Journal of Machine Learning Research 17 (2016) 1-44 Submitted 1/14; Revised 8/16; Published 11/16

Optimal Learning Rates for Localized SVMs

Mona Meister mona.meister@de.bosch.com
Corporate Research
Robert Bosch GmbH
70465 Stuttgart, Germany

Ingo Steinwart ingo.steinwart@mathematik.uni-stuttgart.de
Institute for Stochastics and Applications
University of Stuttgart
70569 Stuttgart, Germany

Editor: Sara van de Geer

Abstract

One of the limiting factors of using support vector machines (SVMs) in large scale applications are their super-linear computational requirements in terms of the number of training samples. To address this issue, several approaches that train SVMs on many small chunks separately have been proposed in the literature. With the exception of random chunks, which is also known as divide-and-conquer kernel ridge regression, however, these approaches have only been empirically investigated. In this work we investigate a spatially oriented method to generate the chunks. For the resulting localized SVM that uses Gaussian kernels and the least squares loss we derive an oracle inequality, which in turn is used to deduce learning rates that are essentially minimax optimal under some standard smoothness assumptions on the regression function. In addition, we derive local learning rates that are based on the local smoothness of the regression function. We further introduce a data-dependent parameter selection method for our local SVM approach and show that this method achieves the same almost optimal learning rates. Finally, we present a few larger scale experiments for our localized SVM showing that it achieves essentially the same test error as a global SVM for a fraction of the computational requirements. In addition, it turns out that the computational requirements for the local SVMs are similar to those of a vanilla random chunk approach, while the achieved test errors are significantly better.

Keywords: least squares regression, support vector machines, localization

1. Introduction

Based on a training set $D := ((x_1, y_1), \dots, (x_n, y_n))$ of i.i.d. input/output observations drawn from an unknown distribution P on $X \times Y$, where $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$, the goal of non-parametric least squares regression is to find a function $f_D : X \to \mathbb{R}$ that is a good estimate of the unknown conditional mean $f^*(x) := \mathbb{E}(Y|x)$, $x \in X$. For this classical estimation problem various methods have been proposed and studied in the literature, see e.g., (Simonoff, 1996) and the book (Györfi et al., 2002) for detailed accounts.

In this paper, we consider kernel-based regularized empirical risk minimizers, also known as support vector machines (SVMs), which solve the regularized problem

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

Here, $\lambda > 0$ is a fixed real number and H is a reproducing kernel Hilbert space (RKHS) over X with reproducing kernel $k : X \times X \to \mathbb{R}$, see e.g., (Aronszajn, 1950; Berlinet and Thomas-Agnan, 2004; Steinwart and Christmann, 2008). The function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is a loss function, where in the following we either consider the least squares loss $L_{LS} : Y \times \mathbb{R} \to [0, \infty)$ defined by $(y, t) \mapsto (y - t)^2$, or variants of it that may depend on $x \in X$. Besides, $\mathcal{R}_{L,\mathrm{D}}(f)$ denotes the empirical risk of a function $f : X \to \mathbb{R}$, that is

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

where D is the empirical measure associated to the data D defined by $\mathrm{D} := \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$ with Dirac measure $\delta_{(x_i,y_i)}$ at (x_i, y_i). Recall that the empirical SVM solution $f_{\mathrm{D},\lambda}$ exists and is unique (cf. Steinwart and Christmann, 2008, Theorem 5.5) whenever the loss L is convex in its last argument, which is true for the least squares loss and its variants that will be considered later on. Moreover, an SVM is L-risk consistent under a few assumptions on the RKHS H and the regularization parameter λ, see (Steinwart and Christmann, 2008, Section 6.4) for more details.

An essential theoretical task, which has attracted many considerations, is the investigation of learning rates for SVMs. For example, such rates for SVMs using the least squares loss and generic kernels can be found in (Cucker and Smale, 2002; De Vito et al., 2005; Smale and Zhou, 2007; Caponnetto and De Vito, 2007; Mendelson and Neeman, 2010; Steinwart et al., 2009) and the references mentioned therein. At this point, we do not want to take a closer look at these results, instead we relegate to (Eberts and Steinwart, 2013), where a detailed discussion can be found. More important for our purposes is the fact that Eberts and Steinwart (2011, 2013) establish (essentially) asymptotically optimal learning rates for least squares SVMs (LS-SVMs) using Gaussian RBF kernels. More precisely, for a domain $X \subset B_{\ell_2^d}$, $Y := [-M, M]$ with $M > 0$, a distribution P on $X \times Y$ such that P_X has a bounded Lebesgue density on X, and for f^* contained in the Sobolev space $W_2^\alpha(\mathrm{P}_X)$, $\alpha \in \mathbb{N}$, or in the Besov-like space $B_{2,\infty}^\alpha(\mathrm{P}_X)$, $\alpha \geq 1$, respectively, the LS-SVM using Gaussian kernels learns for all $\xi > 0$ with rate $n^{-\frac{2\alpha}{2\alpha+d}+\xi}$ with a high probability. In other words, it learns at least with a rate that is arbitrarily close to the optimal learning rate.

Although these rates are essentially asymptotically optimal, they depend on the order of smoothness of the regression function on the *entire* input space X. That is, if the regression function f^* is on some area of X smoother than on another area, the learning rate is determined by the part of X, where the regression function f^* is least smooth. In contrast to this, it would be desirable to achieve a learning rate on every region of X that corresponds with the order of smoothness of f^* on this region. Therefore, one of our goals of this paper is to modify the standard SVM approach such that we achieve local learning rates that are asymptotically optimal.

Our technique to achieve such local learning rates is a special data splitting approach, which first creates a geometrically well-behaved partition of the input space X and then finds a separate SVM on each of the resulting cells with the help of the training samples that fall into these cells. Recall that various other *local* splitting approaches have already been extensively investigated in the literature, but mostly to speed-up the training time, see for instance, the early works (Bottou and Vapnik, 1992; Vapnik and Bottou, 1993).

Here the basic idea of most other local approaches is to *a)* split the training data and just consider a few examples near a testing sample, *b)* train on this small subset of the training data, and *c)* use the solution for a prediction w.r.t. the test sample. Here, many up-to-date investigations use SVMs to train on the local data set but, yet there are different ways to split the whole training data set into smaller, local sets. For example, Chang et al. (2010); Wu et al. (1999); Bennett and Blue (1998) use decision trees while in (Hable, 2013; Segata and Blanzieri, 2010, 2008; Blanzieri and Melgani, 2008; Blanzieri and Bryl, 2007a,b; Zhang et al., 2006) local subsets are built considering k nearest neighbors. The latter approaches further vary, for example, Zhang et al. (2006); Blanzieri and Bryl (2007a); Hable (2013) consider different metrics w.r.t. the input space whereas Segata and Blanzieri (2008); Blanzieri and Melgani (2008); Blanzieri and Bryl (2007b) consider metrics w.r.t. the feature space. Nonetheless, the basic idea of all these articles is that an SVM problem based on k training samples is solved for *each* test sample. Another approach using k nearest neighbors is investigated in (Segata and Blanzieri, 2010). Here, k-neighborhoods consisting of training samples and collectively covering the training data set are constructed and an SVM is calculated on each neighborhood. The prediction for a test sample is then made according to the nearest training sample that is a center of a k-neighborhood. As for the other nearest neighbor approaches, however, the results are mainly experimental. An exception to this rule is (Hable, 2013), where universal consistency for localized versions of SVMs, or more precisely, a large class of regularized kernel methods, is proven. Another article presenting theoretical results for localized versions of learning methods is (Zakai and Ritov, 2009). Here, the authors show that a consistent learning method behaves locally, i.e., the prediction is essentially influenced by close by samples. However, this result is based on a localization technique considering only training samples contained in a neighborhood with a fixed radius and center x when an estimate in x is sought. Probably closest to our approach is the one examined in (Cheng et al., 2010) and (Cheng et al., 2007), where the training data is split into clusters and then an SVM is trained on each cluster. However, the presented results are again only of experimental character.

Unlike in the papers mentioned above, our main goal is to theoretically investigate local SVMs based on local splitting. Namely, we establish both global and local learning rates for our local splitting approach (VP-SVM) that do match the best existing and essentially optimal rates for global SVMs derived by Eberts and Steinwart (2013). In addition, we show that these rates can be obtained without knowing characteristics of P by a simple and well-known hold-out technique. Furthermore, we empirically compare our VP-SVM to another data splitting approach known as random chunking (RC-SVM) or divide-and-conquer kernel ridge regression for which learning rates, at least for generic kernels, have been recently established by Zhang et al. (2015); Lin et al. (2016). In these experiments it turns out that for splittings that lead to comparable training times, our VP-SVM has a significantly smaller test error than RC-SVMs.

Investigating other speed-up schemes for SVMs theoretically has been in the focus of research in the last few years. For example, Zhang et al. (2015); Lin et al. (2016) established optimal learning rates in expectation for RC-SVMs under the assumption that the conditional mean f^* is contained in the used RKHS, or in the image of a fractional integral operator, respectively. Although these results are very interesting they are not very useful for SVMs with Gaussian kernels, since for these kernels the imposed assumptions on f^*

imply $f^* \in C^\infty$, which is usually considered to be too restrictive. For a similar reason the results by Rudi et al. (2015) for the popular Nyström method require too restrictive assumptions when applied to SVMs with Gaussian kernels. On a side note, we like to mention that this difference between generic kernels on the one hand and Gaussian kernels on the other hand already appears for the standard global SVMs. Indeed, in the generic case, one usually addresses the approximation error by assuming the conditional mean to be contained in the image of a fractional integral operator, which can in turn be identified as an interpolation space of the real method, see (Steinwart and Scovel, 2012). For certain kernels, the classical theory of interpolation spaces then identifies the considered interpolation spaces as Besov spaces, so that the approximation error assumption has a clear intuitive meaning. On the other hand, for Gaussian kernels with fixed width it has been shown by Smale and Zhou (2003) that their interpolation spaces consist of C^∞-functions, so that the generic theory would again lead to a too restrictive approximation error assumption. To address this issue, one considers widths that change with the sample size. However, to make this approach successful, one requires both a manual estimation of the approximation error, see (Eberts and Steinwart, 2011), and eigenvalue/entropy number bounds that do depend on the kernel width. For these reasons, learning rates for SVMs with Gaussian kernels under realistic assumptions are, in general, harder to obtain. Nonetheless, they are important, since in practice, Gaussian kernels are by far the most often used kernels.

The rest of this paper is organized as follows: In Section 2 we describe our splitting approach in detail. Section 3 then presents some theoretical results on RKHSs that enable the analysis of our method. After that, Section 4 contains the main results, namely an oracle inequality and learning rates for our localized SVM method. Moreover, a data-dependent parameter selection method is studied that induces the same rates. Section 5 then presents some experimental results w.r.t. the localized SVM technique. Finally, Section 6 collects the proofs for the results of the earlier sections as well as some necessary and important ancillary findings.

2. Description of the Localized SVM Approach

In this section, we introduce some general notations and assumptions. Based on the latter we modify the standard SVM approach. Let us start with the probability measure P on $X \times Y$, where $X \subset \mathbb{R}^d$ is non-empty, $Y := [-M, M]$ for some $M > 0$, and P_X is the marginal distribution of X. Depending on the learning target one chooses a loss function L, i.e., a function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ that is measurable. Then, for a measurable function $f : X \to \mathbb{R}$, the L-risk is defined by

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

and the optimal L-risk, called the Bayes risk with respect to P and L, is given by

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

A measurable function $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ with $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ is called a Bayes decision function. For the commonly used losses such as the least squares loss treated in Section 4 the Bayes decision function $f^*_{L,\mathrm{P}}$ is P_X-almost surely $[-M, M]$-valued, since $Y = [-M, M]$.

In this case, it seems obvious to consider estimators with values in $[-M, M]$ on X. To this end, we introduce the concept of clipping the decision function. Let $\widehat{t}$ be the clipped value of some $t \in \mathbb{R}$ at $\pm M$ defined by

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

Then, a loss is called clippable at $M > 0$ if, for all $(x, y, t) \in X \times Y \times \mathbb{R}$, we have

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

Obviously, the latter implies $\mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) \leq \mathcal{R}_{L,\mathrm{P}}(f)$ for all $f : X \to \mathbb{R}$. In other words, restricting the decision function to the interval $[-M, M]$ containing our labels cannot worsen the risk, in fact, clipping this function typically reduces the risk. Hence, we consider the clipped version $\widehat{f}_D$ of the decision function as well as the risk $\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_D)$ instead of the risk $\mathcal{R}_{L,\mathrm{P}}(f_D)$ of the unclipped decision function. Note that this clipping idea does *not* change the required solver since it is performed *after* the training phase.

To modify the standard SVM approach (1), we assume that $(A_j)_{j=1,\dots,m}$ is a partition of X such that all its cells have non-empty interior, that is $\mathring{A}_j \neq \emptyset$ for every $j \in \{1, \dots, m\}$. Now, the basic idea of our approach is to consider for each cell of the partition an individual SVM. To describe this approach in a mathematically rigorous way, we have to introduce some more definitions and notations. Let us begin with the index set

$$I_j := \big\{ i \in \{1, \dots, n\} : x_i \in A_j \big\}\,, \qquad j = 1, \dots, m\,,$$

indicating the samples of D contained in A_j, as well as the corresponding data set

$$D_j := \{(x_i, y_i) \in D : i \in I_j\}\,, \qquad j = 1, \dots, m\,.$$

Moreover, for every $j \in \{1, \dots, m\}$, we define a (local) loss $L_j : X \times Y \times \mathbb{R} \to [0, \infty)$ by

$$L_j(x, y, t) := \mathbb{1}_{A_j}(x) L(x, y, t)\,,$$

where $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is the loss that corresponds to our learning problem at hand. We further assume that H_j is an RKHS over A_j with kernel $k_j : A_j \times A_j \to \mathbb{R}$. Here, every function $f \in H_j$ is only defined on A_j even though a function $f_D : X \to \mathbb{R}$ is finally sought. To this end, for $f \in H_j$, we define the zero-extension $\hat{f} : X \to \mathbb{R}$ by

$$\hat{f}(x) := \begin{cases} f(x)\,, & x \in A_j\,, \\ 0\,, & x \notin A_j\,. \end{cases}$$

Then, the space $\hat{H}_j := \{\hat{f} : f \in H_j\}$ equipped with the norm

$$\|\hat{f}\|_{\hat{H}_j} := \|f\|_{H_j}\,, \qquad \hat{f} \in \hat{H}_j\,,$$

is an RKHS on X (cf. Lemma 2), which is isometrically isomorphic to H_j. With these preparations we can now formulate our local SVM approach. To this end, for every $j \in \{1, \dots, m\}$, we consider the local SVM optimization problem

$$f_{\mathrm{D}_j,\lambda_j} = \arg\min_{\hat{f} \in \hat{H}_j} \lambda_j \|\hat{f}\|^2_{\hat{H}_j} + \frac{1}{n}\sum_{i=1}^{n} L_j(x_i, y_i, \hat{f}(x_i))\,, \tag{2}$$

where $\lambda_j > 0$ for every $j \in \{1, \dots, m\}$. Based on these empirical SVM solutions, we then define the decision function $f_{\mathrm{D},\boldsymbol{\lambda}} : X \to \mathbb{R}$ by

$$f_{\mathrm{D},\boldsymbol{\lambda}}(x) := \sum_{j=1}^{m} f_{\mathrm{D}_j,\lambda_j}(x) = \sum_{j=1}^{m} \mathbb{1}_{A_j}(x) f_{\mathrm{D}_j,\lambda_j}(x)\,, \tag{3}$$

where $\boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_m)$. Since all $f_{\mathrm{D}_j,\lambda_j}$ in (2) are usual empirical SVM solutions the common properties hold. Moreover, for arbitrary $j \in \{1, \dots, m\}$, $f_{\mathrm{D}_j,\lambda_j}(x_i) = 0$ if $x_i \notin A_j$ for all $i \in \{1, \dots, n\}$. Furthermore, note that the SVM optimization problem (2) equals the SVM optimization problem (1) using H_j, D_j, and the regularization parameter $\tilde{\lambda}_j := \frac{n}{|I_j|}\lambda_j$. That is, $f_{\mathrm{D}_j,\lambda_j}$ as in (2) and $h_{\mathrm{D}_j,\tilde{\lambda}_j} := \arg\min_{f \in H_j} \tilde{\lambda}_j \|f\|^2_{H_j} + \mathcal{R}_{L,\mathrm{D}_j}(f)$ coincide on A_j. Besides, it is easy to show that, whenever a Bayes decision function $f^*_{L,\mathrm{P}}$ w.r.t. P and L exists, it additionally is a Bayes decision function w.r.t. P and L_j.

Let us now briefly discuss the required computing time of our modified SVM. To this end, recall that the costs for solving an usual SVM problem are $\mathcal{O}(n^q)$ where $q \in [2, 3]$. For the new approach we consider m working sets of size $n_1, \dots, n_m$ where for simplicity we assume $n_i \approx \frac{n}{m}$ for all $i \in \{1, \dots, m\}$. Then for each working set an usual SVM problem has to be solved such that, altogether, the modified SVM induces a computational cost of $\mathcal{O}\left(m\left(\frac{n}{m}\right)^q\right)$. Therefore, if $m \approx n^\beta$ for some $\beta > 0$, then our approach is computationally cheaper than a traditional SVM. Note that our strategy using a partition of the input space is a typical way to speed-up SVMs. Other techniques that possess similar properties are, e.g., applied in the articles cited in the introduction. Besides, we refer to (Tsang et al., 2007) and (Tsang et al., 2005) using enclosing ball problems to solve an SVM, to (Graf et al., 2005) presenting an model of multiple filtering SVMs and to (Collobert et al., 2001) investigating a mixture of SVMs based on several subsets of the training set.

To describe the above SVM approach $(A_j)_{j=1,\dots,m}$ only has to be some partition of X. However, for the theoretical investigations concerning learning rates of our new approach, we have to further specify the partition. To this end, we denote the closed unit ball of the d-dimensional Euclidean space ℓ_2^d by $B_{\ell_2^d}$ and we define balls $B_1, \dots, B_m$ with radius $r > 0$ and mutually distinct centers $z_1, \dots, z_m \in B_{\ell_2^d}$ by

$$B_j := B_r(z_j) := \{x \in \mathbb{R}^d : \|x - z_j\|_2 \le r\}\,, \qquad j \in \{1, \dots, m\}\,, \tag{4}$$

where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^d$. Moreover, we choose r and $z_1, \dots, z_m$ such that

$$B_{\ell_2^d} \subset \bigcup_{j=1}^{m} B_j\,,$$

i.e., such that the balls $B_1, \ldots, B_m$ cover $B_{\ell_2^d}$ and, simultaneously, any non-empty set $X \subset B_{\ell_2^d}$ (cf. Figure 1). The following well-known lemma relates the radius of such a cover with the number of centers.

Lemma 1 *For all $c > 0$ and $r \in (0, c]$, there exist balls $(B_r(z_j))_{j=1,\ldots,m}$ with radius r and centers $z_1, \ldots, z_m \in cB_{\ell_2^d}$ such that $\bigcup_{j=1}^m B_r(z_j)$ covers $cB_{\ell_2^d}$ and $r \leq 3cm^{-\frac{1}{d}}$.*

For simplicity of notation, we assume in the following that $X \subset B_{\ell_2^d}$. Thus, according to Lemma 1, there exists a cover $(B_j)_{j=1,\ldots,m}$ of X with

$$r \leq 3m^{-\frac{1}{d}}\,. \tag{5}$$

Let us finally specify the partition $(A_j)_{j=1,\ldots,m}$ of X by the following assumption.

(A) Let $r \in (0, 1]$ and $(A'_j)_{j=1,\ldots,\widetilde{m}}$ be a partition of $B_{\ell_2^d}$ such that $\mathring{A}'_j \neq \emptyset$ as well as $\overline{\mathring{A}'_j} = \overline{A'_j}$ for every $j \in \{1, \ldots, \widetilde{m}\}$ and such that there exist balls $B_j := B_r(z_j) \supset A'_j$ with radius r and mutually distinct centers $z_1, \ldots, z_{\widetilde{m}} \in B_{\ell_2^d}$ satisfying (5). In addition, assume that X is a non-empty, closed subset of $B_{\ell_2^d}$ satisfying $\overline{\mathring{X}} = X$. W.l.o.g. we assume that, for some $m \leq \widetilde{m}$, $A'_j \cap \mathring{X} \neq \emptyset$ for all $j \in \{1, \ldots, m\}$ and $A'_j \cap \mathring{X} = \emptyset$ for all $j \in \{m+1, \ldots, \widetilde{m}\}$. Then we define $A''_j := A'_j \cap \mathring{X}$ for all $j \in \{1, \ldots, m\}$ and assume that $(A_j)_{j=1,\ldots,m}$ is a partition of X satisfying $A''_j \subset A_j \subset \overline{A''_j}$.

Note that the partition $(A_j)_{j=1,\ldots,m}$ of X in Assumption **(A)** satisfies, for every $j \in \{1, \ldots, m\}$, $A_j \subset B_j$ for B_j as in **(A)** and $\mathring{A}_j \neq \emptyset$, where the latter is shown in Lemma 8 in the Appendix. Obviously, for the partition $(A_j)_{j=1,\ldots,m}$, r and m fulfill (5).

In Assumption **(A)** $(A'_j)_{j=1,\ldots,\widetilde{m}}$ is a partition of $B_{\ell_2^d}$ from which we build a partition $(A_j)_{j=1,\ldots,m}$ of $X \subset B_{\ell_2^d}$. However, for the construction of our local SVM approach and the proofs of the belonging learning rates, it will be negligible whether we first consider a partition $(A'_j)_{j=1,\ldots,\widetilde{m}}$ of $B_{\ell_2^d}$ or only a partition $(A_j)_{j=1,\ldots,m}$ of X, since the cells $A'_{m+1}, \ldots A'_{\widetilde{m}}$, which are removed, have zero mass w.r.t. the marginal distribution P_X of X if $P_X(\partial X) = 0$.

In the remaining sections we will frequently refer to Assumption **(A)**. Thus, let us illustrate by the following example that **(A)** is indeed a natural assumption.

Example 1 *For some $r \in (0, 1]$, let us consider an r-net $z_1, \ldots, z_m$ of $B_{\ell_2^d}$, where $z_1, \ldots, z_m$ are mutually distinct. Moreover, we assume that $X \subset B_{\ell_2^d}$ satisfies $\overline{\mathring{X}} = X$. Based on the r-net $z_1, \ldots, z_m$, a Voronoi partition $(A_j)_{j=1,\ldots,m}$ of X is defined by*

$$A_j := \left\{ x \in X : \min \operatorname*{arg\,min}_{k \in \{1,\ldots,m\}} \|x - z_k\|_2 = j \right\}, \tag{6}$$

*cf. Figure 2. That is, A_j contains all $x \in X$ such that the center z_j is the nearest center to x, and in the case of ties the center with the smallest index is taken. Obviously, $(A_j)_{j=1,\ldots,m}$ is a partition of X with $\mathring{A}_j \neq \emptyset$ and $A_j \subset B_r(z_j)$ for all $j \in \{1, \ldots, m\}$, and hence it satisfies condition **(A)**, if r and m fulfill (5).*

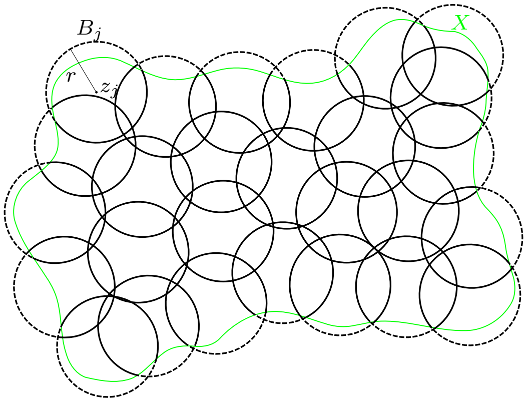


Figure 1: Cover $(B_j)_{j=1,\dots,m}$ of X, where $B_1,\dots,B_m$ are balls with radius r and centers z_j $(j=1,\dots,m)$.

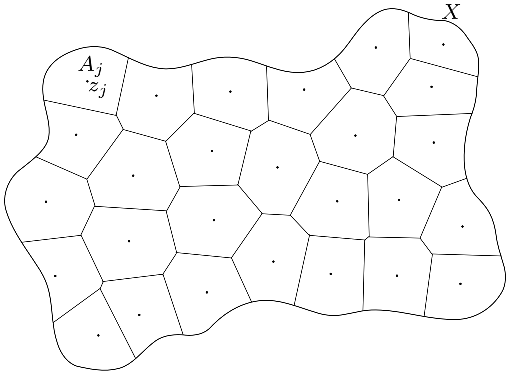


Figure 2: Voronoi partition $(A_j)_{j=1,\dots,m}$ of X defined by (6), where $A_j \subset B_j$ for every $j \in \{1,\dots,m\}$.

Motivated by Example 1, we call the learning method producing $f_{\mathrm{D},\boldsymbol{\lambda}}$ given by (3) a *Voronoi partition support vector machine*, in short VP-SVM. Despite this name, however, we just take a partition $(A_j)_{j=1,\dots,m}$ satisfying **(A)** as basis here instead of requesting $(A_j)_{j=1,\dots,m}$ to be a Voronoi partition.

Recall that our goal is to derive not only global but also local learning rates for this VP-SVM approach. To this end, we additionally consider a $T \subset X$ with $\mathrm{P}_X(T) > 0$. Then we examine the learning rate of the VP-SVM on this subset T of X. To formalize this, it is necessary to introduce some basic notations related to T. Let us define the index set J_T by

$$J_T := \{j \in \{1,\dots,m\} : A_j \cap T \neq \emptyset\} \tag{7}$$

specifying every set A_j that has at least one common point with T. Note that, for every non-empty set $T \subset X$, the index set J_T is also non-empty, i.e., $|J_T| \geq 1$. Besides, deriving local rates on T requires us to investigate the excess risk of the VP-SVM with respect to the distribution P and the loss $L_T : X \times Y \times \mathbb{R} \to [0,\infty)$ defined by

$$L_T(x,y,t) := \mathbb{1}_T(x) L(x,y,t)\,. \tag{8}$$

However, to manage the analysis we additionally need the loss $L_{J_T} : X \times Y \times \mathbb{R} \to [0,\infty)$ given by

$$L_{J_T}(x,y,t) := \mathbb{1}_{\bigcup_{j\in J_T} A_j}(x) L(x,y,t)\,, \tag{9}$$

which may only be nonzero, if x is contained in some set A_j with $j \in J_T$. Note that the risks $\mathcal{R}_{L_T,\mathrm{P}}(f)$ and $\mathcal{R}_{L_{J_T},\mathrm{P}}(f)$ quantify the quality of some function f just on T and

$$A_T := \bigcup_{j\in J_T} A_j \supset T\,,$$

respectively. Hence, examining the excess risks

$$\mathcal{R}_{L_T,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}}) - \mathcal{R}^*_{L_T,\mathrm{P}} \leq \mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}}$$

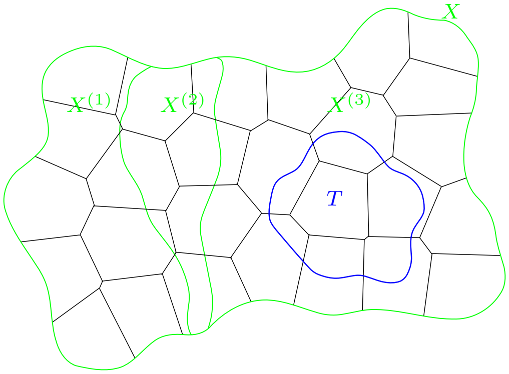


Figure 3: The input space X with the corresponding partition $(A_j)_{j=1,\dots,m}$ and the subset T, where the local learning rate should be examined.

leads to learning rates on A_T and implicitly on T. Recapitulatory, let us declare a set of notations that will be frequently used in the remainder of the paper.

(T) For $T \subset X$, we define an index set J_T by (7), loss functions $L_T, L_{J_T} : X \times Y \times \mathbb{R} \to [0, \infty)$ by (8) and (9), and the set $A_T := \bigcup_{j \in J_T} A_j$.

3. Building Weighted Global Kernels

In this section, we first focus on RKHSs and direct sums of RKHSs. Then, we show that a VP-SVM solution is also the solution of an usual SVM.

Let us begin with some basic notations. For $q \in [1, \infty]$ and a measure ν, we denote by $L_q(\nu)$ the Lebesgue spaces of order q w.r.t. ν and for the Lebesgue measure μ on $X \subset \mathbb{R}^d$ we write $L_q(X) := L_q(\mu)$. In addition, for a measurable space X, the set of all real-valued measurable functions on X is given by $\mathcal{L}_0(X) := \{f : X \to \mathbb{R} \,|\, f \text{ measurable}\}$. Moreover, for a measure ν on X and measurable $\widetilde{X} \subset X$, we define the trace measure $\nu_{|\widetilde{X}}$ of ν in $\widetilde{X}$ by $\nu_{|\widetilde{X}}(A) = \nu(A \cap \widetilde{X})$ for every $A \subset X$.

Our first goal is to show that $f_{\mathrm{D},\boldsymbol{\lambda}}$ in (3) is actually an ordinary SVM solution. To this end, we consider an RKHS on some $A \subsetneq X$ and extend it to an RKHS on X by the following lemma, where we omit the obvious proof.

Lemma 2 *Let $A \subset X$ and H_A be an RKHS on A with corresponding kernel k_A. Denote by $\hat{f}$ the zero-extension of $f \in H_A$ to X defined by*

$$\hat{f}(x) := \begin{cases} f(x)\,, & \text{for } x \in A\,, \\ 0\,, & \text{for } x \in X \backslash A\,. \end{cases}$$

Then, the space $\hat{H}_A := \{\hat{f} : f \in H_A\}$ equipped with the norm $\|\hat{f}\|_{\hat{H}_A} := \|f\|_{H_A}$ is an RKHS on X and its reproducing kernel is given by

$$\hat{k}_A(x, x') := \begin{cases} k_A(x, x')\,, & \text{if } x, x' \in A\,, \\ 0\,, & \text{else.} \end{cases} \tag{10}$$

Based on this lemma, we are now able to construct an RKHS by a direct sum of RKHSs $\hat{H}_A$ and $\hat{H}_B$ with $A, B \subset X$ and $A \cap B = \emptyset$. Here, we skip the proof once more, since the assertion follows immediately using, for example, orthonormal bases of $\hat{H}_A$ and $\hat{H}_B$.

Lemma 3 *For $A, B \subset X$ such that $A \cap B = \emptyset$ and $A \cup B \subset X$, let H_A and H_B be RKHSs of the kernels k_A and k_B over A and B, respectively. Furthermore, let $\hat{H}_A$ and $\hat{H}_B$ be the RKHSs of all functions of H_A and H_B extended to X in the sense of Lemma 2 and let $\hat{k}_A$ and $\hat{k}_B$ given by (10) be the associated reproducing kernels. Then, $\hat{H}_A \cap \hat{H}_B = \{0\}$ and hence the direct sum*

$$H := \hat{H}_A \oplus \hat{H}_B \tag{11}$$

exists. For $\lambda_A, \lambda_B > 0$ and $f \in H$, let $\hat{f}_A \in \hat{H}_A$ and $\hat{f}_B \in \hat{H}_B$ be the unique functions such that $f = \hat{f}_A + \hat{f}_B$. Then, we define the norm $\| \cdot \|_H$ by

$$\|f\|_H^2 := \lambda_A \|\hat{f}_A\|_{\hat{H}_A}^2 + \lambda_B \|\hat{f}_B\|_{\hat{H}_B}^2 \tag{12}$$

and H equipped with the norm $\| \cdot \|_H$ is again an RKHS for which

$$k(x, x') := \lambda_A^{-1} \hat{k}_A(x, x') + \lambda_B^{-1} \hat{k}_B(x, x') \,, \qquad x, x' \in X \,,$$

is the reproducing kernel.

To relate Lemmas 2 and 3 with (3), we have to introduce some more notations. For pairwise disjoint sets $A_1, \ldots, A_m \subset X$, let H_j be an RKHS on A_j for every $j \in \{1, \ldots, m\}$. Then, based on RKHSs $\hat{H}_1, \ldots, \hat{H}_m$ on X defined by Lemma 2, a joined RKHS can be designed analogously to Lemma 3. That is, for an arbitrary index set $J \subset \{1, \ldots, m\}$ and a vector $\boldsymbol{\lambda} = (\lambda_j)_{j \in J} \in (0, \infty)^{|J|}$, the direct sum

$$H_J := \bigoplus_{j \in J} \hat{H}_j = \left\{ f = \sum_{j \in J} f_j : f_j \in \hat{H}_j \text{ for all } j \in J \right\} \tag{13}$$

is again an RKHS equipped with the norm

$$\|f\|_{H_J}^2 = \sum_{j \in J} \lambda_j \|f_j\|_{\hat{H}_j}^2 \,. \tag{14}$$

If $J = \{1, \ldots, m\}$, we simply write $H := H_J$. Note that H contains inter alia $f_{\mathrm{D}, \boldsymbol{\lambda}}$ given by (3).

Let us briefly investigate the regularized empirical risk of $f_{\mathrm{D}, \boldsymbol{\lambda}} = \sum_{j=1}^m \mathbb{1}_{A_j} f_{\mathrm{D}_j, \lambda_j}$, where $f_{\mathrm{D}_j, \lambda_j}$, $j = 1, \ldots, m$, are defined by (2). For an arbitrary $f \in H$, we have

$$\begin{aligned} \|f_{\mathrm{D}, \boldsymbol{\lambda}}\|_H^2 + \mathcal{R}_{L, \mathrm{D}}(\widehat{f}_{\mathrm{D}, \boldsymbol{\lambda}}) &= \sum_{j=1}^m \Big(\lambda_j \left\| f_{\mathrm{D}_j, \lambda_j} \right\|_{\hat{H}_j}^2 + \mathcal{R}_{L_j, \mathrm{D}}(\widehat{f}_{\mathrm{D}, \boldsymbol{\lambda}}) \Big) \\ &\leq \sum_{j=1}^m \Big(\lambda_j \left\| \mathbb{1}_{A_j} f \right\|_{\hat{H}_j}^2 + \mathcal{R}_{L_j, \mathrm{D}}(f) \Big) \end{aligned}$$

$$= \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)\,, \tag{15}$$

where we used $\mathcal{R}_{L,\mathrm{D}}(f) = \sum_{j=1}^m \mathcal{R}_{L_j,\mathrm{D}}(f)$, which immediately follows by Lemma 9 given in the appendix. That is, $f_{\mathrm{D},\boldsymbol{\lambda}}$ is the decision function of an SVM using H and L as well as the regularization parameter $\tilde{\lambda} = 1$. In other words, the latter SVM equals the VP-SVM given by (3). This will be a key insight used in our analysis.

Subsequently, we only consider RKHSs of Gaussian RBF kernels. For this purpose, we summarize some assumptions for the Gaussian case of joined RKHSs in the following assumption set.

(G) For pairwise disjoint subsets $A_1, \dots, A_m$ of X, let $H_j := H_{\gamma_j}(A_j)$, $j \in \{1, \dots, m\}$, be the RKHS of the Gaussian kernel k_{γ_j} with width $\gamma_j \in (0, r]$ over A_j. Consequently, for $\boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_m) \in (0, \infty)^m$, we define the joined RKHS $H := \bigoplus_{j=1}^m \hat{H}_{\gamma_j}(A_j)$ and equip it with the norm (14).

In the following we do not consider SVMs with a fixed kernel, thus, we use a more detailed notation than (2) and (3) specifying the kernel width γ_j of the RKHS $H_{\gamma_j}(A_j)$ at hand. Namely, for all $j \in \{1, \dots, m\}$ and $\boldsymbol{\gamma} := (\gamma_1, \dots, \gamma_m)$, we write

$$f_{\mathrm{D}_j,\lambda_j,\gamma_j} = \operatorname*{arg\,min}_{f \in \hat{H}_{\gamma_j}(A_j)} \lambda_j \|f\|_{\hat{H}_{\gamma_j}(A_j)}^2 + \frac{1}{n} \sum_{i=1}^n L_j(x_i, y_i, f(x_i))\,,$$

and

$$f_{\mathrm{D},\boldsymbol{\lambda},\boldsymbol{\gamma}} := \sum_{j=1}^m f_{\mathrm{D}_j,\lambda_j,\gamma_j}$$

instead of $f_{\mathrm{D}_j,\lambda_j}$ and $f_{\mathrm{D},\boldsymbol{\lambda}}$ in the remainder of this work.

4. Learning Rates for Least Squares VP-SVMs

In this section, the non-parametric least squares regression problem is considered using the least squares loss $L : Y \times \mathbb{R} \to [0, \infty)$ defined by $L(y,t) := (y-t)^2$. It is well known that, in this case, the Bayes decision function $f_{L,\mathrm{P}}^* : \mathbb{R}^d \to \mathbb{R}$ is given by $f_{L,\mathrm{P}}^*(x) = \mathbb{E}_{\mathrm{P}}(Y|x)$ for P_X-almost all $x \in \mathbb{R}^d$. Moreover, this function is unique up to zero-sets. Besides, for the least squares loss the equality

$$\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}^* = \left\|f - f_{L,\mathrm{P}}^*\right\|_{L_2(\mathrm{P}_X)}^2$$

can be shown by some simple, well-known transformations. In the first part of Subsection 4.1 we introduce some tools to describe smoothness properties of $f_{L,\mathrm{P}}^*$, which are then used in the oracle inequalities and learning rates of the second part. In Subsection 4.2 we then investigate a simple parameter selection strategy for which we will show that it is adaptive.

4.1 Basic Oracle Inequalities for LS-VP-SVMs

To formulate oracle inequalities and derive rates for VP-SVMs using the least squares loss, the target function $f^*_{L,\mathrm{P}}$ is assumed to satisfy certain smoothness conditions. To this end, we initially recall the modulus of smoothness, a device to measure the smoothness of functions, see e.g., DeVore and Lorentz, 1993, p. 44; DeVore and Popov, 1988, p. 398; as well as Berens and DeVore, 1978, p. 360. Denote by $\|\cdot\|_2$ the Euclidean norm and let $\Omega \subset \mathbb{R}^d$ be a subset with non-empty interior, ν be an arbitrary measure on Ω, $p \in (0,\infty]$, and $f : \Omega \to \mathbb{R}$ be contained in $L_p(\nu)$. Then, for $s \in \mathbb{N}$, the s-th modulus of smoothness of f is defined by

$$\omega_{s,L_p(\nu)}(f,t) = \sup_{\|h\|_2 \le t} \|\triangle_h^s(f,\cdot\,)\|_{L_p(\nu)}\ , \qquad\qquad t \ge 0\ ,$$

where $\triangle_h^s(f,\cdot\,)$ denotes the s-th difference of f given by

$$\triangle_h^s(f,x) = \begin{cases} \sum_{j=0}^{s} \binom{s}{j} (-1)^{s-j} f(x+jh) & \text{if } x \in \Omega_{s,h} \\ 0 & \text{if } x \notin \Omega_{s,h} \end{cases}$$

for $h = (h_1,\dots,h_d) \in \mathbb{R}^d$ and $\Omega_{s,h} := \{x \in \Omega : x + th \in \Omega \text{ f.a. } t \in [0,s]\}$. Based on the modulus of smoothness, we introduce Besov-like spaces, i.e., function spaces that provide a finer scale of smoothness than the commonly used Sobolev spaces and that will thus be assumed to contain the target function later on. To this end, let $\alpha > 0$, $s := \lfloor \alpha \rfloor + 1$, and ν be an arbitrary measure. Then, the Besov-like space $B^\alpha_{2,\infty}(\nu)$ is defined by

$$B^\alpha_{2,\infty}(\nu) := \left\{ f \in L_2(\nu) : |f|_{B^\alpha_{2,\infty}(\nu)} < \infty \right\}\ ,$$

where the semi-norm $|\cdot|_{B^\alpha_{2,\infty}(\nu)}$ is given by

$$|f|_{B^\alpha_{2,\infty}(\nu)} := \sup_{t>0} \left(t^{-\alpha} \omega_{s,L_2(\nu)}(f,t) \right)$$

and the norm by $\|f\|_{B^\alpha_{2,\infty}(\nu)} := \|f\|_{L_2(\nu)} + |f|_{B^\alpha_{2,\infty}(\nu)}$. Here, note that we defined Besov-like spaces for arbitrary measures ν on $\Omega \subset \mathbb{R}^d$ whereas in the literature Besov spaces are usually defined for the Lebesgue measure. Nevertheless, our definition of Besov-like spaces is well-defined. Moreover, for the proofs it is important to notice that, if $\Omega = \mathbb{R}^d$ and ν is a distribution on Ω with $\operatorname{supp}\nu \subsetneq \Omega$, then $\Omega_{s,h}$ still equals $\mathbb{R}^d$, i.e., $\Omega_{s,h} = \Omega$. Also note that for the Lebesgue measure on Ω, where $\Omega = \mathbb{R}^d$ or Ω is a bounded Lipschitz domain in $\mathbb{R}^d$, our definition of Besov-like spaces actually coincides, up to equivalent norms, to the definition of the classical Besov spaces in the literature, see e.g., (Adams and Fournier, 2003, Section 7), (Triebel, 2006, Section 1), (Triebel, 1992, Section 1), and (Triebel, 2010, Sections 2 and 3), where this classical type of Besov spaces is also defined for $1 \le p,q \le \infty$ and $\alpha > 0$. For more details on the equivalences of our definition of Besov-like spaces and the classical definitions, we refer to (Eberts, 2015, Section 3.1). If ν is the Lebesgue measure on Ω, we write $B^\alpha_{2,\infty}(\Omega) := B^\alpha_{2,\infty}(\nu)$. Additionally, let us briefly consider a few embedding properties for Besov-like spaces $B^\alpha_{2,\infty}(\nu)$ where the corresponding proofs can be found in (Eberts, 2015, Section 3.1). To this end, let ν be a finite measure on $\mathbb{R}^d$ such that $\operatorname{supp}\nu =: \Omega \subset \mathbb{R}^d$ has non-empty interior and ν has a Lebesgue density g on Ω. If g is bounded away from 0 on Ω, then $B^{\alpha}_{2,\infty}(\nu) \subset B^{\alpha}_{2,\infty}(\Omega)$ for $\alpha > 0$. Alternatively, for $g \in L_\infty(\Omega)$ and $\alpha > 0$, we have $B^{\alpha}_{2,\infty}(\mathbb{R}^d) \subset B^{\alpha}_{2,\infty}(\nu)$ and $\big(B^{\alpha}_{2,\infty}(\Omega^{+\delta}) \cap L_\infty(\mathbb{R}^d)\big) \subset B^{\alpha}_{2,\infty}(\nu)$, where $\delta > 0$ and $\Omega^{+\delta} := \{x \in \mathbb{R}^d : \exists\, x' \in \Omega \text{ such that } \|x - x'\|_2 \leq \delta\}$. For the sake of completeness, recall from, e.g., (Adams and Fournier, 2003, Section 3) and (Triebel, 2010, Sections 2 and 3) the scale of Sobolev spaces $W^{\alpha}_2(\nu)$ defined by

$$W_2^{\alpha}\left(\nu\right) := \left\{ f \in L_p\left(\nu\right) : \partial^{(\beta)} f \in L_2\left(\nu\right) \text{ exists for all } \beta \in \mathbb{N}_0^d \text{ with } |\beta| \leq \alpha \right\} ,$$

where $\alpha \in \mathbb{N}_0$, ν is an arbitrary measure, and $\partial^{(\beta)}$ is the β-th weak derivative for a multi-index $\beta = (\beta_1, \dots, \beta_d) \in \mathbb{N}_0^d$ with $|\beta| = \sum_{i=1}^d \beta_i$. That is, $W_2^{\alpha}(\nu)$ is the space of all functions in $L_2(\nu)$ whose weak derivatives up to order α exist and are contained in $L_2(\nu)$. Moreover, the Sobolev space is equipped with the Sobolev norm

$$\|f\|^p_{W_2^{\alpha}(\nu)} := \sum_{|\beta| \leq \alpha} \left\| \partial^{(\beta)} f \right\|^2_{L_2(\nu)} ,$$

(cf. Adams and Fournier, 2003, p. 60). We write $W_2^0(\nu) = L_2(\nu)$ and, for the Lebesgue measure μ on $\Omega \subset \mathbb{R}^d$, we define $W_2^{\alpha}(\Omega) := W_2^{\alpha}(\mu)$. It is well-known, see e.g., (Edmunds and Triebel, 1996, p. 25 and p. 44), that the Sobolev spaces $W_2^{\alpha}(\mathbb{R}^d)$ fall into the scale of Besov spaces, e.g., $W_2^{\alpha}(\mathbb{R}^d) \subset B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ for $\alpha \in \mathbb{N}$. Furthermore, note that functions $f : \Omega \to \mathbb{R}^d$ can be extended to functions $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ such that $\hat{f}$ inherits the smoothness properties of f, whenever $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain. More precisely, in this case Stein's Extension Theorem (cf. Stein, 1970, p. 181) guarantees the existence of a linear extension operator $\mathfrak{E}$ mapping functions $f : \Omega \to \mathbb{R}$ to functions $\mathfrak{E}f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathfrak{E}f_{|\Omega} = f$ and such that $\mathfrak{E}$ continuously maps $W_2^m(\Omega)$ into $W_2^m(\mathbb{R}^d)$ for all integers $m \geq 0$ and $B^{\alpha}_{2,\infty}(\Omega)$ into $B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ for all $\alpha \geq 0$ simultaneously. For more details, we refer to Stein (1970, p. 181), Triebel (2006, Section 1.11.5), and Adams and Fournier (2003, Chapter 5). In this case, Eberts (2015, Corollary 3.4) shows, for a finite measure ν on $\mathbb{R}^d$ such that $\operatorname{supp} \nu =: \widetilde{\Omega} \supset \Omega$ and such that ν has a Lebesgue density g on $\widetilde{\Omega}$ with $g \in L_\infty(\widetilde{\Omega})$, that $f \in B^{\alpha}_{2,\infty}(\Omega)$ implies $\mathfrak{E}f \in B^{\alpha}_{2,\infty}(\nu)$.

Based on the least squares loss and RKHSs using Gaussian kernels over the partition sets A_j, the subsequent theorem presents an oracle inequality for VP-SVMs.

Theorem 4 *Let $Y := [-M, M]$ for $M > 0$, $L : Y \times \mathbb{R} \to [0, \infty)$ be the least squares loss, and P be a distribution on $\mathbb{R}^d \times Y$. We denote the marginal distribution of P onto $\mathbb{R}^d$ by P_X, write $X := \operatorname{supp} \mathrm{P}_X$, and assume $P_X(\partial X) = 0$. Furthermore, let **(A)** and **(G)** be satisfied. In addition, for an arbitrary subset $T \subset X$, we assume **(T)** . Moreover, let $f^*_{L,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ be a Bayes decision function such that $f^*_{L,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ as well as $f^*_{L,\mathrm{P}} \in B^{\alpha}_{2,\infty}(\mathrm{P}_{X|A_T})$ for some $\alpha \geq 1$. Then, for all $p \in (0,1)$, $n \geq 1$, $\tau \geq 1$, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m) \in (0, r]^m$, and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m) > 0$, the VP-SVM given by (3) using $\hat{H}_{\gamma_1}(A_1), \dots, \hat{H}_{\gamma_m}(A_m)$, and the loss L_{J_T} satisfies*

$$\sum_{j=1}^m \lambda_j \|f_{\mathrm{D}_j,\lambda_j,\gamma_j}\|^2_{\hat{H}_{\gamma_j}(A_j)} + \mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}}$$

$$\leq C_{M,\alpha,p}\left(\sum_{j\in J_T}\lambda_j\gamma_j^{-d}+\left(\frac{\max_{j\in J_T}\gamma_j}{\min_{j\in J_T}\gamma_j}\right)^d\max_{j\in J_T}\gamma_j^{2\alpha}+r^{2p}\left(\sum_{j=1}^{m}\lambda_j^{-1}\gamma_j^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p n^{-1}+\tau n^{-1}\right)$$

with probability P^n *not less than* $1-e^{-\tau}$*, where* $C_{M,\alpha,p}>0$ *is a constant only depending on* M*,* α*,* p*,* d*,* $\|f^*_{L,\mathrm{P}}\|_{L_2(\mathbb{R}^d)}$*,* $\|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}$*, and* $\|f^*_{L,\mathrm{P}}\|_{B^\alpha_{2,\infty}(\mathrm{P}_{X|A_T})}$*.*

We like to emphasize that in the theorem above $X := \operatorname{supp}\mathrm{P}_X$ only serves as a notation. Indeed, the partition $(A'_j)_{j=1,\ldots,\widetilde{m}}$ of **(A)** can be found without knowing $\operatorname{supp}\mathrm{P}_X$, and whether we actually remove the cells that do not intersect the interior of $\operatorname{supp}\mathrm{P}_X$ is irrelevant since these cells will neither contain samples nor will they contribute to the overall risk of our decision function $\widehat{f}_{\mathrm{D},\boldsymbol{\lambda},\boldsymbol{\gamma}}$ as we assumed $P_X(\partial X)=0$. Despite from this, the proofs anyway do not require that X exactly corresponds to the support of the distribution P_X. Instead we can as well assume $\operatorname{supp}\mathrm{P}_X\subset X\subset B_{\ell_2^d}$. Moreover, note for the proofs that the considered Besov-like space $B^\alpha_{2,\infty}(\mathrm{P}_{X|A_T})$ is defined w.r.t. $\Omega=\mathbb{R}^d$.

Theorem 4 only focuses on the least squares loss, however, a similar version can be shown under more general assumptions for generic losses and RKHSs, where we refer the interested reader to (Eberts, 2015, Theorem 4.4). Moreover, considering a trivial partition consisting of only one set A_1 the oracle inequalities for VP-SVMs are comparable to the already known ones, see (Eberts, 2015, p. 81) for more details.

Using the oracle inequality of Theorem 4, we derive learning rates w.r.t. the loss L_{J_T} for the learning method described by (2) and (3) in the following theorem.

Theorem 5 *Let* $\tau\geq 1$ *be fixed and* $\beta\geq\frac{2\alpha}{d}+1$*. Under the assumptions of Theorem 4 and with*

$$r_n = c_1 n^{-\frac{1}{\beta d}}\,, \tag{16}$$

$$\lambda_{n,j} = c_2 r^d n^{-1}\,, \tag{17}$$

$$\gamma_{n,j} = c_3 n^{-\frac{1}{2\alpha+d}}\,, \tag{18}$$

for every $j\in\{1,\ldots,m_n\}$*, we have, for all* $n\geq 1$ *and* $\xi>0$*,*

$$\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n})-\mathcal{R}^*_{L_{J_T},\mathrm{P}}\leq C\left(n^{-\frac{2\alpha}{2\alpha+d}+\xi}+\tau n^{-1}\right)$$

with probability P^n *not less than* $1-e^{-\tau}$*, where* $\boldsymbol{\lambda}_n := (\lambda_{n,1},\ldots,\lambda_{n,m_n})$ *as well as* $\boldsymbol{\gamma}_n := (\gamma_{n,1},\ldots,\gamma_{n,m_n})$ *and* C,c_1,c_2,c_3 *are positive constants with* $c_3\leq c_1$*.*

In the latter theorem the condition $\beta\geq\frac{2\alpha}{d}+1$ is required to ensure $\gamma_{n,j}\leq r_n$, $j=1,\ldots,m_n$, which in turn is a prerequisite arising from Theorem 12 and the used entropy estimate. Let us briefly examine the extreme case $\beta=\frac{2\alpha}{d}+1$. Using $r_n\approx n^{-\frac{1}{\beta d}}$ and (5) leads to covering numbers of the form $m_n\approx n^{\frac{d}{2\alpha+d}}$ and computational costs of $\mathcal{O}\big(m_n\big(\frac{n}{m_n}\big)^q\big)=\mathcal{O}\big(n^{\frac{2\alpha q+d}{2\alpha+d}}\big)$ which is actually less than the computational cost of order n^q, $q\in[2,3]$, of an usual SVM. Note that for increasing β the computational costs of an VP-SVM are increasing as well. However, for $\beta>\frac{2\alpha}{d}+1$, $r_n\approx n^{-\frac{1}{\beta d}}$, and $m_n\approx n^{\frac{1}{\beta}}$, a VP-SVM has costs of $\mathcal{O}\big(n^{\frac{1+(\beta-1)q}{\beta}}\big)$ which still is less that $\mathcal{O}\left(n^q\right)$.

Let us finally take a closer look at the VP-SVM given by (3) and the considerations related to (15), where $f_{\mathrm{D},\boldsymbol{\lambda}} \in H = \bigoplus_{j=1}^{m} \hat{H}_j$ solves the minimization problem

$$f_{\mathrm{D},\boldsymbol{\lambda}} = \operatorname*{arg\,min}_{f_1 \in \hat{H}_1, \dots, f_m \in \hat{H}_m} \sum_{j=1}^{m} \lambda_j \, \|f_j\|_{\hat{H}_j}^2 + \mathcal{R}_{L,\mathrm{D}}\Big(\sum_{j=1}^{m} f_j\Big) .$$

Choosing $\lambda_1 = \ldots = \lambda_m$, the VP-SVM problem can be understood as particular ℓ_2-multiple kernel learning (MKL) problem using the RKHSs $\hat{H}_1, \dots, \hat{H}_m$. Learning rates for MKL have been treated, for example, in (Suzuki, 2011) and (Kloft and Blanchard, 2012). Assuming $f_{L,\mathrm{P}}^* \in H$, the learning rate achieved in (Suzuki, 2011) is $mn^{-\frac{1}{1+s}}$ for dense settings, where s is the so-called spectral decay coefficient. In addition, Kloft and Blanchard (2012) obtain essentially the same rates under these assumptions. Let us therefore briefly investigate the above rate of (Suzuki, 2011). For RKHSs that are continuously embedded in a Sobolev space $W_2^{\alpha}(X)$, we have $s = \frac{d}{2\alpha}$ such that the learning rate reduces to $mn^{-\frac{2\alpha}{2\alpha+d}}$. Note that this learning rate is m times the optimal learning rate $n^{-\frac{2\alpha}{2\alpha+d}}$, where the number $m = m_n$ of kernels may increase with the sample size n. In particular, if $m_n \to \infty$ polynomially, then the rates obtained in (Suzuki, 2011) become substantially worse than the optimal rate. In contrast, due to the special choice of the RKHSs, this is not the case for our VP-SVM problem, provided that m_n does not grow faster than $n^{1/\beta}$.

Note that the oracle inequalities and learning rates achieved in Theorems 4 and 5 require $f_{L,\mathrm{P}}^* \in B_{2,\infty}^{\alpha}(\mathrm{P}_{X|\bigcup_{j\in J_T} A_j})$. However, for an increasing sample size n, the sets A_j shrink and the index set J_T, indicating every set A_j such that $A_j \cap T \neq \emptyset$ and $T \subset \bigcup_{j\in J_T} A_j$, increases. In particular, this also involves that the set $\bigcup_{j\in J_T} A_j$ covering T changes in tandem with n. Since this is very inconvenient and since it would be desirable to assume a certain level of smoothness of the target function on a fixed region for all $n \in \mathbb{N}$, we consider the set T enlarged by an δ-tube. To this end, for $\delta > 0$, we define $T^{+\delta}$ by

$$T^{+\delta} := \left\{ x \in X \,\middle|\, \exists t \in T : \|x - t\|_2 \leq \delta \right\} ,$$

which implies $T \subset T^{+\delta} \subset X$, cf. Figure 4. Note that, for every $\delta > 0$, there exists an $n_\delta \in \mathbb{N}$ such that, for every $n \geq n_\delta$, the union of all partition sets A_j, having at least one common point with T, is contained in $T^{+\delta}$, i.e.

$$\forall \delta > 0 \quad \exists n_\delta \in \mathbb{N} \quad \forall n \geq n_\delta \quad : \quad \bigcup_{j\in J_T} A_j \subset T^{+\delta} , \tag{19}$$

where $J_T := \{j \in \{1, \dots, m_n\} : A_j \cap T \neq \emptyset\}$. Collectively, this implies $T \subset \bigcup_{j\in J_T} A_j \subset T^{+\delta}$ for all $n \geq n_\delta$. Furthermore, since every set A_j is contained in a ball with radius $r_n = cn^{-\frac{1}{\beta d}}$ satisfying (5), the lowest sample size n_δ in (19) can be determined by choosing the smallest $n_\delta \in \mathbb{N}$ such that $\delta \geq 2r_{n_\delta}$, that is

$$n_\delta = \left\lceil \left(\frac{2c}{\delta}\right)^{\beta d} \right\rceil .$$

This leads to the following corollary, which presents an oracle inequality and learning rates assuming the smoothness level α of the target function on a fixed region.

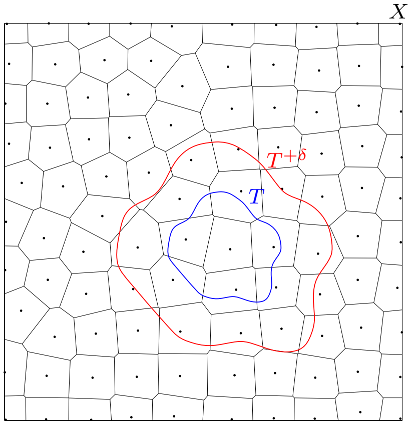


Figure 4: An input space X with a Voronoi partition as well as a subset $T \subset X$ enlarged by an δ-tube to $T^{+\delta}$.

Corollary 6 *Let $Y := [-M, M]$ for $M > 0$, $L : Y \times \mathbb{R} \to [0, \infty)$ be the least squares loss, and P be a distribution on $\mathbb{R}^d \times Y$. We denote the marginal distribution of P onto $\mathbb{R}^d$ by P_X, write $X := \operatorname{supp} \mathrm{P}_X$, and assume $P_X(\partial X) = 0$. Furthermore, let **(A)** and **(G)** be satisfied. In addition, for an arbitrary subset $T \subset X$, we assume **(T)** . Moreover, let $f^*_{L,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ be a Bayes decision function with $f^*_{L,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ as well as*

$$f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X|T^{+\delta}})$$

for $\alpha \geq 1$ and some $\delta > 0$. Then, for all $p \in (0,1)$, $n \geq n_\delta$, $\tau \geq 1$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m) \in (0, r]^m$, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m) > 0$, the VP-SVM given by (3) using $\hat{H}_{\gamma_1}(A_1), \ldots, \hat{H}_{\gamma_m}(A_m)$, and the loss L_T satisfies

$$\sum_{j=1}^m \lambda_j \|f_{\mathrm{D}_j,\lambda_j,\gamma_j}\|^2_{\hat{H}_{\gamma_j}(A_j)} + \mathcal{R}_{L_T,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_T,\mathrm{P}}$$

$$\leq C_{M,\alpha,p} \left(\sum_{j \in J_T} \lambda_j \gamma_j^{-d} + \left(\frac{\max_{j \in J_T} \gamma_j}{\min_{j \in J_T} \gamma_j} \right)^d \max_{j \in J_T} \gamma_j^{2\alpha} + r^{2p} \left(\sum_{j=1}^m \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^p n^{-1} + \tau n^{-1} \right)$$

with probability P^n not less than $1 - e^{-\tau}$, where $C_{M,\alpha,p} > 0$ is the same constant as in Theorem 4.

Additionally, let $\beta \geq \frac{2\alpha}{d} + 1$ as well as, for every $j \in \{1, \ldots, m_n\}$, r_n, $\lambda_{n,j}$, and $\gamma_{n,j}$ be as in (16), (17), and (18), respectively, where c_1, c_2, c_3 are user-specified positive constants with $c_3 \leq c_1$. Then, for all $n \geq n_\delta = \left\lceil \left(\frac{2c_1}{\delta}\right)^{\beta d} \right\rceil$ and $\xi > 0$, we have

$$\mathcal{R}_{L_T,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_T,\mathrm{P}} \leq C \left(n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \tau n^{-1} \right)$$

with probability P^n not less than $1 - e^{-\tau}$, where $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,m_n})$, $\boldsymbol{\gamma}_n := (\gamma_{n,1}, \ldots, \gamma_{n,m_n})$, and C is a positive constant.

Note that the assumption $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X|T^{+\delta}})$ made in Corollary 6 is satisfied if, for example, P_X has a bounded Lebesgue density on $T^{+\delta}$, $f^*_{L,\mathrm{P}} \in L_\infty(T^{+\delta})$, and either $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(T^{+2\delta})$ for $\alpha \geq 1$ or $f^*_{L,\mathrm{P}} \in W^\alpha_2(\widetilde{T}) \subset B^\alpha_{2,\infty}(T^{+2\delta})$ for $\alpha \in \mathbb{N}$ and a bounded Lipschitz domain $\widetilde{T} \subset \mathbb{R}^d$ such that $T^{+2\delta} \subset \widetilde{T}$. Moreover, if this density of P_X is even bounded away from 0, it is well-known that the minmax rate is $n^{-\frac{2\alpha}{2\alpha+d}}$ for $\alpha > d/2$ and target functions $f^*_{L,\mathrm{P}} \in W^\alpha_2(T)$ as well as for $\alpha > d$ and $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(T)$. Modulo ξ, our rate is therefore asymptotically optimal in a minmax sense on T.

Although the obtained learning rates are arbitrary close to the optimal rates, it is needless to say that the results are not fully satisfying. Indeed, an ideal result would not contain a gap of the form n^ξ, and a close to ideal result would at least replace the gap n^ξ by a logarithmic factor. Unfortunately, even for global SVMs using Gaussian kernels, such results seem to be currently out of reach, see (Eberts and Steinwart, 2013) for the latter case. Let us briefly describe the technical obstacles. One key ingredient for both the local and the global approach are estimates on the entropy numbers e_i of the embeddings $\mathrm{id}: H_\gamma \to L_2(P_X)$ or $\mathrm{id}: H_\gamma \to \ell_\infty(X)$, see Section 6 for a definition. Several such estimates do exist. For example, Zhou (2002) and Kühn (2011) proved (optimal) super-polynomial estimates but unfortunately their bounds have a unfavorable dependence on γ, which makes it impossible to get arbitrarily close to the optimal rates, see e.g., (Xiang and Zhou, 2009) for a similar situation in which this problem occurs. For this reason we followed the path of (Eberts and Steinwart, 2013), in which we employ an entropy estimate of the form

$$e_i\big(\mathrm{id}: H_\gamma \to L_2(\mathrm{P}_X)\big) \leq c_{p,d}\, \gamma^{-p} i^{-\frac{p}{d}}\,, \qquad\qquad i \geq 1, \gamma \in (0,1]\,,$$

where $c_{p,d} \geq 1$ is a constant only depending on $p \in \mathbb{N}$ and d. Note that this estimate is clearly sub-optimal in i, but it has a significantly better behavior in γ compared to the above mentioned results. Now, using this entropy estimate, Eberts and Steinwart (2013) obtain an oracle inequality of the form

$$\mathcal{R}_{L,\mathrm{P}}(f_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \leq K_p\bigg(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{c_{p,d}^{d/p}\,\gamma^{-d}}{\lambda^{\frac{d}{2p}} n} + \frac{\tau}{n}\bigg)\,,$$

where the constant K_p is independent of γ, λ, τ, and n, and its dependence on p can be tracked, cf. (Steinwart and Christmann, 2008, p. 267). Note that for the local approach a structurally identical formula is derived implicitly in the proof of Theorem 4. Now, the rates in this paper as well as in (Eberts and Steinwart, 2013) are obtained by optimizing the right hand side with respect to both λ and γ for an arbitrarily large but fixed p. Since the resulting rates become better the larger we pick p it is tempting to consider $p = p_n \to \infty$. Unfortunately, however, this only becomes feasible if we have an explicit expression describing how $c_{p,d}$ may depend on p. For example, some preliminary considerations suggest that we could already replace the gap n^ξ by a logarithmic factor if we had a rough bound of the form $c_{p,d} \leq c_d p^{cp}$. Unfortunately, we neither could derive such a bound for $c_{p,d}$ nor could we find it in the literature. Even worse, we also asked several experts for bounding entropy numbers of function space embeddings without any success. In addition, we are unaware of any other technique that has the potential to fill the gap in either the global or the local case, and therefore we leave this problem as an open question for future research.

4.2 Data-Dependent Parameter Selection for VP-SVMs

In the previous theorems the choice of the regularization parameters $\lambda_{n,1}, \ldots, \lambda_{n,m_n}$ and the kernel widths $\gamma_{n,1}, \ldots, \gamma_{n,m_n}$ requires us to know the smoothness parameter α. Unfortunately, in practice, we usually do know neither this value nor its existence. In this subsection, we thus show that a training/validation approach similar to the one examined in (Steinwart and Christmann, 2008, Chapters 6.5, 7.4, 8.2) and (Eberts and Steinwart, 2013) achieves the same rates adaptively, i.e., without knowing α. For this purpose, let $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be sequences of finite subsets $\Lambda_n \subset (0, r_n^d]$ and $\Gamma_n \subset (0, r_n]$. For a data set $D := ((x_1, y_1), \ldots, (x_n, y_n))$, we define

$$\begin{aligned} D_1 &:= ((x_1, y_1), \ldots, (x_l, y_l)) \,, \\ D_2 &:= ((x_{l+1}, y_{l+1}), \ldots, (x_n, y_n)) \,, \end{aligned}$$

where $l := \lfloor \frac{n}{2} \rfloor + 1$ and $n \geq 4$. We further split these sets in data sets

$$\begin{aligned} D_j^{(1)} &:= \{(x_i, y_i) \in D_1 : x_i \in A_j\} \,, \qquad && j \in \{1, \ldots, m_n\} \,, \\ D_j^{(2)} &:= \{(x_i, y_i) \in D_2 : x_i \in A_j\} \,, \qquad && j \in \{1, \ldots, m_n\} \,, \end{aligned}$$

and define $l_j := |D_j^{(1)}|$ for all $j \in \{1, \ldots, m_n\}$ such that $\sum_{j=1}^{m_n} l_j = l$. For every $j \in \{1, \ldots, m_n\}$, we basically use $D_j^{(1)}$ as a training set, i.e., based on D_1 in combination with the loss function $L_j := \mathbb{1}_{A_j} L$ we compute SVM decision functions

$$f_{\mathrm{D}_j^{(1)}, \lambda_j, \gamma_j} := \operatorname*{arg\,min}_{f \in \hat{H}_{\gamma_j}(A_j)} \lambda_j \|f\|^2_{\hat{H}_{\gamma_j}(A_j)} + \mathcal{R}_{L_j, \mathrm{D}_1}(f) \,, \qquad (\lambda_j, \gamma_j) \in \Lambda_n \times \Gamma_n \,.$$

Note that $f_{\mathrm{D}_j^{(1)}, \lambda_j, \gamma_j} = 0$ if $D_j^{(1)} = \emptyset$. Next, for each j, we use D_2 in tandem with L_j (or essentially $D_j^{(2)}$) to determine a pair $(\lambda_{\mathrm{D}_2,j}, \gamma_{\mathrm{D}_2,j}) \in \Lambda_n \times \Gamma_n$ such that

$$\mathcal{R}_{L_j, \mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_j^{(1)}, \lambda_{\mathrm{D}_2,j}, \gamma_{\mathrm{D}_2,j}}\right) = \min_{(\lambda_j, \gamma_j) \in \Lambda_n \times \Gamma_n} \mathcal{R}_{L_j, \mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_j^{(1)}, \lambda_j, \gamma_j}\right) \,.$$

Finally, combining the decision functions $f_{\mathrm{D}_j^{(1)}, \lambda_{\mathrm{D}_2,j}, \gamma_{\mathrm{D}_2,j}}$ for all $j \in \{1, \ldots, m_n\}$, and defining $\boldsymbol{\lambda}_{\mathrm{D}_2} := (\lambda_{\mathrm{D}_2,1}, \ldots, \lambda_{\mathrm{D}_2,m_n})$ and $\boldsymbol{\gamma}_{\mathrm{D}_2} := (\gamma_{\mathrm{D}_2,1}, \ldots, \gamma_{\mathrm{D}_2,m_n})$, we obtain a function

$$f_{\mathrm{D}_1, \boldsymbol{\lambda}_{\mathrm{D}_2}, \boldsymbol{\gamma}_{\mathrm{D}_2}} := \sum_{j=1}^{m_n} f_{\mathrm{D}_j^{(1)}, \lambda_{\mathrm{D}_2,j}, \gamma_{\mathrm{D}_2,j}} = \sum_{j=1}^{m_n} \mathbb{1}_{A_j} f_{\mathrm{D}_j^{(1)}, \lambda_{\mathrm{D}_2,j}, \gamma_{\mathrm{D}_2,j}} \,,$$

and we call every learning method that produces these resulting decision functions $f_{\mathrm{D}_1, \boldsymbol{\lambda}_{\mathrm{D}_2}, \boldsymbol{\gamma}_{\mathrm{D}_2}}$ a *training validation Voronoi partition support vector machine* (TV-VP-SVM) w.r.t. $\Lambda \times \Gamma$. Moreover, we have, for $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_{m_n})$ and $\boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_{m_n})$,

$$\mathcal{R}_{L, \mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_1, \boldsymbol{\lambda}_{\mathrm{D}_2}, \boldsymbol{\gamma}_{\mathrm{D}_2}}\right) = \sum_{j=1}^{m_n} \mathcal{R}_{L_j, \mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_j^{(1)}, \lambda_{\mathrm{D}_2,j}, \gamma_{\mathrm{D}_2,j}}\right)$$

$$
\begin{aligned}
&= \sum_{j=1}^{m_n} \min_{(\lambda_j,\gamma_j)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L_j,\mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_j^{(1)},\lambda_j,\gamma_j}\right) \\
&= \min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{m_n}} \sum_{j=1}^{m_n} \mathcal{R}_{L_j,\mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_j^{(1)},\lambda_j,\gamma_j}\right) \\
&= \min_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{m_n}} \mathcal{R}_{L,\mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}\right) ,
\end{aligned}
$$

where $f_{\mathrm{D}_1,\boldsymbol{\lambda},\boldsymbol{\gamma}} := \sum_{j=1}^{m_n} f_{\mathrm{D}_j^{(1)},\lambda_j,\gamma_j}$ with $(\lambda_j,\gamma_j)\in\Lambda_n\times\Gamma_n$ for all $j\in\{1,\dots,m_n\}$. In other words, the function $\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda}_{\mathrm{D}_2},\boldsymbol{\gamma}_{\mathrm{D}_2}}$ really minimizes the empirical risk $\mathcal{R}_{L,\mathrm{D}_2}$ w.r.t. the validation data set D_2 and the loss L, where the minimum is taken over all functions $\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}$ with $(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{m_n}$.

Before we analyze the TV-VP-SVM algorithm, let us briefly discuss the computational complexity of the hyper-parameter selection step. To this end, we first note that the parameter selection on, e.g., the j-th cell is *completely independent* of the parameter selection on all other cells. Maybe the easiest way to visualize this is by thinking of having two cells and candidates $\Lambda=(\lambda_1,\dots,\lambda_k)$, only. Naively, this would give the candidate set $\Lambda\times\Lambda$ for the overall hyper-parameter selection procedure. However, inspecting the candidates on the first cell, we see the same results for the candidates in $\Lambda\times\{\lambda_1\}$ and in $\Lambda\times\{\lambda_2\}$ since any decision we make on the second cell does not influence our situation on the first cell. Consequently, we only need to consider the candidates $\Lambda\times\{\lambda_1\}$, that is the candidates in Λ, when performing parameter selection on the first cell, and analogously we only need to consider the candidates $\{\lambda_1\}\times\Lambda$ for the parameter selection on the second cell. Together this gives $2|\Lambda|$ many candidates, instead of $|\Lambda|^2$ many candidates of the naive approach.

Generalizing the reasoning above to m cells and $\Lambda\times\Gamma$, we easily see that our parameter selection strategy leads to the inspection of $m\times|\Lambda|\times|\Gamma|$ many candidates. Moreover, because of the independence of all cells, we could actually perform parameter selection on the cells in parallel. Clearly such a parallel approach would be easy to implement and would have minimal synchronization and communication overhead.

The following theorem presents learning rates for the above described TV-VP-SVM.

Theorem 7 *Let $r_n := cn^{-\frac{1}{\beta d}}$ with constants $c>0$ and $\beta>1$. Under the assumptions of Theorem 4 we fix sequences $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ of finite subsets $\Lambda_n\subset(0,r_n^d]$ and $\Gamma_n\subset(0,r_n]$ such that Λ_n is an $(r_n^d\varepsilon_n)$-net of $(0,r_n^d]$ and Γ_n is a δ_n-net of $(0,r_n]$ with $\varepsilon_n\le n^{-1}$ and $\delta_n\le n^{-\frac{1}{2+d}}$. Furthermore, assume that the cardinalities $|\Lambda_n|$ and $|\Gamma_n|$ grow polynomially in n. Then, for all $\xi>0$, $\tau\ge1$, and $\alpha<\frac{\beta-1}{2}d$, the TV-VP-SVM producing the decision functions $f_{\mathrm{D}_1,\boldsymbol{\lambda}_{\mathrm{D}_2},\boldsymbol{\gamma}_{\mathrm{D}_2}}$ satisfies*

$$
\mathrm{P}^n\left(\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda}_{\mathrm{D}_2},\boldsymbol{\gamma}_{\mathrm{D}_2}})-\mathcal{R}^*_{L_{J_T},\mathrm{P}}\le c\left(n^{-\frac{2\alpha}{2\alpha+d}+\xi}+\tau n^{-1}\right)\right)\ge 1-e^{-\tau} ,
$$

where $c>0$ is a constant independent of n and τ.

Once more, we can replace the assumption $f^*_{L,\mathrm{P}}\in B^\alpha_{2,\infty}(\mathrm{P}_{X|A_T})$ by $f^*_{L,\mathrm{P}}\in B^\alpha_{2,\infty}(\mathrm{P}_{X|T^{+\delta}})$ for some $\delta>0$ and obtain the same learning rate as in Theorem 7 for all $n\ge n_\delta$ although

$T^{+\delta}$ is fixed for all $n \in \mathbb{N}$. Here, recall that $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X|T^{+\delta}})$ whenever P_X has a bounded Lebesgue density on $T^{+\delta}$, $f^*_{L,\mathrm{P}} \in L_\infty(T^{+\delta})$, and either $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(T^{+2\delta})$ for $\alpha \geq 1$ or $f^*_{L,\mathrm{P}} \in W^\alpha_2(\widetilde{T}) \subset B^\alpha_{2,\infty}(T^{+2\delta})$ for $\alpha \in \mathbb{N}$ and a bounded Lipschitz domain $\widetilde{T} \subset \mathbb{R}^d$ such that $T^{+2\delta} \subset \widetilde{T}$. Moreover, let us assume that $\widetilde{T} \supseteq T^{+\delta}$ is a bounded Lipschitz domain in $\mathbb{R}^d$ such that Stein's extension operator $\mathfrak{E}$ exists and that P is a distribution on $\mathbb{R}^d \times Y$ such that P_X has a Lebesgue density g on $T^{+\delta}$ with $g \in L_\infty(T^{+\delta})$. Then, the assumptions $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\widetilde{T})$ and $f^*_{L,\mathrm{P}} \in L_\infty(\widetilde{T})$ yield $\mathfrak{E}f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X|T^{+\delta}})$ and $\mathfrak{E}f^*_{L,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$, see (Eberts, 2015, Corollary 3.4 and Theorem 3.2) for more details. Thus, applying $\mathcal{R}^*_{L_{J_T},\mathrm{P}} = \mathcal{R}_{L_{J_T},\mathrm{P}}(\mathfrak{E}f^*_{L,\mathrm{P}})$ and choosing $f_0 := \sum_{j \in J_T} \mathbb{1}_{A_j} \cdot (K_j * \mathfrak{E}f^*_{L,\mathrm{P}})$, we obtain the same results as in Corollary 6 and Theorem 7 for $n \geq n_\delta$. Obviously, the same is true, if we assume $f^*_{L,\mathrm{P}} \in W^\alpha_2(\widetilde{T})$ instead of $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\widetilde{T})$. For all these cases, note that, if P_X has a Lebesgue density that is bounded away from 0 and ∞ and either $f^*_{L,\mathrm{P}} \in W^\alpha_2(T)$ for $\alpha > d/2$ or $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(T)$ for $\alpha > d$, the achieved learning rate $n^{-\frac{2\alpha}{2\alpha+d}}$ is again asymptotically optimal modulo ξ on T in a minmax sense. Here, we only derived learning rates when using the least squares loss. However, similar rates are shown by Eberts (2015, Section 9) for quantile regression using the pinball loss.

To derive the above learning rates, we need the condition $\alpha < \frac{\beta-1}{2}d$. However, this condition restricts the set of α-values where we obtain learning rates adaptively. To be more precise, there is a trade-off between α and β. On the one hand, for small values of β only a small number of possible values for α is covered. On the other hand, for larger values of β the set of α-values where we achieve rates adaptively is increasing but the savings in terms of computing time is decreasing.

Finally, we note that if we have a fixed computational budget in terms of RAM and/or computing time, this trade-off can be approximately resolved in the following way. First, we consider a couple of candidates for β, or the resulting number of cells m. Then, we pick a suitably sized random subset of the entire training set and build Voronoi partitions of this random subset for the different candidates. For each cell of these partitions we then estimate the computational costs and finally we pick the largest candidate β for which the resulting partition still satisfies our computational budget. This procedure has several benefits: *a)* it is very cheap compared to the subsequent training and parameter selection phase, *b)* the choice of β, or m, has a clear meaning for the user, *c)* it approximately leads to widest adaptivity we can afford by our computational budget, and *d)* our experiments in the next section show that there is no significant risk for the user by focusing on the maximal computational resources.

5. Experimental Results

In this section we report a few experiments for VP-SVMs, which illustrate the influence of the chosen radius and which compare them to standard global SVMs as well as to RC-SVMs in terms of both training time and test error.

In the experiments we report here, we consider the classical COVTYPE data set, which contains 581.012 samples of dimension 54. More experimental results on additional data sets can be found in (Eberts, 2015) and in the earlier arXiv version (Eberts and Steinwart, 2014) of this paper. The code we used was an early version of Steinwart (2016), which provides

Algorithm 1 Determine a Voronoi partition of the input data

```
Require: Input data set $D_X = \{x_1, \dots, x_n\}$ with sample size $n \in \mathbb{N}$ and some radius $r > 0$.
Ensure: Working sets indicating a Voronoi partition of $D_X$.
Pick an arbitrary $z \in D_X$
$Cover_1 \leftarrow z$
$m \leftarrow 1$
while $\max_{x \in D_X} \|x - Cover\|_2 > r$ do
    $z \leftarrow \arg\max_{x \in D_X} \|x - Cover\|_2$
    $m \leftarrow m + 1$
    $Cover_m \leftarrow z$
    $WorkingSet_m \leftarrow \emptyset$
end while
for $i = 1$ to $n$ do
    $k \leftarrow \arg\min_{j \in \{1,\dots,m\}} \|x_i - Cover_j\|_2$
    $WorkingSet_k \leftarrow WorkingSet_k \cup \{x_i\}$
end for
return $WorkingSet_1, \dots, WorkingSet_m$
```

highly efficient SVM solvers for different loss functions based on the ideas developed by (Steinwart et al., 2011). In particular, it is easy to repeat every experiment by the current version of the code.

In order to prepare the data set for the experiments, we first merged the split raw data sets so that we obtained one data set. In a next step, we scaled the data component-wise such that all samples including labels lie in $[-1, 1]^{d+1}$, where d is the dimension of the input data. Finally, we generated random subsets that were afterwards randomly split into a training and a test data set. In this manner, we obtained training sets consisting of $n = 1\,000,\ 2\,500,\ 5\,000,\ 10\,000,\ 25\,000,\ 50\,000,\ 100\,000,\ 250\,000$, and $500\,000$ samples. The test data sets associated to the various training sets consist of $n_{\text{test}} = 50\,000$ random samples, apart from the training sets with $n_{\text{train}} \leq 5\,000$, for which we took $n_{\text{test}} = 10\,000$ test samples. To minimize random effects, we repeated the experiment for each setting several times. Since experiments using large data sets entail long run times, we reran every experiment using a training set of size $n \geq 50\,000$ only three times while for training sets of size $n = 10\,000,\ 25\,000$ we performed ten repetitions and for smaller training sets, namely of size $n = 1\,000,\ 2\,500,\ 5\,000$, even 100 runs.

To train the global SVM for sufficiently large data sets we used a professional compute server equipped with four INTEL XEON E7-4830 (2.13 GHz) 8-core processor, 256 GB RAM. In order to have comparable run times, we ran the experiments for the VP-SVMs and RC-SVMs on this machine, too. In all experiments we used eight cores to pre-compute the kernel matrix and to evaluate the final decision functions on the test set, but only one core for the actual solver.

Let us quickly illustrate the routines of the VP- and the RC-SVM implemented around the LS-solver. For the VP-SVM, we first split the training set by Algorithm 1 in several working sets representing a Voronoi partition w.r.t. the user-specified radius. For this purpose, Algorithm 1 initially determines a cover of the input data applying the farthest first traversal algorithm, see (Dasgupta, 2008) and (Gonzalez, 1985) for more details. Note that this procedure induces working sets whose sizes may be considerably varying. In the case of an RC-SVM the working sets form a random partition of the training samples, where their sizes are basically equal and the number of working sets is predefined by the user. Then, for the VP-SVM- as well as for the RC-SVM-algorithm the implemented LS-solver is applied on every working set. For each working set, we randomly split the respective training data set of size n_{train} in five folds to apply 5-fold cross-validation in order to deal with the hyper-parameters λ and γ taken from an 10 by 10 grid geometrically generated in $[0.001 \cdot n_{\text{train}}^{-1}, 0.1] \times [0.5 \cdot n_{\text{train}}^{-1/d}, 10]$. Finally, we obtain one decision function for each working set. To further process these decision functions the VP-SVM-algorithms picks exactly one decision function depending on the working set affiliation of the input value. On the contrary, the RC-SVM-algorithm simply takes the average of all the decision functions. Moreover, the computed decision functions are clipped at ±1. Altogether, note that the usual LS-SVM-algorithm can be interpreted as special case of both the VP-SVM- and the RC-SVM-algorithm using one working set.

The results, which are displayed in Figure 5, can be quickly summarized: Not surprisingly, smaller radii for the VP-SVM lead to less crowded cells, which in turn reduces the training time significantly. In addition, the VP-SVM is, unlike the global SVM, not affected by the amount of available memory, so that runs with more than 100.000 samples, which would require kernel matrix caching for the global SVM, are still very feasible for the VP-SVM. Despite these advantages in terms of required computational resources, however, the test errors of the VP-SVM are only a bit worse than those of the global SVM. Moreover, the test errors become slightly better with increasing radii, so that there is a clear trade-off between computational resources and test accuracy as discussed in the previous section. When comparing the RC-SVM with the global SVM, we see, not surprisingly, the same computational advantages, but the test errors become significantly worse. As a consequence, the VP-SVM clearly outperforms the RC-SVM in terms of test errors, when both approaches have about the same training time. In this respect we also like to mention that in terms of test time, the VP-SVM was significantly faster than the RC-SVM, simply because for the VP-SVM each decision function evaluation only requires the support vector of the corresponding cell, whereas the final decision function of the RC-SVM requires all support vectors. See (Eberts and Steinwart, 2014) for details.

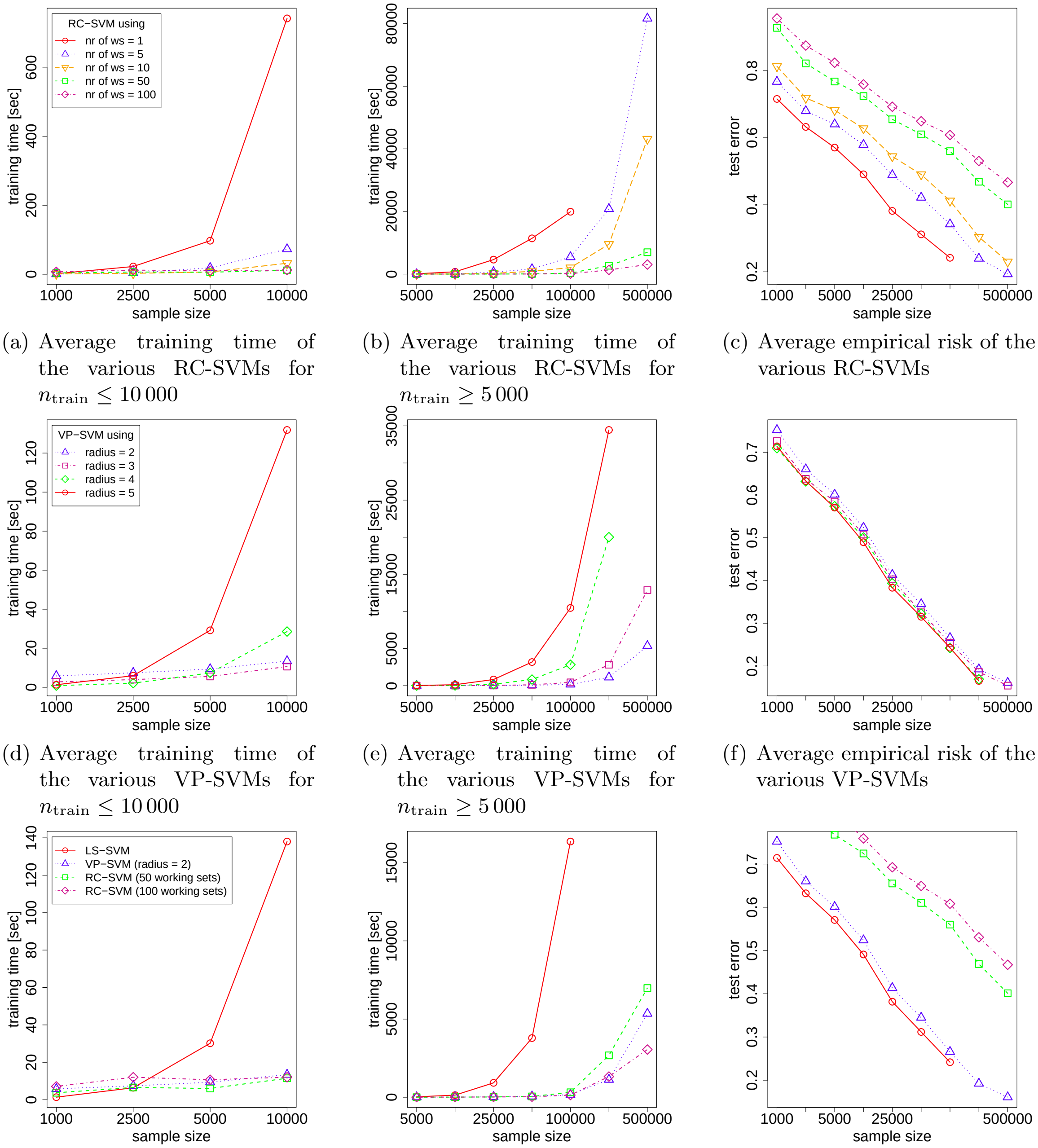
6. Proofs

This section is dedicated to prove the results of the previous sections.

We begin by recalling the definition of entropy and covering numbers. To this end, let (T, d) be a metric space. Then, the i-th (dyadic) entropy number of T is

$$e_i(T, d) := \inf\left\{ \varepsilon > 0 : \exists s_1, \dots, s_{2^{i-1}} \in T \text{ such that } T \subset \bigcup_{j=1}^{2^{i-1}} B(s_j, \varepsilon) \right\},$$

where $B_d(s, \varepsilon) := \{t \in T : d(t, s) \le \varepsilon\}$ and $\inf \emptyset := \infty$. Moreover, if $S : E \to F$ is a bounded linear operator between the normed spaces E and F, then its (dyadic) entropy numbers are defined by $e_i(S : E \to F) := e_i(SB_E, \| \cdot \|_F)$, where B_E denotes the closed unit ball of E.



(a) Average training time of the various RC-SVMs for $n_{\mathrm{train}} \leq 10\,000$

(b) Average training time of the various RC-SVMs for $n_{\mathrm{train}} \geq 5\,000$

(c) Average empirical risk of the various RC-SVMs

(d) Average training time of the various VP-SVMs for $n_{\mathrm{train}} \leq 10\,000$

(e) Average training time of the various VP-SVMs for $n_{\mathrm{train}} \geq 5\,000$

(f) Average empirical risk of the various VP-SVMs

(g) Average training time of LS-, VP-, and RC-SVMs for $n_{\mathrm{train}} \leq 10\,000$

(h) Average training time of LS-, VP-, and RC-SVMs for $n_{\mathrm{train}} \geq 5\,000$

(i) Average empirical risk of LS-, VP-, and RC-SVMs

Figure 5: Average training time and test error of LS-, VP-, and RC-SVMs for the real-world data COVTYPE depending on the training set size $n_{\mathrm{train}} = 1\,000, \ldots, 500\,000$. Subfigures (a)–(c) show the results for RC-SVMs using different numbers of working sets and Subfigures (d)–(f) illustrate the results for VP-SVMs using various radii. At the bottom, Subfigures (g)–(i) contain the average training times and the average test errors of the LS-SVM, one VP-SVM and two RC-SVMs. Here, the VP-SVM is the one which trains fastest for $n_{\mathrm{train}} = 500\,000$ and the two RC-SVMs are those which achieve for $n_{\mathrm{train}} = 500\,000$ roughly the same training time as the chosen VP-SVM. Here, note that, for $n_{\mathrm{train}} = 10\,000$, the RC-SVM using one working set trains substantially slower than the LS-SVM, even though this RC-SVM is basically an LS-SVM. As a reason for this phenomenon, we conjecture that the used compute server was busy because of other influences.

Similarly, the ε-covering number of T is defined by

$$\mathcal{N}(T,d,\varepsilon) := \inf\left\{ n \geq 1 : \exists s_1,\ldots,s_n \in T \text{ such that } T \subset \bigcup_{i=1}^{n} B_d(s_i,\varepsilon) \right\},$$

and again, this definition can be applied to bounded linear operators $S : E \to F$ by considering the set SB_E. Moreover, every subset $S \subset T$ for which for all $t \in T$ there exists an $s \in S$ with $d(s,t) \leq \varepsilon$ is called an ε-net of T. Consequently, $\mathcal{N}(T,d,\varepsilon)$ is the size of the smallest ε-net of T. Recall that entropy and covering numbers are in some sense inverse to each other. To be more precise, for all constants $a > 0$ and $q > 0$, the implication

$$e_i(T,d) \leq a i^{-1/q}, \qquad i \geq 1 \qquad \Longrightarrow \qquad \ln \mathcal{N}(T,d,\varepsilon) \leq \ln(4) \left(\frac{a}{\varepsilon}\right)^q, \qquad \forall\, \varepsilon > 0 \tag{20}$$

holds by (Steinwart and Christmann, 2008, Lemma 6.21). Additionally, (Steinwart and Christmann, 2008, Exercise 6.8) yields the opposite implication, namely

$$\ln \mathcal{N}(T,d,\varepsilon) < \left(\frac{a}{\varepsilon}\right)^q, \qquad \varepsilon > 0 \qquad \Longrightarrow \qquad e_i(T,d) \leq 3^{1/q} a i^{-1/q}, \qquad \forall\, i \geq 1\,. \tag{21}$$

With these preparations, we can now prove Lemma 1, which relates the radius r of a cover $B_r(z_1),\ldots,B_r(z_m)$ of $B_{\ell_2^d} \supset X$ defined by (4) with the number m of centers $z_1,\ldots,z_m$.

Proof [of Lemma 1] It is easy to show that $\mathcal{N}(cB_{\ell_2^d}, \ell_2^d, r) = \mathcal{N}(B_{\ell_2^d}, \ell_2^d, \frac{r}{c})$ holds for all $r, c > 0$. Moreover, applying Proposition 1.1 of (Temlyakov, 2013) yields

$$\tilde{r}^{-d} \leq \mathcal{N}(B_{\ell_2^d}, \ell_2^d, \tilde{r}) \leq \left(1 + \frac{2}{\tilde{r}}\right)^d, \qquad \tilde{r} \in (0,1].$$

Consequently, we can find a cover $(B_r(z_j))_{j=1,\ldots,m}$ of $X \subset cB_{\ell_2^d}$ with centers $z_j \in cB_{\ell_2^d}$ and radius $r \leq c$ such that

$$\left(\frac{r}{c}\right)^{-d} \leq m \leq \left(1 + \frac{2c}{r}\right)^d.$$

Since $r \leq c$, we thus have $r \leq (r + 2c)\, m^{-\frac{1}{d}} \leq 3cm^{-\frac{1}{d}}$ ■

Next, we consider a lemma that is part of our construction of the partition $(A_j)_j$ of X.

Lemma 8 *Let $(A_j')_{j=1,\ldots,m}$ be a partition of $B_{\ell_2^d}$ such that $\mathring{A}_j' \neq \emptyset$ as well as $\overline{\mathring{A}_j'} = \overline{A_j'}$ for every $j \in \{1,\ldots,m\}$. Let X be some closed subset of $B_{\ell_2^d}$ such that $\mathring{X} \neq \emptyset$ and $\overline{\mathring{X}} = X$. Without loss of generality we further assume that there is an $m_0 \leq m$ such that $A_j' \cap \mathring{X} \neq \emptyset$ for all $j \in \{1,\ldots,m_0\}$ and $A_j' \cap \mathring{X} = \emptyset$ for all $j \in \{m_0+1,\ldots,m\}$. Then, we define $A_j'' := A_j' \cap \mathring{X}$ for all $j \in \{1,\ldots,m_0\}$. Moreover, let $(A_j)_{j=1,\ldots,m_0}$ be a partition of X with $A_j'' \subset A_j \subset \overline{A_j''}$. Then, for every $j \in \{1,\ldots,m_0\}$, we have $\mathring{A}_j'' \neq \emptyset$, and thus $\mathring{A}_j \neq \emptyset$.*

Proof Let us assume that there is an $j \in \{1, \dots, m_0\}$ with $\mathring{A}''_j = \emptyset$. By our assumption we then know $A''_j = A'_j \cap \mathring{X} \neq \emptyset$, i.e., there exists some $x \in A'_j \cap \mathring{X}$. Since

$$\emptyset = \mathring{A}''_j = \text{interior}(A'_j \cap \mathring{X}) = \mathring{A}'_j \cap \text{interior}\, \mathring{X} = \mathring{A}'_j \cap \mathring{X}\,,$$

where we used the notation interior $B := \mathring{B}$, it immediately follows that $x \in \partial A'_j \subset \overline{A'_j} = \overline{\mathring{A}'_j}$. Hence, there exists a sequence $(x_n)_n \subset \mathring{A}'_j$ such that $x_n \xrightarrow{n \to \infty} x$. On the other hand, $x \in A''_j \subset \mathring{X}$ together with the fact that $\mathring{X}$ is open, gives $x_n \in \mathring{X}$ for all sufficiently large n. For such an n, we obtain $x_n \in \mathring{A}'_j \cap \mathring{X} = \mathring{A}''_j$, which contradicts the assumed $\mathring{A}''_j = \emptyset$. The second assertion follows from $\mathring{A}''_j \subset \mathring{A}_j$. ■

Next, let us consider a crucial property of the risk of functions contained in a joined RKHS.

Lemma 9 *Let* P *be a distribution on* $X \times Y$ *and* $L : X \times Y \times \mathbb{R} \to [0, \infty)$ *be a loss function. For* $A, B \subset X$ *such that* $A \cup B = X$ *and* $A \cap B = \emptyset$*, define loss functions* $L_A, L_B : X \times Y \times \mathbb{R} \to [0, \infty)$ *by* $L_A(x, y, t) = \mathbb{1}_A(x) L(x, y, t)$ *and* $L_B(x, y, t) = \mathbb{1}_B(x) L(x, y, t)$*, respectively. Furthermore, let* $f_A : X \to \mathbb{R}$ *as well as* $f_B : X \to \mathbb{R}$ *be measurable functions and* $f : X \to \mathbb{R}$ *be defined by* $f(x) = \mathbb{1}_A(x) f_A(x) + \mathbb{1}_B(x) f_B(x)$ *for all* $x \in X$*. Then, we have*

$$\mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}_{L_A,\mathrm{P}}(f_A) + \mathcal{R}_{L_B,\mathrm{P}}(f_B)\,.$$

as well as

$$\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} = \left(\mathcal{R}_{L_A,\mathrm{P}}(f_A) - \mathcal{R}^*_{L_A,\mathrm{P}}\right) + \left(\mathcal{R}_{L_B,\mathrm{P}}(f_B) - \mathcal{R}^*_{L_B,\mathrm{P}}\right)\,.$$

Proof Simple transformations using $A \cup B = X$ and $A \cap B = \emptyset$ show

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f) &= \int_{X \times Y} L\left(x, y, \mathbb{1}_A(x) f_A(x) + \mathbb{1}_B(x) f_B(x)\right)\, d\mathrm{P}(x, y) \\ &= \int_{X \times Y} \mathbb{1}_A(x) L(x, y, f_A(x)) + \mathbb{1}_B(x) L(x, y, f_B(x))\, d\mathrm{P}(x, y) \\ &= \mathcal{R}_{L_A,\mathrm{P}}(f_A) + \mathcal{R}_{L_B,\mathrm{P}}(f_B)\,.\end{aligned}$$

The second assertion follows immediately. ■

6.1 Some General Estimates on Entropy Numbers

To derive an oracle inequality for VP-SVMs we will have to relate the entropy numbers of H_j, $j \in \{1, \dots, m\}$, to those of H. Our first result establishes such a relationship for covering numbers, instead.

Lemma 10 *Let ν be a distribution on X and $A, B \subset X$ with $A \cap B = \emptyset$. Moreover, let H_A and H_B be RKHSs on A and B that are embedded into $L_2(\nu_{|A})$ and $L_2(\nu_{|B})$, respectively. Let the extended RKHSs $\hat{H}_A$ and $\hat{H}_B$ be defined as in Lemma 2 and denote their direct sum by H as in (11), where the norm is given by (12) with $\lambda_A, \lambda_B > 0$. Then, for the ε-covering number of H w.r.t. $\|\cdot\|_{L_2(\nu)}$, we have*

$$\mathcal{N}(B_H, \|\cdot\|_{L_2(\nu)}, \varepsilon) \le \mathcal{N}\left(\lambda_A^{-1/2} B_{\hat{H}_A}, \|\cdot\|_{L_2(\nu_{|A})}, \varepsilon_A\right) \cdot \mathcal{N}\left(\lambda_B^{-1/2} B_{\hat{H}_B}, \|\cdot\|_{L_2(\nu_{|B})}, \varepsilon_B\right) ,$$

where $\varepsilon_A, \varepsilon_B > 0$ and $\varepsilon := \sqrt{\varepsilon_A^2 + \varepsilon_B^2}$.

Proof First of all, we assume that there exist $a, b \in \mathbb{N}$ and functions $\hat{f}_1, \ldots, \hat{f}_a \in \lambda_A^{-\frac{1}{2}} B_{\hat{H}_A}$ and $\hat{h}_1, \ldots, \hat{h}_b \in \lambda_B^{-\frac{1}{2}} B_{\hat{H}_B}$ such that $\{\hat{f}_1, \ldots, \hat{f}_a\}$ is an ε_A-cover of $\lambda_A^{-\frac{1}{2}} B_{\hat{H}_A}$ w.r.t. $\|\cdot\|_{L_2(\nu_{|A})}$, $\{\hat{h}_1, \ldots, \hat{h}_b\}$ is an ε_B-cover of $\lambda_B^{-\frac{1}{2}} B_{\hat{H}_B}$ w.r.t. $\|\cdot\|_{L_2(\nu_{|B})}$,

$$a = \mathcal{N}(\lambda_A^{-\frac{1}{2}} B_{\hat{H}_A}, \|\cdot\|_{L_2(\nu_{|A})}, \varepsilon_A) \qquad \text{and} \qquad b = \mathcal{N}(\lambda_B^{-\frac{1}{2}} B_{\hat{H}_B}, \|\cdot\|_{L_2(\nu_{|B})}, \varepsilon_B) .$$

That is, for every function $\hat{g}_A \in \lambda_A^{-\frac{1}{2}} B_{\hat{H}_A}$, there exists an $i_A \in \{1, \ldots, a\}$ such that

$$\left\| \hat{g}_A - \hat{f}_{i_A} \right\|_{L_2(\nu_{|A})} \le \varepsilon_A , \tag{22}$$

and for every function $\hat{g}_B \in \lambda_B^{-\frac{1}{2}} B_{\hat{H}_B}$, there exists an $i_B \in \{1, \ldots, b\}$ such that

$$\left\| \hat{g}_B - \hat{h}_{i_B} \right\|_{L_2(\nu_{|B})} \le \varepsilon_B . \tag{23}$$

Let us now consider an arbitrary function $g \in B_H$. Then, there exists an $\hat{g}_A \in \lambda_A^{-\frac{1}{2}} B_{\hat{H}_A}$ and an $\hat{g}_B \in \lambda_B^{-\frac{1}{2}} B_{\hat{H}_B}$ such that $g = \hat{g}_A + \hat{g}_B$. Together with (22) and (23), this implies

$$\begin{aligned}
\left\| g - \left(\hat{f}_{i_A} + \hat{h}_{i_B}\right) \right\|_{L_2(\nu)}^2 &= \left\| \left(\hat{g}_A - \hat{f}_{i_A}\right) + \left(\hat{g}_B - \hat{h}_{i_B}\right) \right\|_{L_2(\nu)}^2 \\
&= \left\| \hat{g}_A - \hat{f}_{i_A} \right\|_{L_2(\nu_{|A})}^2 + \left\| \hat{g}_B - \hat{h}_{i_B} \right\|_{L_2(\nu_{|B})}^2 \\
&\le \varepsilon_A^2 + \varepsilon_B^2 \\
&=: \varepsilon^2 .
\end{aligned}$$

With this, we know that

$$\left\{ \hat{f}_{i_A} + \hat{h}_{i_B} \ : \ \hat{f}_{i_A} \in \{\hat{f}_1, \ldots, \hat{f}_a\} \text{ and } \hat{h}_{i_B} \in \{\hat{h}_1, \ldots, \hat{h}_b\} \right\}$$

is an ε-net of H w.r.t. $\|\cdot\|_{L_2(\nu)}$. Concerning the ε-covering number of H, this finally implies

$$\mathcal{N}(B_H, \|\cdot\|_{L_2(\nu)}, \varepsilon) \le a \cdot b = \mathcal{N}\left(\lambda_A^{-1/2} B_{\hat{H}_A}, \|\cdot\|_{L_2(\nu_{|A})}, \varepsilon_A\right) \cdot \mathcal{N}\left(\lambda_B^{-1/2} B_{\hat{H}_B}, \|\cdot\|_{L_2(\nu_{|B})}, \varepsilon_B\right).$$

■

Based on Lemma 10, the following theorem relates entropy numbers of H_A and H_B to those of H.

Theorem 11 *Let P_X be a distribution on X and $A_1, \dots, A_m \subset X$ be pairwise disjoint. Moreover, for $j \in \{1, \dots, m\}$, let H_j be a separable RKHS of a measurable kernel k_j over A_j such that $\|k_j\|^2_{L_2(\mathrm{P}_{X|A_j})} := \int_X k_j(x,x) d\mathrm{P}_{X|A_j}(x) < \infty$. Define RKHSs $\hat{H}_1, \dots, \hat{H}_m$ by Lemma 2 and the joined RKHS H by (13) with the norm (14) and weights $\lambda_1, \dots, \lambda_m > 0$. In addition, assume that there exist constants $p \in (0,1)$ and $a_j > 0$, $j \in \{1, \dots, m\}$, such that for every $j \in \{1, \dots, m\}$*

$$e_i(\mathrm{id} : H_j \to L_2(\mathrm{P}_{X|A_j})) \leq a_j \, i^{-\frac{1}{2p}} \,, \qquad i \geq 1 \,. \tag{24}$$

Then, we have

$$e_i(\mathrm{id} : H \to L_2(\mathrm{P}_X)) \leq 2\sqrt{m} \left(3 \ln(4) \sum_{j=1}^m \lambda_j^{-p} a_j^{2p} \right)^{\frac{1}{2p}} i^{-\frac{1}{2p}} \,, \qquad i \geq 1 \,,$$

and, for the average entropy numbers,

$$\mathbb{E}_{D_X \sim \mathrm{P}_X^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \leq c_p \sqrt{m} \left(\sum_{j=1}^m \lambda_j^{-p} a_j^{2p} \right)^{\frac{1}{2p}} i^{-\frac{1}{2p}} \,, \qquad i, n \geq 1 \,.$$

Proof First of all, note that the restriction operator $\mathcal{I} : B_{\hat{H}_j} \to B_{H_j}$ with $\mathcal{I}\hat{f} = f$ is an isometric isomorphism. Together with (Steinwart and Christmann, 2008, (A.36)) and assumption (24), this yields

$$\begin{aligned} e_i(\lambda_j^{-\frac{1}{2}} B_{\hat{H}_j}, L_2(\mathrm{P}_{X|A_j})) &= 2\lambda_j^{-\frac{1}{2}} e_i(B_{\hat{H}_j}, L_2(\mathrm{P}_{X|A_j})) \\ &\leq 2\lambda_j^{-\frac{1}{2}} \|\mathcal{I} : B_{\hat{H}_j} \to B_{H_j}\| e_i(B_{H_j}, L_2(\mathrm{P}_{X|A_j})) \\ &\leq 2\lambda_j^{-\frac{1}{2}} a_j i^{-\frac{1}{2p}} \,. \end{aligned}$$

Furthermore, we know by (20) that

$$\ln \mathcal{N}\left(\lambda_j^{-\frac{1}{2}} B_{\hat{H}_j}, \|\cdot\|_{L_2(\mathrm{P}_{X|A_j})}, \varepsilon\right) \leq \ln(4) \left(2\lambda_j^{-\frac{1}{2}} a_j\right)^{2p} \varepsilon^{-2p}$$

holds for all $\varepsilon > 0$. With this and $\varepsilon_j := \frac{\varepsilon}{\sqrt{m}}$ for every $j \in \{1, \dots, m\}$, Lemma 10 implies

$$\ln \mathcal{N}(B_H, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon) \leq \ln \left(\prod_{j=1}^m \mathcal{N}\left(\lambda_j^{-\frac{1}{2}} B_{\hat{H}_j}, \|\cdot\|_{L_2(\mathrm{P}_{X|A_j})}, \varepsilon_j\right) \right)$$

$$
\begin{aligned}
&= \sum_{j=1}^{m} \ln \mathcal{N}\left(\lambda_j^{-\frac{1}{2}} B_{\hat{H}_j}, \|\cdot\|_{L_2(\mathrm{P}_{X|A_j})}, \frac{\varepsilon}{\sqrt{m}}\right) \\
&\leq \sum_{j=1}^{m} \ln(4) \left(2\lambda_j^{-\frac{1}{2}} a_j\right)^{2p} \left(\frac{\sqrt{m}}{\varepsilon}\right)^{2p} \\
&= \left(2\ln(4)^{\frac{1}{2p}} \sqrt{m} \left(\sum_{j=1}^{m} \lambda_j^{-p} a_j^{2p}\right)^{\frac{1}{2p}}\right)^{2p} \varepsilon^{-2p} .
\end{aligned}
$$

Using (21), the latter bound for the covering number of B_H finally implies the following entropy estimate

$$
\begin{aligned}
e_i(\mathrm{id} : H \to L_2(\mathrm{P}_X)) &\leq 3^{\frac{1}{2p}} \left(2\ln(4)^{\frac{1}{2p}} \sqrt{m} \left(\sum_{j=1}^{m} \lambda_j^{-p} a_j^{2p}\right)^{\frac{1}{2p}}\right) i^{-\frac{1}{2p}} \\
&\leq 2\,(3\ln(4))^{\frac{1}{2p}} \sqrt{m} \left(\sum_{j=1}^{m} \lambda_j^{-p} a_j^{2p}\right)^{\frac{1}{2p}} i^{-\frac{1}{2p}} .
\end{aligned}
$$

The second assertion immediately follows by (Steinwart and Christmann, 2008, Corollary 7.31). ■

In the following subsections, we first focus on RKHSs using Gaussian RBF kernels and examine the associated entropy numbers to specify (24). Subsequently, we additionally consider the least squares loss to prove Theorem 4.

6.2 Entropy Estimates for Local Gaussian RKHSs

In this subsection, we derive an estimate in terms of assumption (24) for the RKHS $H_\gamma(A)$ over A of the Gaussian RBF kernel k_γ on $A \subset \mathbb{R}^d$ given by

$$
k_\gamma(x, x') := \exp\left(-\gamma^{-2}\|x - x'\|_2^2\right) , \qquad x, x' \in A ,
$$

for some width $\gamma > 0$. More precisely, in the subsequent theorem we determine an upper bound for the entropy numbers of the operator $\mathrm{id} : H_\gamma(A) \to L_2(\mathrm{P}_{X|A})$.

Theorem 12 *Let $X \subset \mathbb{R}^d$, P_X be a distribution on X and $A \subset X$ be such that $\mathring{A} \neq \emptyset$ and such that there exists an Euclidean ball $B \subset \mathbb{R}^d$ with radius $r > 0$ containing A, i.e., $A \subset B$. Moreover, for $0 < \gamma \leq r$, let $H_\gamma(A)$ be the RKHS of the Gaussian RBF kernel k_γ over A. Then, for all $p \in (0,1)$, there exists a constant $c_p > 0$ such that*

$$
e_i(\mathrm{id} : H_\gamma(A) \to L_2(\mathrm{P}_{X|A})) \leq c_p \sqrt{\mathrm{P}_X(A)}\, r^{\frac{d+2p}{2p}} \gamma^{-\frac{d+2p}{2p}} i^{-\frac{1}{2p}} , \qquad i \geq 1 .
$$

Proof First of all, we consider the commutative diagram

$$\begin{tikzcd}
H_\gamma(A) \arrow[r, "\mathrm{id}"] \arrow[d, "\mathcal{I}_B^{-1}\circ\mathcal{I}_A"'] & L_2(\mathrm{P}_{X|A}) \\
H_\gamma(B) \arrow[r, "\mathrm{id}"'] & \ell_\infty(B) \arrow[u, "\mathrm{id}"']
\end{tikzcd}$$

where the extension operator $\mathcal{I}_A : H_\gamma(A) \to H_\gamma(\mathbb{R}^d)$ and the restriction operator $\mathcal{I}_B^{-1} : H_\gamma(\mathbb{R}^d) \to H_\gamma(B)$ given by (Steinwart and Christmann, 2008, Corollary 4.43) are isometric isomorphisms, so that $\|\mathcal{I}_B^{-1} \circ \mathcal{I}_A : H_\gamma(A) \to H_\gamma(B)\| = 1$. Furthermore, for $f \in \ell_\infty(B)$, where $\ell_\infty(B)$ is the space of all bounded functions on B, we have

$$\|f\|_{L_2(\mathrm{P}_{X|A})} = \left(\int_X \mathbb{1}_A(x)|f(x)|^2 d\mathrm{P}_X(x)\right)^{\frac{1}{2}} \leq \|f\|_\infty \left(\int_X \mathbb{1}_A(x) d\mathrm{P}_X(x)\right)^{\frac{1}{2}} = \sqrt{\mathrm{P}_X(A)}\,\|f\|_\infty\,,$$

i.e., $\|\mathrm{id} : \ell_\infty(B) \to L_2(\mathrm{P}_{X|A})\| \leq \sqrt{\mathrm{P}_X(A)}$. Together with (Steinwart and Christmann, 2008, (A.38) and (A.39)) as well as (Steinwart and Christmann, 2008, Theorem 6.27), we obtain for all $i \geq 1$

$$\begin{aligned}
&e_i(\mathrm{id} : H_\gamma(A) \to L_2(\mathrm{P}_{X|A})) \\
&\leq \|\mathcal{I}_B^{-1} \circ \mathcal{I}_A : H_\gamma(A) \to H_\gamma(B)\| \cdot e_i(\mathrm{id} : H_\gamma(B) \to \ell_\infty(B)) \cdot \|\mathrm{id} : \ell_\infty(B) \to L_2(\mathrm{P}_{X|A})\| \\
&\leq \sqrt{\mathrm{P}_X(A)}\, c_{m,d} r^m \gamma^{-m} i^{-\frac{m}{d}}\,,
\end{aligned}$$

where $m \geq 1$ is an arbitrary integer and $c_{m,d}$ a positive constant. For $p \in (0,1)$, the choice $m = \left\lceil \frac{d}{2p} \right\rceil$ finally yields

$$e_i(\mathrm{id} : H_\gamma(A) \to L_2(\mathrm{P}_{X|A})) \leq \sqrt{\mathrm{P}_X(A)}\, c_{m,d} r^m \gamma^{-m} i^{-\frac{m}{d}} \leq c_p \sqrt{\mathrm{P}_X(A)}\, r^{\frac{d+2p}{2p}} \gamma^{-\frac{d+2p}{2p}} i^{-\frac{1}{2p}}\,.$$

■

6.3 Proofs Related to the Least Squares VP-SVMs

In this subsection, we prove the results that are linked with the least squares loss, i.e., the results of Section 4. Before we elaborate on the oracle inequality for VP-SVMs using the least squares loss as well as RKHSs of Gaussian kernels, we have to examine the excess risk

$$\mathcal{R}_{L_{J_T},\mathrm{P}}(f_0) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} = \|f_0 - f^*_{L,\mathrm{P}}\|^2_{L_2\left(\mathrm{P}_{X|A_T}\right)}\,. \tag{25}$$

Let us begin by writing for fixed $\gamma_j > 0$

$$K_j : \mathbb{R}^d \to \mathbb{R}\,, \qquad x \mapsto \sum_{\ell=1}^{s} \binom{s}{\ell} (-1)^{1-\ell} \left(\frac{2}{\ell^2\gamma_j^2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|x\|_2^2}{\ell^2\gamma_j^2}\right)\,, \tag{26}$$

and choosing $f_0 := \sum_{j=1}^m \mathbb{1}_{A_j} \cdot (K_j * f^*_{L,\mathrm{P}})$. Then, (25) can be estimated with the help of the following theorem, which is together with its proof basically a modification of (Eberts and Steinwart, 2013, Theorem 2.2). Indeed, the proofs proceed mainly identically. Note that we use the notation

$$\gamma_{\max} := \max\{\gamma_1, \dots, \gamma_m\} \qquad \text{and} \qquad \gamma_{\min} := \min\{\gamma_1, \dots, \gamma_m\}$$

in the following theorem and the associated proof. For the sake of generality, we do not only consider the Besov-like space $B^\alpha_{2,\infty}(\nu)$ in the following theorem but instead the Besov-like spaces $B^\alpha_{q,\infty}(\nu)$ for arbitrary $q \in [1,\infty)$. These Besov-like spaces are defined analogously to $B^\alpha_{2,\infty}(\nu)$, however, applying the modulus of smoothness for the $L_q(\nu)$-norm instead of the $L_2(\nu)$-norm. For an explicit definition of these spaces we refer to (Eberts, 2015, Section 3.1)

Theorem 13 *Let us fix some $q \in [1,\infty)$. Assume that ν is a finite measure on $\mathbb{R}^d$ with $\operatorname{supp}\nu =: X \subset cB_{\ell_2^d} \subset \mathbb{R}^d$ for some $c > 0$. Let $(A'_j)_{j=1,\dots,m}$ be a partition of $cB_{\ell_2^d}$. Then, $A_j := A'_j \cap X$ for all $j \in \{1,\dots,m\}$ defines a partition $(A_j)_{j=1,\dots,m}$ of X. Furthermore, let $f : \mathbb{R}^d \to \mathbb{R}$ be such that $f \in B^\alpha_{q,\infty}(\nu)$ for some $\alpha \geq 1$. For the functions $K_j : \mathbb{R}^d \to \mathbb{R}$, $j \in \{1,\dots,m\}$, defined by (26), where $s := \lfloor \alpha \rfloor + 1$ and $\gamma_1,\dots,\gamma_m > 0$, we then have*

$$\|\sum_{j=1}^m \mathbb{1}_{A_j} \cdot (K_j * f) - f\|^q_{L_q(\nu)} \leq C_{\alpha,q} \left(\frac{\gamma_{\max}}{\gamma_{\min}}\right)^d \gamma_{\max}^{q\alpha},$$

where $C_{\alpha,q} := \|f\|^q_{B^\alpha_{q,\infty}(\nu)} \left(\frac{d}{2}\right)^{\frac{q\alpha}{2}} \pi^{-\frac{1}{4}} \Gamma\left(q\alpha + \frac{1}{2}\right)^{\frac{1}{2}}$.

Proof In the following, we write $J := \{1,\dots,m\}$. To show

$$\left\|\sum_{j\in J} \mathbb{1}_{A_j} \cdot (K_j * f) - f\right\|^q_{L_q(\nu)} \leq \|f\|^q_{B^\alpha_{q,\infty}(\nu)} \left(\frac{d}{2}\right)^{\frac{q\alpha}{2}} \pi^{-\frac{1}{4}} \Gamma\left(q\alpha + \frac{1}{2}\right)^{\frac{1}{2}} \left(\frac{\gamma_{\max}}{\gamma_{\min}}\right)^d \gamma_{\max}^{q\alpha},$$

we have to proceed in a similar way as in the proof of (Eberts and Steinwart, 2013, Theorem 2.2). First of all, we use the translation invariance of the Lebesgue measure and $\exp\left(-\|u\|_2^2\right) = \exp\left(-\|-u\|_2^2\right)$ $(u \in \mathbb{R}^d)$ to obtain, for $x \in X$ and $j \in J$,

$$\begin{aligned} K_j * f\,(x) &= \int_{\mathbb{R}^d} \sum_{\ell=1}^s \binom{s}{\ell} (-1)^{1-\ell} \frac{1}{\ell^d} \left(\frac{2}{\gamma_j^2 \pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|x-t\|_2^2}{\ell^2\gamma_j^2}\right) f\,(t)\, dt \\ &= \int_{\mathbb{R}^d} \left(\frac{2}{\gamma_j^2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) \left(\sum_{\ell=1}^s \binom{s}{\ell} (-1)^{1-\ell} f\,(x+\ell h)\right) dh\,. \end{aligned}$$

With this we can derive, for $q \geq 1$,

$$\left\|\sum_{j\in J} \mathbb{1}_{A_j} \cdot (K_j * f) - f\right\|^q_{L_q(\nu)}$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^d} \left| \sum_{j\in J} \mathbb{1}_{A_j}(x) \left(K_j * f\right)(x) - f(x) \right|^q \, d\nu(x) \\
&\leq \int_{\mathbb{R}^d} \left(\sum_{j\in J} \mathbb{1}_{A_j}(x) \left| K_j * f(x) - f(x) \right| \right)^q \, d\nu(x) \\
&= \int_{\mathbb{R}^d} \sum_{j\in J} \mathbb{1}_{A_j}(x) \left| K_j * f(x) - f(x) \right|^q \, d\nu(x) \\
&= \sum_{j\in J} \int_{\mathbb{R}^d} \mathbb{1}_{A_j}(x) \left| K_j * f(x) - f(x) \right|^q \, d\nu(x) \\
&= \sum_{j\in J} \int_{\mathbb{R}^d} \mathbb{1}_{A_j}(x) \left| \int_{\mathbb{R}^d} \left(\frac{2}{\gamma_j^2 \pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) \left(\sum_{\ell=0}^{s} \binom{s}{\ell} (-1)^{2s+1-\ell} f(x+\ell h) \right) dh \right|^q \, d\nu(x) \\
&= \sum_{j\in J} \int_{\mathbb{R}^d} \mathbb{1}_{A_j}(x) \left| \int_{\mathbb{R}^d} (-1)^{s+1} \left(\frac{2}{\gamma_j^2 \pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) \triangle_h^s(f,x) \, dh \right|^q \, d\nu(x) \\
&\leq \sum_{j\in J} \int_{\mathbb{R}^d} \mathbb{1}_{A_j}(x) \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma_j^2 \pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) \left|\triangle_h^s(f,x)\right| \, dh \right)^q \, d\nu(x) \,.
\end{aligned}
$$

Then, Hölder's inequality and $\int_{\mathbb{R}^d} \exp\left(-2\gamma_j^{-2}\|h\|_2^2\right) dh = \left(\frac{\gamma_j^2 \pi}{2}\right)^{d/2}$ yield, for $q > 1$,

$$
\begin{aligned}
&\left\| \sum_{j\in J} \mathbb{1}_{A_j} \cdot (K_j * f) - f \right\|_{L_q(\nu)}^q \\
&\leq \sum_{j\in J} \int_{\mathbb{R}^d} \mathbb{1}_{A_j}(x) \left(\left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma_j^2 \pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) dh \right)^{\frac{q-1}{q}} \right. \\
&\qquad \left. \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma_j^2 \pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) \left|\triangle_h^s(f,x)\right|^q dh \right)^{\frac{1}{q}} \right)^q \, d\nu(x) \\
&= \sum_{j\in J} \int_{\mathbb{R}^d} \mathbb{1}_{A_j}(x) \int_{\mathbb{R}^d} \left(\frac{2}{\gamma_j^2 \pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) \left|\triangle_h^s(f,x)\right|^q \, dh \, d\nu(x) \\
&= \sum_{j\in J} \int_{\mathbb{R}^d} \left(\frac{2}{\gamma_j^2 \pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) \int_{\mathbb{R}^d} \mathbb{1}_{A_j}(x) \left|\triangle_h^s(f,x)\right|^q \, d\nu(x) \, dh \\
&\leq \int_{\mathbb{R}^d} \left(\frac{2}{\pi \gamma_{\min}^2}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) \int_{\mathbb{R}^d} \sum_{j\in J} \mathbb{1}_{A_j}(x) \left|\triangle_h^s(f,x)\right|^q \, d\,\nu(x) \, dh
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^d} \left(\frac{2}{\pi\gamma_{\min}^2}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) \|\triangle_h^s(f,\cdot)\|_{L_q(\nu)}^q \, dh \\
&\leq \int_{\mathbb{R}^d} \left(\frac{2}{\pi\gamma_{\min}^2}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) \omega_{s,L_q(\nu)}^q(f,\|h\|_2) \, dh \, .
\end{aligned}
$$

Moreover, for $q = 1$, we have

$$
\begin{aligned}
&\left\| \sum_{j\in J} \mathbb{1}_{A_j} \cdot (K_j * f) - f \right\|_{L_1(\nu)} \\
&\leq \sum_{j\in J} \int_{\mathbb{R}^d} \mathbb{1}_{A_j}(x) \int_{\mathbb{R}^d} \left(\frac{2}{\gamma_j^2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_j^2}\right) |\triangle_h^s(f,x)| \, dh \, d\nu(x) \\
&\leq \int_{\mathbb{R}^d} \left(\frac{2}{\pi\gamma_{\min}^2}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) \int_{\mathbb{R}^d} \sum_{j\in J} \mathbb{1}_{A_j}(x) \, |\triangle_h^s(f,x)| \, d\nu(x) \, dh \\
&\leq \int_{\mathbb{R}^d} \left(\frac{2}{\pi\gamma_{\min}^2}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) \omega_{s,L_1(\nu)}(f,\|h\|_2) \, dh \, .
\end{aligned}
$$

Consequently, we can proceed in the same way for all $q \geq 1$. To this end, note that the assumption $f \in B_{q,\infty}^\alpha(\nu)$ implies $\omega_{s,L_q(\nu)}(f,t) \leq \|f\|_{B_{q,\infty}^\alpha(\nu)} t^\alpha$ for $t > 0$. The latter together with Hölder's inequality yields

$$
\begin{aligned}
&\left\| \sum_{j\in J} \mathbb{1}_{A_j} \cdot (K_j * f) - f \right\|_{L_q(\nu)}^q \\
&\leq \int_{\mathbb{R}^d} \left(\frac{2}{\pi\gamma_{\min}^2}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) \omega_{s,L_q(\nu)}^q(f,\|h\|_2) \, dh \\
&\leq \|f\|_{B_{q,\infty}^\alpha(\nu)}^q \left(\frac{2}{\pi\gamma_{\min}^2}\right)^{\frac{d}{2}} \int_{\mathbb{R}^d} \|h\|_2^{q\alpha} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) dh \\
&\leq \|f\|_{B_{q,\infty}^\alpha(\nu)}^q \left(\frac{2}{\pi\gamma_{\min}^2}\right)^{\frac{d}{2}} \left(\int_{\mathbb{R}^d} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) dh\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \|h\|_2^{2q\alpha} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) dh\right)^{\frac{1}{2}} \\
&= \|f\|_{B_{q,\infty}^\alpha(\nu)}^q \left(\frac{2\gamma_{\max}^2}{\pi\gamma_{\min}^4}\right)^{\frac{d}{4}} \left(\int_{\mathbb{R}^d} \|h\|_2^{2q\alpha} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) dh\right)^{\frac{1}{2}} \, .
\end{aligned}
$$

Using the embedding constant $d^{\frac{q\alpha-1}{2q\alpha}}$ of $\ell_{2q\alpha}^d$ to ℓ_2^d, we obtain

$$
\begin{aligned}
\int_{\mathbb{R}^d} \|h\|_2^{2q\alpha} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) dh &\leq d^{q\alpha-1} \sum_{\ell=1}^d \int_{\mathbb{R}^d} h_\ell^{2q\alpha} \prod_{l=1}^d \exp\left(-\frac{2h_l^2}{\gamma_{\max}^2}\right) d(h_1,\dots,h_d) \\
&= d^{q\alpha-1} \sum_{\ell=1}^d \left(\frac{\gamma_{\max}^2\pi}{2}\right)^{\frac{d-1}{2}} \int_{\mathbb{R}} h_\ell^{2q\alpha} \exp\left(-\frac{2h_\ell^2}{\gamma_{\max}^2}\right) dh_\ell
\end{aligned}
$$

$$= 2d^{q\alpha}\left(\frac{\gamma_{\max}^2\pi}{2}\right)^{\frac{d-1}{2}}\int_0^\infty t^{2q\alpha}\exp\left(-\frac{2t^2}{\gamma_{\max}^2}\right)\,dt\,.$$

for $\gamma > 0$. With the substitution $t = (\frac{1}{2}\gamma_{\max}^2 u)^{\frac{1}{2}}$, the functional equation $\Gamma(t+1) = t\,\Gamma(t)$ of the Gamma function Γ, and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ we further have

$$\begin{aligned}\int_0^\infty t^{2q\alpha}\exp\left(-\frac{2t^2}{\gamma_{\max}^2}\right)\,dt &= \frac{1}{2}\frac{\gamma_{\max}}{\sqrt{2}}\left(\frac{\gamma_{\max}^2}{2}\right)^{q\alpha}\int_0^\infty u^{\left(q\alpha+\frac{1}{2}\right)-1}\exp\left(-u\right)\,du\\ &= \frac{1}{2}\frac{\gamma_{\max}}{\sqrt{2}}\left(\frac{\gamma_{\max}^2}{2}\right)^{q\alpha}\Gamma\left(q\alpha+\frac{1}{2}\right)\,.\end{aligned}$$

Altogether, we finally obtain

$$\begin{aligned}&\left\|\sum_{j\in J}\mathbb{1}_{A_j}\cdot(K_j * f) - f\right\|_{L_q(\nu)}^q\\ &\le \|f\|_{B_{q,\infty}^\alpha(\nu)}^q\left(\frac{2\gamma_{\max}^2}{\pi\gamma_{\min}^4}\right)^{\frac{d}{4}}\left(\int_{\mathbb{R}^d}\|h\|_2^{2q\alpha}\exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right)\,dh\right)^{\frac{1}{2}}\\ &\le \|f\|_{B_{q,\infty}^\alpha(\nu)}^q\left(\frac{2\gamma_{\max}^2}{\pi\gamma_{\min}^4}\right)^{\frac{d}{4}}\left(\left(\frac{d}{2}\right)^{q\alpha}\left(\frac{\pi^{d-1}}{2^d}\right)^{\frac{1}{2}}\gamma_{\max}^{2q\alpha+d}\Gamma\left(q\alpha+\frac{1}{2}\right)\right)^{\frac{1}{2}}\\ &= \|f\|_{B_{q,\infty}^\alpha(\nu)}^q\left(\frac{d}{2}\right)^{\frac{q\alpha}{2}}\pi^{-\frac{1}{4}}\Gamma\left(q\alpha+\frac{1}{2}\right)^{\frac{1}{2}}\left(\frac{\gamma_{\max}}{\gamma_{\min}}\right)^d\gamma_{\max}^{q\alpha}\,.\end{aligned}$$

■

Based on Theorems 11, 12, and 13, we can now show Theorem 4, where we denote by $L\circ f$ the function $(x,y)\mapsto L(x,y,f(x))$.

Proof [of Theorem 4] First of all, since $H_1,\ldots,H_m$ are RKHSs of Gaussian kernels, the joined RKHS H is seperable and its kernel is measurable. Moreover, since Theorem 12 provides $e_i(\mathrm{id}: H_{\gamma_j}(A_j)\to L_2(\mathrm{P}_{X|A_j})) \le a_j i^{-\frac{1}{2p}}$ for $i\ge 1$ with $a_j = \tilde{c}_p\sqrt{\mathrm{P}_X(A_j)}\,r^{\frac{d+2p}{2p}}\gamma_j^{-\frac{d+2p}{2p}}$, Theorem 11 yields

$$\mathbb{E}_{D_X\sim\mathrm{P}_X^n}e_i(\mathrm{id}: H\to L_2(\mathrm{D}_X)) \le c_p\sqrt{m}\left(\sum_{j=1}^m\lambda_j^{-p}a_j^{2p}\right)^{\frac{1}{2p}}i^{-\frac{1}{2p}}\,,\qquad i,n\ge 1\,.$$

Note that, for the least squares loss, which can be clipped at M with $Y=[-M,M]$, the supremum bound

$$L(x,y,t)\le B\,,\qquad \forall\ (x,y)\in X\times Y,\ t\in[-M,M]\tag{27}$$

holds for $B=4M^2$ and the variance bound

$$\mathbb{E}_{\mathrm{P}}\left(L\circ f - L\circ f_{L,\mathrm{P}}^*\right)^2 \le V\cdot\left(\mathbb{E}_{\mathrm{P}}\left(L\circ f - L\circ f_{L,\mathrm{P}}^*\right)\right)^{\vartheta}\,,\qquad \forall\ f: X\to[-M,M]\tag{28}$$

for $V = 16M^2$ and $\vartheta = 1$ (cf. Steinwart and Christmann, 2008, Example 7.3). Actually, (27) immediately yields the supremum bound for L_{J_T}, too. The same holds for the variance bound (28), which can be easily shown by the use of $\widetilde{f}(x) := \mathbb{1}_{\bigcup_{j\in J_T} A_j}(x)f(x) + \mathbb{1}_{X\setminus\left(\bigcup_{j\in J_T} A_j\right)}(x)f^*_{L,\mathrm{P}}(x)$ for all $f : X \to [-M, M]$. Using the constant B, we now have

$$\begin{aligned}
&\left(\max\left\{c_p\sqrt{m}\left(\sum_{j=1}^m \lambda_j^{-p}a_j^{2p}\right)^{\frac{1}{2p}}, B\right\}\right)^{2p}\\
&= \left(\max\left\{c_p\tilde{c}_p\sqrt{m}r^{\frac{d+2p}{2p}}\left(\sum_{j=1}^m \left(\lambda_j^{-1}\gamma_j^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p\right)^{\frac{1}{2p}}, B\right\}\right)^{2p}\\
&\le \left(\max\left\{c_p\tilde{c}_p m^{\frac{1}{2p}}r^{\frac{d+2p}{2p}}\left(\sum_{j=1}^m \lambda_j^{-1}\gamma_j^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^{\frac{1}{2}}, B\right\}\right)^{2p}\\
&\le \left(\max\left\{c_p\tilde{c}_p 3^{\frac{d}{2p}}r\left(\sum_{j=1}^m \lambda_j^{-1}\gamma_j^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^{\frac{1}{2}}, B\right\}\right)^{2p}\\
&\le C_p r^{2p}\left(\sum_{j=1}^m \lambda_j^{-1}\gamma_j^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p + B^{2p}\\
&=: a^{2p}\,,
\end{aligned}$$

where we used $\|\cdot\|_{\ell_p^m} \le m^{\frac{1-p}{p}}\|\cdot\|_{\ell_1^m}$, $mr^d \le 3^d$ by (5), and $C_p := c_p^{2p}\tilde{c}_p^{2p}3^d$. Then, we can apply (Steinwart and Christmann, 2008, Theorem 7.23) using the regularization parameter $\tilde{\lambda} = 1$. That is, for $\lambda_1, \dots, \lambda_m > 0$, all fixed $\tau > 0$, and for an $f_0 \in H$ and a constant $B_0 \ge B$ such that $\|L_J \circ f_0\|_\infty \le B_0$, we obtain

$$\begin{aligned}
&\sum_{j=1}^m \lambda_j \|f_{\mathrm{D}_j,\lambda_j}\|^2_{\hat{H}_j} + \mathcal{R}_{L_J,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}}) - \mathcal{R}^*_{L_J,\mathrm{P}}\\
&= \|f_{\mathrm{D},\boldsymbol{\lambda}}\|^2_H + \mathcal{R}_{L_J,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}}) - \mathcal{R}^*_{L_J,\mathrm{P}}\\
&\le 9\left(\|f_0\|^2_H + \mathcal{R}_{L_J,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_J,\mathrm{P}}\right) + C\left(a^{2p}n^{-1}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}\\
&\le 9\left(\sum_{j=1}^m \lambda_j\|\mathbb{1}_{A_j}f_0\|^2_{\hat{H}_j} + \mathcal{R}_{L_J,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_J,\mathrm{P}}\right) + C\left(a^{2p}n^{-1}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}
\end{aligned} \tag{29}$$

with probability P^n not less than $1 - 3e^{-\tau}$, where $C > 0$ is the constant of (Steinwart and Christmann, 2008, Theorem 7.23) only depending on p, M, V, ϑ, and B. To continue estimate (29), we have to choose a function $f_0 \in H$. To this end, we define functions $K_j : \mathbb{R}^d \to \mathbb{R}$, $j \in \{1, \dots, m\}$, by (26), where $s := \lfloor\alpha\rfloor + 1$ and $\gamma_j > 0$. Then, we define f_0 by convolving each K_j with the Bayes decision function $f^*_{L,\mathrm{P}}$, that is

$$f_0(x) := \sum_{j\in J_T} \mathbb{1}_{A_j}(x)\cdot(K_j * f^*_{L,\mathrm{P}})(x)\,, \qquad\qquad x \in \mathbb{R}^d\,.$$

Now, to show that f_0 is indeed a suitable function to bound the approximation error, we first need to ensure that f_0 is contained in H. In addition, we need to derive bounds for both, the regularization term and the excess risk of f_0. To this end, we apply (Eberts and Steinwart, 2013, Theorem 2.3) and obtain, for every $j \in J_T$,

$$\big(K_j * f^*_{L,\mathrm{P}}\big)_{|A_j} \in H_{\gamma_j}(A_j)$$

with

$$\begin{aligned}\|\mathbb{1}_{A_j} f_0\|_{\hat{H}_{\gamma_j}(A_j)} &= \big\|\mathbb{1}_{A_j}(K_j * f^*_{L,\mathrm{P}})\big\|_{\hat{H}_{\gamma_j}(A_j)} \\ &= \Big\|\big(K_j * f^*_{L,\mathrm{P}}\big)_{|A_j}\Big\|_{H_{\gamma_j}(A_j)} \\ &\leq (\gamma_j\sqrt{\pi})^{-\frac{d}{2}}(2^s-1)\|f^*_{L,\mathrm{P}}\|_{L_2(\mathbb{R}^d)}\,.\end{aligned}$$

This implies

$$f_0 = \sum_{j\in J_T} \underbrace{\mathbb{1}_{A_j}(K_j * f^*_{L,\mathrm{P}})}_{\in \hat{H}_{\gamma_j}(A_j)} \in H_{J_T}\,.$$

Besides, note that $0 \in \hat{H}_{\gamma_j}(A_j)$ for every $j \in \{1,\dots,m\}$ such that f_0 can be written as $f_0 = \sum_{j=1}^m f_j$, where

$$f_j := \begin{cases}\mathbb{1}_{A_j}(K_j * f^*_{L,\mathrm{P}})\,, & j \in J_T\,,\\ 0\,, & j \notin J_T\,.\end{cases}$$

Obviously, the latter implies $f_0 \in H$. Furthermore, for $A_T := \bigcup_{j\in J_T} A_j$, (25) and Theorem 13 yield

$$\begin{aligned}\mathcal{R}_{L_{J_T},\mathrm{P}}(f_0) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} &= \|f_0 - f^*_{L,\mathrm{P}}\|^2_{L_2(\mathrm{P}_{X|A_T})} \\ &= \|\sum_{j\in J_T}\mathbb{1}_{A_j}(K_j * f^*_{L,\mathrm{P}}) - f^*_{L,\mathrm{P}}\|^2_{L_2(\mathrm{P}_{X|A_T})} \\ &\leq C_{\alpha,2}\left(\frac{\max_{j\in J_T}\gamma_j}{\min_{j\in J_T}\gamma_j}\right)^d \max_{j\in J_T}\gamma_j^{2\alpha}\,,\end{aligned}$$

where $C_{\alpha,2}$ is a constant only depending on α, d, and $\|f^*_{L,\mathrm{P}}\|_{B^{\alpha}_{2,\infty}(\mathrm{P}_{X|A_T})}$. Next, we derive a bound for $\|L \circ f_0\|_\infty$ using (Eberts and Steinwart, 2013, Theorem 2.3) which provides, for every $x \in X$, the supremum bound

$$|f_0(x)| = \left|\sum_{j\in J_T}\mathbb{1}_{A_j}(x)\cdot(K_j * f^*_{L,\mathrm{P}})(x)\right| \leq \sum_{j\in J_T}\mathbb{1}_{A_j}(x)\left|K_j * f^*_{L,\mathrm{P}}(x)\right| \leq (2^s-1)\left\|f^*_{L,\mathrm{P}}\right\|_{L_\infty(\mathbb{R}^d)}\,.$$

The latter implies

$$\|L_{J_T}\circ f_0\|_\infty = \sup_{(x,y)\in X\times Y}|L(y, f_0(x))|$$

$$
\begin{aligned}
&\le \sup_{(x,y)\in X\times Y} \left(M^2 + 2M|f_0(x)| + |f_0(x)|^2\right) \\
&\le 4^s \max\left\{M^2, \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\right\} ,
\end{aligned}
$$

i.e., $B_0 := 4^s \max\{M^2, \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\}$. Applying (29) then yields

$$
\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&\le \sum_{j=1}^m \lambda_j \|f_{\mathrm{D}_j,\lambda_j,\gamma_j}\|^2_{\hat{H}_{\gamma_j}(A_j)} + \mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&\le 9 \left(\sum_{j=1}^m \lambda_j \|\mathbb{1}_{A_j} f_0\|^2_{\hat{H}_{\gamma_j}(A_j)} + \mathcal{R}_{L_{J_T},\mathrm{P}}(f_0) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \right) \\
&\quad + C\left(a^{2p} n^{-1}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n} \\
&\le 9 \left(\sum_{j\in J_T} \lambda_j (\gamma_j\sqrt{\pi})^{-d} (2^s-1)^2 \|f^*_{L,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)} + C_{\alpha,2} \left(\frac{\max_{j\in J_T}\gamma_j}{\min_{j\in J_T}\gamma_j}\right)^d \max_{j\in J_T} \gamma_j^{2\alpha} \right) \\
&\quad + CC_p r^{2p} \left(\sum_{j=1}^m \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^p n^{-1} + CB^{2p} n^{-1} + \frac{3456M^2\tau}{n} \\
&\quad + 15\cdot 4^s \max\{M^2, \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\} \frac{\tau}{n} \\
&\le 9(2^s-1)^2 \pi^{-\frac{d}{2}} \|f^*_{L,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)} \sum_{j\in J_T} \lambda_j \gamma_j^{-d} + 9C_{\alpha,2} \left(\frac{\max_{j\in J_T}\gamma_j}{\min_{j\in J_T}\gamma_j}\right)^d \max_{j\in J_T} \gamma_j^{2\alpha} \\
&\quad + CC_p r^{2p} \left(\sum_{j=1}^m \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^p n^{-1} + 16^p C M^{4p} n^{-1} \\
&\quad + \left(3456M^2 + 15\cdot 4^s \max\{M^2, \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\}\right) \frac{\tau}{n}
\end{aligned}
$$

with probability P^n not less than $1-3e^{-\tau}$. Finally, for $\hat{\tau}\ge 1$, a variable transformation implies

$$
\begin{aligned}
&\sum_{j=1}^m \lambda_j \|f_{\mathrm{D}_j,\lambda_j,\gamma_j}\|^2_{\hat{H}_{\gamma_j}(A_j)} + \mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&\le C_{M,\alpha,p} \left(\sum_{j\in J_T} \lambda_j \gamma_j^{-d} + \left(\frac{\max_{j\in J_T}\gamma_j}{\min_{j\in J_T}\gamma_j}\right)^d \max_{j\in J_T} \gamma_j^{2\alpha} + r^{2p} \left(\sum_{j=1}^m \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^p n^{-1} + \hat{\tau} n^{-1} \right)
\end{aligned}
$$

with probability P^n not less than $1-e^{-\hat{\tau}}$, where the constant $C_{M,\alpha,p}$ is defined by

$$
C_{M,\alpha,p} := \max\Bigg\{ 9(2^s-1)^2 \pi^{-\frac{d}{2}} \|f^*_{L,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)} \,,\, 9\|f^*_{L,\mathrm{P}}\|^2_{B^\alpha_{2,\infty}(\mathrm{P}_{X|A_T})} \left(\frac{d}{2}\right)^\alpha \pi^{-\frac{1}{4}} \Gamma\left(2\alpha+\frac{1}{2}\right)^{\frac{1}{2}} ,
$$

$$3^d C c_p^{2p} \tilde{c}_p^{2p}\,,\ 16^p C M^{4p} + \Big(3456 M^2 + 15 \cdot 4^s \max\{M^2, \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\}\Big)\,(1+\ln(3))\Bigg\}\,.$$

■

Next, using the just proven oracle inequality presented in Theorem 4, we show the learning rates of Theorem 5 in only a few steps.

Proof [of Theorem 5] First of all, we define sequences $\tilde{\lambda}_n := c_2 n^{-1}$ and $\tilde{\gamma}_n := c_3 n^{-\frac{1}{2\alpha+d}}$ to simplify the presentation. Then, Theorem 4, $\sum_{j=1}^{m_n} \mathrm{P}_X(A_j) = 1$, and $|J_T| \leq m_n \leq 3^d r_n^{-d}$ together with $\lambda_{n,j} = r_n^d \tilde{\lambda}_n$ and $\gamma_{n,j} = \tilde{\gamma}_n$ for all $j \in \{1, \ldots, m_n\}$ yield

$$\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&\leq C_{M,\alpha,p}\left(\sum_{j\in J_T}\lambda_{n,j}\gamma_{n,j}^{-d} + \left(\frac{\max_{j\in J_T}\gamma_{n,j}}{\min_{j\in J_T}\gamma_{n,j}}\right)^d \max_{j\in J_T}\gamma_{n,j}^{2\alpha} + r_n^{2p}\left(\sum_{j=1}^{m_n}\lambda_{n,j}^{-1}\gamma_{n,j}^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p n^{-1} + \frac{\tau}{n}\right) \\
&= C_{M,\alpha,p}\left(|J_T| r_n^d \tilde{\lambda}_n \tilde{\gamma}_n^{-d} + \tilde{\gamma}_n^{2\alpha} + r_n^{(2-d)p}\tilde{\lambda}_n^{-p}\tilde{\gamma}_n^{-(d+2p)}\left(\sum_{j=1}^{m_n}\mathrm{P}_X(A_j)\right)^p n^{-1} + \tau n^{-1}\right) \\
&\leq 3^d C_{M,\alpha,p}\left(\tilde{\lambda}_n\tilde{\gamma}_n^{-d} + \tilde{\gamma}_n^{2\alpha} + \tilde{\lambda}_n^{-p}\tilde{\gamma}_n^{-(d+2p)} r_n^{(2-d)p} n^{-1} + \tau n^{-1}\right)\,.
\end{aligned}$$

Using the choices $\tilde{\lambda}_n = c_2 n^{-1}$, $\tilde{\gamma}_n = c_3 n^{-\frac{1}{2\alpha+d}}$, as well as $r_n = c_1 n^{-\frac{1}{\beta d}}$ finally implies

$$\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&\leq 3^d C_{M,\alpha,p}\left(\tilde{\lambda}_n\tilde{\gamma}_n^{-d} + \tilde{\gamma}_n^{2\alpha} + \tilde{\lambda}_n^{-p}\tilde{\gamma}_n^{-(d+2p)} r_n^{(2-d)p} n^{-1} + \tau n^{-1}\right) \\
&\leq \hat{C}_{M,\alpha,p}\left(n^{-1} n^{\frac{d}{2\alpha+d}} + n^{-\frac{2\alpha}{2\alpha+d}} + n^p n^{\frac{d+2p}{2\alpha+d}} n^{-\frac{(2-d)p}{\beta d}} n^{-1} + \tau n^{-1}\right) \\
&= \hat{C}_{M,\alpha,p}\left(n^{-\frac{2\alpha}{2\alpha+d}} + n^{-\frac{2\alpha}{2\alpha+d}} + n^{-\frac{2\alpha}{2\alpha+d}+\left(1+\frac{2}{2\alpha+d}+\frac{1}{\beta}-\frac{2}{\beta d}\right)p} + \tau n^{-1}\right) \\
&\leq C\left(n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \tau n^{-1}\right)
\end{aligned}$$

with probability P^n not less than $1 - e^{-\tau}$, where $C > 0$ is a constant and $\xi \geq \left(1 + \frac{2}{2\alpha+d} + \frac{1}{\beta} - \frac{2}{\beta d}\right) p > 0$. ■

Proof [of Corollary 6] For simplicity of notation, we write $\boldsymbol{\lambda}$, λ_j, $\boldsymbol{\gamma}$, and γ_j instead of $\boldsymbol{\lambda}_n$, $\lambda_{n,j}$, $\boldsymbol{\gamma}_n$, and $\gamma_{n,j}$. Since $\bigcup_{j\in J_T} A_j \subset T^{+\delta}$ for all $n \geq n_\delta$, the assumption $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X|T^{+\delta}})$ implies

$$f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X|\bigcup_{j\in J_T} A_j})\,.$$

With this, Theorems 4 and 5 immediately yield

$$\mathcal{R}_{L_T,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_T,\mathrm{P}}$$

$$
\begin{aligned}
&\leq \sum_{j=1}^{m} \lambda_j \| f_{\mathrm{D}_j,\lambda_j,\gamma_j} \|^2_{\hat{H}_{\gamma_j}(A_j)} + \mathcal{R}_{L_T,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_T,\mathrm{P}} \\
&\leq \sum_{j=1}^{m} \lambda_j \| f_{\mathrm{D}_j,\lambda_j,\gamma_j} \|^2_{\hat{H}_{\gamma_j}(A_j)} + \mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&\leq C_{M,\alpha,p} \left(\sum_{j\in J_T} \lambda_j \gamma_j^{-d} + \left(\frac{\max_{j\in J_T} \gamma_j}{\min_{j\in J_T} \gamma_j} \right)^{d} \max_{j\in J_T} \gamma_j^{2\alpha} + r^{2p} \left(\sum_{j=1}^{m} \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^{p} n^{-1} + \frac{\tau}{n} \right) \\
&\leq C \left(n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \tau n^{-1} \right)
\end{aligned}
$$

with probability P^n not less than $1 - e^{-\tau}$, where $\xi \geq \left(1 + \frac{2}{2\alpha+d} + \frac{1}{\beta} - \frac{2}{\beta d}\right) p > 0$. Moreover, the constants $C_{M,\alpha,p} > 0$ and $C > 0$ coincide with those of Theorems 4 and 5. ■

It remains to prove Theorem 7. However, we previously have to consider the following technical lemma.

Lemma 14 *Let $d \geq 1$ and $r_n := cn^{-\frac{1}{\beta d}}$ with $\beta > 1$ and a constant $c > 0$. We fix finite subsets $\Lambda_n \subset (0, r_n^d]$ and $\Gamma_n \subset (0, r_n]$ such that Λ_n is an $(r_n^d \varepsilon_n)$-net of $(0, r_n^d]$ and Γ_n is an δ_n-net of $(0, r_n]$ with $0 < \varepsilon_n \leq n^{-1}$, $\delta_n > 0$, $r_n^d \in \Lambda_n$, and $r_n \in \Gamma_n$. Moreover, let $J \subset \{1, \ldots, m_n\}$ be an arbitrary non-empty index set and $|J| \leq m_n \leq 3^d r_n^{-d}$. Then, for all $0 < \alpha < \frac{\beta-1}{2} d$, $n \geq 1$, and all $p \in (0,1)$ with $p \leq \frac{\beta d - 2\alpha - d}{2\alpha+d+2}$, we have*

$$
\begin{aligned}
&\inf_{(\lambda_j,\gamma_j)_{j=1}^{m_n} \in (\Lambda_n \times \Gamma_n)^{m_n}} \left(\sum_{j\in J} \lambda_j \gamma_j^{-d} + \left(\frac{\max_{j\in J} \gamma_j}{\min_{j\in J} \gamma_j} \right)^{d} \max_{j\in J} \gamma_j^{2\alpha} + r_n^{2p} \left(\sum_{j=1}^{m_n} \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^{p} n^{-1} \right) \\
&\leq C \left(n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \delta_n^{2\alpha} \right) ,
\end{aligned}
$$

where $\xi := \big(\frac{2\alpha(2\alpha+d+2)}{(2\alpha+d)((2\alpha+d)(1+p)+2p)} + \max\{ \frac{d-2}{\beta d}, 0 \} \big) p$ and $C > 0$ is a constant independent of n, Λ_n, ε_n, Γ_n, and δ_n.

Proof Without loss of generality, we may assume that Λ_n and Γ_n are of the form $\Lambda_n = \{\lambda^{(1)}, \ldots, \lambda^{(u)}\}$ and $\Gamma_n = \{\gamma^{(1)}, \ldots, \gamma^{(v)}\}$ with $\lambda^{(u)} = r_n^d$ and $\gamma^{(v)} = r_n$ as well as $\lambda^{(i-1)} < \lambda^{(i)}$ and $\gamma^{(\ell-1)} < \gamma^{(\ell)}$ for all $i = 2, \ldots, u$ and $\ell = 2, \ldots, v$. With $\lambda^{(0)} := 0$ and $\gamma^{(0)} := 0$ it is easy to see that

$$
\lambda^{(i)} - \lambda^{(i-1)} \leq 2 r_n^d \varepsilon_n \qquad \text{and} \qquad \gamma^{(\ell)} - \gamma^{(\ell-1)} \leq 2\delta_n \tag{30}
$$

hold for all $i = 1, \ldots, u$ and $\ell = 1, \ldots, v$. Furthermore, define $\lambda^* := n^{-\frac{2\alpha+d}{(2\alpha+d)(1+p)+2p}}$ and $\gamma^* := cn^{-\frac{1}{(2\alpha+d)(1+p)+2p}}$. Then, there exist indices $i \in \{1, \ldots, u\}$ and $\ell \in \{1, \ldots, v\}$ with $\lambda^{(i-1)} \leq r_n^d \lambda^* \leq \lambda^{(i)}$ and $\gamma^{(\ell-1)} \leq \gamma^* \leq \gamma^{(\ell)}$. Together with (30), this yields

$$
r_n^d \lambda^* \leq \lambda^{(i)} \leq r_n^d \lambda^* + 2 r_n^d \varepsilon_n \qquad \text{and} \qquad \gamma^* \leq \gamma^{(\ell)} \leq \gamma^* + 2\delta_n \, . \tag{31}
$$

Moreover, the definition of λ^* implies $\varepsilon_n \leq \lambda^*$ and the one of γ^* implies $\gamma^* \leq r_n$ for $\alpha < \frac{\beta-1}{2}d$ and $p \in (0, p^*]$, where $p^* := \frac{\beta d - 2\alpha - d}{2\alpha+d+2}$. Additionally, it is easy to check that

$$\lambda^*\left(\gamma^*\right)^{-d} + \left(\gamma^*\right)^{2\alpha} + \left(\lambda^*\right)^{-p}\left(\gamma^*\right)^{-(d+2p)} r_n^{(2-d)p} n^{-1} \leq \hat{c} n^{-\frac{2\alpha}{(2\alpha+d)(1+p)+2p}+\max\left\{\frac{d-2}{\beta d},0\right\}p}, \tag{32}$$

where $\hat{c}$ is a positive constant. Using (31), the bound $|J| \leq m_n \leq 3^d r_n^{-d}$, and (32), we obtain

$$\begin{aligned}
&\inf_{(\lambda_j,\gamma_j)_{j=1}^{m_n} \in (\Lambda_n \times \Gamma_n)^{m_n}} \left(\sum_{j\in J} \lambda_j \gamma_j^{-d} + \left(\frac{\max_{j\in J}\gamma_j}{\min_{j\in J}\gamma_j}\right)^d \max_{j\in J} \gamma_j^{2\alpha} + r_n^{2p}\left(\sum_{j=1}^{m_n} \lambda_j^{-1}\gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j)\right)^p n^{-1}\right)\\
&\leq \sum_{j\in J} \lambda^{(i)}\left(\gamma^{(\ell)}\right)^{-d} + \left(\gamma^{(\ell)}\right)^{2\alpha} + \left(\sum_{j=1}^{m_n}\left(\lambda^{(i)}\right)^{-1}\left(\gamma^{(\ell)}\right)^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p r_n^{2p} n^{-1}\\
&\leq |J|\lambda^{(i)}\left(\gamma^{(\ell)}\right)^{-d} + \left(\gamma^{(\ell)}\right)^{2\alpha} + \left(\lambda^{(i)}\right)^{-p}\left(\gamma^{(\ell)}\right)^{-(d+2p)} r_n^{2p} n^{-1}\\
&\leq |J|\left(r_n^d\lambda^* + 2r_n^d\varepsilon_n\right)\left(\gamma^*\right)^{-d} + \left(\gamma^* + 2\delta_n\right)^{2\alpha} + \left(r_n^d\lambda^*\right)^{-p}\left(\gamma^*\right)^{-(d+2p)} r_n^{2p} n^{-1}\\
&\leq 3^d \cdot 3\lambda^*\left(\gamma^*\right)^{-d} + \left(\gamma^* + 2\delta_n\right)^{2\alpha} + \left(\lambda^*\right)^{-p}\left(\gamma^*\right)^{-(d+2p)} r_n^{(2-d)p} n^{-1}\\
&\leq \tilde{c}\left(\lambda^*\left(\gamma^*\right)^{-d} + \left(\gamma^*\right)^{2\alpha} + \left(\lambda^*\right)^{-p}\left(\gamma^*\right)^{-(d+2p)} r_n^{(2-d)p} n^{-1}\right) + \tilde{c}\delta_n^{2\alpha}\\
&\leq \tilde{c}\hat{c} n^{-\frac{2\alpha}{(2\alpha+d)(1+p)+2p}+\max\left\{\frac{d-2}{\beta d},0\right\}p} + \tilde{c}\delta_n^{2\alpha}\\
&\leq C\left(n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \delta_n^{2\alpha}\right)
\end{aligned}$$

with $\xi := \left(\frac{2\alpha(2\alpha+d+2)}{(2\alpha+d)((2\alpha+d)(1+p)+2p)} + \max\left\{\frac{d-2}{\beta d}, 0\right\}\right)p$ and constants $\tilde{c} > 0$ and $C > 0$ independent of n, Λ_n, ε_n, Γ_n, and δ_n. ■

In the end, we show Theorem 7 using Theorem 4 as well as Lemma 14.

Proof [of Theorem 7] Let l be defined by $l := \left\lfloor \frac{n}{2} \right\rfloor + 1$, i.e., $l \geq \frac{n}{2}$. With this, Theorem 4 yields with probability P^l not less than $1 - |\Lambda_n \times \Gamma_n|^{m_n} e^{-\tau}$ that

$$\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}}\\
&\leq \frac{c_1}{2}\left(\sum_{j\in J_T}\lambda_j\gamma_j^{-d} + \left(\frac{\max_{j\in J_T}\gamma_j}{\min_{j\in J_T}\gamma_j}\right)^d \max_{j\in J_T}\gamma_j^{2\alpha} + r_n^{2p}\left(\sum_{j=1}^{m_n}\lambda_j^{-1}\gamma_j^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p l^{-1} + \tau l^{-1}\right)\\
&\leq c_1\left(\sum_{j\in J_T}\lambda_j\gamma_j^{-d} + \left(\frac{\max_{j\in J_T}\gamma_j}{\min_{j\in J_T}\gamma_j}\right)^d \max_{j\in J_T}\gamma_j^{2\alpha} + r_n^{2p}\left(\sum_{j=1}^{m_n}\lambda_j^{-1}\gamma_j^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p n^{-1} + \tau n^{-1}\right)
\end{aligned} \tag{33}$$

for all $(\lambda_j, \gamma_j) \in \Lambda_n \times \Gamma_n$, $j \in \{1, \ldots, m_n\}$, simultaneously, where $c_1 > 0$ is a constant independent of n, τ, $\boldsymbol{\lambda}$, and $\boldsymbol{\gamma}$. Furthermore, the oracle inequality of (Steinwart and Christmann, 2008, Theorem 7.2) for empirical risk minimization, $n - l \geq \frac{n}{2} - 1 \geq \frac{n}{4}$, and $\tau_n := \tau + \ln(1 + |\Lambda_n \times \Gamma_n|^{m_n})$ yield

$$\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda}_{\mathrm{D}_2},\boldsymbol{\gamma}_{\mathrm{D}_2}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&< 6 \left(\inf_{(\lambda_j,\gamma_j)_{j=1}^{m_n} \in (\Lambda_n \times \Gamma_n)^{m_n}} \mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \right) + 512 M^2 \frac{\tau_n}{n-l} \\
&< 6 \left(\inf_{(\lambda_j,\gamma_j)_{j=1}^{m_n} \in (\Lambda_n \times \Gamma_n)^{m_n}} \mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \right) + 2048 M^2 \frac{\tau_n}{n}
\end{aligned} \tag{34}$$

with probability P^{n-l} not less than $1 - e^{-\tau}$. With (33), (34), and Lemma 14 we can conclude

$$\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda}_{\mathrm{D}_2},\boldsymbol{\gamma}_{\mathrm{D}_2}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&< 6 \left(\inf_{(\lambda_j,\gamma_j)_{j=1}^{m_n} \in (\Lambda_n \times \Gamma_n)^{m_n}} \mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \right) + 2048 M^2 \frac{\tau_n}{n} \\
&\leq 6c_1 \left(\inf_{(\lambda_j,\gamma_j)_{j=1}^{m_n} \in (\Lambda_n \times \Gamma_n)^{m_n}} \left(\sum_{j \in J_T} \lambda_j \gamma_j^{-d} + \left(\frac{\max_{j \in J_T} \gamma_j}{\min_{j \in J_T} \gamma_j} \right)^d \max_{j \in J_T} \gamma_j^{2\alpha} \right.\right. \\
&\qquad\qquad \left.\left. + r_n^{2p} \left(\sum_{j=1}^{m_n} \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^p n^{-1} \right) + \tau n^{-1} \right) + 2048 M^2 \frac{\tau_n}{n} \\
&\leq 6c_1 \left(C \left(n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \delta_n^{2\alpha} \right) + \tau n^{-1} \right) + 2048 M^2 \frac{\tau_n}{n} \\
&\leq 12 c_1 C n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \left(6c_1 \tau + 2048 M^2 \tau_n \right) n^{-1}
\end{aligned}$$

with probability P^n not less than $1 - (1 + |\Lambda_n \times \Gamma_n|^{m_n})\, e^{-\tau}$. Finally, a variable transformation yields

$$\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda}_{\mathrm{D}_2},\boldsymbol{\gamma}_{\mathrm{D}_2}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&< 12 c_1 C n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \big(6c_1 \left(\tau + \ln\left(1 + |\Lambda_n \times \Gamma_n|^{m_n}\right)\right) \\
&\quad + 2048 M^2 \left(\tau + 2 \ln\left(1 + |\Lambda_n \times \Gamma_n|^{m_n}\right)\right) \big) n^{-1} \\
&\leq 12 c_1 C n^{-\frac{2\alpha}{2\alpha+d}+\xi} + (6c_1 + 2048 M^2) \left(\tau + 2 m_n \ln\left(1 + |\Lambda_n \times \Gamma_n|\right)\right) n^{-1} \\
&\leq 12 c_1 C n^{-\frac{2\alpha}{2\alpha+d}+\xi} + (6c_1 + 2048 M^2) \left(\tau + 2 \cdot 3^d r_n^{-d} \ln\left(1 + |\Lambda_n \times \Gamma_n|\right)\right) n^{-1} \\
&= 12 c_1 C n^{-\frac{2\alpha}{2\alpha+d}+\xi} + (6c_1 + 2048 M^2) \left(\tau n^{-1} + 2 \cdot 3^d c^{-d} \ln\left(1 + |\Lambda_n \times \Gamma_n|\right) n^{-\frac{\beta-1}{\beta}} \right) \\
&< \left(12 c_1 C + 2 \cdot 3^d c^{-d} (6c_1 + 2048 M^2) \ln\left(1 + |\Lambda_n \times \Gamma_n|\right)\right) n^{-\frac{2\alpha}{2\alpha+d}+\xi} + (6c_1 + 2048 M^2) \tau n^{-1}
\end{aligned}$$

with probability P^n not less than $1 - e^{-\tau}$, where we used

$$\alpha < \frac{\beta - 1}{2} d \quad \Longleftrightarrow \quad n^{-\frac{\beta-1}{\beta}} < n^{-\frac{2\alpha}{2\alpha+d}}$$

in the last step. ■

Acknowledgements

We would like to thank the Institute for Applied Analysis and Numerical Simulation of the University of Stuttgart for placing their professional compute servers at our disposal. We further like to thank the anonymous reviewers for many valuable comments, which improved the final version of this paper. Finally, we like to thank the action editor, S. van de Geer, for her patience during the revision process.

References

R. A. Adams and J. J. F. Fournier. *Sobolev Spaces.* Academic Press, New York, 2nd edition, 2003.

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

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

H. Berens and R. DeVore. Quantitative Korovkin theorems for positive linear operators on L_p-spaces. *Trans. Amer. Math. Soc.*, 245:349–361, 1978.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer, Boston, 2004.

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

E. Blanzieri and A. Bryl. Evaluation of the highest probability SVM nearest neighbor classifier with variable relative error cost. In *Proceedings of 4th Conference on Email and Anti-Spam, CEAS'2007*, 2007b.

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

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

A. Caponnetto and E. De Vito. Optimal rates for regularized least squares algorithm. *Found. Comput. Math.*, 7:331–368, 2007.

F. Chang, C.-Y. Guo, X.-R. Lin, and C.-J. Lu. Tree decomposition for large-scale SVM problems. *J. Mach. Learn. Res.*, 11:2935–2972, 2010.

H. Cheng, P.-N. Tan, and R. Jin. Localized support vector machine and its efficient algorithm. In *SIAM International Conference on Data Mining*, 2007.

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

R. Collobert, S. Bengio, and Y. Bengio. A parallel mixture of SVMs for very large scale problems. In *Advances in Neural Information Processing Systems*, pages 633–640, 2001.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39:1–49, 2002.

S. Dasgupta. Lecture 1: Clustering in metric spaces. CSE 291: Topics in unsupervised learning, 2008. URL `http://cseweb.ucsd.edu/~dasgupta/291-unsup/lec1.pdf`.

E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.*, 5:59–85, 2005.

R.A. DeVore and G.G. Lorentz. *Constructive Approximation.* Springer-Verlag, Berlin, 1993.

R.A. DeVore and V.A. Popov. Interpolation of Besov spaces. *Trans. Amer. Math. Soc.*, 305:397–414, 1988.

M. Eberts. *Adaptive Rates for Support Vector Machines.* Shaker, Aachen, 2015.

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

M. Eberts and I. Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Statist.*, 7:1–42, 2013.

M. Eberts and I. Steinwart. Optimal learning rates for localized SVMs. 2014. URL `http://arxiv.org/pdf/1507.06615.pdf`.

D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators.* Cambridge University Press, Cambridge, 1996.

T.F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

H.P. Graf, E. Cosatto, L. Bottou, I. Durdanovic, and V. Vapnik. Parallel support vector machines: The cascade SVM. In *Advances in Neural Information Processing Systems*, pages 521–528, 2005.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York, 2002.

R. Hable. Universal consistency of localized versions of regularized kernel methods. *J. Mach. Learn. Res.*, 14, 2013.

M. Kloft and G. Blanchard. On the convergence rate of ℓ_p-norm multiple kernel learning. *J. Mach. Learn. Res.*, 13:2465–2502, 2012.

T. Kühn. Covering numbers of Gaussian reproducing kernel Hilbert spaces. *J. Complexity*, 27:489–499, 2011.

S. Lin, X. Guo, and D.X. Zhou. Distributed learning with regularized least squares. 2016. URL https://arxiv.org/abs/1608.03339.

S. Mendelson and J. Neeman. Regularization in kernel learning. *Ann. Statist.*, 38:526–565, 2010.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1657–1665. 2015.

N. Segata and E. Blanzieri. Empirical assessment of classification accuracy of local SVM. Technical report, University of Trento, Information Engineering and Computer Science, 2008. URL eprints.biblio.unitn.it/1398/1/014.pdf.

N. Segata and E. Blanzieri. Fast and scalable local kernel machines. *J. Mach. Learn. Res.*, 11:1883–1926, 2010.

J.S. Simonoff. *Smoothing Methods in Statistics.* Springer, New York, 1996.

S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Anal. Appl.*, 1:17–41, 2003.

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

E.M. Stein. *Singular Integrals and Differentiability Properties of Functions.* Princeton University Press, Princeton, NJ, 1970.

I. Steinwart. A fast SVM toolbox. 2016. URL http://www.isa.uni-stuttgart.de/software/.

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

I. Steinwart and C. Scovel. Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constr. Approx.*, 35:363–417, 2012.

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

I. Steinwart, D. Hush, and C. Scovel. Training SVMs without offset. *J. Mach. Learn. Res.*, 12:141–202, 2011.

T. Suzuki. Unifying framework for fast learning rate of non-sparse multiple kernel learning. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1575–1583. 2011.

V. Temlyakov. A remark on covering, 2013. URL http://arxiv.org/pdf/1301.3043.pdf.

H. Triebel. *Theory of Function Spaces II.* Springer, Basel, 1992.

H. Triebel. *Theory of function spaces III.* Birkhäuser, Basel, 2006.

H. Triebel. *Theory of Function Spaces.* Birkhäuser, Basel, 2010.

I.W. Tsang, J.T. Kwok, and P.-K. Cheung. Core vector machines: Fast SVM training on very large data sets. *J. Mach. Learn. Res.*, 6:363–392, 2005.

I.W. Tsang, A. Kocsor, and J.T. Kwok. Simpler core vector machines with enclosing balls. In *Proceedings of the 24th international conference on Machine learning*, pages 911–918, 2007.

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

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

D.-H. Xiang and D.-X. Zhou. Classification with Gaussians and convex loss. *J. Mach. Learn. Res.*, 10:1447–1468, 2009.

A. Zakai and Y. Ritov. Consistency and localizability. *J. Mach. Learn. Res.*, 10:827–856, 2009.

H. Zhang, A.C. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2126–2136, 2006.

Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16:3299–3340, 2015.

D.-X. Zhou. The covering number in learning theory. *J. Complexity*, 18:739–767, 2002.