obtain results on consistency of non-localized SVMs. This paper now transfers these results to the case of localized SVMs. The aforementioned generality of also allowing for regionalizations of the input space that change as the size of the data set increases however evokes the challenge that it is not possible to obtain consistency of localized SVMs directly from that of non-localized ones. For this reason, we use from [16] only those results that concern the general connection between risk consistency and L_p-consistency, but not those on consistency of non-localized SVMs. To overcome the added difficulty that comes from the changes in the regionalization, parts of the deviations that are investigated for proving consistency are bounded in a suitable way by according deviations for functions that do *not* depend on the regionalization (see especially the proof of Lemma A.2).

The paper is organized as follows: Section 2 contains some general prerequisites as well as a formal definition of SVMs, whereas the localized approach is described in more detail in Section 3. The main results can be found in Section 4, and finally, Section 5 gives a short summary.

2. Prerequisites

Before introducing localized SVMs in Section 3.2 and stating our results about their consistency in Section 4, we first need to define the underlying (non-localized) SVMs in more detail as well as state some additional prerequisites.

Given a training data set $D_n := ((x_1, y_1), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ consisting of independent and identically distributed (i.i.d.) observations sampled from some unknown probability measure P on a space $\mathcal{X} \times \mathcal{Y}$, we aim at learning a function $f : \mathcal{X} \to \mathbb{R}$. More specifically, we denote by (X, Y) a pair of random variables with values in $\mathcal{X} \times \mathcal{Y}$ distributed according to P, and the goal is to estimate certain characteristics of the conditional distribution $\mathrm{P}(\cdot \mid X)$ of Y given X. We impose the following standard and not very restrictive assumptions on the underlying space $\mathcal{X} \times \mathcal{Y}$ throughout this paper:

Assumption 2.1. Let $\mathcal{X}$ be a complete separable metric space and let $\mathcal{Y} \subseteq \mathbb{R}$ be closed. Let $\mathcal{X}$ and $\mathcal{Y}$ be equipped with their respective Borel σ-algebras $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$. Let $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, where $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ denotes the set of all Borel probability measures on the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X} \times \mathcal{Y}})$.

Notably, $\mathcal{Y} \subseteq \mathbb{R}$ guarantees that the conditional probability $\mathrm{P}(\cdot \mid X)$ does indeed uniquely exist [17, Theorems 10.2.1 and 10.2.2], because $\mathcal{Y}$ is Polish [18, p. 157].

Which exact characteristics of $\mathrm{P}(\cdot \mid X)$ are to be learned is determined by the chosen *loss function*, which is a measurable function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$. For example, estimating the conditional mean function can be approached by using the least squares loss, and conditional quantile functions can be estimated by using the pinball loss. $L(x, y, f(x))$ quantifies the loss associated with predicting $f(x)$ while the true output belonging to x is y, and the goal is to find a predictor whose expected loss is as small as possible. To this end, we call

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

L-risk (or just *risk*) of a measurable function f, and

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

Bayes risk. We call a measurable function $f^*_{L,\mathrm{P}}$ achieving $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ a *Bayes function*.

A sequence $(f_n)_{n \in \mathbb{N}}$ is called *risk consistent* if

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

in probability, and it is called L_p*-consistent* for some $p \in [1, \infty)$ if

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

in probability, where P^X denotes the marginal distribution on $\mathcal{X}$ associated with P. For the latter consistency property, we always assume $f^*_{L,\mathrm{P}}$ to P^X-almost surely (a.s.) uniquely exist. As mentioned in the introduction, the notion of L_p-consistency does directly depend on the difference between the functions instead of on the difference between their risks, which additionally depends on the loss function and the conditional distribution of Y.

As P is unknown, it is not possible to minimize $\mathcal{R}_{L,\mathrm{P}}$ directly and one instead has to use the *empirical risk*

$$\mathcal{R}_{L,\mathrm{D}_n}(f) := \mathbb{E}_{\mathrm{D}_n}\left[L(X, Y, f(X))\right] = \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, f(x_i)),$$

where

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

is the empirical distribution corresponding to D_n, with $\delta_{(x_i, y_i)}$ denoting the Dirac measure in (x_i, y_i). In order to avoid overfitting, a regularization term is added to this empirical risk, which results in the *empirical SVM* being defined as the solution of the minimization problem

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

Here, $\lambda > 0$ controls the amount of regularization and H is a *reproducing kernel Hilbert space* (RKHS) over $\mathcal{X}$. Each such RKHS is associated with a *kernel* on $\mathcal{X}$, which is a symmetric and positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We call k bounded if $\|k\|_\infty := \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty$. We refer to [19–21] for a detailed introduction of kernels, RKHSs and their properties.

By the empirical representer theorem [cf. 2, Theorem 4.2], such empirical SVMs always take the form

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

for some $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. In practice, these unknown coefficients are obtained by solving a suitable convex optimization program, see for example [2, Section 9.2].

The goal of Section 4 is to derive L_p- respectively risk consistency of localized versions of such SVMs as the size n of the data set increases. As an intermediate step in the according proofs, we additionally need the *theoretical SVM*

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

As a last part of these prerequisites, we need to specify some properties of loss functions. We only investigate loss functions which are convex – by which we mean convexity in the last argument of L – and additionally distance-based. The latter is a property that is satisfied by most of the typical loss functions for regression tasks, but not necessarily by those used in classification tasks. However, some distance-based losses are also popular choices in classification tasks, like for example the least squares loss, cf. [22, Section 1.4].

Definition 2.2. A loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called *distance-based* if there exists a representing function $\psi : \mathbb{R} \to [0, \infty)$ satisfying $\psi(0) = 0$ and $L(x, y, t) = \psi(y - t)$ for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$.

Let $p \in (0, \infty)$. A distance-based loss $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ with representing function ψ is of

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

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

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

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

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

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

For example, the aforementioned least squares loss and pinball loss are of growth type 2 and 1 respectively. Depending on the growth type p, our results require that the *averaged pth moment* of P is finite, which guarantees that there exists a function in H that has finite risk. This averaged pth moment is defined as

$$|P|_p := \left(\int_{\mathcal{X}} \int_Y |y|^p \, dP(y \mid x) \, dP^X(x) \right)^{1/p} = \left(\int_{\mathcal{X}} |P(\cdot \mid x)|_p^p \, dP^X(x) \right)^{1/p},$$

thus making the moment condition $|P|_p < \infty$ slightly more restrictive when dealing with loss functions of a higher growth type. In the definition of the averaged pth moment, $|P(\cdot \mid x)|_p$ denotes the *pth moment* of $P(\cdot \mid x)$, where for an arbitrary distribution Q on $\mathcal{Y}$ this pth moment is defined by

$$|Q|_p := \left(\int_{\mathcal{Y}} |y|^p \, dQ(y) \right)^{1/p}.$$

3. Localized approach

As mentioned in the introduction, SVMs, while possessing many desirable theoretical properties, suffer from their super-linear (with respect to the size of the training data set) computational requirements when dealing with large data sets. There exist different approaches to reduce this computational complexity, one of them being localization. Section 3.1 gives a quick overview of some existing approaches as well as an introduction of the idea behind and the additional advantages of the localization approach. Section 3.2 formally defines localized SVMs and states requirements which the underlying structure, like the regions and the applied kernels, need to satisfy.

3.1. Overview of localized and other approaches

Approaches to reduce the computational complexity of SVMs include using twin SVMs [23–27], online learning approaches such as stochastic gradient descent [28–32], as well as algorithms approximating the kernel matrix via column subsampling [33–36], and random feature approximations of the kernel [37–41], with [42] comparing the last two approaches. Additionally, there are also methods combining multiple of these approaches [43,44].

Closer to the localized approach are methods that decompose the available data set into $m \in \mathbb{N}$ subsets and train m "small" SVMs on these subsets instead of a single "large" one on all of D_n, which can substantially reduce the training time as well as required storage space because of the aforementioned super-linear computational requirements of SVMs. This can for example be done by means of distributed learning [45–50], which randomly splits D_n into subsets, trains an SVM on each such subset, and then averages the resulting m SVMs in order to obtain the final predictor.

In the localized approach, one also trains SVMs on subsets of D_n, but the split of D_n is now obtained in a spatial way – based on some regionalization of the input space $\mathcal{X}$ – instead of randomly. Following early theoretical investigations of such localized approaches [51,52], different methods for obtaining the required regions have been examined. These include decision trees [53–56], k-nearest neighbors (kNN) methods [14,57–60], as well as variants of k-means [61,62]. In comparison to distributed learning, this has the disadvantage that, no matter which method of regionalization is chosen, the process of regionalizing the input space clearly also takes some time for large data sets – albeit considerably less time than just training an SVM on the whole data set –, thus making the computational gain of such a localized approach in the training phase smaller than that of distributed learning. On the other hand, the evaluation of the resulting predictor for a test sample can actually be *significantly faster* in localized approaches than it is in distributed ones: Whereas one has to evaluate each of the m different SVMs (and then average the results) in distributed learning, it suffices to evaluate the one SVM belonging to the region of the test sample in localized learning (if the regions do not overlap).

Several of the referenced publications on localized SVMs also include experimental analyses of how much the respective methods of localization reduce the computation time in comparison to regular SVMs. For example, Chang et al. [54] compared their decision tree based localized SVMs `DTSVM` (among some other approaches) to regular SVMs on different medium-size data sets (ca. 10,000 to 500,000 samples, using about two thirds of them for training) and observed drastic reductions in training time, especially for the larger ones among these data sets. Depending on how the training of SVMs was implemented, the training time for the largest data set was reduced by a factor of almost 3,700 or even 5,800 [54, Figures 5 and 10]. At the same time, `DTSVM` exhibited comparable or even increased test accuracy over regular SVMs [54, Figures 6 and 11] and also drastically reduced testing time [54, Table 4]. Additionally, they also took a look at some large-size data sets (ca. 600,000 to 5,000,000 samples), for which they could not perform a comparison with regular SVMs because of them requiring an excessive amount of training time and memory, but showed that `DTSVM` is able to still perform well on such large-size data sets. Similarly, Segata and Blanzieri [59] compared different variants of their kNN based localized SVMs with regular SVMs on different data sets containing ca. 50,000 to 1,000,000 data samples and also observed decreased training and testing times (by factors ranging up to ca. 100 for the variant `FaLK-SVM`, which was the one observing the highest test accuracy) as well as comparable or even increased test accuracy [59, Tables 5–7]. Gu and Han [62] perform similar comparisons for their k-means based localized SVMs `CSVM` (as well as for some other approaches), however only for data sets containing only ca. 3,000 to 60,000 training samples – which were however very high-dimensional, consisting of up to 784 features. Even for these rather small data sets, `CSVM` exhibits training times that are considerably lower than those of regular SVMs (called "kernel SVM" in their tables), by factors of up to almost 130, while at the same time yielding comparable test accuracy [62, Tables 2 and 3]. In all of these three publications, the library `LIBSVM` [63] was used for computing SVMs. Thomann et al. [11] on the other hand, used their own library `liquidSVM` [64] and looked at very large training data sets of up to almost 10,000,000 samples. They showed that localized SVMs based on the Voronoi partition approach that is built into `liquidSVM` can be computed in few hours and yield good test accuracy (as with the large-size data sets used by Chang et al. [54], it was of course not feasible to also compute regular SVMs for such large data sets in order to compare them). Additionally, by using not only a single but instead multiple machines, they succeeded in obtaining good results in just a little over one day of combined training and testing time even for an enormous training data set consisting of 32,000,000 samples in 631 dimensions – whereas among the other data sets used in the four papers mentioned in this paragraph, there was none that had more than 54 dimensions and at the same time more than 240,000 samples.

Even though the exact training and testing time depends on the method chosen for localization as well as on the exact implementation and therefore differs between these publications, they all observed a drastic reduction compared to the computation time of regular SVMs. In addition to this computational gain, localizing the SVM approach can also yield advantages regarding the *quality of prediction* – compared to distributed learning as well as regular SVMs: Whereas the underlying true function, which one aims to estimate, can of course exhibit discontinuities, SVMs based on a continuous and bounded kernel such as the commonly used Gaussian RBF kernel are always continuous (and bounded) themselves [3, Lemma 4.28]. This can lead to SVMs not accurately modeling the true function near such discontinuities, but instead greatly oscillating and overshooting – an effect that is also

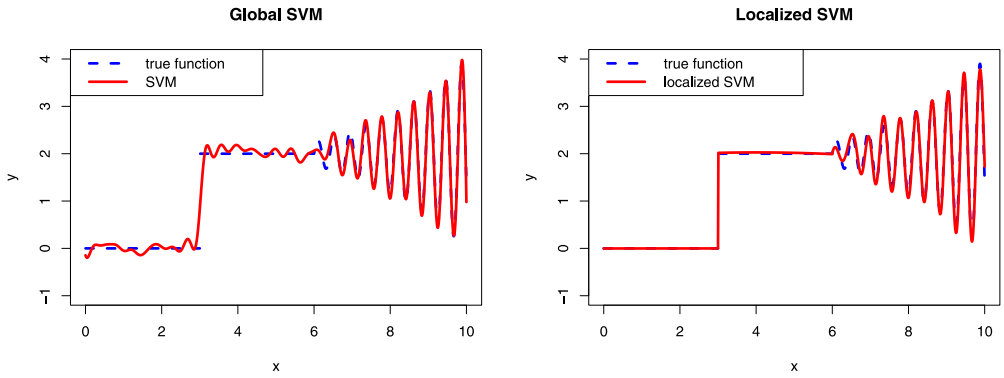


Fig. 3.1. A global SVM (left plot) and a localized SVM (right plot; splits between the regions at $x = 3$ and $x = 6$) fitted to the same data which was generated according to the plotted true function and some normally distributed error. The global SVM (slightly) overshoots at the discontinuity at $x = 3$ and oscillates too much for $x \leq 6$ because the underlying hyperparameters have to be chosen in a way that also allows for a reasonably good fit for $x > 6$, where the true function oscillates very quickly. The localized SVM does not exhibit these problems and yields a considerably better fit overall.

known from Fourier series, where it is called the Gibbs phenomenon, cf. [65]. Additionally, in global learning approaches like SVMs, the complexity of the predictor is usually controlled globally by a very small amount of hyperparameters. Hence, an accurate prediction can be difficult for such global approaches if the complexity and variability of the true function, or that of the conditional distributions $\mathrm{P}(Y \mid X = x)$, greatly differ between different areas of the input space $\mathcal{X}$, even if the true function does not exhibit any discontinuities. Both of these problems can be overcome by the use of localized methods, as a good regionalization can split the input space into separate regions at (or at least close to) discontinuities and such that the complexity and variability do not change too much throughout the individual regions, see also Fig. 3.1.

This intuition of localized SVMs also being able to improve regular SVMs with regard to the quality of prediction gets affirmed by Blaschzyk and Steinwart [12], who, in the case of using the hinge loss for classification, derived learning rates exceeding those known for regular SVMs. Whereas most of the papers on localized SVMs mentioned in the preceding paragraphs focus on the experimental analysis of a specific method of localization, [12] constitutes an example of a paper deriving theoretical results and additionally not requiring any special method of localization (instead only requiring the resulting regionalization to satisfy some conditions which are often quite mild). There are several papers taking a similar approach and also deriving learning rates for such localized SVMs, with [11] also using the hinge loss and [15,66] investigating least squares regression.

Whereas learning rates of course also imply (risk) consistency, they always require additional assumptions regarding the unknown probability measure P because of the no-free-lunch theorem [67], and most of the mentioned papers for example additionally require $\mathcal{X}$ to be contained in some ball and $\mathcal{Y}$ to be bounded as well. We however take an approach similar to Dumpert and Christmann [13], Dumpert [68], Köhler and Christmann [69], who allowed for even more general regionalizations as well as more general kernels and loss functions and did not impose any restrictive assumptions regarding P, and who then proved that localized SVMs are risk consistent (which we in some aspects considerably generalize in Section 4), statistically robust with respect to the maxbias as well as the influence function, and totally stable with respect to simultaneous changes in not only the probability measure but also the regularization parameter, the kernel and the regionalization. We derive results on L_p- as well as risk consistency in Section 4.

3.2. Prerequisites regarding localized SVMs

Before stating our results in Section 4, we first have to formally define localized SVMs as well as to specify the mild assumptions which we need to impose upon the regionalizations in order to be able to then derive our results.

As already mentioned, we actually allow for regionalizations that change with n. For $n \in \mathbb{N}$, we define the regionalization $\boldsymbol{\mathcal{X}}_n$ as $\boldsymbol{\mathcal{X}}_n := \{\mathcal{X}_{n,1}, \ldots, \mathcal{X}_{n,m_n}\}$ for sets $\mathcal{X}_{n,1}, \ldots, \mathcal{X}_{n,m_n}$. We further denote $\boldsymbol{\mathcal{X}}_n(x) := \{\tilde{\mathcal{X}} \in \boldsymbol{\mathcal{X}}_n \mid x \in \tilde{\mathcal{X}}\}$ for all $x \in \mathcal{X}$ and $n \in \mathbb{N}$, and assume the following three conditions to hold true:

(R1) $\mathcal{X}_{n,1}, \ldots, \mathcal{X}_{n,m_n} \subseteq \mathcal{X}$ complete (as metric spaces) and measurable such that $\mathcal{X} = \bigcup_{i=1}^{m_n} \mathcal{X}_{n,i}$ for all $n \in \mathbb{N}$.

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

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

Remark 3.1. Condition **(R3)** might seem restrictive at first glance because it seemingly constitutes a restriction to only using regionalizations whose construction does not take the observed data into account. However, one can easily circumvent this restriction by randomly partitioning the whole data set into not only the usual three parts – namely a training data set D_n, a validation data set and a test data set – but four parts instead, where the fourth part is a regionalization data set. This way, the regionalizations can be chosen data-dependently without violating **(R3)**. By putting only a relatively small part of the available data into the regionalization data set – because one reason for regionalizing is to just reduce the subsequent training time of the SVMs, for which no "perfect" regionalization is necessary –, this procedure does not substantially reduce the amount of data available for training, validating and testing.

Note that **(R1)** tells us that, for every $n \in \mathbb{N}$, the regions need not necessarily be pairwise disjoint but can instead also overlap – as long as **(R2)** is satisfied, that is, as long as the number of regions overlapping does not exceed some global constant $s_{\max}$ in any point $x \in \mathcal{X}$. If the regionalization does not change with n, then **(R2)** is trivially satisfied for $s_{\max} = m_1$.

Remark 3.2. By [70, Lemma I.6.4 and Theorem I.6.12], any subset of a separable metric space is a separable metric space again if it is equipped with the metric of the original space. Hence, Assumption 2.1 being satisfied for $\mathcal{X}$ implies it also being satisfied for the regions $\mathcal{X}_{n,i}$, $n \in \mathbb{N}$ and $i \in \{1, \ldots, m_n\}$.

In order to define local SVMs on the different regions, we need to have a probability measure on each of these regions. It suggests itself

to define these measures by restricting P. For $n \in \mathbb{N}$ and $i \in \{1, \dots, m_n\}$, we define the local measure $P_{n,i}$ on $\mathcal{X}_{n,i} \times \mathcal{Y}$ by

$$P_{n,i} := \begin{cases} \frac{1}{P(\mathcal{X}_{n,i} \times \mathcal{Y})} \cdot P\big|_{\mathcal{X}_{n,i} \times \mathcal{Y}} & , \text{if } P(\mathcal{X}_{n,i} \times \mathcal{Y}) > 0 \\ 0 & , \text{else.} \end{cases}$$

This obviously only is a probability measure if $P(\mathcal{X}_{n,i} \times \mathcal{Y}) > 0$, but we will see that we can mostly ignore the regions with $P(\mathcal{X}_{n,i} \times \mathcal{Y}) = 0$ for our results. We denote

$$I_{\boldsymbol{\mathcal{X}}_n, P} := \big\{ i \in \{1, \dots, m_n\} \mid P(\mathcal{X}_{n,i} \times \mathcal{Y}) > 0 \big\}$$

and $\tilde{m}_n := |I_{\boldsymbol{\mathcal{X}}_n, P}|$ for $n \in \mathbb{N}$. Similarly, we define the local empirical measures $D_{n,i}$ by

$$D_{n,i} := \begin{cases} \frac{1}{D_n(\mathcal{X}_{n,i} \times \mathcal{Y})} \cdot D_n\big|_{\mathcal{X}_{n,i} \times \mathcal{Y}} & , \text{if } D_n(\mathcal{X}_{n,i} \times \mathcal{Y}) > 0 \\ 0 & , \text{else,} \end{cases}$$

such that (if $D_n(\mathcal{X}_{n,i} \times \mathcal{Y}) > 0$) they are the empirical probability measures associated with the subsets $D_{n,i} := D_n \cap (\mathcal{X}_{n,i} \times \mathcal{Y})$ of D_n, for which we denote $d_{n,i} := |D_{n,i}|$.

As mentioned before, one of the goals behind this localized approach is to increase the method's capability to accurately learn a function whose complexity and variability differ between different areas of the input space, by separating these areas into different regions. Since a principal mechanism for controlling the complexity of an SVM is the choice of the regularization parameter and of the kernel (respectively the hyperparameters of the kernel), one should therefore also be allowed to choose different regularization parameters and kernels in the different regions. We hence have, for each $n \in \mathbb{N}$, a vector of regularization parameters $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \dots, \lambda_{n,m_n})$, with $\lambda_{n,i} > 0$ for all $i \in \{1, \dots, m_n\}$, and a vector of kernels $\boldsymbol{k}_n := (k_{n,1}, \dots, k_{n,m_n})$, where $k_{n,i}$ is a kernel on $\mathcal{X}_{n,i}$ for each $i \in \{1, \dots, m_n\}$.

Based on the regularization parameters, kernels and a loss function L, one obtains from (2) SVMs

$$f_{L,P_{n,i},\lambda_{n,i},k_{n,i}} : \mathcal{X}_{n,i} \to \mathbb{R}, \qquad n \in \mathbb{N},\ i \in \{1, \dots, m_n\},$$

which we call *local SVMs* on $\mathcal{X}_{n,i}$. If $P_{n,i}$ is the zero measure, the above SVM is undefined and we just define it as the zero function, $f_{L,P_{n,i},\lambda_{n,i},k_{n,i}} \equiv 0$, in this case. Analogously, we define the local empirical SVMs

$$f_{L,D_{n,i},\lambda_{n,i},k_{n,i}} : \mathcal{X}_{n,i} \to \mathbb{R}, \qquad n \in \mathbb{N},\ i \in \{1, \dots, m_n\},$$

as in (1), with $f_{L,D_{n,i},\lambda_{n,i},k_{n,i}} \equiv 0$ if $D_{n,i}$ is the zero measure.

Since we want to combine these local SVMs in order to obtain 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 all functions $g : \tilde{\mathcal{X}} \to \mathbb{R}$ on $\tilde{\mathcal{X}} \subseteq \mathcal{X}$, we define the zero-extension $\hat{g} : \mathcal{X} \to \mathbb{R}$ by

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

Now, all that is left to do in order to obtain our global predictors, is to equip the local SVMs with weight functions which pointwisely control the influence of each local SVM in areas where two or more regions overlap. We only impose the following three standard assumptions for weight functions on them:

(W1) $w_{n,i} : \mathcal{X} \to [0,1]$ measurable for all $i \in \{1, \dots, m_n\}$ and $n \in \mathbb{N}$.

(W2) $\sum_{i=1}^{m_n} w_{n,i}(x) = 1$ for all $x \in \mathcal{X}$ and $n \in \mathbb{N}$.

(W3) $w_{n,i}(x) = 0$ for all $x \notin \mathcal{X}_{n,i}$ and all $i \in \{1, \dots, m_n\}$ and $n \in \mathbb{N}$.

Our global predictor $f_{L,P,\boldsymbol{\lambda}_n,\boldsymbol{k}_n,\boldsymbol{\mathcal{X}}_n}$, which we call *localized SVM* even though it is not necessarily an SVM itself, is then defined by

$$f_{L,P,\boldsymbol{\lambda}_n,\boldsymbol{k}_n,\boldsymbol{\mathcal{X}}_n} : \mathcal{X} \to \mathbb{R},\ x \mapsto \sum_{i=1}^{m_n} w_{n,i}(x) \cdot \hat{f}_{L,P_{n,i},\lambda_{n,i},k_{n,i}}(x) \tag{3}$$

for $n \in \mathbb{N}$. Analogously, we define the *empirical localized SVM*

$$f_{L,D_n,\boldsymbol{\lambda}_n,\boldsymbol{k}_n,\boldsymbol{\mathcal{X}}_n} : \mathcal{X} \to \mathbb{R},\ x \mapsto \sum_{i=1}^{m_n} w_{n,i}(x) \cdot \hat{f}_{L,D_{n,i},\lambda_{n,i},k_{n,i}}(x) \tag{4}$$

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

Finally, before stating the consistency results for localized SVMs in Section 4, we introduce the concept of *families of kernels of type* $\boldsymbol{\beta}$ which will be needed in those results.

Definition 3.3. Let I be an index set such that $0 \in I$. For kernels $k^{(r)}$ and constants $\beta^{(r)} \in (0, \infty)$, $r \in I$, we say that $\boldsymbol{k} := (k^{(r)})_{r \in I}$ is a *family of kernels of type* $\boldsymbol{\beta} := (\beta^{(r)})_{r \in I}$ if, for all $r \in I$,

(i) $H^{(r)} \supseteq H^{(0)}$, where $H^{(r)}$ and $H^{(0)}$ are the RKHSs associated with $k^{(r)}$ and $k^{(0)}$ respectively, and
(ii) $\|f\|_{H^{(r)}} \leq \beta^{(r)} \cdot \|f\|_{H^{(0)}}$ for all $f \in H^{(0)}$.

Remark 3.4. By [21, Theorem 2.17] (see also [19, Part I.7] and [20, Section 4.5] for related considerations), condition (i) from Definition 3.3 already implies that there exists some $\beta^{(r)} \in (0, \infty)$ such that (ii) is satisfied as well. Hence, every family of kernels satisfying (i) will also be a family of kernels of type $\boldsymbol{\beta}$ for suitable $\boldsymbol{\beta}$. Furthermore, the same theorem also yields that the two conditions from Definition 3.3 are equivalent to

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

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

Example 3.5. Let $d \in \mathbb{N}$, $\mathcal{X} \subseteq \mathbb{R}^d$ non-empty and I be an index set such that $0 \in I$. For $r \in I$, define $k^{(r)}$ as the Gaussian kernel with bandwidth $\gamma^{(r)} \in (0, \infty)$, that is,

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

By [3, Proposition 4.46], the conditions from Definition 3.3 are satisfied with $\beta^{(r)} := (\gamma^{(0)}/\gamma^{(r)})^{d/2}$ if $\gamma^{(0)} \geq \sup_{r \in I \setminus \{0\}} \gamma^{(r)}$.

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

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

Note that the concept of families of kernels of type $\boldsymbol{\beta}$ also allows for infinite index sets (see also Example 3.5). This will lead to the kernels $k_{n,i}$, $n \in \mathbb{N}$, $i \in \{1, \dots, m_n\}$, being allowed to be chosen from an possibly infinite set of kernels.

4. Consistency of localized SVMs

In the following, we first derive L_p-consistency (Section 4.1) and afterwards risk consistency (Section 4.2) of localized SVMs as defined

in Section 3.2. To our knowledge, there do not exist any results on L_p-consistency of localized SVMs so far, and whereas there do exist results on their risk consistency, our result significantly generalizes those in several ways. Before stating the results, we impose the following assumptions, which we assume to hold true throughout this section:

Assumption 4.1.

- Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss function of growth type $p \in [1,\infty)$.
- Let $\mathcal{X}_n := \{\mathcal{X}_{n,1}, \dots, \mathcal{X}_{n,m_n}\}$, $n \in \mathbb{N}$, be regionalizations satisfying **(R1)**, **(R2)**, **(R3)**, and let $\omega_{n,i}$, $n \in \mathbb{N}$ and $i = 1, \dots, m_n$, be weight functions satisfying **(W1)**, **(W2)**, **(W3)**.
- Let $\ell \in \mathbb{N}$ and let, for $j = 1, \dots, \ell$, $\boldsymbol{k}^{(j)} := (k^{(j,r)})_{r \in I^{(j)}}$ be a family of uniformly bounded and measurable kernels of type $\boldsymbol{\beta}^{(j)} := (\beta^{(j,r)})_{r \in I^{(j)}}$ on $\mathcal{X}$ with separable RKHSs $(H^{(j,r)})_{r \in I^{(j)}}$ such that $H^{(j,0)} \subseteq L_p(\mathrm{P}^X)$ dense. Let, for all $n \in \mathbb{N}$ and $i \in \{1, \dots, m_n\}$,
$$k_{n,i} \in \left\{ k^{(j,r)}\big|_{\mathcal{X}_{n,i} \times \mathcal{X}_{n,i}} : j \in \{1, \dots, \ell\}, r \in I^{(j)} \right\}.$$
- Assume $|\mathrm{P}|_p < \infty$ and $\sup_{n \in \mathbb{N}, i \in I_{\mathcal{X}_n, \mathrm{P}}} |\mathrm{P}_{n,i}|_p < \infty$.

Remark 4.2. The condition $\sup_{n \in \mathbb{N}, i \in I_{\mathcal{X}_n, \mathrm{P}}} |\mathrm{P}_{n,i}|_p < \infty$ is disadvantageous in that it requires knowledge about all regionalizations $\mathcal{X}_n$, $n \in \mathbb{N}$. Because
$$|\mathrm{P}_{n,i}|_p^p = \int_{\mathcal{X}_{n,i}} |\mathrm{P}(\cdot \mid x)|_p^p \, \mathrm{dP}_{n,i}^X(x) \le \sup_{x \in \mathcal{X}} |\mathrm{P}(\cdot \mid x)|_p^p$$
for all $n \in \mathbb{N}$ and $i \in I_{\mathcal{X}_n, \mathrm{P}}$ (and analogously also $|\mathrm{P}|_p^p \le \sup_{x \in \mathcal{X}} |\mathrm{P}(\cdot \mid x)|_p^p$), it however suffices if $\sup_{x \in \mathcal{X}} |\mathrm{P}(\cdot \mid x)|_p < \infty$.

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

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

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

4.1. L_p-Consistency of localized SVMs

The subsequent theorem shows that localized SVMs are indeed L_p-consistent under Assumption 4.1.

Theorem 4.3. *Let Assumptions 2.1 and 4.1 be satisfied. Let $f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}$, $n \in \mathbb{N}$, be defined as in (4) and assume that $f_{L,\mathrm{P}}^*$ is P^X-a.s. unique. Define $p_1^* := \max\{p+1, p(p+1)/2\}$. Further choose $p_2^* := \max\{2(p-1)/p, p-1\}$ if $p > 1$ and $p_2^* \in (0,\infty)$ arbitrary if $p = 1$. If the regularization parameters satisfy $\lambda_{n,i} \in (0, C)$ for all $n \in \mathbb{N}$ and $i \in \{1, \dots, m_n\}$ for some $C \in (0,\infty)$, as well as*
$$\max_{i \in I_{\mathcal{X}_n, \mathrm{P}}} \beta_{n,i}^2 \lambda_{n,i} \to 0 \tag{5}$$
and
$$\min_{i \in I_{\mathcal{X}_n, \mathrm{P}}} \frac{\lambda_{n,i}^{p_1^*} d_{n,i}}{\tilde{m}_n^{p_2^*}} \to \infty \tag{6}$$
as $n \to \infty$, then
$$\lim_{n \to \infty} \left\| f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n} - f_{L,\mathrm{P}}^* \right\|_{L_p(\mathrm{P}^X)} = 0 \qquad \text{in probability } \mathrm{P}^\infty.$$

Theorem 4.3 yields that localized SVMs can indeed be guaranteed to asymptotically achieve the goal of learning the Bayes function $f_{L,\mathrm{P}}^*$, i.e. the best predictor as measured by the applied loss function, arbitrarily well in the sense of the L_p-norm. This guarantee is given under weak assumptions, which for the most part only concern entities that can be chosen by the person computing the localized SVM, and which can therefore be ensured to hold true.[1] Notably, the conditions that – directly or indirectly – concern the regionalization and therefore also the way this regionalization is obtained, can be ensured for most of the methods mentioned in Section 3.1, with the exception of some of the kNN methods, which do not satisfy **(R2)**.

In the following, Example 4.4 and Remark 4.5 take a look at how conditions (5) and (6) simplify in certain situations that might occur in practical applications of localized SVMs. Afterwards, Example 4.6 empirically shows how the convergence guaranteed by Theorem 4.3 takes place. For this, we look at simulated data rather than real world data sets because for real world data sets the Bayes function would be unknown and it would therefore not be possible to accurately estimate $\left\| f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n} - f_{L,\mathrm{P}}^* \right\|_{L_p(\mathrm{P}^X)}$.

Example 4.4. If $p = 2$, like for the popular least squares loss, we have $p_1^* = 3$ and $p_2^* = 1$ and condition (6) therefore becomes
$$\min_{i \in I_{\mathcal{X}_n, \mathrm{P}}} \frac{\lambda_{n,i}^3 d_{n,i}}{\tilde{m}_n} \to \infty.$$
If $p = 1$, like for the pinball loss or the ε-insensitive loss, we have $p_1^* = 2$ and p_2^* can be chosen arbitrarily small. Hence, condition (6) relaxes even further in this case, becoming
$$\min_{i \in I_{\mathcal{X}_n, \mathrm{P}}} \frac{\lambda_{n,i}^2 d_{n,i}}{\tilde{m}_n^\delta} \to \infty$$
for an arbitrarily small $\delta > 0$.

Remark 4.5. In some special cases, we can slightly simplify the conditions regarding the regularization parameters in Theorem 4.3:

If one only allows for a finite amount of kernels to choose from (instead of a finite amount of families of kernels of type $\boldsymbol{\beta}$), it is obviously possible to view each of these kernels as its own family of kernels with index set $I^{(j)} = \{0\}$ and $\beta^{(j,0)} = 1$ for all $j \in \{1, \dots, \ell\}$, and thus simplify (5) by eliminating $\beta_{n,i}$ from it.

Additionally, if the regionalization $\mathcal{X}_n$ does not change with n, then $\tilde{m}_n$ is constant and we can erase it from (6).

Hence, if both of these hold true (finite amount of kernels and constant regionalization), the conditions regarding the regularization parameters are exactly the same as in [16], where L_p-consistency of non-localized SVMs was derived, with the only difference being that the conditions obviously need to hold true for each region now instead of only globally.

[1] The most notable exception to this is the so-called moment condition from the last part of Assumption 4.1, which is however fundamentally necessary for the existence of a function $f \in L_p(\mathrm{P}^X)$ with finite risk [cf. 3, Lemma 2.38(iii)] and therefore also necessary for even having a P^X-a.s. unique Bayes function that lies in $L_p(\mathrm{P}^X)$.

Table 1
Estimated values of $\left\| f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n} - f^*_{L,\mathrm{P}} \right\|_{L_1(\mathrm{P}^X)}$ for the regression problem Friedman 1 with different training set sizes n and different amounts of underlying regions.

n \ #Reg.	1	3	5	10	20	40	100
600	1.14	1.34	1.39	1.57	1.76	–	–
2,000	0.95	1.02	1.08	1.08	1.25	1.39	–
6,000	0.87	0.88	0.88	0.85	0.93	1.00	1.16
20,000	0.79	0.74	0.74	0.68	0.71	0.72	0.81
60,000	0.70	0.66	0.63	0.58	0.59	0.57	0.62
200,000	–	–	0.54	0.50	0.50	0.46	0.48
600,000	–	–	–	0.44	0.44	0.40	0.40
2,000,000	–	–	–	–	–	0.34	0.33

Example 4.6. We used R Statistical Software [71, v4.2.2] to perform median regression (that is, we used the 0.5-pinball loss function in our SVMs) on synthetic data generated according to the regression problem Friedman 1 from the library `mlbench` [72] as described by Friedman [73]. Here, the input space $\mathcal{X}$ is 10-dimensional and each component of the input is uniformly distributed on $[0,1]$, with however only 5 of these components actually influencing the output y via the function

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

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

We proceeded by generating a regionalization data set of size 10,000, based on which we used a k-means approach to partition $\mathcal{X}$ into 3, 5, 10, 20, 40 and 100 regions. For each of these regionalization choices, we then used `liquidSVM` [64] with the 0.5-pinball loss function to compute according localized SVMs for different training set sizes ranging from $n = 600$ to $n = 2{,}000{,}000$. Additionally, we did the same computations for a regular SVM (i.e. one based on a single global region). We used fixed Gaussian RBF kernels not changing with n. By Example 4.4, Theorem 4.3 then guarantees convergence of $\left\| f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n} - f^*_{L,\mathrm{P}} \right\|_{L_1(\mathrm{P}^X)}$ to 0 whenever $\lambda_{n,i}$ tends to 0 slower than $d_{n,i}^{-1/2}$ (because $\tilde{m}_n$ is constant for each fixed regionalization). For this reason, we chose some constant $c_i > 0$ on each region $\mathcal{X}_i$ (for each regionalization) and then used $\lambda_{n,i} = c_i \cdot d_{n,i}^{-1/3}$.

To empirically verify the postulated convergence, we collected estimations of the resulting values of $\left\| f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n} - f^*_{L,\mathrm{P}} \right\|_{L_1(\mathrm{P}^X)}$ (based on 1,000,000 test data points generated according to Friedman 1 without the random errors in order to obtain evaluations of the Bayes function $f^*_{L,\mathrm{P}}$) in Table 1.[2] It can be seen that – no matter the number of regions – the convergence does indeed seem to take place. For small n, having only a small amount of regions yields better results than having more regions. This is plausible because the number of training points in the individual regions gets too small in the latter case. As n however increases, the localized SVMs that are based on larger numbers of regions quickly catch up or even overtake those that are based on fewer regions, which supports the idea of increasing the number of regions as n increases in practice.

This idea also gets supported by Table 2, where we collected the according computation times (training and testing times combined).[3] The table shows that computation times indeed drastically decrease as the number of regions increases, also recall the analyses of computation times referenced in Section 3.1.

[2] The missing values in the table are due to the according localized SVMs not having been computed – either because of having too few data points in the different regions or because of having so many data points in a single region that the calculations become exceedingly memory-intensive.

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

4.2. Risk consistency of localized SVMs

Now, we can turn our attention to risk consistency of localized SVMs. To our knowledge, the only existing results which explicitly examine risk consistency of localized SVMs are those by Hable [14, Theorem 1] and Dumpert and Christmann [13, Theorem 3.1], both of which are in certain aspects considerably less general than the subsequent Theorem 4.7: Dumpert and Christmann [13] only considered Lipschitz continuous (shifted) loss functions, whereas we take a look at distance-based loss functions, thus covering a different subset of all loss functions, notably also including the popular and not Lipschitz continuous least squares loss. Additionally, Dumpert and Christmann [13] assumed a fixed regionalization and fixed kernels on the different regions, which stay the same independently of the size n of the underlying data set. We however also allow for regionalizations which change with n (cf. Section 3.2), since the regionalization is oftentimes not predefined in practice but instead might change when new data points are added to the data set – for example, becoming finer when n grows. We also allow for kernels that change with n and that are chosen from an possibly infinite set of kernels – for example, Gaussian kernels whose bandwidth decreases as n increases (cf. Example 3.5). Thus, we significantly generalize the investigations by Dumpert and Christmann [13] in these aspects. Hable [14] on the other hand only allows for a bounded output space $\mathcal{Y}$ and only considers the special case of the regionalization stemming from some k-nearest neighbor method. Whereas this approach implicitly also allows for regionalizations which change with n, this makes our Theorem 4.7 applicable to a much wider array of localization methods – even though the k-nearest neighbor approach described by Hable [14] is not one of them because it can lead to condition **(R2)** from Section 3.2 being violated, thus making our result and that of Hable [14] applicable to different situations.

Apart from that, the oracle inequalities by Meister and Steinwart [15], Thomann et al. [11], Mücke [66], Blaschzyk and Steinwart [12] of course also imply risk consistency if the different parameters in these results are chosen accurately. However, these oracle inequalities are only valid for the least squares respectively the hinge loss, whereas we aim at deriving a much more general result which is applicable for the considerably larger class of convex, distance-based loss functions. Additionally, these oracle inequalities require stricter conditions than our consistency results, like for example $\mathcal{X}$ being contained in a ball of fixed radius, $\mathcal{Y}$ being bounded, the kernels all being Gaussian kernels, and also additional requirements regarding the regionalization.

In the subsequent theorem, we derive such a general result on the risk consistency of localized SVMs. Condition (7) in that theorem is slightly more restrictive and complicated than its counterpart (6) in the result on L_p-consistency. However, the additional factor $\lambda_{n,j}^{p_3^*}$ can be eliminated from (7) in several important special cases, thus weakening and simplifying this condition again: If the loss function is of growth type $p = 1$, one directly obtains $p_3^* = 0$, and if the regionalizations underlying the localized SVMs partition $\mathcal{X}$ or $f^*_{L,\mathrm{P}}$ is P^X-a.s. unique, the special cases (i) and (ii) of the theorem also yield similar relaxations.

Table 2
Computation times (training plus testing) in seconds for the regression problem Friedman 1 with different training set sizes n and different amounts of underlying regions.

n \ #Reg.	1	3	5	10	20	40	100
600	12	7	7	5	6	–	–
2,000	29	13	11	8	7	8	–
6,000	85	31	21	14	10	9	12
20,000	537	95	61	33	21	15	14
60,000	1,481	466	167	92	50	29	24
200,000	–	–	1,028	398	172	100	54
600,000	–	–	–	2,410	743	388	175
2,000,000	–	–	–	–	–	3,821	1,103

Theorem 4.7. *Let Assumptions 2.1 and 4.1 be satisfied. Let $f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}$, $n \in \mathbb{N}$, be defined as in (4). Define $p_1^* := \max\{p+1, p(p+1)/2\}$ and $p_3^* := \max\{p-1, p(p-1)/2\}$. Further choose $p_2^* := \max\{2(p-1)/p, p-1\}$ if $p > 1$ and $p_2^* \in (0,\infty)$ arbitrary if $p = 1$. If the regularization parameters satisfy $\lambda_{n,i} \in (0, C)$ for all $n \in \mathbb{N}$ and $i \in \{1, \dots, m_n\}$ for some $C \in (0,\infty)$, as well as $\max_{i \in I_{\mathcal{X}_n,\mathrm{P}}} \beta_{n,i}^2 \lambda_{n,i} \to 0$ and*

$$\min_{i,j \in I_{\mathcal{X}_n,\mathrm{P}}} \frac{\lambda_{n,j}^{p_3^*} \lambda_{n,i}^{p_1^*} d_{n,i}}{\tilde{m}_n^{p_2^*}} \to \infty \tag{7}$$

as $n \to \infty$, then

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

If some additional conditions are satisfied, it is possible to slightly relax assumption (7) regarding the regularization parameters:

(i) If, for all $n \in \mathbb{N}$, the regionalization $\mathcal{X}_n$ is a partition of $\mathcal{X}$, then it suffices if (7) is satisfied for $p_1^* := \max\{2p, p^2\}$ and $p_3^* := 0$.
(ii) If $f_{L,\mathrm{P}}^*$ is P^X-a.s. unique, then it suffices if (7) is satisfied for $p_3^* := 0$.

If $p = 1$, the cases (i) and (ii) can be ignored since they do not yield an actual relaxation because $p_3^* = 0$ then also holds true in the general case. Furthermore, the possible relaxations mentioned in Remark 4.5 are obviously also valid for Theorem 4.7.

Example 4.8. We look at the regression problem Friedman 1 the same way as we did in Example 4.6, notably also choosing the regularization parameters as $\lambda_{n,i} = c_i \cdot d_{n,i}^{-1/3}$ for constants c_i. Theorem 4.7 yields that $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}) - \mathcal{R}_{L,\mathrm{P}}^*$ converges to zero because the special case (ii) of that theorem tells us that condition (7) coincides with (6) in the situation of this example, and the latter condition was explained to be satisfied for this choice of $\lambda_{n,i}$ in Example 4.6. Table 3 shows the resulting values of $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}) - \mathcal{R}_{L,\mathrm{P}}^*$, from which it can be seen that the postulated convergence does indeed take place and that it does so considerably faster than that of $\left\| f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n} - f_{L,\mathrm{P}}^* \right\|_{L_1(\mathrm{P}^X)}$ in Example 4.6.[4]

5. Discussion

In this paper, the L_p- and risk consistency of localized SVMs has been investigated, as localized SVMs can offer reduced computational requirements as well as advantages regarding the quality of the predictions over non-localized SVMs (cf. Section 3.1). We saw that it is possible to derive both types of consistency of localized SVMs under very mild conditions on the underlying probability distribution as well as the applied regionalization and the kernels used in the different local SVMs. Notably, we even allowed for regionalizations which change as the size n of the data set increases – in contrast to [13], where risk consistency of localized SVMs had already been examined, but only for non-changing regionalizations and kernels and for a different subset of loss functions. Hence, we added another entry to the list of properties that localized SVMs inherit from non-localized ones. This further justifies applying localized SVMs to learning problems, especially to those in which non-localized methods struggle, like in big data scenarios or if the function which one wishes to estimate contains discontinuities or exhibits greatly differing complexity and variability across different areas of the input space.

[4] The missing values in the table are due to the according localized SVMs not having been computed – either because of having too few data points in the different regions or because of having so many data points in a single region that the computation becomes exceedingly memory-intensive.

CRediT authorship contribution statement

Hannes Köhler: Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

I would like to thank my PhD supervisor Andreas Christmann for helpful discussions on this topic. The work described in this paper was partially supported by grant CH291/3-1 of the Deutsche Forschungsgemeinschaft.

Appendix A. Auxiliary results

In this section, we prove auxiliary results that are needed in the proofs of Theorems 4.3 and 4.7. In both these results, the difference between $f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}$ and $f_{L,\mathrm{P}}^*$ is examined – the L_p-norm of the difference in the former and the difference between the risks in the latter. In both cases, we do not examine this difference directly, but instead plug in the theoretical localized SVM $f_{L,\mathrm{P},\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}$ as an intermediate step and then examine the difference between $f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}$ and $f_{L,\mathrm{P},\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}$ as well as that between $f_{L,\mathrm{P},\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}$ and $f_{L,\mathrm{P}}^*$. The lemmas from this section deal with these differences.

As the assumptions needed for these lemmas are slightly weaker than those needed in the theorems from Section 4 (and additionally differ between these lemmas), Assumption 4.1 is *not* assumed to hold true in this section, but we will instead explicitly list the required assumptions in the lemmas.

Table 3
Estimated values of $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}) - \mathcal{R}^*_{L,\mathrm{P}}$ for the regression problem Friedman 1 with different training set sizes n and different amounts of underlying regions.

n \ #Reg.	1	3	5	10	20	40	100
600	0.31	0.39	0.41	0.49	0.58	–	–
2,000	0.24	0.26	0.28	0.28	0.36	0.41	–
6,000	0.21	0.21	0.21	0.20	0.23	0.25	0.31
20,000	0.18	0.16	0.16	0.14	0.15	0.15	0.18
60,000	0.15	0.14	0.13	0.11	0.12	0.10	0.12
200,000	–	–	0.10	0.09	0.09	0.08	0.08
600,000	–	–	–	0.07	0.07	0.06	0.06
2,000,000	–	–	–	–	–	0.05	0.04

Lemma A.1. *Let Assumption 2.1 be satisfied. Let $L : \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss function of upper growth type $p \in [1,\infty)$. Let $f_{L,\mathrm{P},\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}$ and $f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}$, $n \in \mathbb{N}$, be defined as in (3) and (4) such that the underlying regionalizations and weight functions satisfy **(R1)**, **(R3)**, **(W1)**, **(W2)**, **(W3)** and $\sup_{n\in\mathbb{N},i\in I_{\mathcal{X}_n,\mathrm{P}}} |\mathrm{P}_{n,i}|_p < \infty$. Assume that, for all $n \in \mathbb{N}$ and $i \in \{1,\dots,m_n\}$, $k_{n,i}$ is a bounded and measurable kernel on $\mathcal{X}_{n,i}$ with separable RKHS $H_{n,i}$, such that $\sup_{n\in\mathbb{N},i\in I_{\mathcal{X}_n,\mathrm{P}}} \|k_{n,i}\|_\infty < \infty$. Define $p_1^* := \max\{p+1, p(p+1)/2\}$. Further choose $p_2^* := \max\{2(p-1)/p, p-1\}$ if $p > 1$ and $p_2^* \in (0,\infty)$ arbitrary if $p = 1$. If the regularization parameters satisfy $\lambda_{n,i} \in (0,C)$ for all $n \in \mathbb{N}$ and $i \in \{1,\dots,m_n\}$ for some $C \in (0,\infty)$, as well as*

$$\min_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \frac{\lambda_{n,i}^{p_1^*} d_{n,i}}{\tilde{m}_n^{p_2^*}} \to \infty \tag{A.1}$$

as $n \to \infty$, then

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

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

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

$$\left| f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}(x) - f_{L,\mathrm{P},\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n}(x) \right| = \left| \sum_{i=1}^{m_n} w_{n,i}(x) \cdot \left(\hat{f}_{\mathrm{D}_n,n,i}(x) - \hat{f}_{\mathrm{P},n,i}(x) \right) \right|$$

$$\le \sum_{i=1}^{m_n} w_{n,i}(x) \cdot \left| \hat{f}_{\mathrm{D}_n,n,i}(x) - \hat{f}_{\mathrm{P},n,i}(x) \right| \le \max_{i\in\{1,\dots,m_n\}} \left| \hat{f}_{\mathrm{D}_n,n,i}(x) - \hat{f}_{\mathrm{P},n,i}(x) \right|$$

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

$$\left\| f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n} - f_{L,\mathrm{P},\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n} \right\|_{L_\infty(\mathrm{P}^X)} \le \max_{i\in\{1,\dots,m_n\}} \left\| \hat{f}_{\mathrm{D}_n,n,i} - \hat{f}_{\mathrm{P},n,i} \right\|_{L_\infty(\mathrm{P}^X)}$$

$$= \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left\| f_{\mathrm{D}_n,n,i} - f_{\mathrm{P},n,i} \right\|_{L_\infty(\mathrm{P}^X_{n,i})} \le \kappa \cdot \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left\| f_{\mathrm{D}_n,n,i} - f_{\mathrm{P},n,i} \right\|_{H_{n,i}} \tag{A.2}$$

for all $n \in \mathbb{N}$, with the last inequality holding true because of [3, Lemma 4.23]. Hence, we start by fixing an $n \in \mathbb{N}$ and an $i \in I_{\mathcal{X}_n,\mathrm{P}}$ and investigating the corresponding difference on the right hand side of (A.2).

First, note that employing [3, Lemma 4.23, equation (5.4) and Lemma 2.38(i)] yields

$$\|f_{\mathrm{P},n,i}\|_\infty \le \|k_{n,i}\|_\infty \cdot \|f_{\mathrm{P},n,i}\|_{H_{n,i}} \le \|k_{n,i}\|_\infty \cdot \mathcal{R}_{\mathrm{P}_{n,i}}(0)^{1/2} \cdot \lambda_{n,i}^{-1/2} \le c_{p,L,\rho,\kappa} \cdot \lambda_{n,i}^{-1/2} \tag{A.3}$$

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

Assume now without loss of generality that $d_{n,i} > 0$ (which by (A.1) has to be satisfied for n sufficiently large), i.e. that $f_{\mathrm{D}_n,n,i}$ is indeed an empirical SVM and not just defined as the zero function. We know from [3, Corollary 5.11] that there exists a function $h_{n,i} : \mathcal{X}_{n,i} \times \mathcal{Y} \to \mathbb{R}$ such that

$$\left\| f_{\mathrm{D}_n,n,i} - f_{\mathrm{P},n,i} \right\|_{H_{n,i}} \le \frac{1}{\lambda_{n,i}} \cdot \left\| \mathbb{E}_{\mathrm{D}_{n,i}} \left[h_{n,i} \boldsymbol{\Phi}_{n,i} \right] - \mathbb{E}_{\mathrm{P}_{n,i}} \left[h_{n,i} \boldsymbol{\Phi}_{n,i} \right] \right\|_{H_{n,i}} \tag{A.4}$$

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

$$\|h_{n,i}\|_{L_s(\mathrm{P}_{n,i})} \le 8^p \cdot c_L \cdot \left(1 + |\mathrm{P}_{n,i}|_p^{p-1} + \|f_{\mathrm{P},n,i}\|_\infty \right)$$

$$\le 8^p \cdot c_L \cdot \left(1 + \rho^{p-1} + c_{p,L,\rho,\kappa}^{p-1} \cdot \lambda_{n,i}^{-(p-1)/2} \right)$$

$$\le \tilde{c}_{p,L,\rho,\kappa} \cdot \lambda_{n,i}^{-(p-1)/2}, \tag{A.5}$$

where we employed (A.3) in the second and $\lambda_{n,i} \le C$ in the third step, and where $c_L \in (0,\infty)$ and $\tilde{c}_{p,L,\rho,\kappa} \in (0,\infty)$ denote constants depending only on L respectively p, L, ρ and κ.

Assume without loss of generality that $p_2^* \le 1$ if $p = 1$. Then, we can apply [3, Lemma 9.2] with $q := p/(p-1)$ if $p > 1$ and $q := 2/p_2^*$ if $p = 1$, which leads to $q^* := \min\{1/2, 1-1/q\} = \min\{1/2, 1/p\} = (p+1)/(2p_1^*)$, to the functions $h_{n,i}\boldsymbol{\Phi}_{n,i}$, $n \in \mathbb{N}$: First of all, with the help of (A.5) we obtain

$$\|h_{n,i}\boldsymbol{\Phi}_{n,i}\|_q := \left(\mathbb{E}_{\mathrm{P}_{n,i}} \left[\|h_{n,i}\boldsymbol{\Phi}_{n,i}\|_{H_{n,i}}^q \right] \right)^{1/q}$$

$$\le \|k_{n,i}\|_\infty \cdot \|h_{n,i}\|_{L_q(\mathrm{P}_{n,i})} \le \kappa \cdot \tilde{c}_{p,L,\rho,\kappa} \cdot \lambda_{n,i}^{-(p-1)/2} < \infty,$$

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

$$\|h_{n,i}(x,y)\boldsymbol{\Phi}_{n,i}(x)\|_{H_{n,i}}^q = |h_{n,i}(x,y)|^q \cdot \|\boldsymbol{\Phi}_{n,i}(x)\|_{H_{n,i}}^q$$

$$= |h_{n,i}(x,y)|^q \cdot k_{n,i}(x,x)^{q/2} \le |h_{n,i}(x,y)|^q \|k_{n,i}\|_\infty^q$$

by the reproducing property, cf. for example [2, Definition 2.9]. Hence, we obtain for all $\varepsilon > 0$, by combining this Lemma 9.2 with (A.4),

$$\mathrm{P}_{n,i}^{d_{n,i}} \left(D_{n,i} \in (\mathcal{X}_{n,i} \times \mathcal{Y})^{d_{n,i}} : \left\| f_{\mathrm{D}_n,n,i} - f_{\mathrm{P},n,i} \right\|_{H_{n,i}} \ge \frac{\varepsilon}{\kappa} \right)$$

$$\le \mathrm{P}_{n,i}^{d_{n,i}} \left(D_{n,i} \in (\mathcal{X}_{n,i} \times \mathcal{Y})^{d_{n,i}} : \left\| \mathbb{E}_{\mathrm{D}_{n,i}} \left[h_{n,i}\Phi n, i \right] - \mathbb{E}_{\mathrm{P}_{n,i}} \left[h_{n,i}\boldsymbol{\Phi}_{n,i} \right] \right\|_{H_{n,i}} \ge \frac{\lambda_{n,i}\varepsilon}{\kappa} \right)$$

$$\le c_q \cdot \left(\frac{\kappa \|h_{n,i}\boldsymbol{\Phi}_{n,i}\|_q}{\lambda_{n,i}\varepsilon d_{n,i}^{q^*}} \right)^q \le c_{q,p,L,\rho,\kappa} \cdot \left(\frac{1}{\lambda_{n,i}^{(p+1)/2}\varepsilon d_{n,i}^{q^*}} \right)^q$$

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

With this, we can now return to investigating the whole global predictors with the help of (A.2): For all $\varepsilon > 0$ and $n \in \mathbb{N}$, we have

$$\mathrm{P}^n \Bigg(D_n \in (\mathcal{X} \times \mathcal{Y})^n : \left\| f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n} - f_{L,\mathrm{P},\lambda_n,\boldsymbol{k}_n,\mathcal{X}_n} \right\|_{L_\infty(\mathrm{P}^X)} \ge \varepsilon$$

$$\left| |D_{n,1}| = d_{n,1}, \dots, |D_{n,m_n}| = d_{n,m_n} \right)$$

$$\le \mathrm{P}^n \Bigg(D_n \in (\mathcal{X} \times \mathcal{Y})^n : \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left\| f_{\mathrm{D}_n,n,i} - f_{\mathrm{P},n,i} \right\|_{H_{n,i}} \ge \frac{\varepsilon}{\kappa}$$

$$\left| |D_{n,1}| = d_{n,1}, \dots, |D_{n,m_n}| = d_{n,m_n} \right)$$

$$\le \sum_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \mathrm{P}_{n,i}^{d_{n,i}} \left(D_{n,i} \in (\mathcal{X}_{n,i} \times \mathcal{Y})^{d_{n,i}} : \left\| f_{\mathrm{D}_n,n,i} - f_{\mathrm{P},n,i} \right\|_{H_{n,i}} \ge \frac{\varepsilon}{\kappa} \right)$$

$$\le c_{q,p,L,\rho,\kappa} \cdot \tilde{m}_n \cdot \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left(\frac{1}{\lambda_{n,i}^{(p+1)/2}\varepsilon d_{n,i}^{q^*}} \right)^q, \tag{A.6}$$

and it remains to further investigate the right hand side:

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

$$\tilde{m}_n \cdot \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left(\frac{1}{\lambda_{n,i}^{(p+1)/2} d_{n,i}^{q^*}} \right)^q = \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left(\frac{\tilde{m}_n^{1/(qq^*)}}{\lambda_{n,i}^{(p+1)/(2q^*)} d_{n,i}} \right)^{qq^*} = \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left(\frac{\tilde{m}_n^{p_2^*}}{\lambda_{n,i}^{p_1^*} d_{n,i}} \right)^{qq^*},$$

which by assumption converges to 0 as $n \to \infty$. Hence, the whole right hand side of (A.6) converges to 0, which completes the proof. □

Lemma A.2. *Let Assumption 2.1 be satisfied. Let $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex, distance-based loss function of upper growth type $p \in [1, \infty)$. Let $\ell \in \mathbb{N}$ and let, for $j = 1, \dots, \ell$, $\boldsymbol{k}^{(j)} := (k^{(j,r)})_{r \in I^{(j)}}$ be a family of measurable kernels of type $\boldsymbol{\beta}^{(j)} := (\beta^{(j,r)})_{r \in I^{(j)}}$ on $\mathcal{X}$ with RKHSs $(H^{(j,r)})_{r \in I^{(j)}}$ such that $H^{(j,0)} \subseteq L_p(\mathrm{P}^X)$ dense. Assume that $|\mathrm{P}|_p < \infty$. Let $f_{L,\mathrm{P},\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\mathcal{X}_n}$, $n \in \mathbb{N}$, be defined as in (3) such that the underlying regionalizations and weight functions satisfy **(R1)**, **(R2)**, **(W1)**, **(W2)** and **(W3)**, and such that*

$$k_{n,i} \in \{k^{(j,r)}\big|_{\mathcal{X}_{n,i} \times \mathcal{X}_{n,i}} : j \in \{1, \dots, \ell\}, r \in I^{(j)}\}$$

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

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

Proof. Define the inner risk $C_{L,\mathrm{P}(\cdot|x)}$ as

$$C_{L,\mathrm{P}(\cdot|x)}(t) := \int_{\mathcal{Y}} L(y,t)\,\mathrm{dP}(y \mid x) \qquad \forall\, x \in \mathcal{X}, t \in \mathbb{R}$$

and denote by

$$C^*_{L,\mathrm{P}(\cdot|x)} := \inf_{t \in \mathbb{R}} C_{L,\mathrm{P}(\cdot|x)}(t) \qquad \forall\, x \in \mathcal{X}$$

the minimal inner risk at x. We will use these in order to split the risk of a given function (and the Bayes risk) into an outer integral with respect to P^X and the inner risk.

First, we however show that all risks appearing in the assertion are finite: [3, Lemma 2.38(i)] yields $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$ as well as $\mathcal{R}_{L,\mathrm{P}_{n,i}}(0) < \infty$ for all $n \in \mathbb{N}$ and $i \in I_{\mathcal{X}_n,\mathrm{P}}$ (with the latter holding true because $|\mathrm{P}_{n,i}|_p < \infty$ by Remark 4.2). Since $\mathcal{R}^*_{L,\mathrm{P}} \le \mathcal{R}_{L,\mathrm{P}}(0)$ by definition, we obtain the finiteness of $\mathcal{R}^*_{L,\mathrm{P}}$. Furthermore,

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

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

With this, we can now write

$$\begin{aligned}
&\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\mathcal{X}_n}) - \mathcal{R}^*_{L,\mathrm{P}} \\
&= \int_{\mathcal{X}} \left(C_{L,\mathrm{P}(\cdot|x)}(f_{L,\mathrm{P},\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\mathcal{X}_n}(x)) - C^*_{L,\mathrm{P}(\cdot|x)} \right) \mathrm{dP}^X(x) \\
&\le \int_{\mathcal{X}} \sum_{i=1}^{m_n} w_{n,i}(x) \cdot \left(C_{L,\mathrm{P}(\cdot|x)}(\hat{f}_{L,\mathrm{P}_{n,i},\lambda_{n,i},k_{n,i}}(x)) - C^*_{L,\mathrm{P}(\cdot|x)} \right) \mathrm{dP}^X(x) \\
&\le \sum_{i=1}^{m_n} \int_{\mathcal{X}_{n,i}} \left(C_{L,\mathrm{P}(\cdot|x)}(f_{L,\mathrm{P}_{n,i},\lambda_{n,i},k_{n,i}}(x)) - C^*_{L,\mathrm{P}(\cdot|x)} \right) \mathrm{dP}^X(x) \\
&= \sum_{i \in I_{\mathcal{X}_n,\mathrm{P}}} \left(\mathrm{P}(\mathcal{X}_{n,i} \times \mathcal{Y}) \cdot \mathcal{R}_{L,\mathrm{P}_{n,i}}(f_{L,\mathrm{P}_{n,i},\lambda_{n,i},k_{n,i}}) - \int_{\mathcal{X}_{n,i}} C^*_{L,\mathrm{P}(\cdot|x)}\,\mathrm{dP}^X(x) \right),
\end{aligned} \tag{A.7}$$

where we applied [3, Lemma 3.4] in the first, **(W1)**, **(W2)** and the convexity of L in the second, and **(W1)**, **(W3)** and $C_{L,\mathrm{P}(\cdot|x)}(f_{L,\mathrm{P}_{n,i},\lambda_{n,i},k_{n,i}}) - C^*_{L,\mathrm{P}(\cdot|x)} \ge 0$ for all $x \in \mathcal{X}$ (by the definition of $C^*_{L,\mathrm{P}(\cdot|x)}$) in the third step. In the final step, we once more used that $\mathrm{P}(\mathcal{X}_{n,i} \times \mathcal{Y}) = 0$ for $i \notin I_{\mathcal{X}_n,\mathrm{P}}$.

If we define $\tilde{\lambda}_n := \max_{i \in I_{\mathcal{X}_n,\mathrm{P}}} \beta_{n,i}^2 \lambda_{n,i}$ as well as $\tilde{k}_{n,i} \in \{k^{(j,r)} : j \in \{1, \dots, \ell\}, r \in I^{(j)}\}$ such that $\tilde{k}_{n,i}\big|_{\mathcal{X}_{n,i}\times\mathcal{X}_{n,i}} = k_{n,i}$ and analogously $\tilde{k}^{(0)}_{n,i} \in \{k^{(j,0)} : j \in \{1, \dots, \ell\}\}$ such that $\tilde{k}^{(0)}_{n,i}\big|_{\mathcal{X}_{n,i}\times\mathcal{X}_{n,i}} = k^{(0)}_{n,i}$, we can further analyze the right hand side of (A.7) by noting that, for all $n \in \mathbb{N}$ and $i \in I_{\mathcal{X}_n,\mathrm{P}}$,

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

Here, we employed the definition of $f_{L,\mathrm{P}_{n,i},\lambda_{n,i},k_{n,i}}$ and $f_{L,\mathrm{P}_{n,i},\beta^2_{n,i}\lambda_{n,i},k^{(0)}_{n,i}}$ as the minimizers of the respective regularized risks (combined with the fact that $f_{L,\mathrm{P}_{n,i},\beta_{n,i}{}^2\lambda_{n,i},k^{(0)}_{n,i}} \in H^{(0)}_{n,i} \subseteq H_{n,i}$ and that $f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,i}}\big|_{\mathcal{X}_{n,i}} \in H^{(0)}_{n,i}$ by [20, Theorem 6]) in the second and in the fourth step, and again [20, Theorem 6] in the last step. Furthermore, the third step holds true because

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

for all $f \in H^{(0)}_{n,i}$, where we once more applied [20, Theorem 6] and that $\tilde{H}^{(0)}_{n,i} \subseteq \tilde{H}_{n,i}$.

Plugging this into the right hand side of (A.7), we obtain

$$\begin{aligned}
&\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\mathcal{X}_n}) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\le \sum_{i \in I_{\mathcal{X}_n,\mathrm{P}}} \left(\mathrm{P}(\mathcal{X}_{n,i} \times \mathcal{Y}) \cdot \left(\mathcal{R}_{L,\mathrm{P}_{n,i}}(f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,i}}\big|_{\mathcal{X}_{n,i}}) + \tilde{\lambda}_n \cdot \left\| f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,i}} \right\|^2_{\tilde{H}^{(0)}_{n,i}} \right) \right. \\
&\qquad \left. - \int_{\mathcal{X}_{n,i}} C^*_{L,\mathrm{P}(\cdot|x)}\,\mathrm{dP}^X(x) \right) \\
&= \sum_{i \in I_{\mathcal{X}_n,\mathrm{P}}} \left(\mathrm{P}(\mathcal{X}_{n,i} \times \mathcal{Y}) \cdot \tilde{\lambda}_n \cdot \left\| f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,i}} \right\|^2_{\tilde{H}^{(0)}_{n,i}} \right. \\
&\qquad \left. + \int_{\mathcal{X}_{n,i}} \left(C_{L,\mathrm{P}(\cdot|x)}(f_{L,\mathrm{P},\tilde{\lambda}_n,\tilde{k}^{(0)}_{n,i}}(x)) - C^*_{L,\mathrm{P}(\cdot|x)} \right) \mathrm{dP}^X(x) \right) \\
&\le \sum_{j=1}^{\ell} \sum_{i=1}^{m_n} \left(\mathrm{P}(\mathcal{X}_{n,i} \times \mathcal{Y}) \cdot \tilde{\lambda}_n \cdot \left\| f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}} \right\|^2_{H^{(j,0)}} \right. \\
&\qquad \left. + \int_{\mathcal{X}_{n,i}} \left(C_{L,\mathrm{P}(\cdot|x)}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}(x)) - C^*_{L,\mathrm{P}(\cdot|x)} \right) \mathrm{dP}^X(x) \right) \\
&\le \sum_{j=1}^{\ell} s_{\max} \cdot \left(\tilde{\lambda}_n \cdot \left\| f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}} \right\|^2_{H^{(j,0)}} + \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}) - \mathcal{R}^*_{L,\mathrm{P}} \right),
\end{aligned} \tag{A.8}$$

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

$$\sum_{i=1}^{m_n} \int_{\mathcal{X}_{n,i}} \left(C_{L,\mathrm{P}(\cdot|x)}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}(x)) - C^*_{L,\mathrm{P}(\cdot|x)} \right) \mathrm{dP}^X(x)$$

$$= \int_{\mathcal{X}} \sum_{i=1}^{m_n} \mathbb{1}_{\mathcal{X}_{n,i}}(x) \cdot \left(C_{L,\mathrm{P}(\cdot|x)}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}(x)) - C^*_{L,\mathrm{P}(\cdot|x)} \right) \mathrm{d}\mathrm{P}^X(x)$$
$$\leq s_{\max} \cdot \left(\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\tilde{\lambda}_n,k^{(j,0)}}) - \mathcal{R}^*_{L,\mathrm{P}} \right)$$

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

Now, by [3, Lemma 2.38(i)], L is a P-integrable Nemitski loss of order p. Hence, for all $j \in \{1, \ldots, l\}$, we know from [3, Theorem 5.31] that

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

and [3, Lemma 5.15] then yields that

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

because $\tilde{\lambda}_n \to 0$ as $n \to \infty$. Thus, the whole right hand side of (A.8) converges to 0 as $n \to \infty$ and we obtain the assertion because $\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda_n,\boldsymbol{k}_n,\boldsymbol{\mathcal{X}}_n}) - \mathcal{R}^*_{L,\mathrm{P}} \geq 0$ by the definition of $\mathcal{R}^*_{L,\mathrm{P}}$. □

Lemma A.3. *Let Assumption 2.1 be satisfied. Let $L : \mathcal{Y} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss function of upper growth type $p \in [1,\infty)$. Assume that $|\mathrm{P}|_p < \infty$. Let $f_{L,\mathrm{P},\lambda_n,\boldsymbol{k}_n,\boldsymbol{\mathcal{X}}_n}$ and $f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\boldsymbol{\mathcal{X}}_n}$, $n \in \mathbb{N}$, be defined as in (3) and (4) such that the underlying regionalizations and weight functions satisfy **(R1)**, **(R3)**, **(W1)**, **(W2)**, **(W3)** and $\sup_{n\in\mathbb{N}, i\in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} |\mathrm{P}_{n,i}|_p < \infty$. Assume that, for all $n \in \mathbb{N}$ and $i \in \{1,\ldots,m_n\}$, $k_{n,i}$ is a bounded and measurable kernel on $\mathcal{X}_{n,i}$ with separable RKHS $H_{n,i}$, such that $\sup_{n\in\mathbb{N}, i\in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} \|k_{n,i}\|_\infty < \infty$. Define $p_1^* := \max\{p+1, p(p+1)/2\}$ and $p_3^* := \max\{p-1, p(p-1)/2\}$. Further choose $p_2^* := \max\{2(p-1)/p, p-1\}$ if $p > 1$ and $p_2^* \in (0,\infty)$ arbitrary if $p = 1$. If the regularization parameters satisfy $\lambda_{n,i} \in (0, C)$ for all $n \in \mathbb{N}$ and $i \in \{1,\ldots,m_n\}$ for some $C \in (0,\infty)$, as well as*

$$\min_{i,j\in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} \frac{\lambda_{n,j}^{p_3^*} \lambda_{n,i}^{p_1^*} d_{n,i}}{\tilde{m}_n^{p_2^*}} \to \infty \tag{A.9}$$

as $n \to \infty$, then

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

If additionally, the regionalizations $\boldsymbol{\mathcal{X}}_n$, $n \in \mathbb{N}$, are partitions of $\mathcal{X}$, then it suffices if (A.9) is satisfied for $p_1^* := \max\{2p, p^2\}$ and $p_3^* := 0$.

Proof. Assume, for all $n \in \mathbb{N}$ and $i \in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}$, without loss of generality that $d_{n,i} > 0$ (which by (A.9) has to be satisfied for n sufficiently large), such that the respective local empirical SVM $f_{L,\mathrm{D}_{n,i},\lambda_{n,i},k_{n,i}}$ is indeed an empirical SVM and not just defined as the zero function. To shorten the notation, we denote $f_{\mathrm{P},n} := f_{L,\mathrm{P},\lambda_n,\boldsymbol{k}_n,\boldsymbol{\mathcal{X}}_n}$, $f_{\mathrm{D}_n,n} := f_{L,\mathrm{D}_n,\lambda_n,\boldsymbol{k}_n,\boldsymbol{\mathcal{X}}_n}$, $f_{\mathrm{P},n,i} := f_{L,\mathrm{P}_{n,i},\lambda_{n,i},k_{n,i}}$ and $f_{\mathrm{D}_n,n,i} := f_{L,\mathrm{D}_{n,i},\lambda_{n,i},k_{n,i}}$ for all $n \in \mathbb{N}$ and $i \in \{1,\ldots,m_n\}$, as well as $\kappa := \sup_{n\in\mathbb{N}, i\in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} \|k_{n,i}\|_\infty$, $\rho := |\mathrm{P}|_p \vee \sup_{n\in\mathbb{N}, i\in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} |\mathrm{P}_{n,i}|_p$ and $\tilde{\lambda}_n := \min_{i\in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} \lambda_{n,i}$ throughout this proof. Additionally, note that Lemma A.1 is applicable in the situation of this lemma (in the base case as well as in the special case of the regionalizations being partitions of $\mathcal{X}$) as (A.9) in combination with $\lambda_{n,j} \in (0, C)$ for all $n \in \mathbb{N}$ and $j \in \{1,\ldots,m_n\}$ implies the validity of (A.1).

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

By applying [3, Lemma 2.38(ii)] with $q := p$, we obtain

$$|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},n})|$$
$$\leq c_{p,L} \cdot \left(|\mathrm{P}|_p^{p-1} + \|f_{\mathrm{P},n}\|^{p-1}_{L_p(\mathrm{P}^X)} + \left\|f_{\mathrm{D}_n,n}\right\|^{p-1}_{L_p(\mathrm{P}^X)} + 1 \right) \cdot \left\| f_{\mathrm{D}_n,n} - f_{\mathrm{P},n} \right\|_{L_p(\mathrm{P}^X)}, \tag{A.10}$$

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

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

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

with the first inequality following from **(W1)** and **(W2)**, similarly to (A.2), and the last one analogously to (A.3), with $c_{p,L,\rho,\kappa} \in (0,\infty)$ denoting a constant depending only on p, L, ρ and κ. Hence,

$$\|f_{\mathrm{P},n}\|^{p-1}_{L_p(\mathrm{P}^X)} \leq \|f_{\mathrm{P},n}\|^{p-1}_{L_\infty(\mathrm{P}^X)} \leq \max_{i\in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} c^{p-1}_{p,L,\rho,\kappa} \cdot \lambda_{n,i}^{-(p-1)/2} = c^{p-1}_{p,L,\rho,\kappa} \cdot \tilde{\lambda}_n^{-(p-1)/2}. \tag{A.11}$$

Similarly, we obtain

$$\left\|f_{\mathrm{D}_n,n}\right\|^{p-1}_{L_p(\mathrm{P}^X)} \leq \left(\|f_{\mathrm{P},n}\|_{L_p(\mathrm{P}^X)} + \left\|f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}\right\|_{L_p(\mathrm{P}^X)} \right)^{p-1}$$
$$\leq 2^{p-1} \cdot c^{p-1}_{p,L,\rho,\kappa} \cdot \tilde{\lambda}_n^{-(p-1)/2} + 2^{p-1} \cdot \left\|f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}\right\|^{p-1}_{L_p(\mathrm{P}^X)}, \tag{A.12}$$

where we applied (A.11) in the last step.

Plugging (A.11) and (A.12) into (A.10) then yields

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

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

We know from Lemma A.1 that the second summand on the right hand side converges to 0 in probability as $n \to \infty$. Hence, we only need to further investigate the first summand. For this, we can proceed in exactly the same way as in the proof of Lemma A.1 and only need to additionally consider the factor $\tilde{\lambda}_n^{-(p-1)/2}$. By doing this, we obtain for all $\varepsilon > 0$

$$\mathrm{P}^n \left(D_n \in (\mathcal{X} \times \mathcal{Y})^n : \tilde{\lambda}_n^{-(p-1)/2} \cdot \left\|f_{\mathrm{D}_n,n} - f_{\mathrm{P},n}\right\|_{L_\infty(\mathrm{P}^X)} \geq \varepsilon \;\middle|\; |D_{n,1}| = d_{n,1}, \ldots, |D_{n,m_n}| = d_{n,m_n} \right)$$
$$\leq \mathrm{P}^n \left(D_n \in (\mathcal{X} \times \mathcal{Y})^n : \max_{i\in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} \left\|f_{\mathrm{D}_n,n,i} - f_{\mathrm{P},n,i}\right\|_{H_{n,i}} \geq \frac{\varepsilon \tilde{\lambda}_n^{(p-1)/2}}{\kappa} \;\middle|\; |D_{n,1}| = d_{n,1}, \ldots, |D_{n,m_n}| = d_{n,m_n} \right)$$
$$\leq \sum_{i\in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} \mathrm{P}^{d_{n,i}}_{n,i} \left(D_{n,i} \in (\mathcal{X}_{n,i} \times \mathcal{Y})^{d_{n,i}} : \left\|f_{\mathrm{D}_n,n,i} - f_{\mathrm{P},n,i}\right\|_{H_{n,i}} \geq \frac{\varepsilon \tilde{\lambda}_n^{(p-1)/2}}{\kappa} \right)$$
$$\leq c_{q,p,L,\rho,\kappa} \cdot \tilde{m}_n \cdot \max_{i\in I_{\boldsymbol{\mathcal{X}}_n,\mathrm{P}}} \left(\frac{1}{\tilde{\lambda}_n^{(p-1)/2} \lambda_{n,i}^{(p+1)/2} \varepsilon d_{n,i}^{q^*}} \right)^q, \tag{A.13}$$

analogously to (A.6), with $c_{q,p,L,\rho,\kappa} \in (0,\infty)$ denoting a constant depending only on q, p, L, ρ and κ. Here, as in the proof of Lemma A.1, $q := p/(p-1)$ if $p > 1$, $q := 2/p_2^*$ if $p = 1$, and $q^* := \min\{1/2, 1-1/q\} = \min\{1/2, 1/p\} = (p+1)/(2p_1^*) = (p-1)/(2p_3^*)$.

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

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

by assumption. Hence, the whole right hand side of (A.13) converges to 0, which yields the main assertion.

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

$$|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},n})|$$

$$
\begin{aligned}
&= \left| \int_{\mathcal{X}\times\mathcal{Y}} L\left(y, \sum_{i=1}^{m_n} \mathbb{1}_{\mathcal{X}_{n,i}}(x) \hat{f}_{\mathrm{D}_n,n,i}(x)\right) \mathrm{dP}(x,y) \right. \\
&\qquad \left. - \int_{\mathcal{X}\times\mathcal{Y}} L\left(y, \sum_{i=1}^{m_n} \mathbb{1}_{\mathcal{X}_{n,i}}(x) \hat{f}_{\mathrm{P},n,i}(x)\right) \mathrm{dP}(x,y) \right| \\
&= \left| \sum_{i=1}^{m_n} \left(\int_{\mathcal{X}_{n,i}\times\mathcal{Y}} L(y, f_{\mathrm{D}_n,n,i}(x)) \,\mathrm{dP}(x,y) - \int_{\mathcal{X}_{n,i}\times\mathcal{Y}} L(y, f_{\mathrm{P},n,i}(x)) \,\mathrm{dP}(x,y) \right) \right| \\
&\le \sum_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \mathrm{P}(\mathcal{X}_{n,i}\times\mathcal{Y}) \cdot \left| \mathcal{R}_{L,\mathrm{P}_{n,i}}(f_{\mathrm{D}_n,n,i}) - \mathcal{R}_{L,\mathrm{P}_{n,i}}(f_{\mathrm{P},n,i}) \right| \\
&\le \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left| \mathcal{R}_{L,\mathrm{P}_{n,i}}(f_{\mathrm{D}_n,n,i}) - \mathcal{R}_{L,\mathrm{P}_{n,i}}(f_{\mathrm{P},n,i}) \right| \qquad \text{(A.14)}
\end{aligned}
$$

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

The argument of the maximum on the right hand side of (A.14) can, for each $i \in I_{\mathcal{X}_n,\mathrm{P}}$, be examined in the same way as we previously examined the difference on the left hand side for proving the main assertion. A difference appears in (A.11), where we now have

$$
\left\|f_{\mathrm{P},n,i}\right\|_{L_p(\mathrm{P}^X_{n,i})}^{p-1} \le \left\|f_{\mathrm{P},n,i}\right\|_{L_\infty(\mathrm{P}^X_{n,i})}^{p-1} \le c_{p,L,\rho,\kappa}^{p-1} \cdot \lambda_{n,i}^{-(p-1)/2} .
$$

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

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

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

where the second summand on the right hand side converges to 0 in probability by Lemma A.1.

As for the first summand, we can derive

$$
\begin{aligned}
&\mathrm{P}^n\left(D_n \in (\mathcal{X}\times\mathcal{Y})^n : \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left(\lambda_{n,i}^{-(p-1)/2} \cdot \left\|f_{\mathrm{D}_n,n,i} - f_{\mathrm{P},n,i}\right\|_{L_\infty(\mathrm{P}^X_{n,i})} \right) \ge \epsilon \right. \\
&\qquad \left. \Big| \, |D_{n,1}| = d_{n,1}, \dots, |D_{n,m_n}| = d_{n,m_n} \right) \\
&\le c_{q,p,L,\rho,\kappa} \cdot \tilde{m}_n \cdot \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left(\frac{1}{\lambda_{n,i}^{p} \epsilon d_{n,i}^{q^*}} \right)^{q} ,
\end{aligned}
$$

analogously to (A.13). Finally, we obtain convergence to 0 of the right hand side, and thus the assertion, because

$$
\tilde{m}_n \cdot \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left(\frac{1}{\lambda_{n,i}^{p} d_{n,i}^{q^*}} \right)^{q} = \max_{i\in I_{\mathcal{X}_n,\mathrm{P}}} \left(\frac{\tilde{m}_n^{p_2^*}}{\lambda_{n,i}^{p_1^*} d_{n,i}} \right)^{qq^*} \to 0 , \qquad n\to\infty ,
$$

by assumption, where we applied that $p/q^* = p_1^*$ since $p_1^* = \max\{2p, p^2\}$ now. □

Appendix B. Proofs

Proof of Theorem 4.3. We can split up the difference, which we wish to investigate, as

$$
\begin{aligned}
&\left\|f_{L,\mathrm{D}_n,\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n} - f^*_{L,\mathrm{P}}\right\|_{L_p(\mathrm{P}^X)} \\
&\le \left\|f_{L,\mathrm{D}_n,\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n} - f_{L,\mathrm{P},\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n}\right\|_{L_p(\mathrm{P}^X)} + \left\|f_{L,\mathrm{P},\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n} - f^*_{L,\mathrm{P}}\right\|_{L_p(\mathrm{P}^X)} . \qquad \text{(B.1)}
\end{aligned}
$$

Because $\left\|f_{L,\mathrm{D}_n,\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n} - f_{L,\mathrm{P},\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n}\right\|_{L_p(\mathrm{P}^X)} \le \left\|f_{L,\mathrm{D}_n,\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n} - f_{L,\mathrm{P},\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n}\right\|_{L_\infty(\mathrm{P}^X)}$, we know from Lemma A.1 that the first summand on the right hand side converges to 0 in probability as $n\to\infty$.

Thus, only the second summand remains to be examined: From Lemma A.2, we obtain

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

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

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

by **(W1)**, **(W2)** and [3, Lemma 4.23], similarly to (A.2). Employing [16, Theorem 3.2 and Remark 3.3] then yields convergence to 0 (as $n\to\infty$) of the second summand on the right hand side of (B.1), which completes the proof. □

Proof of Theorem 4.7. We start by proving the main assertion and the special case (i): We can split up the difference, which we wish to investigate, as

$$
\begin{aligned}
&|\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n}) - \mathcal{R}^*_{L,\mathrm{P}}| \\
&\le |\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D}_n,\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n}) - \mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n})| + |\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{P},\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n}) - \mathcal{R}^*_{L,\mathrm{P}}| . \qquad \text{(B.2)}
\end{aligned}
$$

The assertions then follow directly by applying Lemma A.3 to the first and Lemma A.2 to the second summand on the right hand side.

As for the special case (ii): If $f^*_{L,\mathrm{P}}$ is P^X-a.s. unique, the assertion follows directly from Theorem 4.3 and [16, Theorem 3.4], which is applicable because $f^*_{L,\mathrm{P}} \in L_p(\mathrm{P}^X)$ (cf. [16, Remark 3.3]) and $f_{L,\mathrm{D}_n,\lambda_{\boldsymbol{n}},\boldsymbol{k}_{\boldsymbol{n}},\boldsymbol{\mathcal{X}}_n} \in L_p(\mathrm{P}^X)$ for all $n\in\mathbb{N}$ (cf. proof of Theorem 4.3). □

References

[1] F. Cucker, D.-X. Zhou, Learning theory: An approximation theory viewpoint, in: Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2007.

[2] B. Schölkopf, A.J. Smola, Learning with kernels, in: Adaptive Computation and Machine Learning, MIT Press, Cambridge, Massachusetts, 2002.

[3] I. Steinwart, A. Christmann, Support vector machines, in: Information Science and Statistics, Springer, New York, 2008.

[4] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[5] V.N. Vapnik, Statistical learning theory, in: Adaptive and Learning Systems for Signal Processing, Communications and Control, Wiley, New York, 1998.

[6] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.

[7] G. Klambauer, T. Unterthiner, A. Mayr, S. Hochreiter, Self-normalizing neural networks, Adv. Neural Inf. Process. Syst. 30 (2017) 971–980.

[8] M.E. Paoletti, J.M. Haut, J. Plaza, A. Plaza, Deep learning classifiers for hyperspectral imaging: A review, ISPRS J. Photogramm. Remote Sens. 158 (2019) 279–317.

[9] T. Joachims, Making large-scale SVM learning practical, in: B. Schölkopf, C. Burges, A.J. Smola (Eds.), Kernel Methods: Support Vector Learning, MIT Press, 1998.

[10] J.C. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Schölkopf, C. Burges, A.J. Smola (Eds.), Kernel Methods: Support Vector Learning, MIT Press, 1998.

[11] P. Thomann, I. Blaschzyk, M. Meister, I. Steinwart, Spatial decompositions for large scale SVMs, Artif. Intell. Stat. (2017) 1329–1337.

[12] I. Blaschzyk, I. Steinwart, Improved classification rates for localized SVMs, J. Mach. Learn. Res. 23 (2022) 1–59.

[13] F. Dumpert, A. Christmann, Universal consistency and robustness of localized support vector machines, Neurocomputing 315 (2018) 96–106.

[14] R. Hable, Universal consistency of localized versions of regularized kernel methods, J. Mach. Learn. Res. 14 (2013) 153–186.

[15] M. Meister, I. Steinwart, Optimal learning rates for localized SVMs, J. Mach. Learn. Res. 17 (2016) 1–44.

[16] H. Köhler, On the connection between Lp- and risk consistency and its implications on regularized kernel methods, 2023, arXiv preprint arXiv:2303.15210.

[17] R.M. Dudley, Real Analysis and Probability, Cambridge University Press, Cambridge, 2004.

[18] H. Bauer, Measure and Integration Theory, de Gruyter Studies in Mathematics, de Gruyter, Berlin, Boston, 2001.
[19] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (3) (1950) 337–404.
[20] A. Berlinet, C. Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics, Springer Science+Business Media, New York, 2004.
[21] S. Saitoh, Y. Sawano, Theory of reproducing kernels and applications, in: volume 44 of Developments in Mathematics, Springer Science+Business Media, Singapore, 2016.
[22] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, in: Springer Series in Statistics, Springer, New York, 2002.
[23] P. Borah, D. Gupta, Robust twin bounded support vector machines for outliers and imbalanced data, Appl. Intell. 51 (2021) 5314–5343.
[24] D. Gupta, U. Gupta, On robust asymmetric Lagrangian ν-twin support vector regression using pinball loss function, Appl. Soft Comput. 102 (2021) 107099.
[25] U. Gupta, D. Gupta, Lagrangian twin-bounded support vector machine based on L2-norm, in: J. Kalita, V.E. Balas, S. Borah, R. Pradhan (Eds.), Recent Developments in Machine Learning and Data Analytics, in: number 740 in Advances in Intelligent Systems and Computing, Springer, Singapore, 2019.
[26] U. Gupta, D. Gupta, On regularization based twin support vector regression with huber loss, Neural Process. Lett. 53 (2021) 459–515.
[27] Jayadeva, R. Khemchandani, S. Chandra, Twin support vector machines for pattern classification, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 905–910.
[28] A. Dieuleveut, F. Bach, Nonparametric stochastic approximation with large step-sizes, Ann. Statist. 44 (4) (2016) 1363–1399.
[29] J. Lin, L. Rosasco, Optimal rates for multi-pass stochastic gradient methods, J. Mach. Learn. Res. 18 (2017) 1–47.
[30] J. Lin, L. Rosasco, D.-X. Zhou, Iterative regularization for learning with convex loss functions, J. Mach. Learn. Res. 17 (2016) 1–38.
[31] S. Smale, Y. Yao, Online learning algorithms, Found. Comput. Math. 6 (2) (2006) 145–170.
[32] Y. Ying, D.-X. Zhou, Online regularized classification algorithms, IEEE Trans. Inform. Theory 52 (11) (2006) 4775–4788.
[33] A. Alaoui, M.W. Mahoney, Fast randomized kernel ridge regression with statistical guarantees, in: Advances in Neural Information Processing Systems, vol. 28, 2015, pp. 775–783.
[34] F. Bach, Sharp analysis of low-rank kernel matrix approximations, in: Conference on Learning Theory, PMLR, 2013, pp. 185–209.
[35] A. Rudi, R. Camoriano, L. Rosasco, Less is more: Nyström computational regularization, Adv. Neural Inf. Process. Syst. 28 (2015) 1657–1665.
[36] C. Williams, M. Seeger, Using the Nyström method to speed up kernel machines, Adv. Neural Inf. Process. Syst. 13 (2001) 682–688.
[37] F. Liu, X. Huang, Y. Chen, J.A.K. Suykens, Random features for kernel approximation: A survey on algorithms, theory, and beyond, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2022) 7128–7148.
[38] S. Mei, T. Misiakiewicz, A. Montanari, Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration, Appl. Comput. Harmon. Anal. 59 (2022) 3–84.
[39] A. Rahimi, B. Recht, Random features for large-scale kernel machines, Adv. Neural Inf. Process. Syst. 20 (2008) 1177–1184.
[40] A. Rudi, L. Rosasco, Generalization properties of learning with random features, Adv. Neural Inf. Process. Syst. 30 (2017) 3215–3225.
[41] B. Sriperumbudur, Z. Szabó, Optimal rates for random Fourier features, Adv. Neural Inf. Process. Syst. 28 (2015) 1144–1152.
[42] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, Z.-H. Zhou, Nyström method vs random fourier features: A theoretical and empirical comparison, Adv. Neural Inf. Process. Syst. 25 (2012) 476–484.
[43] G. Meanti, L. Carratino, L. Rosasco, A. Rudi, Kernel methods through the roof: Handling billions of points efficiently, Adv. Neural Inf. Process. Syst. 33 (2020) 14410–14422.
[44] A. Rudi, L. Carratino, L. Rosasco, FALKON: An optimal large scale kernel method, Adv. Neural Inf. Process. Syst. 30 (2017).
[45] A. Christmann, I. Steinwart, M. Hubert, Robust learning from bites for data mining, Comput. Statist. Data Anal. 52 (1) (2007) 347–361.
[46] Z.-C. Guo, S.-B. Lin, D.-X. Zhou, Learning theory of distributed spectral algorithms, Inverse Problems 33 (7) (2017) 074009.
[47] S.-B. Lin, X. Guo, D.-X. Zhou, Distributed learning with regularized least squares, J. Mach. Learn. Res. 18 (2017) 3202–3232.
[48] S.-B. Lin, D. Wang, D.-X. Zhou, Distributed kernel ridge regression with communications, J. Mach. Learn. Res. 21 (2020) 1–38.
[49] N. Mücke, G. Blanchard, Parallelizing spectrally regularized kernel algorithms, J. Mach. Learn. Res. 19 (2018) 1–29.
[50] Y. Zhang, J. Duchi, M. Wainwright, Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates, J. Mach. Learn. Res. 16 (1) (2015) 3299–3340.
[51] L. Bottou, V. Vapnik, Local learning algorithms, Neural Comput. 4 (6) (1992) 888–900.
[52] V.N. Vapnik, L. Bottou, Local algorithms for pattern recognition and dependencies estimation, Neural Comput. 5 (6) (1993) 893–909.
[53] K.P. Bennett, J.A. Blue, A support vector machine approach to decision trees, in: 1998 IEEE International Joint Conference on Neural Networks Proceedings, in: IEEE World Congress on Computational Intelligence, vol. 3, 1998, pp. 2396–2401.
[54] F. Chang, C.-Y. Guo, X.-R. Lin, C.-J. Lu, Tree decomposition for large-scale SVM problems, J. Mach. Learn. Res. 11 (2010) 2935–2972.
[55] R. Tibshirani, T. Hastie, Margin trees for high-dimensional classification, J. Mach. Learn. Res. 8 (2007) 637–652.
[56] D. Wu, K.P. Bennett, N. Cristianini, J. Shawe-Taylor, Large margin trees for induction and transduction, in: Proceedings of the 17th International Conference on Machine Learning, 1999, pp. 474–483.
[57] E. Blanzieri, A. Bryl, Instance-based spam filtering using SVM nearest neighbor classifier, in: Proceedings of FLAIRS Conference, 2007, pp. 441–442.
[58] E. Blanzieri, F. Melgani, Nearest neighbor classification of remote sensing images with the maximal margin principle, IEEE Trans. Geosci. Remote Sens. 46 (6) (2008) 1804–1811.
[59] N. Segata, E. Blanzieri, Fast and scalable local kernel machines, J. Mach. Learn. Res. 11 (2010) 1883–1926.
[60] H. Zhang, A.C. Berg, M. Maire, J. Malik, SVM-KNN: Discriminative nearest neighbor classification for visual category recognition, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006.
[61] H. Cheng, P.-N. Tan, R. Jin, Efficient algorithm for localized support vector machine, IEEE Trans. Knowl. Data Eng. 22 (4) (2010) 537–549.
[62] Q. Gu, J. Han, Clustered support vector machines, Artif. Intell. Stat. (2013) 307–315.
[63] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. (2011) 2:27:1–27:27, Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.
[64] I. Steinwart, P. Thomann, liquidSVM: A fast and versatile SVM package, 2017, ArXiv e-prints arXiv:1702.06899. Software available at http://pnp.mathematik.uni-stuttgart.de/isa/steinwart/software/liquidSVM.html.
[65] E. Hewitt, R.E. Hewitt, The Gibbs-Wilbraham phenomenon: An episode in fourier analysis, Arch. Hist. Exact Sci. 21 (2) (1979) 129–160.
[66] N. Mücke, Reducing training time by efficient localized kernel regression, in: The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 2603–2610.
[67] L. Devroye, Any discrimination rule can have an arbitrarily bad probability of error for finite sample size, IEEE Trans. Pattern Anal. Mach. Intell. 4 (2) (1982) 154–157.
[68] F. Dumpert, Quantitative robustness of localized support vector machines, Commun. Pure Appl. Anal. 19 (8) (2020) 3947–3956.
[69] H. Köhler, A. Christmann, Total stability of SVMs and localized SVMs, J. Mach. Learn. Res. 23 (100) (2022) 1–41.
[70] N. Dunford, J.T. Schwartz, Linear operators, part I: General theory, in: volume 7 of Pure and Applied Mathematics. A Series of Texts and Monographs, John Wiley & Sons, 1957.
[71] R. Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2022.
[72] Leisch. F., E. Dimitriadou, Mlbench: Machine learning benchmark problems, 2021, R package version 2.1-3.1.
[73] J.H. Friedman, Multivariate adaptive regression splines, Ann. Statist. 19 (1) (1991) 1–141.

Hannes Köhler received his M.Sc. degree in mathematics from the University of Bayreuth, Germany. Currently, he is pursuing a Ph.D. degree at the Department of Mathematics of the University of Bayreuth. His research interests include statistical machine learning, particularly localized learning and kernel methods such as support vector machines.