Universität Stuttgart

Fachbereich Mathematik

Optimal Learning Rates for Localized SVMs

Mona Eberts, Ingo Steinwart

Preprint 2014/002

Fachbereich Mathematik
Fakultät Mathematik und Physik
Universität Stuttgart
Pfaffenwaldring 57
D-70 569 Stuttgart

E-Mail: preprints@mathematik.uni-stuttgart.de
WWW: http://www.mathematik.uni-stuttgart.de/preprints

ISSN **1613-8309**

Journal of Machine Learning Research ? (????) ?-?? Submitted ?/??; Published ??/??

Optimal Learning Rates for Localized SVMs

Mona Eberts MONA.EBERTS@MATHEMATIK.UNI-STUTTGART.DE
Ingo Steinwart INGO.STEINWART@MATHEMATIK.UNI-STUTTGART.DE
Institute for Stochastics and Applications
University of Stuttgart
70569 Stuttgart, Germany

Editor:

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 of large data sets separately have been proposed in the literature. So far, however, almost all these approaches have only been empirically investigated. In addition, their motivation was always based on computational requirements. In this work, we consider a localized SVM approach based upon a partition of the input space. For this local SVM, we derive a general oracle inequality. Then we apply this oracle inequality to least squares regression using Gaussian kernels and deduce local learning rates that are essentially minimax optimal under some standard smoothness assumptions on the regression function. Moreover, we show that unlike the rates for standard global SVMs, our rates do not change if the smoothness assumption is violated on sets of measure zero. This gives the first motivation for using local SVMs that is not based on computational requirements but on theoretical predictions on the generalization performance. We further introduce a data-dependent parameter selection method for our local SVM approach and show that this method achieves the same learning rates as before. Finally, we present some larger scale experiments for our localized SVM showing that it achieves essentially the same test performance 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 $D := ((x_1, y_1), \ldots, (x_n, y_n))$ 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 regression is to find a function $f_D : X \to \mathbb{R}$ such that important characteristics of the conditional distribution $\mathrm{P}(Y|x)$, $x \in X$, can be recovered. For instance, an f_D approximating the conditional mean $\mathbb{E}(Y|x)$, $x \in X$, is sought in the non-parametric least squares regression. This classical non-parametric regression problem has been extensively studied in the literature, where a general reference is the book (Györfi et al., 2002), presenting plenty of results concerning the non-parametric least squares regression.

In the literature, there are many learning methods that solve the non-parametric regression problems, some of them are e.g. described in (Györfi et al., 2002; Koenker, 2005;

Simonoff, 1996). In this paper, we utilize some kernel-based regularized empirical risk minimizers, also known as support vector machines (SVMs), which solve the regularized problem

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

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, 2008a). 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). Note that the empirical SVM solution $f_{\mathrm{D},\lambda}$ exists and is unique (cf. Steinwart and Christmann, 2008a, Theorem 5.5) whenever the loss L is convex in its last argument. Moreover, an SVM is L-risk consistent under a few assumptions on the RKHS H and the regularization parameter λ, see (Steinwart and Christmann, 2008a, Section 6.4) for more details. Besides, it is worth mentioning that the ability to choose the RKHS H as well as the loss function L in (1) provides the possibility to flexibly apply SVMs to various learning problems. Namely, the learning target is modeled by the loss function, e.g. the least squares loss is used to estimate the conditional mean. Moreover, since RKHSs be defined on arbitrary X, data types that are not $\mathbb{R}^d$-valued can be handled, too. Furthermore, SVMs are enjoying great popularity, since they can be implemented and applied in a relatively simple way and only have a few free parameters that can usually be determined by cross validation.

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

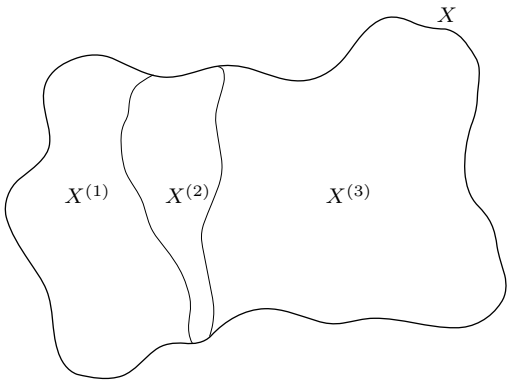


Figure 1: The input space X is partitioned by $X^{(1)}$, $X^{(2)}$, and $X^{(3)}$ such that the regression function f^* is less smooth on $X^{(2)}$ compared to $X^{(1)}$ and $X^{(3)}$. However, it is desirable to achieve locally optimal learning rates.

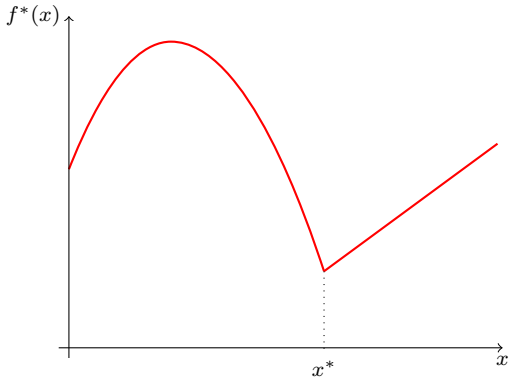


Figure 2: The target function f^* is not smooth in x^*. Is it anyway possible to globally achieve an optimal learning rate?

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 therein, while similar rates for SVMs using the pinball loss can be found in (Steinwart and Christmann, 2008b, 2011). 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 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. 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 (cf. Figure 1). 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. Moreover, this approach will enable us to globally achieve the optimal learning rate even if there is a P_X-zero set where the regression function is not smooth, cf. Figure 2 where the target function has a sharp kink at x^*. Our technique to achieve such local learning rates is a special local SVM approach. Local SVMs have been extensively investigated in the literature to speed-up the training time, see for instance, the early works (Bottou and Vapnik, 1992; Vapnik and Bottou, 1993). The basic idea of many 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 splitted into clusters and then an SVM is trained on each cluster. However, the presented results are only of experimental character.

In this article, we partition the input space X according to a cover of X with radius r_n and build an SVM model for each partition cell. The following section is dedicated to the detailed description of this method. Section 3 then presents some theoretical results that enable the analysis of this new method. For example, we examine extensions and direct sums of RKHSs. At the end of Section 3, we finally present a first oracle inequality for the localized SVM. In Section 4, we focus on RKHSs using Gaussian RBF kernels and, in conjunction with that, we study some entropy estimates. After that, Section 5 concentrates on the least squares loss and introduces an oracle inequality and learning rates for our localized SVM method using Gaussian kernels. Moreover, a data-dependent parameter selection method is studied that induces the same rates. Section 6 then presents some experimental results w.r.t. the localized SVM technique. All proofs can be found in Section 7, and the appendix contains various tables displaying detailed results of our experiments.

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 and $Y := [-M, M]$ for some $M > 0$. 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

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

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 5 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 now 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, this clipping idea does *not* change the learning method since it is performed *after* the training phase.

To modify the standard SVM approach (1), we assume that $(A_j)_{j=1,\ldots,m}$ is a partition of X such that $\mathring{A}_j \neq \emptyset$ for every $j \in \{1, \ldots, m\}$. Obviously, this implies $A_{j_1} \cap A_{j_2} = \emptyset$ for all $j_1, j_2 \in \{1, \ldots, m\}$ with $j_1 \neq j_2$ and

$$X = \bigcup_{j=1}^{m} A_j\,.$$

Now, the basic idea of the approach developed in this paper is to consider for each set of the partition $(A_j)_{j=1,\ldots,m}$ 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

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

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

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

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

$$L_j(x, y, t) := \mathbb{1}_{A_j}(x) L(x, y, t)\,, \tag{2}$$

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 $k_j : A_j \times A_j \to \mathbb{R}$. Note that 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 a function $\hat{f} : X \to \mathbb{R}$ 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

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

is an RKHS on X (cf. Lemma 2). That is, $\hat{H}_j$ is an isometrically isomorphic extension of the RKHS H_j on A_j to an RKHS on X. After all, we are now able to formulate a modified SVM approach. To this end, for every $j \in \{1, \ldots, m\}$, consider the local SVM optimization problem

$$f_{\mathrm{D}_j,\lambda_j} = \arg\min_{\hat{f}\in\hat{H}_j} \lambda_j \|\hat{f}\|^2_{\hat{H}_j} + \frac{1}{n}\sum_{i=1}^{n} L_j(x_i, y_i, \hat{f}(x_i))\,, \tag{3}$$

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

$$f_{\mathrm{D},\boldsymbol{\lambda}}(x) := \sum_{j=1}^{m} f_{\mathrm{D}_j,\lambda_j}(x) = \sum_{j=1}^{m} \mathbb{1}_{A_j}(x) f_{\mathrm{D}_j,\lambda_j}(x)\,, \tag{4}$$

where $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_m)$. Here, clipping $f_{\mathrm{D},\boldsymbol{\lambda}}$ at M yields

$$\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}}(x) = \sum_{j=1}^{m} \mathbb{1}_{A_j}(x) \widehat{f}_{\mathrm{D}_j,\lambda_j}(x)$$

for every $x \in X$. Note that the empirical SVM solutions $f_{\mathrm{D}_j,\lambda_j}$ in (3) exist and are unique by (Steinwart and Christmann, 2008a, Theorem 5.5) and that, for arbitrary $j \in \{1, \ldots, m\}$, $f_{\mathrm{D}_j,\lambda_j} = 0$ if $x_i \notin A_j$ for all $i \in \{1, \ldots, n\}$. In addition, the SVM optimization problem (3) equals the SVM optimization problem (1) using H_j, D_j, and the regularization parameter $\tilde{\lambda}_j := \frac{n}{|I_j|}\lambda_j$, since, for $\hat{f} \in \hat{H}_j$ and $f := \hat{f}_{|A_j}$, we have

$$\begin{aligned}
\lambda_j \|\hat{f}\|^2_{\hat{H}_j} + \frac{1}{n}\sum_{i=1}^{n} L_j(x_i, y_i, \hat{f}(x_i)) &= \lambda_j \|f\|^2_{H_j} + \frac{1}{n}\sum_{i\in I_j}^{n} L(x_i, y_i, f(x_i)) \\
&= \frac{|I_j|}{n}\left(\tilde{\lambda}_j \|f\|^2_{H_j} + \mathcal{R}_{L,\mathrm{D}_j}(f)\right).
\end{aligned}$$

That is, $f_{\mathrm{D}_j,\lambda_j}$ as in (3) and $h_{\mathrm{D}_j,\tilde{\lambda}_j} := \arg\min_{f\in H_j} \tilde{\lambda}_j \|f\|^2_{H_j} + \mathcal{R}_{L,\mathrm{D}_j}(f)$ satisfy

$$h_{\mathrm{D}_j,\tilde{\lambda}_j} = f_{\mathrm{D}_j,\lambda_j|A_j}\,.$$

For the sake of completeness, we briefly examine the Bayes risks w.r.t. P and L_j. To this end, let $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$, $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function and P be a distribution on $X \times Y$ such that a Bayes decision function $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ exists. Then, for all $j \in \{1, \ldots, m\}$ and losses L_j defined by (2), it is easy to show

$$\mathcal{R}_{L_j,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L_j,\mathrm{P}}\,,$$

whenever $f^*_{L,\mathrm{P}}$ exists. In other words, a Bayes decision function $f^*_{L,\mathrm{P}}$ w.r.t. P and L additionally is a Bayes decision function w.r.t. P and L_j. Moreover, for function spaces $\mathcal{F}_1, \ldots, \mathcal{F}_m$ over X, we have

$$\sum_{j=1}^{m} \min_{f_j\in\mathcal{F}_j} \mathcal{R}_{L_j,\mathrm{D}}(f_j) = \min_{f_1\in\mathcal{F}_1,\ldots,f_m\in\mathcal{F}_m} \sum_{j=1}^{m} \mathcal{R}_{L_j,\mathrm{D}}(f_j) \tag{5}$$

by the construction of the loss L_j.

Let us now present an advantageous characteristic of our modified SVM, namely the required computing time. Solving an usual SVM problem has a computational cost of $\mathcal{O}(n^q)$ where $q \in [2,3]$ and n is the sample size. For the new approach we consider m working sets of size $n_1, \ldots, n_m$ where $n_i \approx n_j$ for all $i, j \in \{1, \ldots, m\}$, i.e. $n_i \approx \frac{n}{m}$. Then for each working set an usual SVM problem has to be solved such that, altogether, the modified SVM induces a computational cost of $\mathcal{O}\left(m \left(\frac{n}{m}\right)^q\right)$. That is, for some $\beta > 0$ and $m \approx n^\beta$ 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 algorithms and handle large data sets. 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 $(A_j)_{j=1,\ldots,m}$ 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 by $B_{\ell_2^d}$ the closed unit ball in the d-dimensional Euclidean space ℓ_2^d and we define balls $B_1, \ldots, B_m$ with radius $r > 0$ and mutually distinct centers $z_1, \ldots, z_m \in X$ by

$$B_j := B_r(z_j) := \{x \in X : \|x - z_j\|_2 \leq r\} \qquad j \in \{1, \ldots, m\}\,, \tag{6}$$

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

$$\bigcup_{j=1}^{m} B_j = X\,,$$

i.e. such that the balls $B_1, \ldots, B_m$ cover X (cf. Figure 3). The following well-known Lemma relates the radius of such a cover with the number of centers.

Lemma 1 *Let $X \subset \mathbb{R}^d$ be a bounded subset, i.e. $X \subset cB_{\ell_2^d}$ for some constant $c > 0$. Then there exist balls $(B_j)_{j=1,\ldots,m}$ with radius $r > 0$ covering X such that*

$$r \leq 8cm^{-\frac{1}{d}}\,.$$

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

$$r \leq 8m^{-\frac{1}{d}}\,. \tag{7}$$

Finally, we can specify the partition $(A_j)_{j=1,\ldots,m}$ of X by the following assumption.

(A) Let $(A_j)_{j=1,\ldots,m}$ be a partition of $X \subset B_{\ell_2^d}$ such that $\mathring{A}_j \neq \emptyset$ for every $j \in \{1, \ldots, m\}$ and such that there exist mutually distinct $z_1, \ldots, z_m \in X$ with $A_j \subset B_r(z_j) =: B_j$, where $(B_j)_{j=1,\ldots,m}$ is a cover of X satisfying (7).

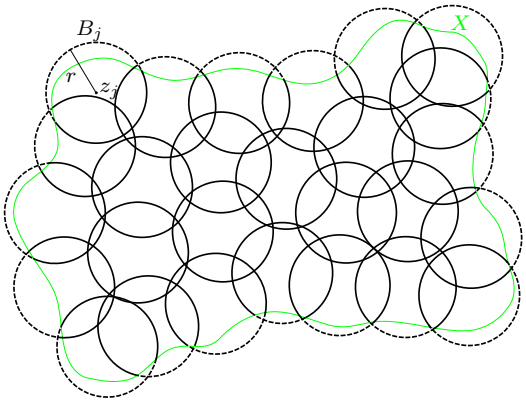


Figure 3: Cover $(B_j)_{j=1,\ldots,m}$ of X, where $B_1, \ldots, B_m$ are balls with radius r and centers z_j $(j = 1, \ldots, m)$.

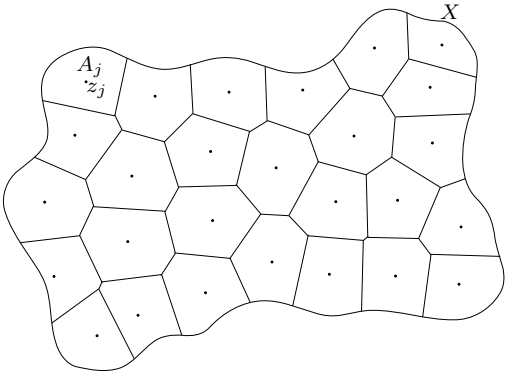


Figure 4: Voronoi partition $(A_j)_{j=1,\ldots,m}$ of X defined by (8), where $A_j \subset B_j$ for every $j \in \{1, \ldots, m\}$.

In the remaining sections we will frequently refer to Assumption (A). However, the results hold as well if we merely assume $z_1, \ldots, z_m \in B_{\ell_2^d}$ instead of $z_1, \ldots, z_m \in X \subset B_{\ell_2^d}$ in (A). The following example illustrates that (A) is indeed a natural assumption.

Example 1 *For some $r > 0$, let us consider an r-net $z_1, \ldots, z_m$ of X, where $z_1, \ldots, z_m$ are mutually distinct. Based on these $z_1, \ldots, z_m$, a Voronoi partition $(A_j)_{j=1,\ldots,m}$ of X is defined by*

$$A_j := \left\{ x \in X : j = \min\{ \operatorname*{arg\,min}_{k \in \{1,\ldots,m\}} \|x - z_k\|_2 \} \right\}, \tag{8}$$

cf. Figure 4. That is, A_j contains all $x \in X$ such that the center z_j is the nearest center to x, and if there exist j_1 and j_2 with $j_1 < j_2$ and

$$\|x - z_{j_1}\|_2 = \|x - z_{j_2}\|_2 < \|x - z_k\|_2$$

for all $k \in \{1, \ldots, m\} \backslash \{j_1, j_2\}$, then $x \in A_{j_1}$ since $j_1 < j_2$. In other words, they are resolved in favor of the smallest index of the involved centers. Moreover, it is obvious that $\mathring{A}_j \neq \emptyset$, $A_j \subset B_r(z_j)$ for all $j \in \{1, \ldots, m\}$, $A_{j_1} \cap A_{j_2} = \emptyset$ for all $j_1, j_2 \in \{1, \ldots, m\}$ with $j_1 \neq j_2$, and $X = \bigcup_{j=1}^m A_j$. In other words, a Voronoi partition based on an r-net $z_1, \ldots, z_m$ of X satisfies condition (A), if r and m fullfil (7).

Following Example 1, we call the learning method producing $f_{\mathrm{D},\boldsymbol{\lambda}}$ given by (4) a *Voronoi partition support vector machine*, in short VP-SVM. Nevertheless, we just take a partition $(A_j)_{j=1,\ldots,m}$ satisfying (A) as basis here instead of requesting $(A_j)_{j=1,\ldots,m}$ 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 an arbitrary measurable set $T \subset X$ such that $\mathrm{P}_X(T) > 0$. 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

$$J_T := \{j \in \{1, \ldots, m\} : A_j \cap T \neq \emptyset\} \tag{9}$$

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 non-empty, too, 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

$$L_T(x, y, t) := \mathbb{1}_T(x) L(x, y, t)\,. \tag{10}$$

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

$$L_{J_T}(x, y, t) := \mathbb{1}_{\bigcup_{j \in J_T} A_j}(x) L(x, y, t) \tag{11}$$

which may only be nonzero, if x is contained in some set A_j with $j \in J_T$. Note that the risiks $\mathcal{R}_{L_T,\mathrm{P}}(f)$ 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

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

leads to learning rates on A_T and implicitly on T. Recapitulatory, let us declare a second set of assumptions.

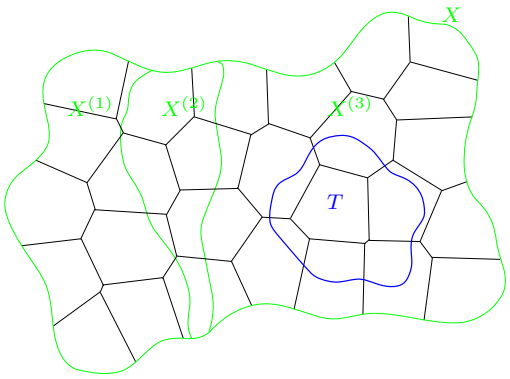


Figure 5: The input space X with the corresponding partition $(A_j)_{j=1,\ldots,m}$ and the subset T, where the local learning rate should be examined.

(T) For $T \subset X$, we define an index set J_T by (9), loss functions $L_T, L_{J_T} : X \times Y \times \mathbb{R} \to [0, \infty)$ by (10) and (11), and the set $A_T := \bigcup_{j \in J_T} A_j$.

3. An Oracle Inequality for VP-SVMs

In this section, we first focus on RKHSs and direct sums of RKHSs. Then we present a lemma that relates the risk of a function w.r.t. the general loss L to the risks w.r.t. the losses L_j. Finally, we establish a first oracle inequality for VP-SVMs.

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 (4) 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 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

$$\hat{k}_A(x, x') := \begin{cases} k_A(x, x')\,, & \text{if } x, x' \in A\,, \\ 0\,, & \text{else.} \end{cases} \tag{12}$$

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 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 (12) be the associated reproducing kernels. Then $\hat{H}_A \cap \hat{H}_B = \{0\}$ and hence the direct sum*

$$H := \hat{H}_A \oplus \hat{H}_B \tag{13}$$

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

$$\|f\|_H^2 := \lambda_A \|\hat{f}_A\|_{\hat{H}_A}^2 + \lambda_B \|\hat{f}_B\|_{\hat{H}_B}^2 \tag{14}$$

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

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

is the reproducing kernel.

To relate Lemmas 2 and 3 with (4), 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, the joined RKHSs 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

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

is again an RKHS equipped with the norm

$$\|f\|_{H_J}^2 = \sum_{j\in J} \lambda_j \|f_j\|_{\hat{H}_j}^2 \,. \tag{15}$$

If $J = \{1, \ldots, m\}$ we simply write

$$H := H_J \tag{16}$$

Note that H contains inter alia $f_{\mathrm{D},\boldsymbol{\lambda}}$ given by (4). Summarizing, we can define another assumption set.

(H) For every $j \in \{1, \ldots, m\}$, let H_j be a separable RKHS of measurable kernels k_j over A_j, where $A_1, \ldots, A_m \subset X$ are pairwise disjoint and

$$\|k_j\|_{L_2(\mathrm{P}_{X|A_j})}^2 := \int_X k_j(x,x) d\mathrm{P}_{X|A_j}(x) < \infty \,.$$

Then we define RKHSs $\hat{H}_1, \ldots, \hat{H}_m$ by Lemma 2 and the joined RKHS H by (16) equipped with the norm (15) for fixed $\lambda_1, \ldots, \lambda_m > 0$.

Having designed a joined RKHS as above, a crucial property of its function's risks is expressed by the following lemma.

Lemma 4 *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

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

Note that Lemma 4 can be transferred to finite, pairwise disjoint unions. To be more precise, let us consider an arbitrary index set $J \subset \{1, \ldots, m\}$ and define the corresponding loss function $L_J : X \times Y \times \mathbb{R} \to [0,\infty)$ by

$$L_J(x,y,t) := \mathbb{1}_{\bigcup_{j\in J} A_j}(x) L(x,y,t) \,.$$

Now, it is straightforward to show

$$\mathcal{R}_{L_J,\mathrm{P}}(f) = \sum_{j\in J} \mathcal{R}_{L_j,\mathrm{P}}(f)$$

for every function $f : X \to \mathbb{R}$. Based on this generalization and the whole index set $J = \{1, \ldots, m\}$, let us briefly consider Lemma 4 for the empirical measure D and for

$f_{\mathrm{D},\boldsymbol{\lambda}} = \sum_{j=1}^{n} \mathbb{1}_{A_j} f_{\mathrm{D}_j,\lambda_j}$, where $f_{\mathrm{D}_j,\lambda_j}$, $j = 1,\ldots,m$, are defined by (3). Then, for an arbitrary $f \in H$, it immediately follows

$$\begin{aligned}
\|f_{\mathrm{D},\boldsymbol{\lambda}}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}}) &= \sum_{j=1}^{m} \left(\lambda_j \left\|f_{\mathrm{D}_j,\lambda_j}\right\|_{\hat{H}_j}^2 + \mathcal{R}_{L_j,\mathrm{D}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}})\right) \\
&\leq \sum_{j=1}^{m} \left(\lambda_j \left\|\mathbb{1}_{A_j} f\right\|_{\hat{H}_j}^2 + \mathcal{R}_{L_j,\mathrm{D}}(f)\right) \\
&= \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)\,. \qquad (17)
\end{aligned}$$

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 (4). This will be a key insight used in our analysis.

To derive an oracle inequality, i.e. an appropriate upper bound for the excess risk $\mathcal{R}_{L_J,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}}) - \mathcal{R}^*_{L_J,\mathrm{P}}$ for some index set $J \subset \{1,\ldots,m\}$, we have to introduce a few more notations. Let P be a distribution on $X \times Y$ such that a Bayes decision function $f^*_{L,\mathrm{P}} : X \to [-M, M]$ exists, for some constant $M > 0$ at which L can be clipped. Moreover, we denote by $L \circ f$ the function $(x, y) \mapsto L(x, y, f(x))$. If there exist constants $B > 0$, $\vartheta \in [0, 1]$, and $V \geq B^{2-\vartheta}$ such that we have

$$L(x, y, t) \leq B\,, \qquad (18)$$

$$\mathbb{E}_{\mathrm{P}} \left(L \circ f - L \circ f^*_{L,\mathrm{P}}\right)^2 \leq V \cdot \left(\mathbb{E}_{\mathrm{P}} \left(L \circ f - L \circ f^*_{L,\mathrm{P}}\right)\right)^{\vartheta}\,, \qquad (19)$$

for all $(x, y) \in X \times Y$, $t \in [-M, M]$, and $f : X \to [-M, M]$, we say that the supremum bound (18) and the variance bound (19), respectively, is fulfilled. Actually, (18) immediately yields

$$L_J(x, y, t) = \mathbb{1}_{\bigcup_{j\in J} A_j}(x) L(x, y, t) \leq L(x, y, t) \leq B$$

for all $(x, y) \in X \times Y$ and $t \in [-M, M]$, i.e. the supremum bound is also satisfied for L_J. Moreover, if (19) holds for all $f : X \to [-M, M]$, the variance bound using the loss L_J is satisfied, too. Indeed, by the use of $\widetilde{f}(x) := \mathbb{1}_{\bigcup_{j\in J} A_j}(x) f(x) + \mathbb{1}_{X\setminus(\bigcup_{j\in J} A_j)}(x) f^*_{L,\mathrm{P}}(x)$ for all $x \in X$, we have

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}} \left(L_J \circ f - L_J \circ f^*_{L,\mathrm{P}}\right)^2 &= \mathbb{E}_{\mathrm{P}} \left(L_J \circ \widetilde{f} - L_J \circ f^*_{L,\mathrm{P}}\right)^2 \\
&= \mathbb{E}_{\mathrm{P}} \left(L \circ \widetilde{f} - L \circ f^*_{L,\mathrm{P}}\right)^2 \\
&\leq V \cdot \left(\mathbb{E}_{\mathrm{P}} \left(L \circ \widetilde{f} - L \circ f^*_{L,\mathrm{P}}\right)\right)^{\vartheta} \\
&\leq V \cdot \left(\mathbb{E}_{\mathrm{P}} \left(L_J \circ f - L_J \circ f^*_{L,\mathrm{P}}\right)\right)^{\vartheta}
\end{aligned}$$

for all $f : X \to [-M, M]$. Let us quickly define a third assumption set.

(P) Let P be a distribution on $X \times Y$ such that the variance bound (19) is satisfied for constants $\vartheta \in [0, 1]$, $V \geq B^{2-\vartheta}$, and all functions $f : X \to [-M, M]$.

Up to now, there is still missing a classical tool that is used to derive learning rates, namely entropy numbers, see (Carl and Stephani, 1990) or (Steinwart and Christmann, 2008a, Definition A.5.26). Recall that, for normed spaces $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$ as well as an integer $i \geq 1$, the i-th (dyadic) entropy number of a bounded, linear operator $S : E \to F$ is defined by

$$\begin{aligned} e_i(S : E \to F) &:= e_i(SB_E, \|\cdot\|_F) \\ &:= \inf\left\{\varepsilon > 0 : \exists s_1, \ldots, s_{2^{i-1}} \in SB_E \text{ such that } SB_E \subset \bigcup_{j=1}^{2^{i-1}} (s_j + \varepsilon B_F)\right\}, \end{aligned}$$

where we use the convention $\inf \emptyset := \infty$, and B_E as well as B_F denote the closed unit balls in E and F, respectively. Finally, we present a first oracle inequality involving an upper bound for the excess risk $\mathcal{R}_{L_J,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}}) - \mathcal{R}^*_{L_J,\mathrm{P}}$, where $J \subset \{1, \ldots, m\}$ is an arbitrary index set.

Theorem 5 *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a locally Lipschitz continuous loss that can be clipped at $M > 0$ and that satisfies the supremum bound (18) for some $B > 0$. Based on a partition $(A_j)_{j=1,\ldots,m}$ of X, where $\mathring{A}_j \neq \emptyset$ for every $j \in \{1, \ldots, m\}$, we assume (H). Furthermore, for an arbitrary index set $J \subset \{1, \ldots, m\}$, we suppose (P). Assume that, for fixed $n \geq 1$, there exist constants $p \in (0, 1)$ and $a_1, \ldots, a_m > 0$ such that for all $j \in \{1, \ldots, m\}$*

$$e_i(\mathrm{id} : H_j \to L_2(\mathrm{P}_{X|A_j})) \leq a_j\, i^{-\frac{1}{2p}}, \qquad i \geq 1. \tag{20}$$

Finally, fix an $f_0 \in H$ and a constant $B_0 \geq B$ such that $\|L_J \circ f_0\|_\infty \leq B_0$. Then, for all fixed $\tau > 0$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m) > 0$, and

$$a := \max\left\{c_p \sqrt{m} \left(\sum_{j=1}^m \lambda_j^{-p} a_j^{2p}\right)^{\frac{1}{2p}}, B\right\},$$

the VP-SVM given by (4) using $\hat{H}_1, \ldots, \hat{H}_m$ and L_J satisfies

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

with probability P^n not less than $1 - 3e^{-\tau}$, where $C > 0$ is a constant only depending on p, M, V, ϑ, and B.

The above theorem deals with the case of a partition with quite a few sets A_j, $j \in \{1, \ldots, m\}$. However, if we consider a partition consisting of just one set A_1, i.e. $A_1 = X$, Theorem 5 is supposed to provide an oracle inequality that is comparable to the already

known ones. To make that sure, let us briefly consider the case $m = 1$ and hence $A_1 := X$, $\lambda_1 := \boldsymbol{\lambda}$ as well as RKHSs $H_1 = \hat{H}_1 = H$ over X with $\|\cdot\|_H^2 = \boldsymbol{\lambda}\|\cdot\|_{H_1}^2$. Note that in this case we have $f_{\mathrm{D},\boldsymbol{\lambda}} = f_{\mathrm{D}_1,\lambda_1}$. If (20) holds for H_1, Theorem 5 yields that an SVM using H and $L_J = L$ satisfies

$$\begin{aligned}
&\boldsymbol{\lambda}\|f_{\mathrm{D},\boldsymbol{\lambda}}\|_{H_1}^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}}) - \mathcal{R}_{L,\mathrm{P}}^* \\
&\leq 9\left(\boldsymbol{\lambda}\|f_0\|_{H_1}^2 + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}_{L,\mathrm{P}}^*\right) + C\left(\tilde{c}_p \frac{a_1^{2p}}{\boldsymbol{\lambda}^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}
\end{aligned}$$

with probability P^n not less than $1 - 3e^{-\tau}$ for fixed $\tau > 0$. Note that this oracle inequality indeed matches with the one stated in (Steinwart and Christmann, 2008a, Theorem 7.23) apart from the constant $\tilde{c}_p$, which is, however, only depending on p, ϑ, and B.

In the folllowing section, we focus on RKHSs using Gaussian RBF kernels and examine the associated entropy numbers to specify (20). Subsequently in Section 5, we additionally consider the least squares loss and adapt the oracle inequality of Theorem 5.

4. Entropy Estimates for Local Gaussian RKHSs

In this section, we refine assumption (20). More precisely, in the subsequent theorem we determine an upper bound for the entropy numbers of the operator $\mathrm{id} : H_\gamma(A) \to L_2(\mathrm{P}_{X|A})$, where $H_\gamma(A)$ is the RKHS over A of the Gaussian RBF kernel k_γ on $A \subset \mathbb{R}^d$ defined by

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

for some width $\gamma > 0$.

Theorem 6 *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*

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

Obviously, this theorem specifies assumption (20). Now, for the Gaussian case we elaborate assumption (H) and introduce the following additional set of assumptions.

(G) Let $A_1, \ldots, A_m$ be pairwise disjoint subsets of X with non-empty interior such that, for some fixed $r > 0$ and every $j \in \{1, \ldots, m\}$, $\sup_{x,x' \in A_j} \|x - x'\|_2 \leq 2r$ is satisfied. Furthermore, for every $j \in \{1, \ldots, m\}$, let $H_j := H_{\gamma_j}(A_j)$ be the RKHS of the Gaussian kernel k_{γ_j} with width $\gamma_j \in (0, r]$ over A_j. Consequently, for $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_m) \in (0, \infty)^m$, we define the joined RKHS $H := \bigoplus_{j=1}^m \hat{H}_{\gamma_j}(A_j)$ by (16) equipped with the norm (15).

Since we do not consider SVMs with a fixed kernel, we use a more detailed notation than (3) and (4) in the following specifying the kernel width γ_j of the RKHS $H_{\gamma_j}(A_j)$ at hand. For all $j \in \{1, \ldots, m\}$ and $\boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_m)$, we thus write

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

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.

In the subsequent section, we consider the least squares loss which, together with Assumption (G) and Theorem 6, allows us to elaborate the oracle inequality stated in Theorem 5 so that we finally obtain learning rates.

5. 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

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

can be shown by some simple, well-known transformations. Recall that T is a non-empty subset of X, where the index set J_T defined by (9) indicates every set A_j of the partition $(A_j)_{j=1,\ldots,m}$ of X that shares at least one point with T. The associated loss function $L_{J_T} : X \times Y \times \mathbb{R} \to [0,\infty)$ is defined by (11).

5.1 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), and (Berens and DeVore, 1978, p. 360)). Denote by $\|\cdot\|_2$ the Euclidean norm and let $X \subset \mathbb{R}^d$ be a subset with non-empty interior, ν be an arbitrary measure on X, $p \in (0,\infty]$, and $f : X \to \mathbb{R}$ be contained in $L_p(\nu)$. Then, for $s \in \mathbb{N}$, the s-th modulus of smoothness of f is defined by

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

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

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

for $h = (h_1,\ldots,h_d) \in [0,\infty)^d$ and $X_{s,h} := \{x \in X : x + th \in X \text{ f.a. } t \in [0,s]\}$. Based on the modulus of smoothness, we introduce Besov 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 $1 \leq p, q \leq \infty$, $\alpha > 0$, $s := \lfloor \alpha \rfloor + 1$, and ν be an arbitrary measure. Then the Besov space $B^\alpha_{p,q}(\nu)$ is defined by

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

where the seminorm $|\cdot|_{B^\alpha_{p,q}(\nu)}$ is given by

$$|f|_{B^\alpha_{p,q}(\nu)} := \left(\int_0^\infty \left(t^{-\alpha} \omega_{s,L_p(\nu)}(f,t) \right)^q \frac{dt}{t} \right)^{\frac{1}{q}}, \qquad 1 \leq q < \infty,$$

or

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

see e.g. (Adams and Fournier, 2003, Section 7) and (Triebel, 2010, Sections 2 and 3). Note that $\|f\|_{B^\alpha_{p,q}(\nu)} := \|f\|_{L_p(\nu)} + |f|_{B^\alpha_{p,q}(\nu)}$ actually describes a norm of $B^\alpha_{p,q}(\nu)$ for all $q \in [1, \infty]$, see e.g. (DeVore and Lorentz, 1993, pp. 54/55) and (DeVore and Popov, 1988, p. 398). Again, if ν is the Lebesgue measure on X, we write $B^\alpha_{p,q}(X) := B^\alpha_{p,q}(\nu)$. 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 $W^\alpha_p(\nu)$ defined by

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

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

$$\|f\|^p_{W^\alpha_p(\nu)} := \sum_{|\beta| \leq \alpha} \left\| \partial^{(\beta)} f \right\|^p_{L_p(\nu)},$$

(cf. Adams and Fournier, 2003, page 60). We write $W^0_p(\nu) = L_p(\nu)$ and, for the Lebesgue measure μ on $X \subset \mathbb{R}^d$, we define $W^\alpha_p(X) := W^\alpha_p(\mu)$. It is well-known, see e.g. (Edmunds and Triebel, 1996, p. 25 and p. 44), that the Sobolev spaces $W^\alpha_p(\mathbb{R}^d)$ fall into the scale of Besov spaces, namely

$$W^\alpha_p(\mathbb{R}^d) \subset B^\alpha_{p,q}(\mathbb{R}^d)$$

for $\alpha \in \mathbb{N}$, $p \in (1, \infty)$, and $\max\{p, 2\} \leq q \leq \infty$. Moreover, for $p = q = 2$ we actually have equality, that is $W^\alpha_2(\mathbb{R}^d) = B^\alpha_{2,2}(\mathbb{R}^d)$ with equivalent norms.

Based on the least squares loss and RKHSs using Gaussian kernels over the partition sets A_j, the subsequent theorem refines the oracle inequality stated in Theorem 5.

Theorem 7 *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 write $X := \operatorname{supp} \mathrm{P}_X$. 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 $p \in (0,1)$, $n \geq 1$, $\tau \geq 1$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m) \in (0, r]^m$, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m) > 0$, the VP-SVM given by (4) using $\hat{H}_{\gamma_1}(A_1), \ldots, \hat{H}_{\gamma_m}(A_m)$, 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 $\|f^*_{L,\mathrm{P}}\|_{B^\alpha_{2,\infty}(\mathrm{P}_{X|A_T})}$.

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

Theorem 8 *Let $\tau \geq 1$ be fixed and $\beta \geq \frac{2\alpha}{d} + 1$. Under the assumptions of Theorem 7 and with*

$$r_n = c_1 n^{-\frac{1}{\beta d}}, \tag{21}$$
$$\lambda_{n,j} = c_2 r^d n^{-1}, \tag{22}$$
$$\gamma_{n,j} = c_3 n^{-\frac{1}{2\alpha+d}}, \tag{23}$$

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

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

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 6 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 (7) 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 cost of an VP-SVM is 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 (4) and the considerations related to (17), where $f_{\mathrm{D},\boldsymbol{\lambda}} \in H = \bigoplus_{j=1}^m \hat{H}_j$ solves the minimization problem

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

Choosing $\lambda_1 = \ldots = \lambda_m$, the VP-SVM problem can be understood as ℓ_2-multiple kernel learning (MKL) problem using the RKHSs $\hat{H}_1, \ldots, \hat{H}_m$. 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) obtainessentially 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 7 and 8 require $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X|\bigcup_{j\in J_T} A_j})$. 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

$$T^{+\delta} := \{x \in X : \exists t \in T \text{ such that } \|x - t\|_2 \leq \delta\} \,, \tag{24}$$

which implies $T \subset T^{+\delta} \subset X$, cf. Figure 6. 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.

$$\forall \delta > 0 \quad \exists n_\delta \in \mathbb{N} \quad \forall n \geq n_\delta \quad : \quad \bigcup_{j\in J_T} A_j \subset T^{+\delta} \,, \tag{25}$$

where $J_T := \{j \in \{1, \ldots, m_n\} : A_j \cap T \neq \emptyset\}$. 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 $r_n = cn^{-\frac{1}{\beta d}}$, the lowest sample size n_δ in (25) can be determined by choosing the smallest $n_\delta \in \mathbb{N}$ such that $\delta \geq 2r_{n_\delta}$ with r_{n_δ} as in (7), that is

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

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

Corollary 9 *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 write $X := \operatorname{supp} \mathrm{P}_X$. 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*

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

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 (4) 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 7.

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 (21), (22), and (23), 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

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

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 9 is satisfied if, for example, $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(T^{+2\delta})$ and P_X has a bounded Lebesgue density on $T^{+\delta}$. Moreover, if this density 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 $f^*_{L,\mathrm{P}} \in W^\alpha_2(T)$. Modulo ξ, our rate is therefore asymptotically optimal in a minmax sense on T. In addition, for $\alpha > d$, the learning rates obtained for $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(T)$ are again asymptotically optimal modulo ξ on T.

Finally, let us consider the case that the regression function $f^*_{L,\mathrm{P}}$ satisfies a certain smoothness level on the whole space X apart from some zero set $Z \subset X$. In this setting we cannot simply use the above results. In fact, we have to take into account that on Z the regression function is less smooth. Fortunately, the subsequent results, which deal with the described setting, establish the same rates as above. For their formulation, however, we need to introduce some notations first. For a subset $Z \subset X$, which is supposed to be fixed and known from the context, and $\delta > 0$, we denote the complement of $Z^{+\delta}$ by

$$A_\delta := X \backslash Z^{+\delta}\,,$$

cf. Figure 7. In addition, we use the abbreviations

$$\gamma_{\max} := \max\{\gamma_1, \ldots, \gamma_m\}\,,$$

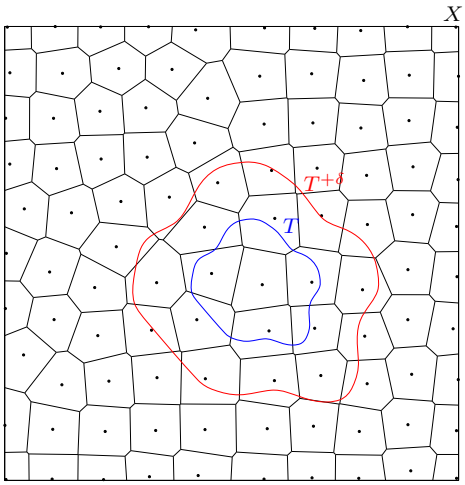


Figure 6: An input space X with the corresponding Voronoi partition as well as a subset $T \subset X$ enlarged by an δ-tube to $T^{+\delta}$.

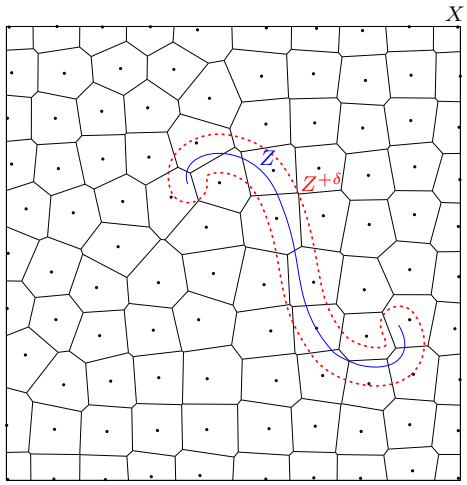


Figure 7: An input space X with the corresponding Voronoi partition as well as a zero set $Z \subset X$ enlarged by an δ-tube to $Z^{+\delta}$.

$$\gamma_{\min} := \min\{\gamma_1, \ldots, \gamma_m\}$$

in the following theorem.

Theorem 10 *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 write $X := \operatorname{supp} \mathrm{P}_X$. Furthermore, let (A) and (G) be satisfied. Moreover, assume that $Z \subset X$ is a P_X-zero set with $\mathrm{P}_X(Z^{+\delta}) > 0$ for all $\delta > 0$. In addition, 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 $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X|X\setminus Z^{+\delta}})$ for fixed $\alpha \geq 1$ and $\delta > 0$. Then, for all $p \in (0,1)$, there exists a constant $C > 0$ such that for all $n \geq 1$, $\tau \geq 1$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m) \in (0, r]^m$, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m) > 0$, the VP-SVM given by (4) using $\hat{H}_{\gamma_1}(A_1), \ldots, \hat{H}_{\gamma_m}(A_m)$, and the loss L satisfies*

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

$$\leq C \left(\sum_{j=1}^m \lambda_j \gamma_j^{-d} + \left(\frac{\gamma_{\max}}{\gamma_{\min}}\right)^d \gamma_{\max}^{2\alpha} + r^{2p} \left(\sum_{j=1}^m \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^p n^{-1} + \frac{\tau}{n} + \mathrm{P}_X\left(Z^{+\delta}\right) \right)$$

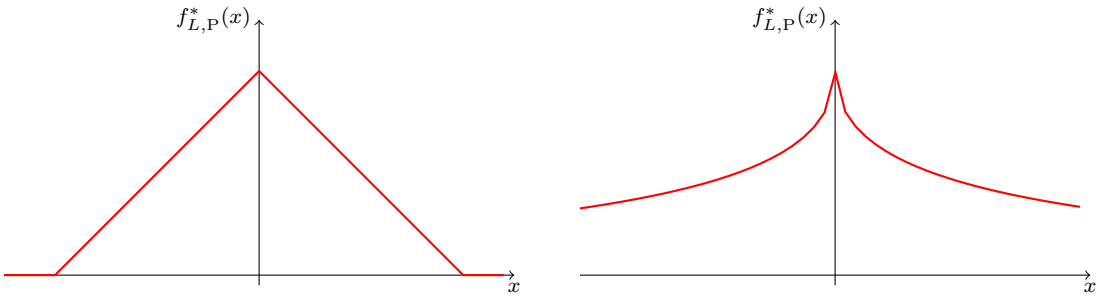
with probability P^n not less than $1 - e^{-\tau}$, where

$$C := \widetilde{C} \cdot \max\left\{ \|f^*_{L,\mathrm{P}}\|^2_{B^\alpha_{2,\infty}(\mathrm{P}_{X|A_\delta})}, 1 \right\}$$

and $\widetilde{C} > 0$ is a constant only depending on M, α, p, d, $\|f^*_{L,\mathrm{P}}\|_{L_2(\mathbb{R}^d)}$, and $\|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}$.

Note that the constant C in Theorem 10 depends on $\|f^*_{L,\mathrm{P}}\|_{B^\alpha_{2,\infty}(\mathrm{P}_{X|A_\delta})}$. However, $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X|A_\delta})$ does not necessarily mean that the corresponding norm is bounded by some

constant independent of δ for all $\delta > 0$. In fact, Figure 8 illustrates that there exist functions $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X|A_\delta})$ such that their norm is bounded by some constant or such that the corresponding norm increases for decreasing δ. In the following corollary we will derive the same learning rates as in Theorem 8 only if we exclude case (b) in Figure 8 and assume that $\|f^*_{L,\mathrm{P}}\|_{B^\alpha_{2,\infty}(\mathrm{P}_{X|A_\delta})}$ is bounded by some constant for arbitrarily small $\delta > 0$. Other scenarios, which consider, for example, certain polynomial behavior of $\|f^*_{L,\mathrm{P}}\|_{B^\alpha_{2,\infty}(\mathrm{P}_{X|A_\delta})}$ in δ, can be treated analogously.



(a) $f^*_{L,\mathrm{P}}(x) = \mathbb{1}_{[-1,1]}(x) \cdot (1 - |x|)$ (b) $f^*_{L,\mathrm{P}}(x) = e^{-\sqrt{|x|}}$

Figure 8: Assuming that P_X is the uniform distribution on $[-1,1]$ and $Z^{+\delta} = [-\delta, \delta]$ for some $\delta \in (0,1)$ yields $|f^*_{L,\mathrm{P}}|_{B^1_{p,\infty}(\mathrm{P}_{X|A_\delta})} \leq 2$ for (a) and $|f^*_{L,\mathrm{P}}|_{B^1_{p,\infty}(\mathrm{P}_{X|A_\delta})} \leq c\delta^{-1}$ for (b) where c is a positive constant.

Corollary 11 *Let the assumptions of Theorem 10 be satisfied and assume that there exists a constant $c > 0$ such that, for all $\delta > 0$, we have*

$$0 < \mathrm{P}_X(Z^{+\delta}) \leq c\delta\,. \tag{26}$$

Moreover, let $\tau \geq 1$ be fixed, $\beta \geq \frac{2\alpha}{d} + 1$, and $\|f^*_{L,\mathrm{P}}\|_{B^\alpha_{2,\infty}(\mathrm{P}_{X|A_\delta})} < \hat{c}$ for all $\delta > 0$ and some constant $\hat{c} \in (0,\infty)$. Then, for $r_n := c_1 n^{-\frac{1}{\beta d}}$, $\lambda_{n,j} := c_2 r^d n^{-1}$ as well as $\gamma_{n,j} := c_3 n^{-\frac{1}{2\alpha+d}}$ for every $j \in \{1,\ldots,m_n\}$, and $\delta_n := c_4 n^{-\frac{2\alpha}{2\alpha+d}}$, we have, for all $n \geq 1$ and $\xi > 0$,

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

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 $\widetilde{C}, c_1, c_2, c_3, c_4$ are positive constants with $c_3 \leq c_1$.

To illustrate the above assumption (26) let us assume that Z is closed and m-rectifiable in the sense of (Federer, 1969, 3.2.14). Then (Federer, 1969, 3.2.27) together with (Federer, 1969, 3.2.39) shows that

$$\mu^d(Z^{+\delta}) \leq 2\sigma_{d-m}\mathcal{H}^m(Z)\delta^{d-m}$$

for all sufficiently small $\delta > 0$, where σ_{d-m} denotes the μ^{d-m}-volume of $B_{\ell_2^{d-m}}$, and $\mathcal{H}^m$ the m-dimensional Hausdorff measure. From the latter it is easy to conclude that (26) holds.

5.2 Data-Dependent Parameter Selection for VP-SVMs

Note that 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, 2008a, 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

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

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

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

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

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

Again, 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, using (5) 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 $(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in (\Lambda_n \times \Gamma_n)^{m_n}$.

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

Theorem 12 *Let $r_n := cn^{-\frac{1}{\beta d}}$ with constants $c > 0$ and $\beta > 1$. Under the assumptions of Theorem 7 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 $\xi > 0$, $\tau \ge 1$, 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*

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

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 12 for all $n \ge n_\delta$ although $T^{+\delta}$ is fixed for all $n \in \mathbb{N}$. Note that, if P_X has a Lebesgue density that is bounded away from 0 and ∞ and either $f^*_{L,\mathrm{P}} \in W_2^\alpha(T)$ for $\alpha > d/2$ or $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(T)$ for $\alpha > d$, these learning rates are again asymptotically optimal modulo ξ on T in a minmax sense. However, the condition $\alpha < \frac{\beta-1}{2}d$ 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.

Let us now consider TV-VP-SVMs in the case of a regression function $f^*_{L,\mathrm{P}}$ which is less smooth on some zero set $Z \subset X$ than on the remaining part of the space X.

Theorem 13 *Let $r_n := cn^{-\frac{1}{\beta d}}$ with constants $c > 0$ and $\beta > 1$ as well as $\|f^*_{L,\mathrm{P}}\|_{B^\alpha_{2,\infty}(\mathrm{P}_{X|A_\vartheta})} < \bar{c}$ for all $\vartheta > 0$ and some constant $\bar{c} \in (0,\infty)$. Under the assumptions of Theorem 10 let $Z \subset X$ be a P_X-zero set such that there exist an $\epsilon \ge 1$ and a constant $\tilde{c} > 0$ such that, for all $\vartheta > 0$, we have $0 < \mathrm{P}_X(Z^{+\vartheta}) \le \tilde{c}\vartheta^\epsilon$. 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*

$\xi > 0$, $\tau \geq 1$, $\vartheta \leq \hat{c}n^{-1}$, *and* $\alpha < \frac{\beta-1}{2}d$, *the TV-VP-SVM producing the decision functions* $f_{\mathrm{D}_1,\boldsymbol{\lambda}_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}$ *satisfies*

$$\mathrm{P}^n \left(\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda}_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}) - \mathcal{R}^*_{L,\mathrm{P}} \leq c\tau n^{-\frac{2\alpha}{2\alpha+d}+\xi} \right) \geq 1 - e^{-\tau} \,,$$

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

6. Experimental Results

In the previous sections we defined VP-SVMs and derived essentially optimal learning rates even if the regression function has jumps. So far, it is, however, not clear if the theoretical results suggesting a generalization performance not worse than that of global SVMS can be empirically confirmed and if the predicted advantages of VP-SVMs in terms of computational costs are preserved in practice. Note that the latter is not as obvious as it may seem to be, since VP-SVMs create an overhead when generating the working sets, and the working sets themselves do not need to be as balanced as we assumed in our naïve analysis. In this section, we thus investigate the performance of VP-SVMs empirically. Namely, we carry out some experiments using the least squares loss with the objective to answer the subsequent questions:

(1) How do different radii affect the performance of VP-SVMs? In particular, what is the impact on the training time and the VP-SVM's test error?

(2) How do the VP-LS-SVMs perform compared to the usual LS-SVMs in terms of the test error? What is the speed-up?

(3) How does the performance of VP-SVMs compare to vanilla data splitting approaches such as random chunking *(RC-SVM)*, in which the data set is devided into a random partition with equally sized subsets, and the final decision function is the average of the SVMs computed on each subset?

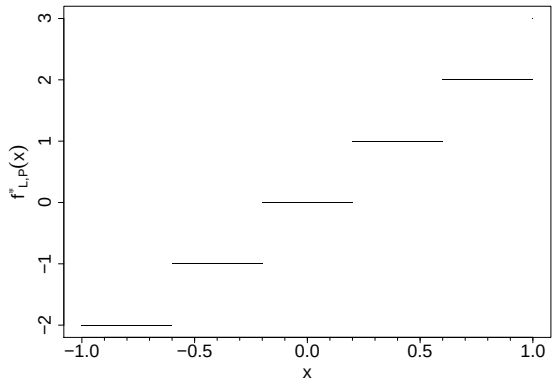
(4) The theoretical results suggests an advantage of VP-SVMs over global LS-SVMs, if the regression function has interruptions of its smoothness on zero sets. Can this be empirically validated?

To address these questions we utilize two kinds of data sets. On the one hand, to answer questions (1), (2), and (3), we examine the three real data sets COVTYPE, IJCNN1, and COD-RNA, which we obtained from LIBSVM's homepage, see (Chang and Lin, 2011). Table 1 summarizes some characteristics of these data sets. On the other hand, we generated

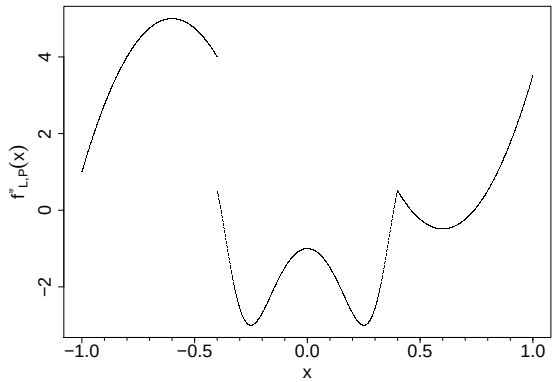
data set type	full data set size	dimension	number of labels
COVTYPE	581 012	54	2
COD-RNA	488 565	8	2
IJCNN1	141 691	22	2

Table 1: Characteristics of the considered LIBSVM data sets.

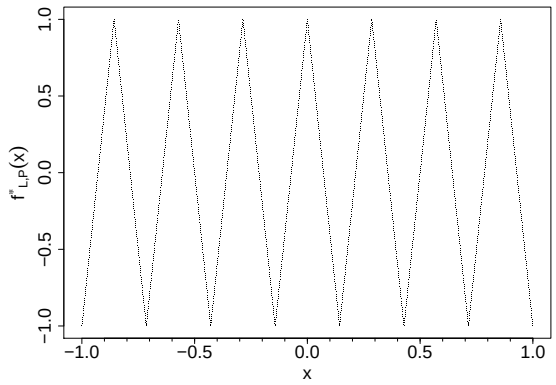
several artificial data sets to address the last question. In order to prepare the data sets



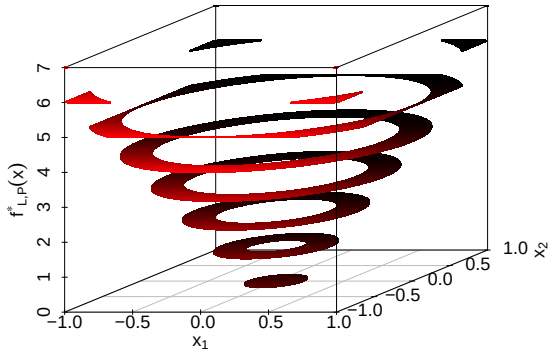
(a) Artificial data Type I: step function



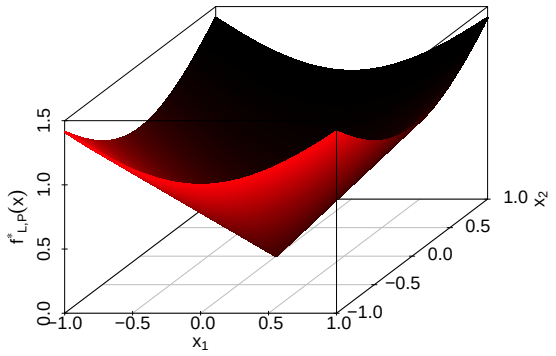
(b) Artificial data Type II: cracked function



(c) Artificial data Type III: jagged function



(d) Artificial data Type IV: circular step function



(e) Artificial data Type V: 2-dimensional Euclidean norm

Figure 9: Unscaled basic functions used to generate the artificial data sets.

for the experiments, we edited the data sets from LIBSVM in the following manner. If for a real-world data set type the raw data set was already split, we first merged these sets so that we obtained one data set for each data set type. In a next step, we scaled the data componentwise such that all samples including labels lie in $[-1,1]^{d+1}$, where d is the dimension of the input data. Finally, for each data set type, we generated random subsets that were afterwards randomly splitted into a training and a test data set. In this manner, we obtained, for each of the three LIBSVM data set types, training sets consisting of $n = 1\,000,\ 2\,500,\ 5\,000,\ 10\,000,\ 25\,000,\ 50\,000,\ 100\,000$ samples. Additionally, for the data sets COVTYPE and COD-RNA, we created training sets of sizes 250 000 and 500 000, and of sizes 250 000 and 400 000, respectively. 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.

For the artificial data, we proceeded in a slightly different way. To generate the data sets we took as fundament the five regression functions pictured in Figure 9 and as noise, the sum of two uniform distributions on $[-c(x), c(x)]$, where $c(x) = \frac{1}{4}\left(3\sin\left(\frac{\pi}{2}|x|\right) + 1\right)$ for the one-dimensional data sets and $c(x) = \frac{1}{4}\left(\sin\left(\frac{\pi}{4}\left(|x_1| + |x_2|\right)\right) + 1\right)$ for the two-dimensional data sets. Thus, we produced five different types of artificial data sets, where the various data set types are named according to their type numbers as in Figure 9. Initially, we created two sets, namely one training and one test data set, each consisting of 10 000 random input samples contained in $[-1,1]$ and $[-1,1]^2$, respectively. Then, for each artificial data set type, we determined the labels belonging to the input data as sum of the corresponding functional value and the noise and, finally, scaled all 20 000 labels to $[-1,1]$. In a last step, we randomly built subsets of the training sets of size $n = 1\,000,\ 2\,500,\ 5\,000$. In this way, we altogether obtained, for each type of artificial data, four training data sets of size $n = 1\,000,\ 2\,500,\ 5\,000,\ 10\,000$ and a corresponding test data set of size $n_{\text{test}} = 10\,000$. Based on the test data sets the Bayes risks can be determined, see Table 2 where the Bayes risks are summarized for the various artificial data set types.

	Type I	Type II	Type III	Type IV	Type V
Bayes risk	0.0254	0.0137	0.0529	0.0083	0.0634

Table 2: Bayes risks w.r.t. test data sets for the various artificial data set types.

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. An exception are the experiments using artificial training sets of size $n = 10\,000$, where we realized 100 repetitions for the sake of uniformity.

To approach the above problems we used the least squares loss and Gaussian kernels for all experiments. We implemented an LS-SVM-solver in C++ similar to the one in (Steinwart et al., 2011). Around this solver, we then built the routines for the VP-SVM and the RC-SVM. The compilation of the three programmes was executed by LINUX's gcc. To produce comparable results in terms of run time, all real-world data experiments were realized by the

Algorithm 1 Determine a Voronoi partition of the input data

Require: Input data set $D_X = \{x_1, \ldots, x_n\}$ 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 ∈ D_X
Cover_1 ← z
m ← 1
while max_{x∈D_X} ‖x − Cover‖_2 > r do
    z ← arg max_{x∈D_X} ‖x − Cover‖_2
    m ← m + 1
    Cover_m ← z
    WorkingSet_m ← ∅
end while
for i = 1 to n do
    k ← arg min_{j∈{1,...,m}} ‖x_i − Cover_j‖_2
    WorkingSet_k ← WorkingSet_k ∪ {x_i}
end for
return WorkingSet_1, ..., WorkingSet_m
```

same professional compute server[1] equipped with four INTEL XEON E7-4830 (2.13 GHz) 8-core processor, 256 GB RAM, and a 64 bit version of Debian GNU/Linux 6.0.7. In order that we can indeed compare their run time, we used eight cores to pre-compute the kernel matrix and to evaluate the final decision functions on the test set, and one core for the subsequent solver for every real data experiment. Since the artificial data sets consist of at most 10 000 samples we performed the according experiments by a computer equipped with one INTEL CORE i7-3770K (3.50 GHz) quad core processor, 16 GB RAM, and a 64 bit version of Debian GNU/Linux 6.0.7. For all artificial data experiments we used four cores to pre-compute the kernel matrix and to evaluate the final decision functions on the test set, and again one core for the solver. Even with pre-computed kernel matrices, our experiments on the real-world data altogether required almost 810 hours (approximately 34 days) for training and additionally almost 4 days for testing. Moreover, the experiments on the artificial data took nearly 43 hours for training and 168 minutes for testing. Without pre-computing the kernel matrices, e.g. by applying a standard caching approach, preliminary experiments suggested a multiplcation of the training time, which would have rendered the experiments infeasible. Besides, our experiments will show that the available amount of RAM does not restrict the size of the training sets used by an VP-SVMs as severely as the ones used by LS-SVMs.

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

1. On this occasion, we would like to thank the Institute for Applied Analysis and Numerical Simulation of the University of Stuttgart, who placed the above mentioned compute server at our disposal and, thus, enabled us to realize our experiments on large real-world data sets. In consequence, the overall time available for our experiments was limited.

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 are created randomly, 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, since we scaled the labels of all data sets to $[-1, 1]$, 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 experimental results for the three real data sets are summarized in Tables 3 to 6. These tables as well as Tables 7 to 11, containing the results for the experiments on the artificial data sets, can be found in the Appendix. In addition to the average run times of the training and test phases, these tables reflect inter alia the average test errors of the empirical SVM solutions. Additionally, the L_2-errors of the empirical SVM solutions, i.e. the value of

$$\sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(\widehat{f}_{\text{D},\lambda,\gamma}(x_{\text{test}\,i}) - f^*_{L,\text{P}}(x_{\text{test}\,i}) \right)^2},$$

is determined for the artificial data sets. Moreover, note that some of the result tables are incomplete for very large real-world training data sets. In these cases, the kernel matrix, whose size depends on the training set size, did not fit into the RAM of the used computer and, thus, these experiments were left out.

6.1 Experiments on Real-World Data

In this subsection, we adress questions (1), (2), and (3) by examining the results for the real-world data sets covtype, cod-rna, and ijcnn1, which are composed in Figures 10–12 and Tables 3–6.

6.1.1 Comparison of VP-SVMs Using Different Radii

In the following, we focus on the VP-SVMs using four different radii for the various real-world data sets, where the experimental results are summarized in Tables 3–6 as well as in Subfigures (d)–(f) of Figures 10–12. Examining the achieved training times for each data set type, we observe that, for increasing training set sizes, the radius that leads to the shortest training time typically decreases. More precisely, for the real data sets with sample size $n_{\text{train}} > 10\,000$, the VP-SVMs using the smallest radius always train fastest, while for the data sets with $n_{\text{train}} \leq 10\,000$, we can not make a uniform statement. Clearly, this finding is not surprising, since an SVM for a small data set trains considerably faster than an SVM

for a large data set, such that splitting the large data set and running an SVM for each of the small data sets may altogether still be faster. Recall additionally the considerations in terms of the computational cost made in Section 2.

Let us now consider the VP-SVM results in terms of the realized test errors. As expected, for the real-world data sets, the test errors achieved by VP-SVMs with fixed radii decrease with increasing sample size of the used data sets except twice. In addition, for the real data sets COVTYPE and COD-RNA, the test errors decrease for increasing radii, cf. Figures 10(f) and 11(f). Here, however, the test errors achieved for the various radii get close to each other with increasing training sample size. The same behavior of the test errors appears for the IJCNN1 data set, though, for $n_{\text{train}} \geq 5\,000$, both intermediate radii yield even smaller empirical risks than the largest radius, see Figure 12(f). In consequence, it is not straightforward to draw any conclusion on the relation between radius and test error. Nonetheless, we can say that VP-SVMs using small radii enjoy test errors that are never significantly larger and somethimes even smaller than those of VP-SVMs using the largest of the applied radii.

Besides, Tables 3 and 5 or Figures 10(e),(f) and 11(e),(f) contain an additional finding. For large data sets, namely for the COVTYPE data set of size 500 000 or for the COD-RNA data sets of size 250 000 and 400 000, the VP-SVMs with large radii did not yield any solution, since they failed due to the technical requirements caused by the used computer. More precisely, in these cases, there was at least one working set such that its kernel matrix did not fit into the RAM any more. Fortunately, the working sets of VP-SVMs using smaller radii were small enough such that we still received an outcome. What is more, these VP-SVMs yielded a better empirical risk in partially less training time compared to VP-SVMs with large radii and training sample sizes that still allowed a successful performance. That is, using a small radius for the VP-SVM and a training set that is oversized for VP-SVMs with a larger radius reduces the test error. More precisely, a large training set is crucial for a small empirical risk, where the possibly arising computational restrictions can be eluded by a VP-SVM with an appropriate radius.

All in all, localized SVMs using some small radius lead in substantially less training time to either negligble worse or even better test errors than VP-SVMs with large radii, if the training sample size is adequate, i.e. $n_{\text{train}} \geq 5\,000$. In addition, the real data sets with a large sample size demonstrate that VP-SVMs with small radii are able to conquer the technical restrictions caused by the used computer and thus yield a better empirical risk than VP-SVMs with bigger radii can attain at all.

6.1.2 Comparing VP-SVMs with Global LS-SVMs

In the following, we compare the results of the VP-SVM using different radii to the standard LS-SVM. For the real-world data sets COD-RNA and IJCNN1, the VP-SVMs, based on the largest of the applied radii, use only one working set. Thus, they coincide with the standard LS-SVM modulo different values generated by the random number generator. To verify this fact, we compare for the real data sets COD-RNA and IJCNN1 the results of the VP-SVMs using one working set to the results of the standard LS-SVMs, see Tables 5 and 6. Here, we note that the LS-SVM test errors typically decrease with increasing training sample size. The same holds for the VP-SVMs using one working set. Moreover, the latter VP-SVM and

the LS-SVMs perform equally well in terms of training time and empirical risk, however, for $n_{\text{train}} \geq 25\,000$ the VP-SVMs train slower.

In practice, a crucial problem is caused by the run time required by an algorithm. Hence, for each data set type, we compare hereafter the LS-SVM to the VP-SVM that trains fastest for the largest training data set. The required average training times and the average test errors of these SVMs are illustrated in Subfigures (g)–(i) of Figures 10–12. First, we notice that the selected VP-SVM uses the smallest of the applied radii for each data set type. Besides, the LS-SVM's test errors are lower than those of the VP-SVMs. However, with increasing training set size the VP-SVM's test errors get close to the ones of the LS-SVM. Moreover, for the IJCNN1 data set of size 100 000, their empirical risks even coincide, cf. Figure 12(i). Besides, the VP-SVMs train considerably faster than the LS-SVMs. In particular, for $n_{\text{train}} = 100\,000$ the VP-SVMs require at most 8.5% of the LS-SVM's training times, see Figures 10(h), 11(h), and 12(h). Finally, recall that, for data sets of size $n_{\text{train}} \geq 250\,000$, the LS-SVM problem is infeasible with our computer, just like the VP-SVMs using the largest of the applied radii. In contrast, for $n_{\text{train}} \geq 250\,000$, VP-SVMs using small radii usually train considerably faster and achieve lower test errors than the LS-SVMs for $n_{\text{train}} = 100\,000$, cf. Figures 10(h)–(i) and 11(h)–(i).

Concluding, we have seen that the application of a VP-SVM using a small radius instead of the standard LS-SVM reduces the run time considerably entailing at most a negligible worsening or even an improvement of the test errors. Moreover, applying VP-SVMs with sufficiently small radii enables us to use large data sets and, thus, to elude the computational restrictions to sufficiently small data sets. As a result, handling really large data sets with the help of suitable VP-SVMs can lead to significantly improved test errors compared to an LS-SVM setting with memory constraints.

6.1.3 Comparison of VP-SVMs with RC-SVMs

First of all, let us investigate the RC-SVM results that are composed in Tables 4–6 as well as in Subfigures (a)–(c) of Figures 10–12. For the real data sets COVTYPE we considered ten, for the data sets COD-RNA nine, and for the data sets IJCNN1 eight different numbers of working sets. In each case, we started with an RC-SVM using one working set, i.e. with an RC-SVM that corresponds to the global LS-SVM modulo different values generated by the random generator, cf. Tables 5 and 6. Comparing for every data set the RC-SVMs using various numbers of working sets, we observe that the number of working sets, minimizing the RC-SVMs training time, increases in tandem with the sample size. Moreover, the RC-SVM using one working set never trains fastest compared to the other RC-SVMs using more than one working set. Furthermore, the average test errors for the applied RC-SVMs usually decrease for a decreasing number of working sets and, hence, are minimized by the smallest possible number of working sets. Of course, all these findings are not surprising, since RC-SVMs are typically used to reduce the training time.

Let us now compare the results of VP- and RC-SVMs using roughly the same number of working sets, cf. Tables 3–6. Initially note that, even though we consider VP- and RC-SVM based on the same number of working sets, the RC-SVM working sets are about the same size whereas the VP-SVM working sets may have different sizes with a large range. That is, the VP-SVMs often deal with a few substantially larger working sets than the

RC-SVMs. Consequently, the RC-SVMs often perform faster than the VP-SVMs, which require up to five times the RC-SVM's training time for $n_{\text{train}} = 100\,000$. Contrarily, the average empirical risks achieved by the VP-SVMs are substantially lower than those of the RC-SVMs. Besides, in a few cases the VP-SVMs possess at least one working set which is oversized for the computer's RAM, so that these VP-SVM problems are infeasible, whereas the comparable RC-SVMs avoid this conflict. Here, consider e.g. the RC-SVM using seven working sets and the VP-SVM with radius $r = 4$ for the COVTYPE data set of size 500 000.

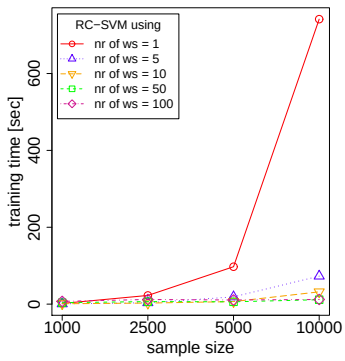
In Section 6.1.2, we compared for each data set type the LS-SVM with the VP-SVM that trains fastest for the largest training data set. Here, we additionally compare this VP-SVM to the RC-SVMs. To be able to draw a fair comparison in terms of the achieved test errors, we choose those RC-SVMs that train roughly as fast as the VP-SVM for the largest training set, i.e. the slowest RC-SVM training faster and the fastest RC-SVM training slower than the above VP-SVM. Subfigures (g)–(i) of Figures 10–12 illustrate the average training times and the average test errors of these RC-SVMs, the above VP-SVM, and the LS-SVM. Considering the RC-SVMs, the faster of the two requires for $n_{\text{train}} = 100\,000$ between 51% and 83% of the VP-SVMs training time and trains at most seven minutes faster than the VP-SVM. However, at least for $n_{\text{train}} \geq 5\,000$, both considered RC-SVMs induce substantially higher test errors than the VP- and LS-SVM. Finally, note that VP-SVMs for $n_{\text{train}} \geq 250\,000$ considerably outperform LS-SVMs for $n_{\text{train}} = 100\,000$, while RC-SVMs for $n_{\text{train}} \geq 250\,000$ lead to even worse test errors than the considered LS-SVMs.

Summarizing, we record that RC-SVMs using as few as possible working sets achieve the smallest RC-SVM test errors, however, those using more working sets perform faster. Furthermore, compared to VP-SVMs using roughly the same number of working sets as the RC-SVMs, the latter ones may learn faster though not as good as the VP-SVMs. Moreover, considering RC-SVMs that require roughly the same training time as the fastest VP-SVM, we saw that the RC-SVMs lead to much higher empirical risks. That is, if the required training time is a hard constraint, then the VP-SVM that satisfies this constraint achieves a better test error than a RC-SVM that also trains fast enough.

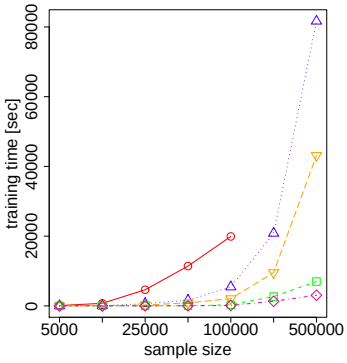
6.2 Experiments on Artificial Data

It remains to address the last question. To this end, we consider the results on the various artificial data sets, on the one hand, for the LS-SVM and, on the other hand, for the VP-SVM performing fastest for $n_{\text{train}} = 10\,000$. Moreover, for the sake of comparability, we again add to this selection the two RC-SVMs training roughly as fast as the VP-SVM for the artificial data sets of size $n_{\text{train}} = 10\,000$. However, for the artificial data sets of Type I, II, and III, none of the executed RC-SVMs trained faster for $n_{\text{train}} = 10\,000$ than the VP-SVM with the smallest radius, so that we only consider one RC-SVM in these cases. The required average training times and average test errors of the selected SVMs are illustrated in Subfigures (g)–(i) of Figures 13–17 and summarized in Tables 7–11. Here, we note that, for the artificial data of Type I, II, and III, the VP-SVM using the smallest of the applied radii trains fastest for $n_{\text{train}} = 10\,000$, while for the artificial data of Type IV and V it is the VP-SVM using the second smallest radius.

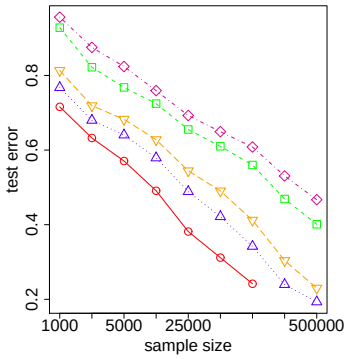
Expectedly, we detect an evident improvement of the various average empirical SVM solutions using 10 000 training samples instead of 1 000 samples. Besides, the considered



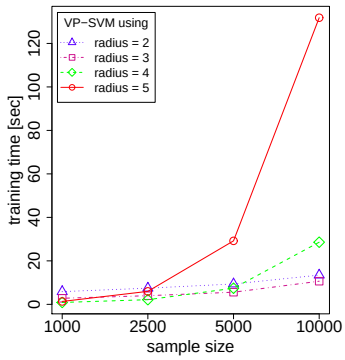
(a) Average training time of the various RC-SVMs for $n_{\text{train}} \leq 10\,000$



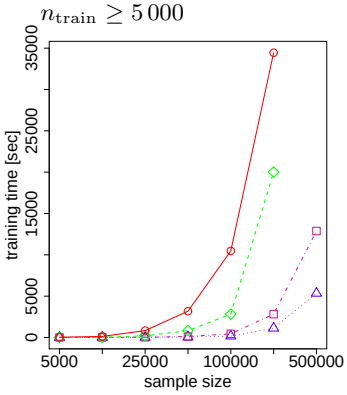
(b) Average training time of the various RC-SVMs for $n_{\text{train}} \geq 5\,000$



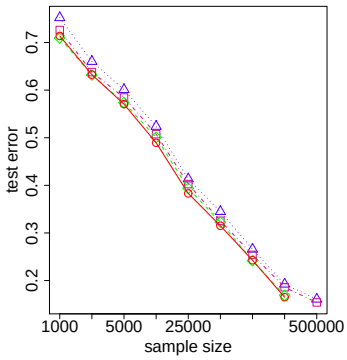
(c) Average empirical risk of the various RC-SVMs



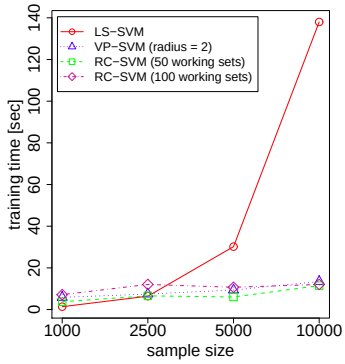
(d) Average training time of the various VP-SVMs for $n_{\text{train}} \leq 10\,000$



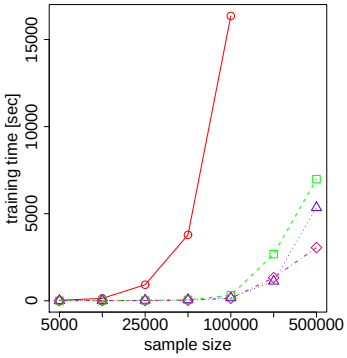
(e) Average training time of the various VP-SVMs for $n_{\text{train}} \geq 5\,000$



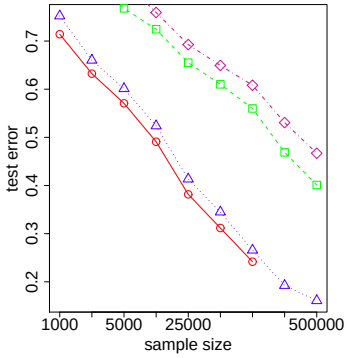
(f) Average empirical risk of the various VP-SVMs



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

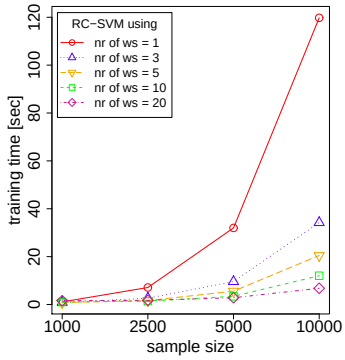


(h) Average training time of LS-, VP-, and RC-SVMs for $n_{\text{train}} \geq 5\,000$

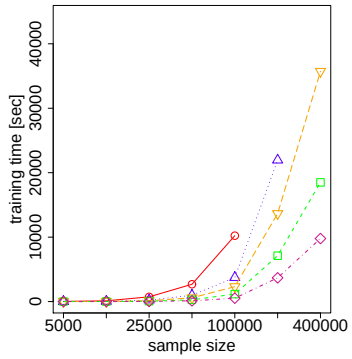


(i) Average empirical risk of LS-, VP-, and RC-SVMs

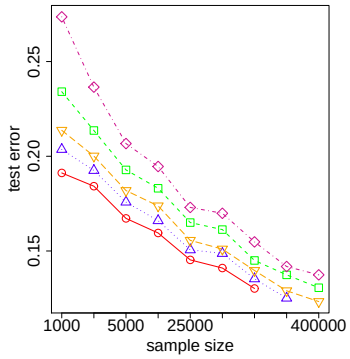
Figure 10: Average training time and test error of LS-, VP-, and RC-SVMs for the real-world data COVTYPE depending on the training set size $n_{\text{train}} = 1\,000, \ldots, 500\,000$. Subfigures (a)–(c) show the results for RC-SVMs using different numbers of working sets and Subfigures (d)–(f) illustrate 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 $n_{\text{train}} = 500\,000$ and the two RC-SVMs are those which achieve for $n_{\text{train}} = 500\,000$ roughly the same training time as the chosen VP-SVM. Here, note that, for $n_{\text{train}} = 10\,000$, 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.



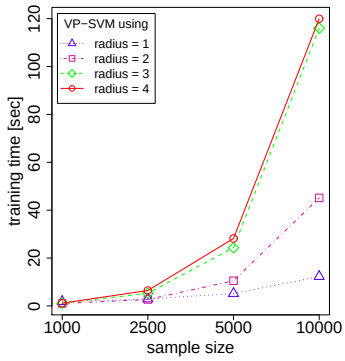
(a) Average training time of the various RC-SVMs for $n_{\text{train}} \leq 10\,000$



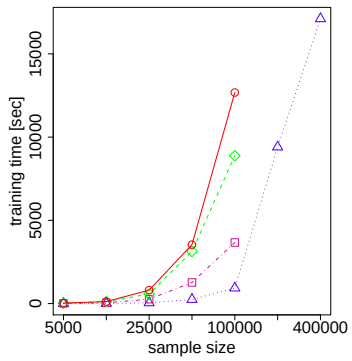
(b) Average training time of the various RC-SVMs for $n_{\text{train}} \geq 5\,000$



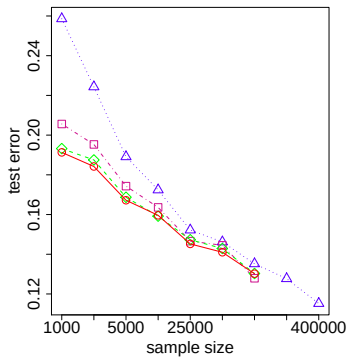
(c) Average empirical risk of the various RC-SVMs



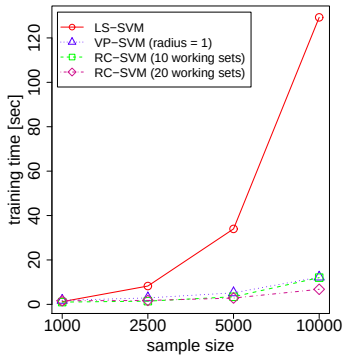
(d) Average training time of the various VP-SVMs for $n_{\text{train}} \leq 10\,000$



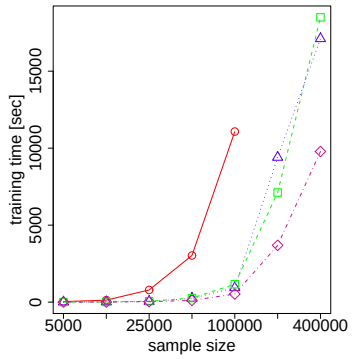
(e) Average training time of the various VP-SVMs for $n_{\text{train}} \geq 5\,000$



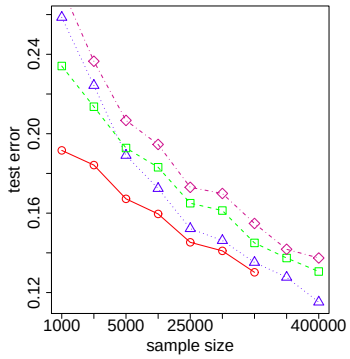
(f) Average empirical risk of the various VP-SVMs



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

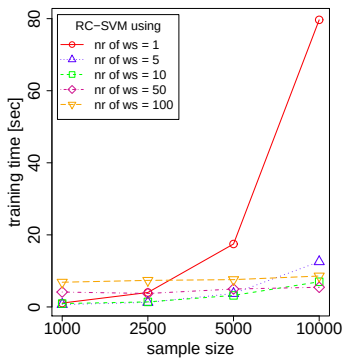


(h) Average training time of LS-, VP-, and RC-SVMs for $n_{\text{train}} \geq 5\,000$

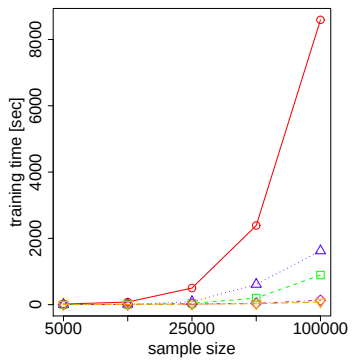


(i) Average empirical risk of LS-, VP-, and RC-SVMs

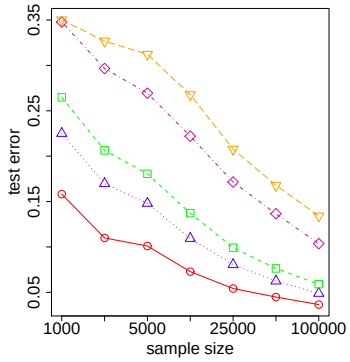
Figure 11: Average training time and test error of LS-, VP-, and RC-SVMs for the real-world data COD-RNA depending on the training set size $n_{\text{train}} = 1\,000, \ldots, 400\,000$. Subfigures (a)–(c) show the results for RC-SVMs using different numbers of working sets and Subfigures (d)–(f) illustrate 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 $n_{\text{train}} = 400\,000$ and the two RC-SVMs are those which achieve for $n_{\text{train}} = 400\,000$ roughly the same training time as the chosen VP-SVM.



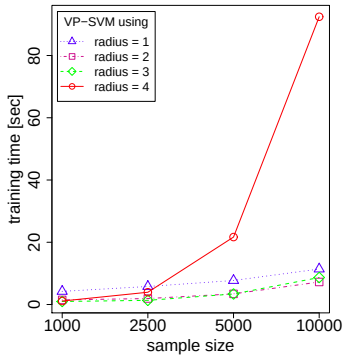
(a) Average training time of the various RC-SVMs for $n_{\text{train}} \leq 10\,000$



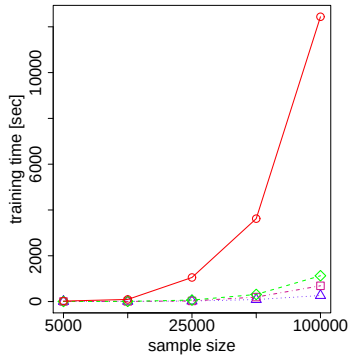
(b) Average training time of the various RC-SVMs for $n_{\text{train}} \geq 5\,000$



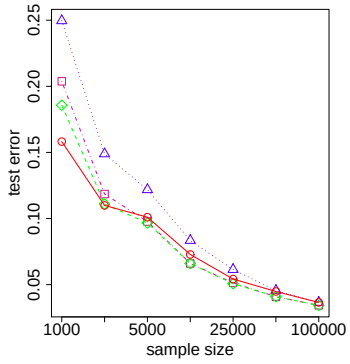
(c) Average empirical risk of the various RC-SVMs



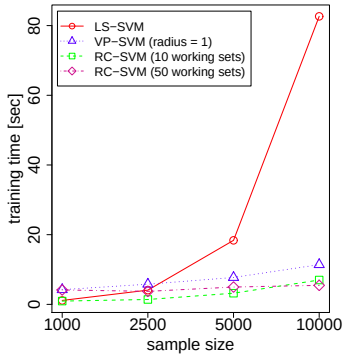
(d) Average training time of the various VP-SVMs for $n_{\text{train}} \leq 10\,000$



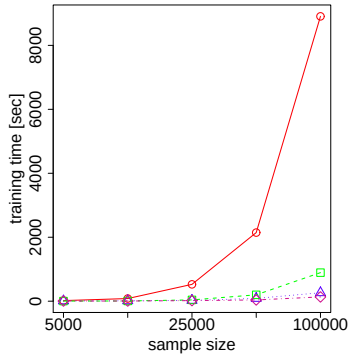
(e) Average training time of the various VP-SVMs for $n_{\text{train}} \geq 5\,000$



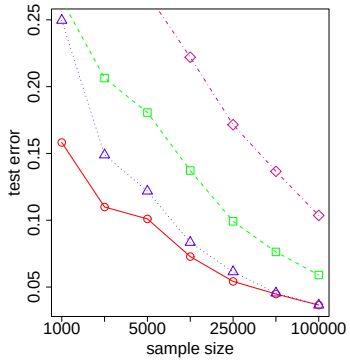
(f) Average empirical risk of the various VP-SVMs



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

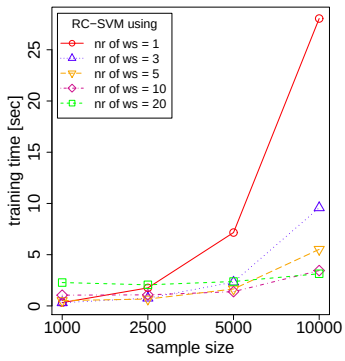


(h) Average training time of LS-, VP-, and RC-SVMs for $n_{\text{train}} \geq 5\,000$

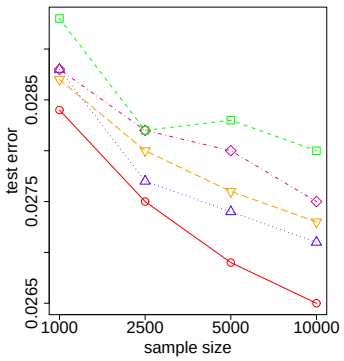


(i) Average empirical risk of LS-, VP-, and RC-SVMs

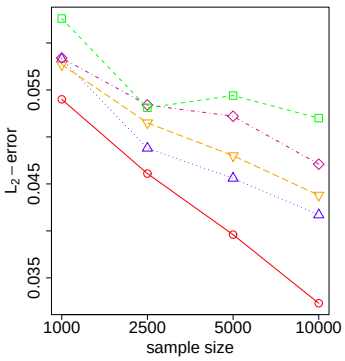
Figure 12: Average training time and test error of LS-, VP-, and RC-SVMs for the real-world data IJCNN1 depending on the training set size $n_{\text{train}} = 1\,000, \ldots, 100\,000$. Subfigures (a)–(c) show the results for RC-SVMs using different numbers of working sets and Subfigures (d)–(f) illustrate 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 $n_{\text{train}} = 100\,000$ and the two RC-SVMs are those which achieve for $n_{\text{train}} = 100\,000$ roughly the same training time as the chosen VP-SVM.



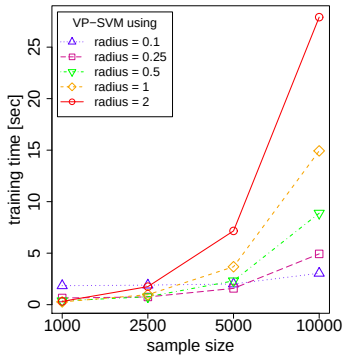
(a) Average training time of the various RC-SVMs



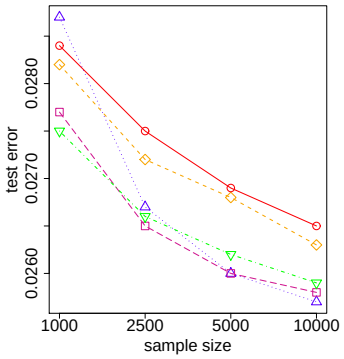
(b) Average empirical risk of the various RC-SVMs



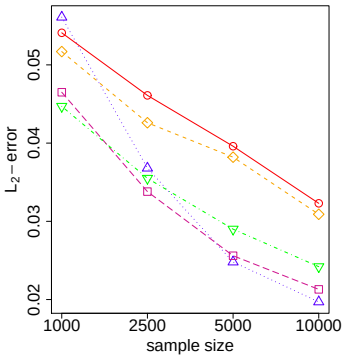
(c) Average empirical L_2-error of the various RC-SVMs



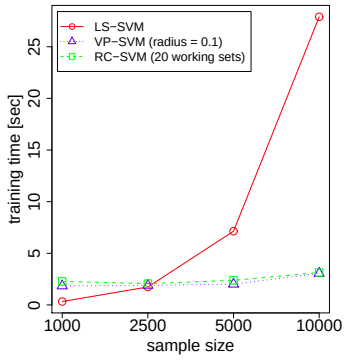
(d) Average training time of the various VP-SVMs



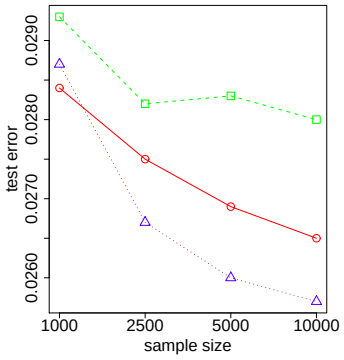
(e) Average empirical risk of the various VP-SVMs



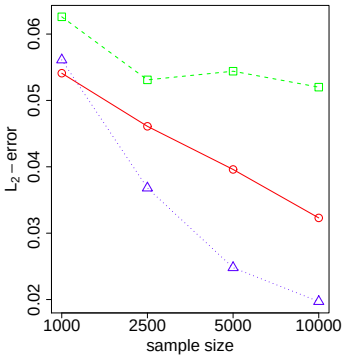
(f) Average empirical L_2-error of the various VP-SVMs



(g) Average training time of LS-, VP-, and RC-SVMs

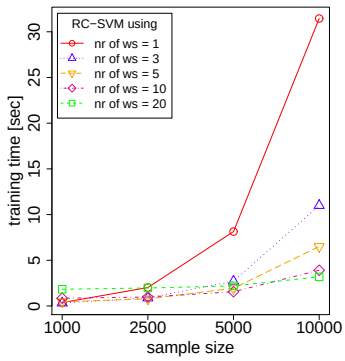


(h) Average empirical risk of LS-, VP-, and RC-SVMs

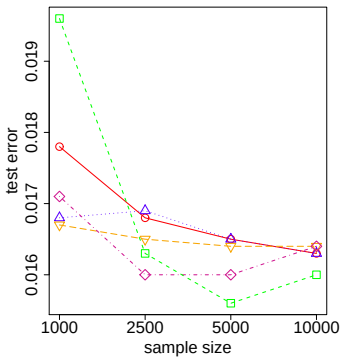


(i) Average empirical L_2-error of LS-, VP-, and RC-SVMs

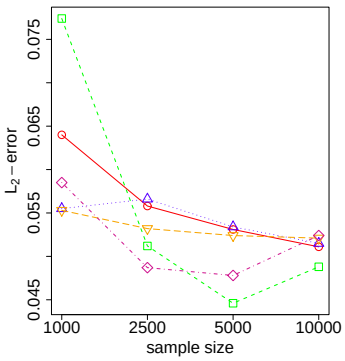
Figure 13: Average training time and test error of LS-, VP-, and RC-SVMs for the artificial data Type I depending on the training set size $n_{\text{train}} = 1\,000, \ldots, 10\,000$. Subfigures (a)–(c) show the results for RC-SVMs using different numbers of working sets and Subfigures (d)–(f) illustrate 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 one RC-SVM. Here, the VP-SVM and the RC-SVM are those which train fastest for $n_{\text{train}} = 10\,000$. Note that in the case at hand none of the considered RC-SVMs performs faster than the fastest VP-SVM for $n_{\text{train}} = 10\,000$.



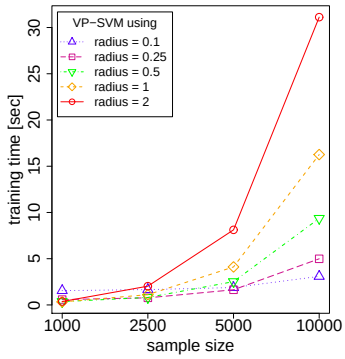
(a) Average training time of the various RC-SVMs



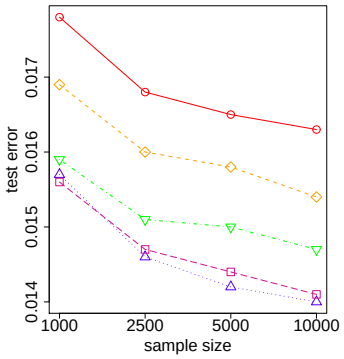
(b) Average empirical risk of the various RC-SVMs



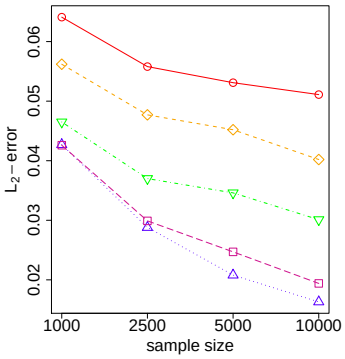
(c) Average empirical L_2-error of the various RC-SVMs



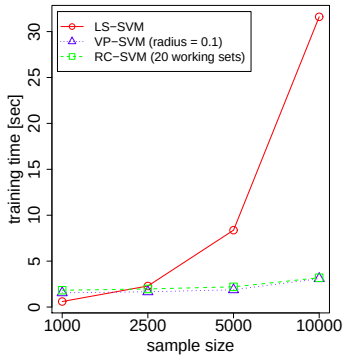
(d) Average training time of the various VP-SVMs



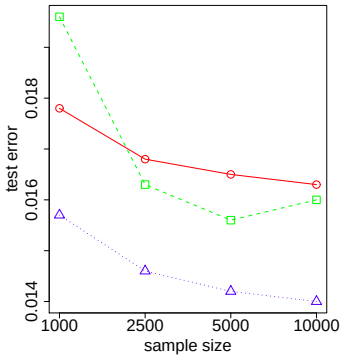
(e) Average empirical risk of the various VP-SVMs



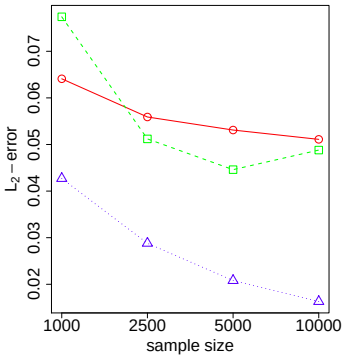
(f) Average empirical L_2-error of the various VP-SVMs



(g) Average training time of LS-, VP-, and RC-SVMs

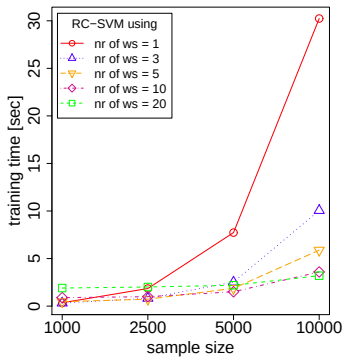


(h) Average empirical risk of LS-, VP-, and RC-SVMs

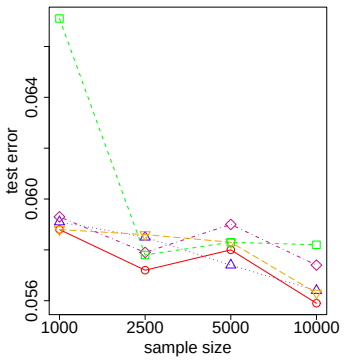


(i) Average empirical L_2-error of LS-, VP-, and RC-SVMs

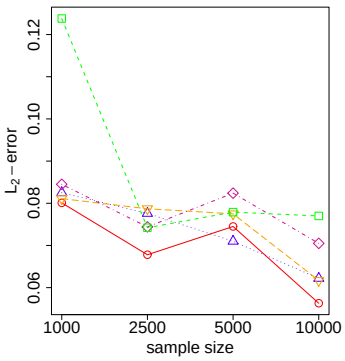
Figure 14: Average training time and test error of LS-, VP-, and RC-SVMs for the artificial data Type II depending on the training set size $n_{\text{train}} = 1\,000, \ldots, 10\,000$. Subfigures (a)–(c) show the results for RC-SVMs using different numbers of working sets and Subfigures (d)–(f) illustrate 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 one RC-SVM. Here, the VP-SVM and the RC-SVM are those which train fastest for $n_{\text{train}} = 10\,000$. Note that in the case at hand none of the considered RC-SVMs performs faster than the fastest VP-SVM for $n_{\text{train}} = 10\,000$.



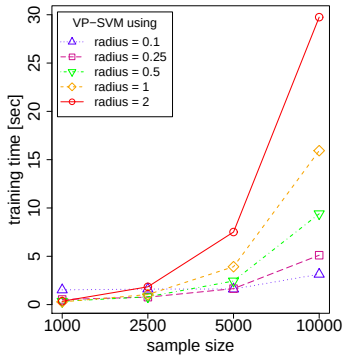
(a) Average training time of the various RC-SVMs



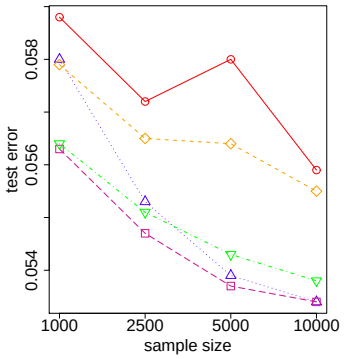
(b) Average empirical risk of the various RC-SVMs



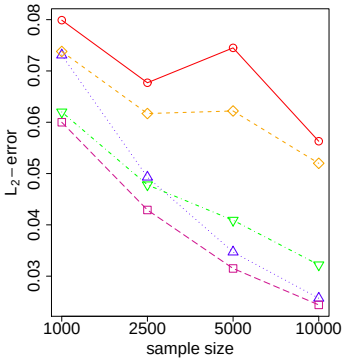
(c) Average empirical L_2-error of the various RC-SVMs



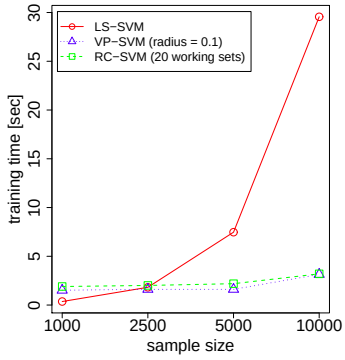
(d) Average training time of the various VP-SVMs



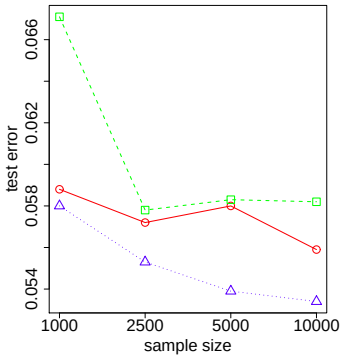
(e) Average empirical risk of the various VP-SVMs



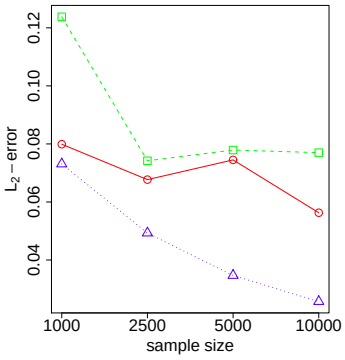
(f) Average empirical L_2-error of the various VP-SVMs



(g) Average training time of LS-, VP-, and RC-SVMs

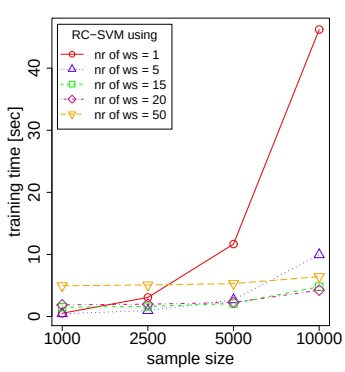


(h) Average empirical risk of LS-, VP-, and RC-SVMs

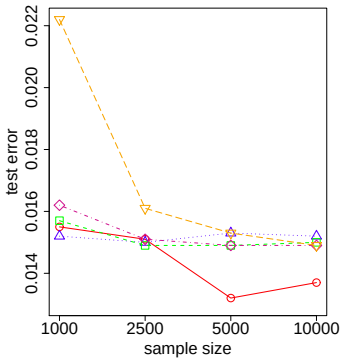


(i) Average empirical L_2-error of LS-, VP-, and RC-SVMs

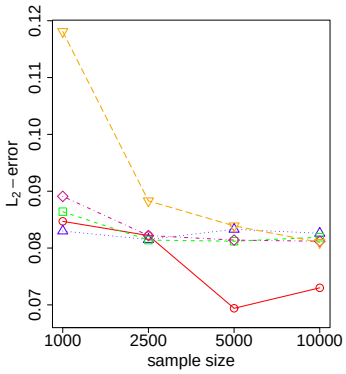
Figure 15: Average training time and test error of LS-, VP-, and RC-SVMs for the artificial data Type III depending on the training set size $n_{\text{train}} = 1\,000, \ldots, 10\,000$. Subfigures (a)–(c) show the results for RC-SVMs using different numbers of working sets and Subfigures (d)–(f) illustrate 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 one RC-SVM. Here, the VP-SVM and the RC-SVM are those which train fastest for $n_{\text{train}} = 10\,000$. Note that in the case at hand none of the considered RC-SVMs performs faster than the fastest VP-SVM for $n_{\text{train}} = 10\,000$.



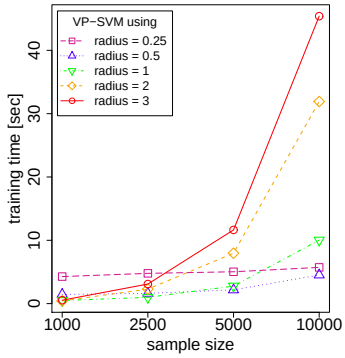
(a) Average training time of the various RC-SVMs



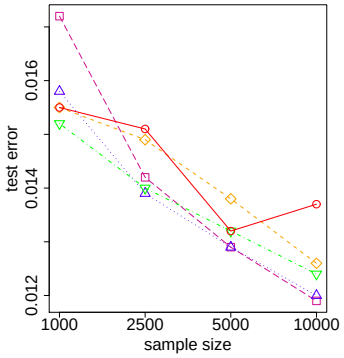
(b) Average empirical risk of the various RC-SVMs



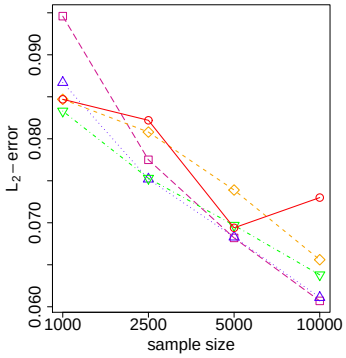
(c) Average empirical L_2-error of the various RC-SVMs



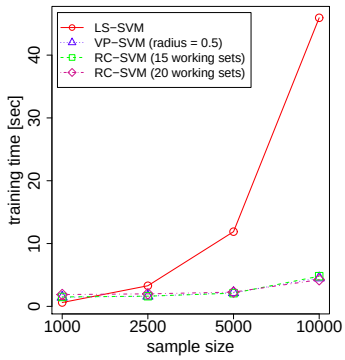
(d) Average training time of the various VP-SVMs



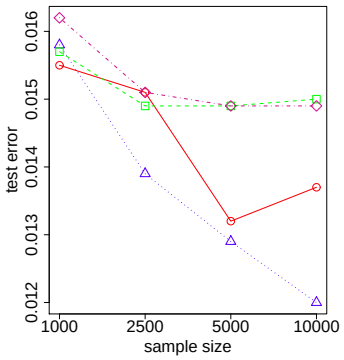
(e) Average empirical risk of the various VP-SVMs



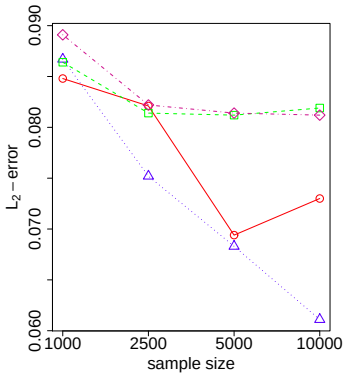
(f) Average empirical L_2-error of the various VP-SVMs



(g) Average training time of LS-, VP-, and RC-SVMs

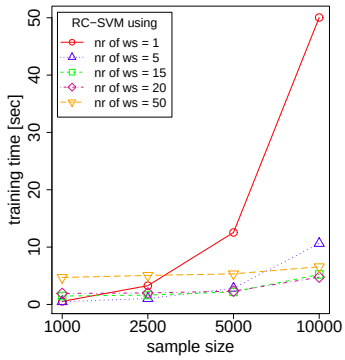


(h) Average empirical risk of LS-, VP-, and RC-SVMs

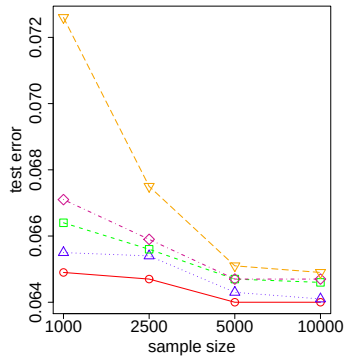


(i) Average empirical L_2-error of LS-, VP-, and RC-SVMs

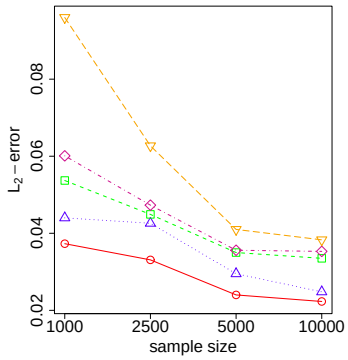
Figure 16: Average training time and test error of LS-, VP-, and RC-SVMs for the artificial data Type IV depending on the training set size $n_{\text{train}} = 1\,000, \ldots, 10\,000$. Subfigures (a)–(c) show the results for RC-SVMs using different numbers of working sets and Subfigures (d)–(f) illustrate 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 $n_{\text{train}} = 10\,000$ and the two RC-SVMs are those which achieve for $n_{\text{train}} = 10\,000$ roughly the same training time as the chosen VP-SVM.



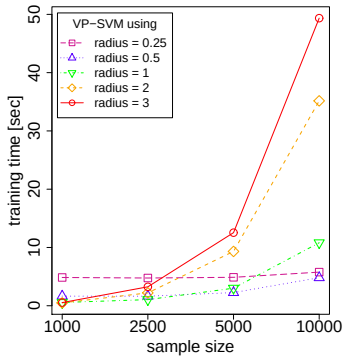
(a) Average training time of the various RC-SVMs



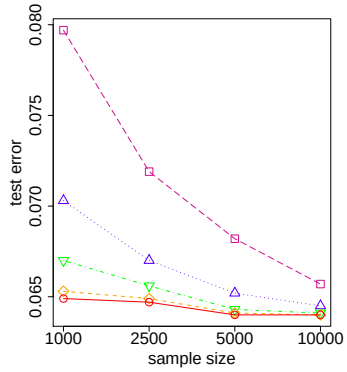
(b) Average empirical risk of the various RC-SVMs



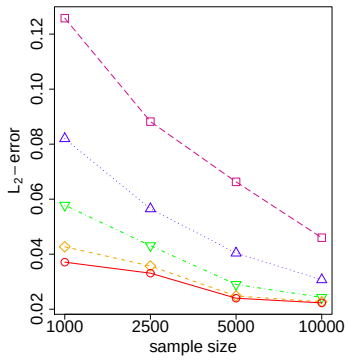
(c) Average empirical L_2-error of the various RC-SVMs



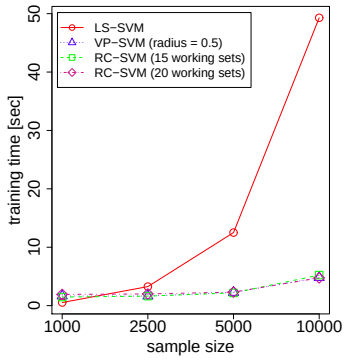
(d) Average training time of the various VP-SVMs



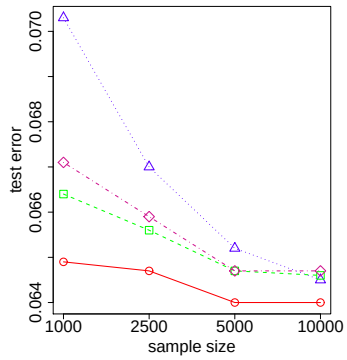
(e) Average empirical risk of the various VP-SVMs



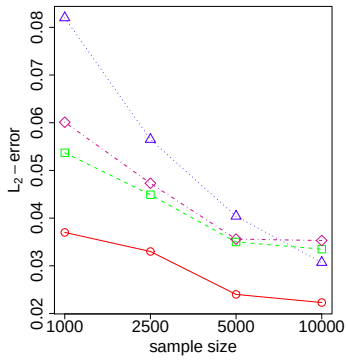
(f) Average empirical L_2-error of the various VP-SVMs



(g) Average training time of LS-, VP-, and RC-SVMs



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



(i) Average empirical L_2-error of LS-, VP-, and RC-SVMs

Figure 17: Average training time and test error of LS-, VP-, and RC-SVMs for the artificial data Type V depending on the training set size $n_{\text{train}} = 1\,000, \ldots, 10\,000$. 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 $n_{\text{train}} = 10\,000$ and the two RC-SVMs are those which achieve for $n_{\text{train}} = 10\,000$ roughly the same training time as the chosen VP-SVM.

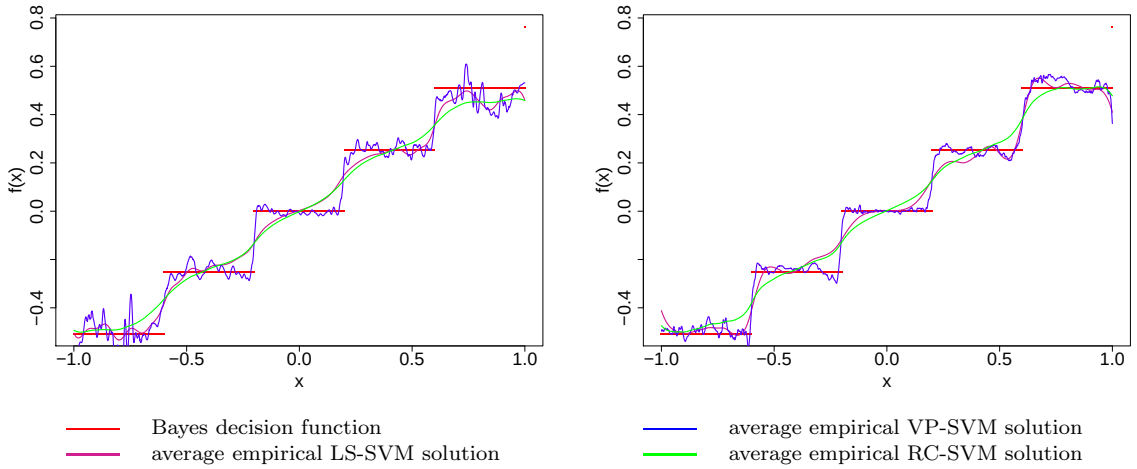


Figure 18: Predictions for the artificial data sets of Type I, drawn from the step function in Figure 9(a) with noise depending on x. The left graphic shows the predictions for the data set of size $n_{\text{train}} = 1\,000$ and the right graphic for the data set of size $n_{\text{train}} = 10\,000$. Here, note that the VP-SVM solutions are not necessarily continuous, nevertheless we continuously connected its predicted values in the above plots.

VP-SVM trains substantially faster than the standard LS-SVM with less than 11% of the LS-SVM's training time for $n_{\text{train}} = 10\,000$. Additionally, the VP-SVM's test errors are usually considerably lower than the test errors of the LS-SVM. Regarding the test errors of the RC-SVMs, we note that, in the majority of cases, they are higher than the VP-SVM's and the LS-SVM's test errors.

So far, we examined the behavior of LS-, VP-, and RC-SVMs in terms of training time and test error. Let us finally compare the three different kinds of SVMs w.r.t. their optical appearance. To this end, the average empirical SVM solutions are plotted in Figures 18–23 for the different artificial data sets of size $n_{\text{train}} = 1\,000$ and $10\,000$. Here, note that, for the artificial data of Type IV and V, we do not consider both RC-SVMs training roughly as fast as the selected VP-SVM but only the one of the both RC-SVMs with the lower test error.

The observation that, for the artificial data of Type I, II, and III, the VP-SVMs perform best, is reinforced by the average empirical VP-SVM solutions illustrated in Figures 18–21. More precisely, Figure 18 shows that only the VP-SVMs exhaust the widths of the steps of $f^*_{L,\mathrm{P}}$ almost completely. Moreover, in Figure 19 the smoothness interruptions of $f^*_{L,\mathrm{P}}$ are again best illustrated by the VP-SVMs, which becomes even more evident in Figure 20. Besides, Figure 21 illustrates that the peaks of $f^*_{L,\mathrm{P}}$ are best reproduced by the VP-SVMs. Considering the LS- and the RC-SVMs, we can not draw an universally valid conclusion, which one performs worse. In particular, Figures 18 and 19 show that, for the artificial data sets of Type I and II, both, the average empirical LS- and RC-SVM solutions, are not very well suited to the Bayes decision function. Considering the data sets of Type III, the LS-SVMs dominate the RC-SVMs in terms of the better test errors, though both kinds of SVMs do not reproduce the peaks of $f^*_{L,\mathrm{P}}$, especially for small values of $|x|$.

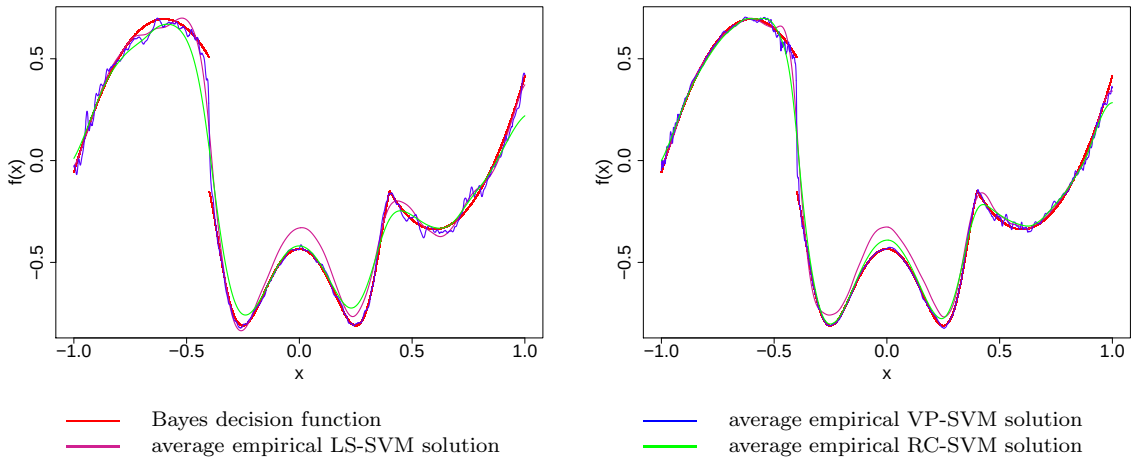


Figure 19: Predictions for the artificial data sets of Type II, drawn from the cracked function in Figure 9(b) with noise depending on x. The left graphic shows the predictions for the data set of size $n_{\text{train}} = 1\,000$ and the right graphic for the data set of size $n_{\text{train}} = 10\,000$.

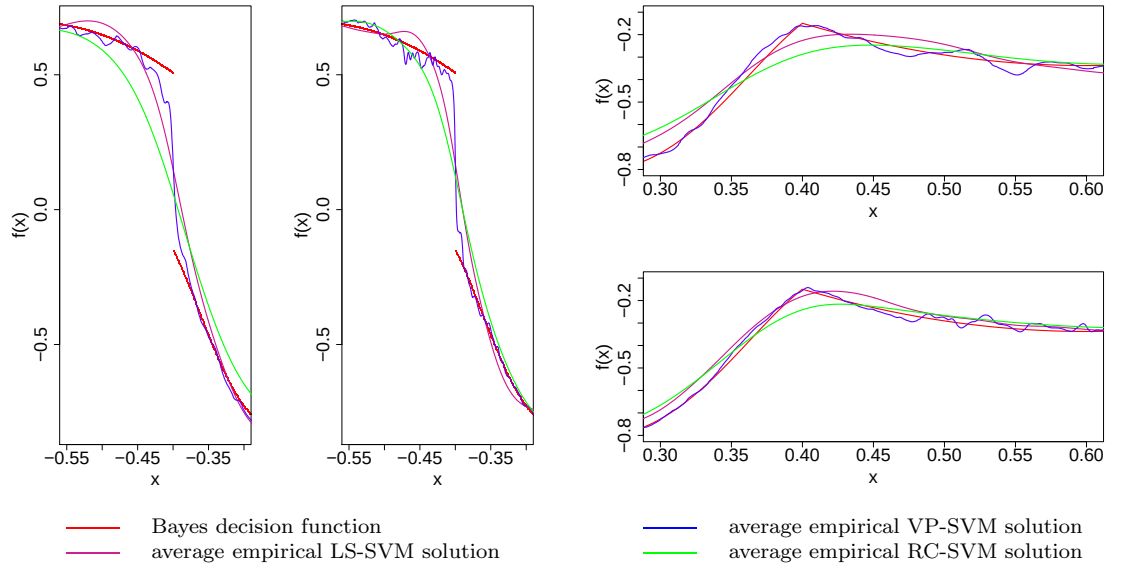


Figure 20: Predictions for the artificial data sets of Type II. The left graphic shows the predictions for $x \in [-0.55, -0.4]$ and the data set of size $n_{\text{train}} = 1\,000$, while the graphic on its right-hand side pictures the predictions for the same interval for x and the data set of size $n_{\text{train}} = 10\,000$. The two graphics on the right-hand side illustrate the predictions for $x \in [0.3, 0.6]$, the upper one for the data set of size $n_{\text{train}} = 1\,000$ and the lower one for the data set of size $n_{\text{train}} = 10\,000$.

It remains to optically analyze the results of the two-dimensional data sets in the following. For the artificial data sets of Type IV, the VP-SVM using 10 000 training samples achieves the best test error. Moreover, ensuing the optical impression, this VP-SVM is the

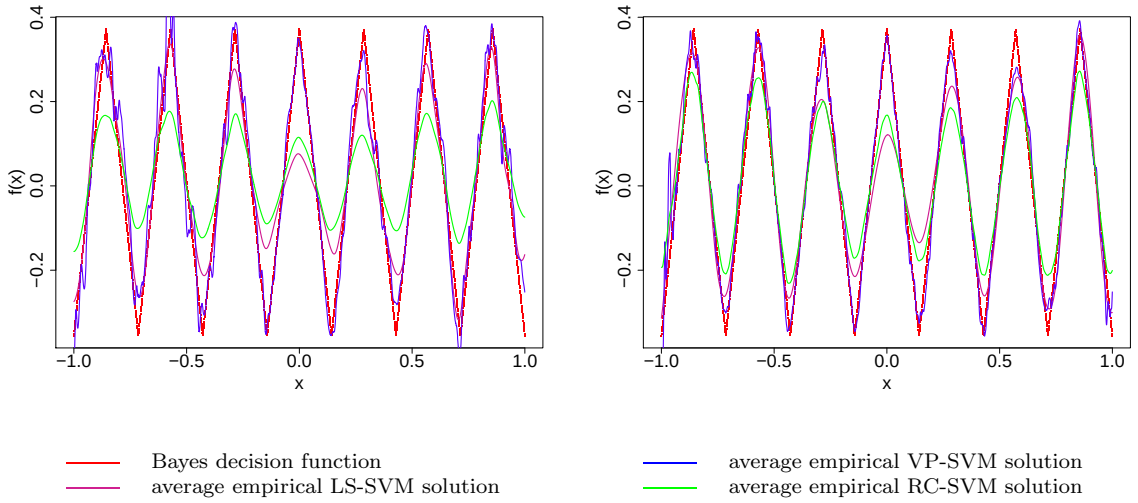


Figure 21: Predictions for the artificial data sets of Type III, drawn from the jagged function in Figure 9(c) with noise depending on x. The left graphic shows the predictions for the data set of size $n_{\text{train}} = 1\,000$ and the right graphic for the data set of size $n_{\text{train}} = 10\,000$.

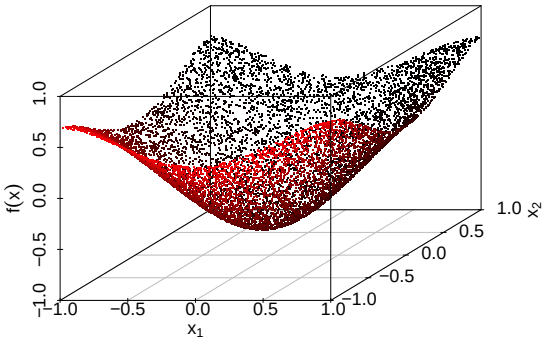
only one of the considered SVM types that reflects the circular steps of the Bayes decision function as in Figure 9(d), cf. Figure 22. Finally, for the data sets of Type V, it is always the LS-SVM which performs best in terms of the test errors, cf. Table 11. This observation is also substantiated optically. To be more precise, for $n_{\text{train}} = 1\,000$, the uneven average empirical decision function induced by the VP-SVM (cf. Figure 23) shows that the RC-SVM even performs better than the VP-SVM. However, for $n_{\text{train}} = 10\,000$, the VP-SVM results are substantially improved such that the RC-SVM is now outperformed by the VP-SVM.

Recapulatory, we realize that the VP-SVMs possess the most distinctive ability to handle smoothness interruptions of the Bayes decision function in most of our artificial data cases, especially if $n_{\text{train}} = 10\,000$. This finding agrees with the theoretical results of Section 5, that already promised a stable behavior of the VP-SVM for Bayes decision functions that are not smooth on zero sets. For the sake of completeness, we point out that the worst performance was induced by the RC-SVMs in almost all cases, in particular for a training sample size amounting to 10 000.

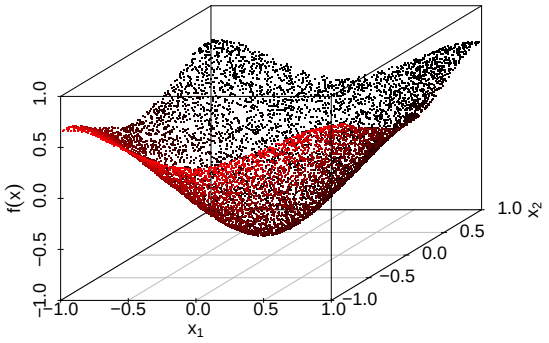
6.3 Conclusions

Finally, we summarize the essential findings of the previous subsections, where we considered standard LS-SVMs and two kinds of localized SVMs, namely VP-SVMs and RC-SVMs. As just analyzed in Subsection 6.2, VP-SVMs have the evident advantage that they manage smoothness interruptions of the Bayes decision function better than LS- and RC-SVMs.

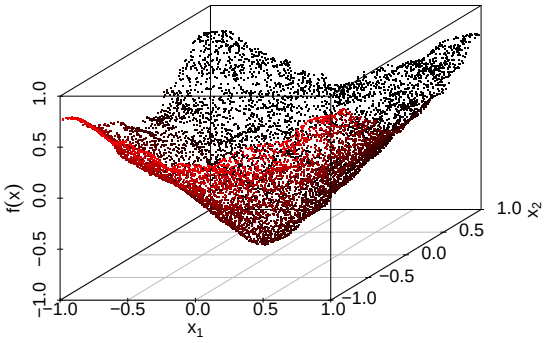
The real-world data sets demonstrated that the RC-SVMs perform considerably worse than the LS-SVMs and the VP-SVMs, while the performance of VP-SVMs using small radii is improved for increasing sample sizes. To be more precise, VP-SVMs outperform LS-SVMs or at most leads to a negligible worsening compared to LS-SVMs for a fraction of the training time and without memory constraints on the large data sets. For very



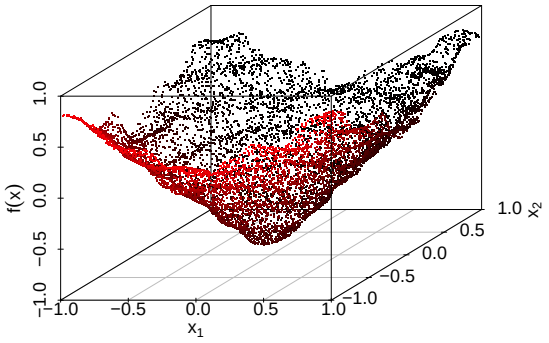
(a) LS-SVM using 1 000 training samples



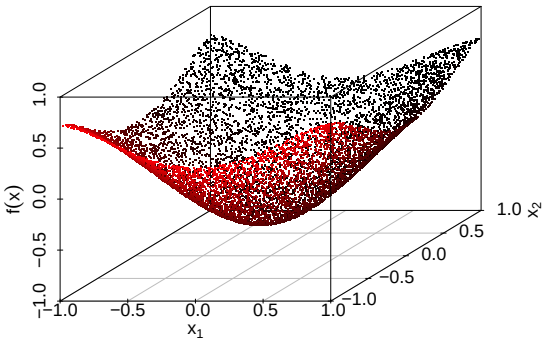
(b) LS-SVM using 10 000 training samples



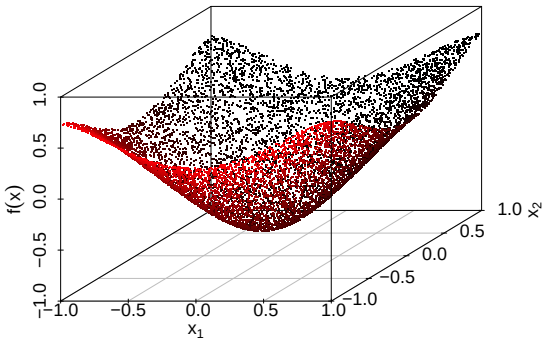
(c) VP-SVM using radius $r = 0.5$ and 1 000 training samples



(d) VP-SVM using radius $r = 0.5$ and 10 000 training samples

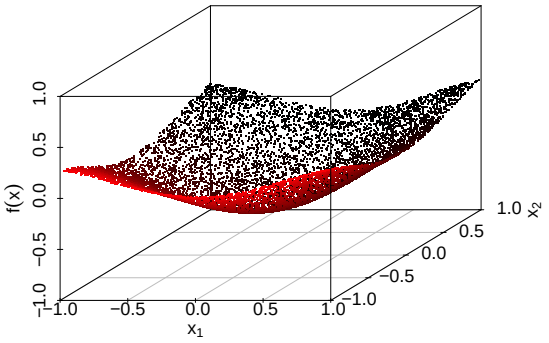


(e) RC-SVM using 20 working sets and 1 000 training samples

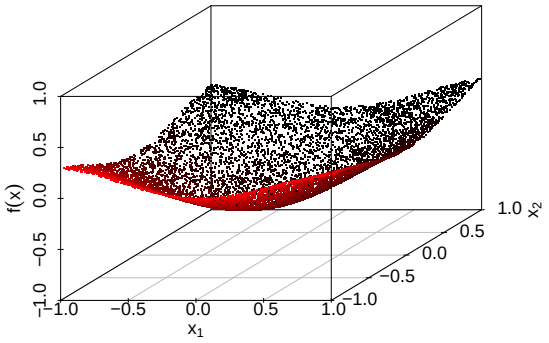


(f) RC-SVM using 20 working sets and 10 000 training samples

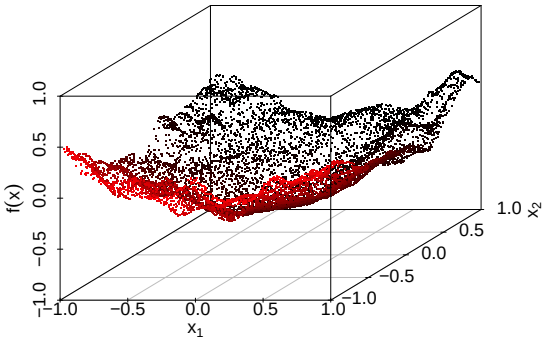
Figure 22: Predictions for the artificial data sets of Type IV, drawn from the circular step function in Figure 9(d) with noise independent of x.



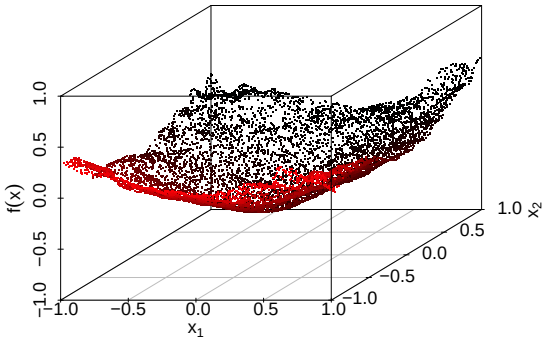
(a) LS-SVM using 1 000 training samples



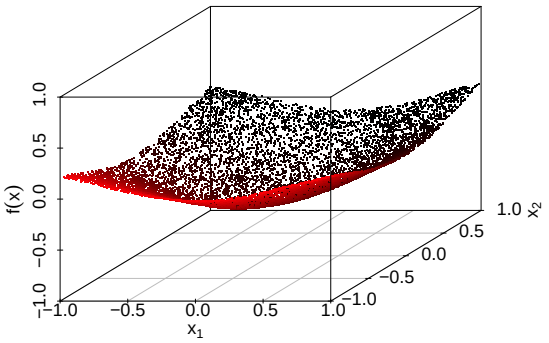
(b) LS-SVM using 10 000 training samples



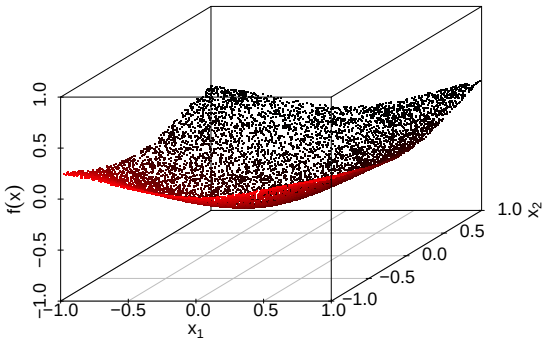
(c) VP-SVM using radius $r = 0.5$ and 1 000 training samples



(d) VP-SVM using radius $r = 0.5$ and 10 000 training samples



(e) RC-SVM using 15 working sets and 1 000 training samples



(f) RC-SVM using 15 working sets and 10 000 training samples

Figure 23: Predictions for the artificial data sets of Type V, drawn from the 2-dimensional Euclidean norm in Figure 9(e) with noise independent of x.

small data sets, however, LS-SVMs actually train faster than VP-SVMs and, hence, are preferable. What is more, for data sets of size $n_{\text{train}} \leq 2\,500$ all LS-SVMs require less than 9 s to train, so that there are probably no reasons to apply a VP-SVM. Besides, really small training sample sizes involve considerably smaller working sets for a VP-SVM using a small radius, so that it is hard to find a well suited prediction.

Furthermore, despite a faster training procedure, a VP-SVM using a sufficiently small radius induces considerably lower test errors for sample sizes $n_{\text{train}} > 100\,000$ than a LS-SVM for training data sets that still enable computational feasibility.

7. Proofs

This section is dedicated to prove the results of the previous sections. We begin with the proof of Lemma 1 relating the radius r of a cover $B_r(z_1), \ldots, B_r(z_m)$ of X defined by (6) with the number m of centers $z_1, \ldots, z_m$.

Proof [Proof of Lemma 1] First of all, let us recall the m-th entropy number of X defined by

$$\varepsilon_m(X) := \inf \left\{ \varepsilon > 0 : \exists z_1, \ldots, z_m \in X \text{ such that } X \subset \bigcup_{j=1}^{m} (z_j + \varepsilon B_{\ell_2^d}) \right\} .$$

Since $X \subset cB_{\ell_2^d}$, the m-th entropy number of X can be upper bounded by

$$\varepsilon_m(X) \leq 2\varepsilon_m(cB_{\ell_2^d}) \leq 2c\varepsilon_m(B_{\ell_2^d}) .$$

Additionally, we know by (Carl and Stephani, 1990, Section 1.1) that

$$m^{-\frac{1}{d}} \leq \varepsilon_m(B_{\ell_2^d}) \leq 4m^{-\frac{1}{d}} ,$$

so that we can find a cover $(B_j)_{j=1,\ldots,m}$ of $X \subset cB_{\ell_2^d}$ satisfying

$$r \leq 8cm^{-\frac{1}{d}} .$$

■

7.1 Proofs of Section 3

In Section 3 we presented a lemma that related the risk w.r.t. the loss L to the risk w.r.t. the restricted loss L_j and also transferred this result to the excess risk. Hereafter, the proof of this lemma can be found.

Proof [Proof of Lemma 4] Simple transformations using $A \cup B = X$ and $A \cap B = \emptyset$ show

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

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

The second assertion follows immediately. ∎

To derive the new oracle inequality of Theorem 5 we first have to relate the entropy numbers of H_j, $j \in \{1,\ldots,m\}$, to those of H. To this end, we consider a similar concept to entropy numbers, namely covering numbers, cf. (Györfi et al., 2002, Definition 9.3) or (Steinwart and Christmann, 2008a, Definition 6.19).

Definition 14 *Let (T,d) be a metric space and $\varepsilon > 0$. A subset $S \subset T$ is called an ε-net of T if for all $t \in T$ there exists an $s \in S$ such that $d(s,t) \leq \varepsilon$. Furthermore, we define the ε-covering number of T by*

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

where $\inf\emptyset := \infty$ and $B_d(s,\varepsilon) := \{t \in T : d(t,s) \leq \varepsilon\}$.

Note that an upper bound on entropy numbers involves a bound on covering numbers. To be more precise, for a metric space (T,d) and constants $a > 0$ and $q > 0$, the implication

$$e_i(T,d) \leq a i^{-1/q}\,, \qquad i \geq 1 \qquad \Longrightarrow \qquad \ln \mathcal{N}(T,d,\varepsilon) \leq \ln(4)\left(\frac{a}{\varepsilon}\right)^q\,, \qquad \forall\, \varepsilon > 0 \tag{27}$$

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

$$\ln \mathcal{N}(T,d,\varepsilon) < \left(\frac{a}{\varepsilon}\right)^q\,, \qquad \varepsilon > 0 \qquad \Longrightarrow \qquad e_i(T,d) \leq 3^{1/q} a i^{-1/q}\,, \qquad \forall\, i \geq 1\,. \tag{28}$$

Recall that we pursue the target to estimate $e_i(\mathrm{id} : H \to L_2(\mathrm{P}_X))$. In fact, the equivalence of entropy and covering numbers enables us to estimate the covering number $\mathcal{N}(B_H, \|\cdot\|_{L_2(\mathrm{P}_X)}, \varepsilon)$ of H instead.

Lemma 15 *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* (13)*, where the norm is given by* (14) *with $\lambda_A, \lambda_B > 0$. Then, for the ε-covering number of H w.r.t. $\|\cdot\|_{L_2(\nu)}$, we have*

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

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

$$\left\|\hat{g}_A - \hat{f}_{i_A}\right\|_{L_2(\nu_{|A})} \leq \varepsilon_A\,, \tag{29}$$

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

$$\left\|\hat{g}_B - \hat{h}_{i_B}\right\|_{L_2(\nu_{|B})} \leq \varepsilon_B\,. \tag{30}$$

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 (29) and (30), this implies

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

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

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

■

Based on Lemma 15, the following theorem relates entropy numbers of H_A and H_B to those of H.

Theorem 16 *Let P_X be a distribution on X and $A_1, \ldots, A_m \subset X$ be pairwise disjoint. Moreover, we assume (H) with weights $\lambda_1, \ldots, \lambda_m > 0$. In addition, assume that there exist constants $p \in (0,1)$ and $a_j > 0$, $j \in \{1, \ldots, m\}$, such that for every $j \in \{1, \ldots, m\}$*

$$e_i(\mathrm{id} : H_j \to L_2(\mathrm{P}_{X|A_j})) \leq a_j\, i^{-\frac{1}{2p}}\,, \qquad i \geq 1\,. \tag{31}$$

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 [Proof of Theorem 16] 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, 2008a, (A.36)) and assumption (31), 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 (27) 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 $j \in \{1, \ldots, m\}$, Lemma 15 implies

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

Using (28), the latter bound for the covering number of B_H finally implies the following entropy estimate

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

$$\leq 2\,(3\ln(4))^{\frac{1}{2p}}\,\sqrt{m}\left(\sum_{j=1}^{m}\lambda_j^{-p}a_j^{2p}\right)^{\frac{1}{2p}} i^{-\frac{1}{2p}}\,.$$

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

Applying Theorem 16, we now prove Theorem 5 and thus an oracle inequality for VP-SVMs using an ordinary type of losses.

Proof [Proof of Theorem 5] Since $H_1,\ldots,H_m$ are seperable RKHS of mesurable kernels $k_1,\ldots,k_m$, H is a seperable RKHS and its kernel k is measurable, too. Furthermore, Theorem 16 yields

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

That is, we can apply (Steinwart and Christmann, 2008a, Theorem 7.23) for a regularization parameter $\tilde{\lambda}=1$ and, for all fixed $\tau>0$ and $\lambda_j>0$, $j\in\{1,\ldots,m\}$, we obtain

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

with probability P^n not less than $1-3e^{-\tau}$, where $C>0$ is the constant of (Steinwart and Christmann, 2008a, Theorem 7.23) only depending on p, M, V, ϑ, and B. Moreover,

$$a:=\max\left\{c_p\sqrt{m}\left(\sum_{j=1}^{m}\lambda_j^{-p}a_j^{2p}\right)^{\frac{1}{2p}},B\right\}\,,$$

where we need $a\geq B$ since it is a condition of (Steinwart and Christmann, 2008a, Theorem 7.23). ∎

7.2 Proofs Related to the Entropy Estimates of Section 4

In this subsection, just as in Section 4, we focus on Gaussian RBF kernels and the associated RKHSs. To be more precise, we derive a bound for the entropy numbers of $H_\gamma(A)$, where $\gamma>0$ and $A\subset\mathbb{R}^d$ with $\mathring{A}\neq\emptyset$.

Proof [Proof of Theorem 6] First of all, we consider the commutative diagram

$$\begin{array}{ccc} H_\gamma(A) & \xrightarrow{\quad\mathrm{id}\quad} & L_2(\mathrm{P}_{X|A}) \\ {\scriptstyle \mathcal{I}_B^{-1}\circ\mathcal{I}_A}\downarrow & & \uparrow{\scriptstyle \mathrm{id}} \\ H_\gamma(B) & \xrightarrow[\quad\mathrm{id}\quad]{} & \ell_\infty(B) \end{array}$$

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, 2008a, 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 $f \in \ell_\infty(B)$, we have

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

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, 2008a, (A.38) and (A.39)) as well as (Steinwart and Christmann, 2008a, 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

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

■

7.3 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 5. 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

$$\mathcal{R}_{L_{J_T},\mathrm{P}}(f_0) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} = \|f_0 - f^*_{L,\mathrm{P}}\|^2_{L_2\left(\mathrm{P}_{X|A_T}\right)}\,. \tag{32}$$

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

$$K_j : \mathbb{R}^d \to \mathbb{R}\,, \qquad x \mapsto \sum_{\ell=1}^{s} \binom{s}{\ell} (-1)^{1-\ell} \left(\frac{2}{\ell^2\gamma_j^2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|x\|_2^2}{\ell^2\gamma_j^2}\right)\,, \tag{33}$$

and choosing $f_0 := \sum_{j=1}^m \mathbb{1}_{A_j} \cdot (K_j * f^*_{L,\mathrm{P}})$. Then (32) 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 in we use the notation

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

in the following theorem and the associated proof.

Theorem 17 *Let us fix some $q \in [1, \infty)$. Assume that ν is a finite measure on $\mathbb{R}^d$ with $\operatorname{supp}\nu =: X \subset \mathbb{R}^d$ and let $(A_j)_{j=1,\ldots,m}$ be a partition 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 $K_j : \mathbb{R}^d \to \mathbb{R}$, $j \in \{1, \ldots, m\}$, defined by (33), where $s := \lfloor \alpha \rfloor + 1$ and $\gamma_1, \ldots, \gamma_m > 0$, we then have*

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

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 $J := \{1, \ldots, m\}$. To show

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

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 $q \geq 1$,

$$\begin{aligned} &\left\| \sum_{j \in J} \mathbb{1}_{A_j} \cdot (K_j * f) - f \right\|^q_{L_q(\nu)} \\ &= \int_{\mathbb{R}^d} \left| \sum_{j \in J} \mathbb{1}_{A_j}(x)\, (K_j * f)\,(x) - f\,(x) \right|^q d\nu\,(x) \\ &\leq \int_{\mathbb{R}^d} \left(\sum_{j \in J} \mathbb{1}_{A_j}(x)\, |K_j * f\,(x) - f\,(x)| \right)^q d\nu\,(x) \end{aligned}$$

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

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 $q > 1$,

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

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 $f \in B^\alpha_{q,\infty}(\nu)$ implies $\omega_{s,L_q(\nu)}(f,t) \leq \|f\|_{B^\alpha_{q,\infty}(\nu)} \, t^\alpha$ for $t > 0$. The latter together with Hölder's inequality yields

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

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

$$\begin{aligned}
\int_{\mathbb{R}^d} \|h\|_2^{2q\alpha} \exp\left(-\frac{2\|h\|_2^2}{\gamma_{\max}^2}\right) dh &\leq d^{q\alpha-1} \sum_{\ell=1}^d \int_{\mathbb{R}^d} h_\ell^{2q\alpha} \prod_{l=1}^d \exp\left(-\frac{2h_l^2}{\gamma_{\max}^2}\right) d(h_1,\ldots,h_d) \\
&= d^{q\alpha-1} \sum_{\ell=1}^d \left(\frac{\gamma_{\max}^2\pi}{2}\right)^{\frac{d-1}{2}} \int_{\mathbb{R}} h_\ell^{2q\alpha} \exp\left(-\frac{2h_\ell^2}{\gamma_{\max}^2}\right) dh_\ell \\
&= 2d^{q\alpha} \left(\frac{\gamma_{\max}^2\pi}{2}\right)^{\frac{d-1}{2}} \int_0^\infty t^{2q\alpha} \exp\left(-\frac{2t^2}{\gamma_{\max}^2}\right) dt \,.
\end{aligned}$$

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

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

■

Based on Theorems 5 and 17, we can now show Theorem 7.

Proof [Proof of Theorem 7] First, we have to choose a function $f_0 \in H$. To this end, we define functions $K_j : \mathbb{R}^d \to \mathbb{R}$, $j \in \{1, \ldots, m\}$, by (33), 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

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

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 $j \in J_T$,

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

with

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

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 $j \in \{1,\ldots,m\}$ 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 $A_T := \bigcup_{j \in J_T} A_j$, (32) and Theorem 17 yield

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

where $C_{\alpha,2}$ is a constant only depending on α, d, and $\|f^*_{L,\mathrm{P}}\|_{B^\alpha_{2,\infty}(\mathrm{P}_{X|A_T})}$. To utilize Theorem 5, it remains to examine the constants B, V, ϑ, and B_0. Since we consider the least squares loss, which can be clipped at M with $Y = [-M, M]$, the supremum bound (18) holds for $B = 4M^2$ and the variance bound (19) for $V = 16M^2$ and $\vartheta = 1$ (cf. Steinwart and Christmann, 2008a, Example 7.3). 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

$$|f_0(x)| = \left|\sum_{j\in J_T} \mathbb{1}_{A_j}(x)\cdot(K_j * f^*_{L,\mathrm{P}})(x)\right| \le \sum_{j\in J_T} \mathbb{1}_{A_j}(x)\left|K_j * f^*_{L,\mathrm{P}}(x)\right| \le (2^s-1)\left\|f^*_{L,\mathrm{P}}\right\|_{L_\infty(\mathbb{R}^d)}. \tag{34}$$

The latter implies

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

i.e. $B_0 := 4^s \max\{M^2, \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\}$. Moreover, since Theorem 6 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 $a_j = \tilde{c}_p \sqrt{\mathrm{P}_X(A_j)}\, r^{\frac{d+2p}{2p}} \gamma_j^{-\frac{d+2p}{2p}}$, we have

$$\left(\max\left\{c_p\sqrt{m}\left(\sum_{j=1}^m \lambda_j^{-p} a_j^{2p}\right)^{\frac{1}{2p}}, B\right\}\right)^{2p}$$

$$
\begin{aligned}
&= \left(\max\left\{ c_p \tilde{c}_p \sqrt{m} r^{\frac{d+2p}{2p}} \left(\sum_{j=1}^m \left(\lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^p \right)^{\frac{1}{2p}}, B \right\} \right)^{2p} \\
&\leq \left(\max\left\{ c_p \tilde{c}_p m^{\frac{1}{2p}} r^{\frac{d+2p}{2p}} \left(\sum_{j=1}^m \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^{\frac{1}{2}}, B \right\} \right)^{2p} \\
&\leq \left(\max\left\{ c_p \tilde{c}_p 8^{\frac{d}{2p}} r \left(\sum_{j=1}^m \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^{\frac{1}{2}}, B \right\} \right)^{2p} \\
&\leq C_p r^{2p} \left(\sum_{j=1}^m \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^p + B^{2p} \\
&=: a^{2p} \,,
\end{aligned}
$$

where we used the concavity of the function $t \mapsto t^p$ for $t \geq 0$, $mr^d \leq 8^d$ by (7), and $C_p := c_p^{2p} \tilde{c}_p^{2p} 8^d$. Finally, applying Theorem 5 yields

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

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

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

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

$$\begin{aligned}
C_{M,\alpha,p}:=\max\Bigg\{&9(2^s-1)^2\pi^{-\frac{d}{2}}\|f^*_{L,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)}\,,\,9\|f^*_{L,\mathrm{P}}\|^2_{B^\alpha_{2,\infty}(\mathrm{P}_{X\,|A_T})}\left(\frac{d}{2}\right)^\alpha\pi^{-\frac{1}{4}}\Gamma\left(2\alpha+\frac{1}{2}\right)^{\frac{1}{2}}\,,\\
&8^dCc_p^{2p}\tilde{c}_p^{2p}\,,\,16^pCM^{4p}+\left(3456M^2+15\cdot 4^s\max\{M^2,\|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\}\right)(1+\ln(3))\Bigg\}.
\end{aligned}$$

■

Next, using the just proven oracle inequality presented in Theorem 7, we show the learning rates of Theorem 8 in only a few steps.

Proof [Proof of Theorem 8] First of all, we define sequences $\tilde{\lambda}_n:=c_2n^{-1}$ and $\tilde{\gamma}_n:=c_3n^{-\frac{1}{2\alpha+d}}$ to simplify the presentation. Then Theorem 7, $\sum_{j=1}^{m_n}\mathrm{P}_X(A_j)=1$, and $|J_T|\leq m_n\leq 8^dr_n^{-d}$ 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

$$\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n})-\mathcal{R}^*_{L_{J_T},\mathrm{P}}\\
&\leq C_{M,\alpha,p}\left(\sum_{j\in J_T}\lambda_{n,j}\gamma_{n,j}^{-d}+\left(\frac{\max_{j\in J_T}\gamma_{n,j}}{\min_{j\in J_T}\gamma_{n,j}}\right)^d\max_{j\in J_T}\gamma_{n,j}^{2\alpha}+r_n^{2p}\left(\sum_{j=1}^{m_n}\lambda_{n,j}^{-1}\gamma_{n,j}^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p n^{-1}+\frac{\tau}{n}\right)\\
&=C_{M,\alpha,p}\left(|J_T|r_n^d\tilde{\lambda}_n\tilde{\gamma}_n^{-d}+\tilde{\gamma}_n^{2\alpha}+r_n^{(2-d)p}\tilde{\lambda}_n^{-p}\tilde{\gamma}_n^{-(d+2p)}\left(\sum_{j=1}^{m_n}\mathrm{P}_X(A_j)\right)^p n^{-1}+\tau n^{-1}\right)\\
&\leq 8^dC_{M,\alpha,p}\left(\tilde{\lambda}_n\tilde{\gamma}_n^{-d}+\tilde{\gamma}_n^{2\alpha}+\tilde{\lambda}_n^{-p}\tilde{\gamma}_n^{-(d+2p)}r_n^{(2-d)p}n^{-1}+\tau n^{-1}\right).
\end{aligned}$$

Using the choices $\tilde{\lambda}_n=c_2n^{-1}$, $\tilde{\gamma}_n=c_3n^{-\frac{1}{2\alpha+d}}$, as well as $r_n=c_1n^{-\frac{1}{\beta d}}$ finally implies

$$\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n})-\mathcal{R}^*_{L_{J_T},\mathrm{P}}\\
&\leq 8^dC_{M,\alpha,p}\left(\tilde{\lambda}_n\tilde{\gamma}_n^{-d}+\tilde{\gamma}_n^{2\alpha}+\tilde{\lambda}_n^{-p}\tilde{\gamma}_n^{-(d+2p)}r_n^{(2-d)p}n^{-1}+\tau n^{-1}\right)\\
&\leq \hat{C}_{M,\alpha,p}\left(n^{-1}n^{\frac{d}{2\alpha+d}}+n^{-\frac{2\alpha}{2\alpha+d}}+n^pn^{\frac{d+2p}{2\alpha+d}}n^{-\frac{(2-d)p}{\beta d}}n^{-1}+\tau n^{-1}\right)\\
&=\hat{C}_{M,\alpha,p}\left(n^{-\frac{2\alpha}{2\alpha+d}}+n^{-\frac{2\alpha}{2\alpha+d}}+n^{-\frac{2\alpha}{2\alpha+d}+\left(1+\frac{2}{2\alpha+d}+\frac{1}{\beta}-\frac{2}{\beta d}\right)p}+\tau n^{-1}\right)
\end{aligned}$$

$$\leq C\tau n^{-\frac{2\alpha}{2\alpha+d}+\xi}$$

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 [Proof of Corollary 9] 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 7 and 8 immediately yield

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

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 7 and 8. ∎

Proof [Proof of Theorem 10] First of all, let us examine the approximation error $\mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}$ of the function $f_0 := \sum_{j=1}^m \mathbb{1}_{A_j}\cdot(K_j * f^*_{L,\mathrm{P}})$, where the functions $K_1,\ldots,K_m$ are defined by (33). We start with

$$\begin{aligned}
&\mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}} \\
&= \int_X \left(f_0(x) - f^*_{L,\mathrm{P}}(x)\right)^2 d\mathrm{P}_X(x) \\
&= \int_X \mathbb{1}_{A_\delta}(x)\left(f_0(x) - f^*_{L,\mathrm{P}}(x)\right)^2 d\mathrm{P}_X(x) + \int_X \mathbb{1}_{Z^{+\delta}}(x)\left(f_0(x) - f^*_{L,\mathrm{P}}(x)\right)^2 d\mathrm{P}_X(x) \\
&= \left\|f_0 - f^*_{L,\mathrm{P}}\right\|^2_{L_2(\mathrm{P}_{X|A_\delta})} + \int_X \mathbb{1}_{Z^{+\delta}}(x)\left(f_0(x) - f^*_{L,\mathrm{P}}(x)\right)^2 d\mathrm{P}_X(x)\,,
\end{aligned}$$

where (34) implies

$$\begin{aligned}
&\int_X \mathbb{1}_{Z^{+\delta}}(x)\left(f_0(x) - f^*_{L,\mathrm{P}}(x)\right)^2 d\mathrm{P}_X(x) \\
&\leq \int_X \mathbb{1}_{Z^{+\delta}}(x)\left(|f_0(x)| + |f^*_{L,\mathrm{P}}(x)|\right)^2 d\mathrm{P}_X(x)
\end{aligned}$$

$$\leq \int_X \mathbb{1}_{Z^{+\delta}}(x) \left((2^s-1)\|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)} + \|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)} \right)^2 d\mathrm{P}_X(x)$$
$$= 4^s \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)} \mathrm{P}_X\left(Z^{+\delta}\right) .$$

To esimate the approximation error on A_δ, we initially investigate $\left\| f_0 - f^*_{L,\mathrm{P}} \right\|^q_{L_q(\mathrm{P}_{X\,|\,A_\delta})}$ for arbitrary $q \geq 1$. To this end, we apply Theorem 17 for $\nu = \mathrm{P}_{X\,|\,A_\delta}$ and for the partition $(\hat{A}_j)_{j=1,\ldots,m}$, where $\hat{A}_j = A_j \cap A_\delta$ for every $j \in \{1,\ldots,m\}$. Since we actually have $\operatorname{supp}\nu = A_\delta = \bigcup_{j=1}^m (A_j \cap A_\delta)$ and since $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_{X\,|\,A_\delta})$ for $\alpha \geq 1$, we obtain

$$\left\| f_0 - f^*_{L,\mathrm{P}} \right\|^2_{L_2(\mathrm{P}_{X\,|\,A_\delta})} = \left\| \sum_{j=1}^m \mathbb{1}_{A_j \cap A_\delta} \cdot (K_j * f^*_{L,\mathrm{P}}) - f^*_{L,\mathrm{P}} \right\|^2_{L_2\left(\mathrm{P}_{X\,|\,A_\delta}\right)} \leq C_{\alpha,2} \left(\frac{\gamma_{\max}}{\gamma_{\min}} \right)^d \gamma^{2\alpha}_{\max} ,$$

where $C_{\alpha,2} := \|f^*_{L,\mathrm{P}}\|^2_{B^\alpha_{2,\infty}(\mathrm{P}_{X\,|\,A_\delta})} \left(\frac{d}{2}\right)^\alpha \pi^{-\frac{1}{4}} \Gamma\left(2\alpha + \frac{1}{2}\right)^{\frac{1}{2}}$. Altogether, we can finally bound the approximation error by

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}} &= \left\| f_0 - f^*_{L,\mathrm{P}} \right\|^2_{L_2(\mathrm{P}_{X\,|\,A_\delta})} + \int_X \mathbb{1}_{Z^{+\delta}}(x) \left(f_0(x) - f^*_{L,\mathrm{P}}(x) \right)^2 d\mathrm{P}_X(x) \\ &\leq C_{\alpha,2} \left(\frac{\gamma_{\max}}{\gamma_{\min}} \right)^d \gamma^{2\alpha}_{\max} + 4^s \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)} \mathrm{P}_X\left(Z^{+\delta}\right) .\end{aligned}$$

Analogously to the proof of Theorem 7, applying Theorem 5 yields

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

with probability P^n not less than $1-e^{-\tau}$ and

$$\begin{aligned}
\widehat{C} &:= \max\Bigg\{ 9(2^s-1)^2\pi^{-\frac{d}{2}}\|f^*_{L,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)}\,,\; 9\|f^*_{L,\mathrm{P}}\|^2_{B^\alpha_{2,\infty}(\mathrm{P}_{X\,|\,A_\delta})}\left(\frac{d}{2}\right)^\alpha \pi^{-\frac{1}{4}}\Gamma\left(2\alpha+\frac{1}{2}\right)^{\frac{1}{2}},\; CC_p,\\
&\qquad 16^p CM^{4p} + \Big(3456M^2 + 15\cdot 4^s \max\{M^2, \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\}\Big)(1+\ln 3), 9\cdot 4^s\|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\Bigg\}\\
&\leq \widetilde{C}\max\left\{\|f^*_{L,\mathrm{P}}\|^2_{B^\alpha_{2,\infty}(\mathrm{P}_{X\,|\,A_\delta})}\,,\,1\right\},
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{C} &:= \max\Bigg\{ 9(2^s-1)^2\pi^{-\frac{d}{2}}\|f^*_{L,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)}\,,\; 9\left(\frac{d}{2}\right)^\alpha \pi^{-\frac{1}{4}}\Gamma\left(2\alpha+\frac{1}{2}\right)^{\frac{1}{2}},\; CC_p\,,\\
&\qquad 16^p CM^{4p} + \Big(3456M^2 + 15\cdot 4^s \max\{M^2, \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\}\Big)(1+\ln 3),\; 9\cdot 4^s\|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)}\Bigg\}.
\end{aligned}$$

■

Proof [Proof of Corollary 11] Using the sequences $r_n = c_1 n^{-\frac{1}{\beta d}}$, $\lambda_{n,j} = c_2 r_n^d n^{-1}$ as well as $\gamma_{n,j} = c_3 n^{-\frac{1}{2\alpha+d}}$ for every $j\in\{1,\ldots,m_n\}$ and $\delta_n = c_4 n^{-\frac{2\alpha}{2\alpha+d}}$, we deduce with probability P^n not less than $1-e^{-\tau}$

$$\begin{aligned}
&\sum_{j=1}^{m_n}\lambda_{n,j}\|f_{\mathrm{D}_j,\lambda_{n,j},\gamma_{n,j}}\|^2_{\mathring{H}_{\gamma_{n,j}}(A_j)} + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L,\mathrm{P}}\\
&\leq \widehat{C}\left(\sum_{j=1}^{m_n}\lambda_{n,j}\gamma_{n,j}^{-d} + \left(\frac{\gamma_{\max}}{\gamma_{\min}}\right)^d \gamma_{\max}^{2\alpha} + r_n^{2p}\left(\sum_{j=1}^{m_n}\lambda_{n,j}^{-1}\gamma_{n,j}^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p n^{-1} + \frac{\tau}{n} + \mathrm{P}_X\left(Z^{+\delta_n}\right)\right)\\
&\leq \widehat{C}\left(c_2c_3^{-d}m_n r_n^d n^{-1} n^{\frac{d}{2\alpha+d}} + c_3^{2\alpha}n^{-\frac{2\alpha}{2\alpha+d}} + c_2^{-p}c_3^{-(d+2p)}r_n^{(2-d)p}n^p n^{\frac{d+2p}{2\alpha+d}}n^{-1} + \tau n^{-1} + c\delta_n\right)\\
&\leq \widehat{C}\left(8^d c_2 c_3^{-d} n^{-\frac{2\alpha}{2\alpha+d}} + c_3^{2\alpha}n^{-\frac{2\alpha}{2\alpha+d}} + c_2^{-p}c_3^{-(d+2p)}r_n^{(2-d)p}n^{-\frac{2\alpha}{2\alpha+d}+\frac{2\alpha+d+2}{2\alpha+d}p} + \tau n^{-1} + c\delta_n\right)\\
&\leq \widehat{C}\Big(8^d c_2 c_3^{-d} n^{-\frac{2\alpha}{2\alpha+d}} + c_3^{2\alpha}n^{-\frac{2\alpha}{2\alpha+d}} + c_1^{(2-d)p}c_2^{-p}c_3^{-(d+2p)}n^{-\frac{2-d}{\beta d}p}n^{-\frac{2\alpha}{2\alpha+d}+\frac{2\alpha+d+2}{2\alpha+d}p}\\
&\quad + \tau n^{-1} + cc_4 n^{-\frac{2\alpha}{2\alpha+d}}\Big)\\
&\leq \widehat{C}\Bigg(8^d c_2 c_3^{-d} n^{-\frac{2\alpha}{2\alpha+d}} + c_3^{2\alpha}n^{-\frac{2\alpha}{2\alpha+d}} + c_1^{(2-d)p}c_2^{-p}c_3^{-(d+2p)}n^{-\frac{2\alpha}{2\alpha+d}+\left(1+\frac{2}{2\alpha+d}+\frac{1}{\beta}-\frac{2}{\beta d}\right)p}\\
&\quad + \tau n^{-1} + cc_4 n^{-\frac{2\alpha}{2\alpha+d}}\Bigg)\\
&\leq \widetilde{C}\tau n^{-\frac{2\alpha}{2\alpha+d}+\xi},
\end{aligned}$$

where $\widetilde{C} := \widehat{C}\left(8^d c_2 c_3^{-d} + c_3^{2\alpha} + \max\{c_1^{2p},1\}c_1^{-dp}c_2^{-p}c_3^{-(d+2p)} + 1 + cc_4\right)$ and $\xi \geq \left(1+\frac{2}{2\alpha+d} + \frac{1}{\beta} - \frac{2}{\beta d}\right)p > 0$.

■

It remains to prove Theorem 12. However, we previously have to consider the following technical lemma.

Lemma 18 *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 $|J| \leq m_n \leq 8^d r_n^{-d}$. 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*

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

where $\xi := \left(\frac{2\alpha(2\alpha+d+2)}{(2\alpha+d)((2\alpha+d)(1+p)+2p)} + \max\left\{ \frac{d-2}{\beta d}, 0 \right\} \right) p$ and $C > 0$ 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 $\Lambda_n = \left\{ \lambda^{(1)}, \ldots, \lambda^{(u)} \right\}$ and $\Gamma_n = \left\{ \gamma^{(1)}, \ldots, \gamma^{(v)} \right\}$ 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

$$\lambda^{(i)} - \lambda^{(i-1)} \leq 2 r_n^d \varepsilon_n \qquad \text{and} \qquad \gamma^{(\ell)} - \gamma^{(\ell-1)} \leq 2\delta_n \tag{35}$$

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 (35), this yields

$$r_n^d \lambda^* \leq \lambda^{(i)} \leq r_n^d \lambda^* + 2 r_n^d \varepsilon_n \qquad \text{and} \qquad \gamma^* \leq \gamma^{(\ell)} \leq \gamma^* + 2\delta_n \,. \tag{36}$$

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

$$\lambda^* (\gamma^*)^{-d} + (\gamma^*)^{2\alpha} + (\lambda^*)^{-p} (\gamma^*)^{-(d+2p)} r_n^{(2-d)p} n^{-1} \leq \hat{c} n^{-\frac{2\alpha}{(2\alpha+d)(1+p)+2p} + \max\left\{ \frac{d-2}{\beta d}, 0 \right\} p}, \tag{37}$$

where $\hat{c}$ is a positive constant. Using (36), the bound $|J| \leq m_n \leq 8^d r_n^{-d}$, and (37), we obtain

$$\inf_{(\lambda_j,\gamma_j)_{j=1}^{m_n} \in (\Lambda_n \times \Gamma_n)^{m_n}} \left(\sum_{j \in J} \lambda_j \gamma_j^{-d} + \left(\frac{\max_{j \in J} \gamma_j}{\min_{j \in J} \gamma_j} \right)^d \max_{j \in J} \gamma_j^{2\alpha} + r_n^{2p} \left(\sum_{j=1}^{m_n} \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^p n^{-1} \right)$$
$$\leq \sum_{j \in J} \lambda^{(i)} \left(\gamma^{(\ell)} \right)^{-d} + \left(\gamma^{(\ell)} \right)^{2\alpha} + \left(\sum_{j=1}^{m_n} \left(\lambda^{(i)} \right)^{-1} \left(\gamma^{(\ell)} \right)^{-\frac{d+2p}{p}} \mathrm{P}_X(A_j) \right)^p r_n^{2p} n^{-1}$$

$$
\begin{aligned}
&\leq |J|\lambda^{(i)}\left(\gamma^{(\ell)}\right)^{-d} + \left(\gamma^{(\ell)}\right)^{2\alpha} + \left(\lambda^{(i)}\right)^{-p}\left(\gamma^{(\ell)}\right)^{-(d+2p)} r_n^{2p} n^{-1} \\
&\leq |J|\left(r_n^d\lambda^* + 2r_n^d\varepsilon_n\right)(\gamma^*)^{-d} + (\gamma^* + 2\delta_n)^{2\alpha} + \left(r_n^d\lambda^*\right)^{-p}(\gamma^*)^{-(d+2p)} r_n^{2p} n^{-1} \\
&\leq 8^d \cdot 3\lambda^*(\gamma^*)^{-d} + (\gamma^* + 2\delta_n)^{2\alpha} + (\lambda^*)^{-p}(\gamma^*)^{-(d+2p)} r_n^{(2-d)p} n^{-1} \\
&\leq \tilde{c}\left(\lambda^*(\gamma^*)^{-d} + (\gamma^*)^{2\alpha} + (\lambda^*)^{-p}(\gamma^*)^{-(d+2p)} r_n^{(2-d)p} n^{-1}\right) + \tilde{c}\delta_n^{2\alpha} \\
&\leq \tilde{c}\hat{c} n^{-\frac{2\alpha}{(2\alpha+d)(1+p)+2p} + \max\left\{\frac{d-2}{\beta d}, 0\right\}p} + \tilde{c}\delta_n^{2\alpha} \\
&\leq C\left(n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \delta_n^{2\alpha}\right)
\end{aligned}
$$

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 12 using Theorem 7 as well as Lemma 18.

Proof [Proof of Theorem 12] Let l be defined by $l := \left\lfloor \frac{n}{2} \right\rfloor + 1$, i.e. $l \geq \frac{n}{2}$. With this, Theorem 7 yields with probability P^l not less than $1 - |\Lambda_n \times \Gamma_n|^{m_n} e^{-\tau}$ that

$$
\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&\leq \frac{c_1}{2}\left(\sum_{j\in J_T}\lambda_j\gamma_j^{-d} + \left(\frac{\max_{j\in J_T}\gamma_j}{\min_{j\in J_T}\gamma_j}\right)^d \max_{j\in J_T}\gamma_j^{2\alpha} + r_n^{2p}\left(\sum_{j=1}^{m_n}\lambda_j^{-1}\gamma_j^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p l^{-1} + \tau l^{-1}\right) \\
&\leq c_1\left(\sum_{j\in J_T}\lambda_j\gamma_j^{-d} + \left(\frac{\max_{j\in J_T}\gamma_j}{\min_{j\in J_T}\gamma_j}\right)^d \max_{j\in J_T}\gamma_j^{2\alpha} + r_n^{2p}\left(\sum_{j=1}^{m_n}\lambda_j^{-1}\gamma_j^{-\frac{d+2p}{p}}\mathrm{P}_X(A_j)\right)^p n^{-1} + \tau n^{-1}\right)
\end{aligned} \tag{38}
$$

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, 2008a, 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

$$
\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda}_{\mathrm{D}_2},\boldsymbol{\gamma}_{\mathrm{D}_2}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \qquad (39) \\
&< 6\left(\inf_{(\lambda_j,\gamma_j)_{j=1}^{m_n}\in(\Lambda_n\times\Gamma_n)^{m_n}} \mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}}\right) + 512M^2\frac{\tau_n}{n-l} \\
&< 6\left(\inf_{(\lambda_j,\gamma_j)_{j=1}^{m_n}\in(\Lambda_n\times\Gamma_n)^{m_n}} \mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}}\right) + 2048M^2\frac{\tau_n}{n} \qquad (40)
\end{aligned}
$$

with probability P^{n-l} not less than $1 - e^{-\tau}$. With (38), (40) and Lemma 18 we can conclude

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

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

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

$$
\begin{aligned}
&\mathcal{R}_{L_{J_T},\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\boldsymbol{\lambda}_{\mathrm{D}_2},\boldsymbol{\gamma}_{\mathrm{D}_2}}) - \mathcal{R}^*_{L_{J_T},\mathrm{P}} \\
&< 12 c_1 C n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \left(6c_1 \left(\tau + \ln\left(1 + |\Lambda_n \times \Gamma_n|^{m_n}\right)\right) \right. \\
&\quad \left. + 2048 M^2 \left(\tau + 2 \ln\left(1 + |\Lambda_n \times \Gamma_n|^{m_n}\right)\right) \right) n^{-1} \\
&\leq 12 c_1 C n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \left(6c_1 + 2048 M^2 \right) \left(\tau + 2 m_n \ln\left(1 + |\Lambda_n \times \Gamma_n|\right)\right) n^{-1} \\
&\leq 12 c_1 C n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \left(6c_1 + 2048 M^2 \right) \left(\tau + 2 \cdot 8^d r_n^{-d} \ln\left(1 + |\Lambda_n \times \Gamma_n|\right)\right) n^{-1} \\
&= 12 c_1 C n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \left(6c_1 + 2048 M^2 \right) \left(\tau n^{-1} + 2 \cdot 8^d c^{-d} \ln\left(1 + |\Lambda_n \times \Gamma_n|\right) n^{-\frac{\beta-1}{\beta}}\right) \\
&< \left(12 c_1 C + \left(6c_1 + 2048 M^2 \right) \left(\tau + 2 \cdot 8^d c^{-d} \ln\left(1 + |\Lambda_n \times \Gamma_n|\right)\right)\right) n^{-\frac{2\alpha}{2\alpha+d}+\xi}
\end{aligned}
$$

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.

■

Proof [Proof of Theorem 13] The assertion follows analogously to the proof of Theorem 12 using the oracle inequality of Theorem 10. ■

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 *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 3, pages 2396–2401 vol.3, 1998. doi: 10.1109/IJCNN.1998.687237.

H. Berens and R. DeVore. Quantitative Korovkin Theorems for Positive Linear Operators on L_p- Spaces. *Trans. Amer. Math. Soc.*, 245:pp. 349–361, 1978. ISSN 00029947. URL http://www.jstor.org/stable/1998871.

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

E. Blanzieri and F. Melgani. Nearest Neighbor Classification of Remote Sensing Images With the Maximal Margin Principle. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(6):1804–1811, 2008. ISSN 0196-2892. doi: 10.1109/TGRS.2008.916090.

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

Enrico Blanzieri and Anton Bryl. Evaluation of the Highest Probability SVM Nearest Neighbor Classifier with Variable Relative Error Cost. In *In Proceedings of Fourth Conference on Email and Anti-Spam, CEAS'2007*, 2007b.

Léon Bottou and Vladimir 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.

B. Carl and I. Stephani. *Entropy, Compactness, and the Approximation of Operators*. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1990. ISBN 9780521330114. URL http://books.google.de/books?id=xlBNHvgP7PIC.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Fu Chang, Chien-Yang Guo, Xiao-Rong Lin, and Chi-Jen Lu. Tree Decomposition for Large-Scale SVM Problems. *J. Mach. Learn. Res.*, 11:2935–2972, December 2010. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1756006.1953027.

Haibin Cheng, Pang-Ning Tan, and Rong Jin. Localized Support Vector Machine and Its Efficient Algorithm. In *SIAM International Conference on Data Mining*, 2007.

Haibin Cheng, Pang-Ning Tan, and Rong Jin. Efficient Algorithm for Localized Support Vector Machine. *Knowledge and Data Engineering, IEEE Transactions on*, 22(4):537–549, 2010. ISSN 1041-4347. doi: 10.1109/TKDE.2009.116.

Ronan Collobert, Samy Bengio, and Yoshua Bengio. A Parallel Mixture of SVMs for Very Large Scale Problems. In *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.

Sanjoy Dasgupta. Lecture 1: Clustering in metric spaces. CSE 291: Topics in unsupervised learning, 2008. URL http://cseweb.ucsd.edu/~dasgupta/291/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 V.A. Popov. Interpolation of Besov Spaces. *Trans. Amer. Math. Soc.*, 305(1):pp. 397–414, 1988.

Ronald Alvin DeVore and George G. Lorentz. *Constructive Approximation*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1993. ISBN 3-540-50627-6. URL http://books.google.de/books?id=cDqNW6k7_ZwC.

Mona Eberts and Ingo 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.

Mona Eberts and Ingo Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Statist.*, 7:1–42, 2013. ISSN 1935-7524. doi: 10.1214/12-EJS760.

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

H. Federer. *Geometric Measure Theory*. Springer, Berlin, 1969.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38(0):293–306, 1985. ISSN 0304-3975. doi: http://dx.doi.org/10.1016/0304-3975(85)90224-5. URL http://www.sciencedirect.com/science/article/pii/0304397585902245.

Hans Peter Graf, Eric Cosatto, Leon Bottou, Igor Durdanovic, and Vladimir Vapnik. Parallel Support Vector Machines: The Cascade SVM. In *In Advances in Neural Information Processing Systems*, pages 521–528. MIT Press, 2005.

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

Robert Hable. Universal consistency of localized versions of regularized kernel methods. *J. Mach. Learn. Res.*, 14(1):153–186, January 2013. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2502581.2502586.

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

R. Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 1st edition, 2005. ISBN 9780521608275. URL http://books.google.de/books?id=hdkt7V4NXsgC.

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

Nicola Segata and Enrico 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.

Nicola Segata and Enrico Blanzieri. Fast and Scalable Local Kernel Machines. *J. Mach. Learn. Res.*, 11:1883–1926, August 2010. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1756006.1859915.

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

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

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

I. Steinwart and A. Christmann. How SVMs can estimate quantiles and the median. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 305–312. MIT Press, Cambridge, MA, 2008b. URL ./2008/StCh08c.pdf.

I. Steinwart and A. Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17:211–225, 2011. URL ./2011/StCh11a.pdf.

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. URL ./2009/StHuSc09a.pdf.

I. Steinwart, D. Hush, and C. Scovel. Training SVMs without offset. *J. Mach. Learn. Res.*, 12:141–202, 2011. URL ./2011/StHuSc11a.pdf.

Taiji 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.

Hans Triebel. *Theory of Function Spaces*. Birkhäuser, Basel, repr. of the 1983 edition, 2010. ISBN 978-3-0346-0415-4.

Ivor W. Tsang, James T. Kwok, and Pak-Ming Cheung. Core Vector Machines: Fast SVM Training on Very Large Data Sets. *J. Mach. Learn. Res.*, 6:363–392, December 2005. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1046920.1058114.

Ivor W. Tsang, Andras Kocsor, and James T. Kwok. Simpler core vector machines with enclosing balls. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 911–918, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273611. URL http://doi.acm.org/10.1145/1273496.1273611.

Vladimir Vapnik and Léon Bottou. Local Algorithms for Pattern Recognition and Dependencies Estimation. *Neural Computation*, 5:893–909, 1993. doi: 10.1162/neco.1993.5.6.893.

Donghui Wu, Kristin P. Bennett, Nello Cristianini, and John Shawe-Taylor. Large Margin Trees for Induction and Transduction. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 474–483, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-612-2. URL `http://dl.acm.org/citation.cfm?id=645528.657622`.

Alon Zakai and Ya'acov Ritov. Consistency and Localizability. *J. Mach. Learn. Res.*, 10:827–856, June 2009. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1577069.1577099`.

Hao Zhang, A.C. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2126–2136, 2006. doi: 10.1109/CVPR.2006.301.

Appendix A

For the sake of completeness, we present in the following some tables containing the computational results achieved by the LS-, VP-, and RC-SVMs for all real and artificial data set types. Here, the training and test times, given in seconds, are averaged over all successful runs. Moreover, for the test and L_2-errors, we also stated the mean of all runs plus/minus the standard deviation. The same is true for the number of working sets (# of ws), except for the LS-SVMs, where we always have one working set by construction. The last two columns contain median, minimum, and maximum of the working set sizes appearing during the various runs.

	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
LS-SVM	1 000 (10 000)	100	1.38	0.48	0.7142 ±0.0097	1.00		
	2 500 (10 000)	100	6.42	0.99	0.6323 ±0.0074	1.00		
	5 000 (10 000)	100	30.19	1.55	0.5707 ±0.0070	1.00		
	10 000 (50 000)	10	138.01	12.14	0.4909 ±0.0089	1.00		
	25 000 (50 000)	10	922.90	34.19	0.3816 ±0.0042	1.00		
	50 000 (50 000)	3	3788	176.68	0.3117 ±0.0012	1.00		
	100 000 (50 000)	3	16 353	507.38	0.2417 ±0.0102	1.00		

	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
VP-SVM radius = 2	1 000 (10 000)	100	5.85	0.73	0.7521 ±0.0108	83.73 ±1.86	6	[1 , 156]
	2 500 (10 000)	100	7.44	1.15	0.6602 ±0.0082	101.92 ±2.04	10	[1 , 481]
	5 000 (10 000)	100	9.37	1.83	0.6011 ±0.0079	106.72 ±1.95	24	[1 , 987]
	10 000 (50 000)	10	13.45	13.68	0.5238 ±0.0076	120.20 ±1.99	40	[1 , 952]
	25 000 (50 000)	10	26.73	31.07	0.4134 ±0.0057	131.00 ±2.26	80	[1 , 3204]
	50 000 (50 000)	3	57.41	183.71	0.3449 ±0.0047	139.67 ±4.16	155	[3 , 4334]
	100 000 (50 000)	3	171.12	493.44	0.2658 ±0.0028	154.33 ±5.51	271	[3 , 8121]
	250 000 (50 000)	3	1128	1169	0.1924 ±0.0016	166.33 ±3.79	533	[3 , 22633]
	500 000 (50 000)	3	5349	3020	0.1608 ±0.0024	178.33 ±0.58	987	[2 , 44585]
VP-SVM radius = 3	1 000 (10 000)	100	2.79	0.59	0.7262 ±0.0092	37.39 ±1.85	15	[1 , 193]
	2 500 (10 000)	100	3.97	1.09	0.6379 ±0.0072	44.40 ±1.84	25	[1 , 493]
	5 000 (10 000)	100	5.58	1.74	0.5845 ±0.0062	47.59 ±1.66	50	[1 , 1008]
	10 000 (50 000)	10	10.64	13.43	0.5075 ±0.0042	51.20 ±1.62	85	[1 , 1918]
	25 000 (50 000)	10	31.83	32.38	0.4026 ±0.0044	58.60 ±1.65	142	[3 , 4971]
	50 000 (50 000)	3	120.10	208.34	0.3256 ±0.0021	61.33 ±2.08	274	[4 , 10005]
	100 000 (50 000)	3	444.39	523.39	0.2539 ±0.0034	64.67 ±1.53	484	[18 , 20082]
	250 000 (50 000)	3	2825	1274	0.1859 ±0.0018	68.33 ±1.15	1169	[21 , 49558]
	500 000 (50 000)	3	12 882	3786	0.1540 ±0.0010	72.00 ±1.00	2366	[54 , 74787]
VP-SVM radius = 4	1 000 (10 000)	100	0.87	0.53	0.7095 ±0.0102	4.40 ±0.49	165	[42 , 489]
	2 500 (10 000)	100	2.14	1.05	0.6317 ±0.0062	5.02 ±0.49	382	[83 , 1248]
	5 000 (10 000)	100	7.41	1.74	0.5737 ±0.0073	4.94 ±0.55	723	[229 , 2401]
	10 000 (50 000)	10	28.58	13.62	0.5007 ±0.0070	5.20 ±0.42	1518	[488 , 4514]
	25 000 (50 000)	10	201.17	33.16	0.3937 ±0.0049	5.70 ±0.67	2533	[566 , 11113]
	50 000 (50 000)	3	827.41	218.85	0.3225 ±0.0028	5.33 ±0.58	7530	[2870 , 24509]
	100 000 (50 000)	3	2823	548.96	0.2418 ±0.0007	7.00 ±0.00	6356	[2621 , 40018]
	250 000 (50 000)	3	20 010	1434	0.1689 ±0.0041	6.33 ±0.58	17 721	[6061 , 110464]
	500 000 (50 000)	0 (3)	NA	NA	NA ±NA	7.33 ±1.53	68 182	[12469 , 233675]
VP-SVM radius = 5	1 000 (10 000)	100	1.33	0.48	0.7138 ±0.0100	1.06 ±0.24	1000	[208 , 1000]
	2 500 (10 000)	100	5.97	1.01	0.6326 ±0.0080	1.16 ±0.37	2500	[375 , 2500]
	5 000 (10 000)	100	29.20	1.58	0.5705 ±0.0071	1.06 ±0.24	5000	[1810 , 5000]
	10 000 (50 000)	10	131.83	12.48	0.4895 ±0.0066	1.10 ±0.32	10 000	[4338 , 10000]
	25 000 (50 000)	10	832.08	33.09	0.3830 ±0.0045	1.20 ±0.42	25 000	[8524 , 25000]
	50 000 (50 000)	3	3182	185.13	0.3151 ±0.0066	1.33 ±0.58	40 062	[19875 , 50000]
	100 000 (50 000)	3	10 472	527.18	0.2427 ±0.0030	1.67 ±0.58	55 873	[42218 , 100000]
	250 000 (50 000)	1 (3)	34 449	1445	0.1650 ±0.0000	3.00 ±0.00	86 655	[46661 , 116684]
	500 000 (50 000)	0 (3)	NA	NA	NA ±NA	1.33 ±0.58	375 000	[186510 , 500000]

Table 3: LS- and VP-SVM results relating to the COVTYPE data sets

	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
RC-SVM # of ws = 1	1 000 (10 000)	100	1.76	0.30	0.7159 ±0.0101	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	22.55	2.19	0.6324 ±0.0073	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	97.30	9.16	0.5707 ±0.0070	1.00 ±0.00	5000	{5000}
	10 000 (50 000)	10	741.46	111.75	0.4909 ±0.0089	1.00 ±0.00	10 000	{10000}
	25 000 (50 000)	10	4629	291.75	0.3816 ±0.0042	1.00 ±0.00	25 000	{25000}
	50 000 (50 000)	3	11 416	521.84	0.3117 ±0.0012	1.00 ±0.00	50 000	{50000}
	100 000 (50 000)	3	19 921	622.99	0.2417 ±0.0102	1.00 ±0.00	100 000	{100000}
	250 000 (50 000)	0	NA	NA	NA ±NA	NA ±NA	NA	{NA}
	500 000 (50 000)	0	NA	NA	NA ±NA	NA ±NA	NA	{NA}
RC-SVM # of ws = 4	1 000 (10 000)	100	0.59	0.29	0.7527 ±0.0143	4.00 ±0.00	250	{250}
	2 500 (10 000)	100	4.76	2.56	0.6676 ±0.0077	4.00 ±0.00	625	{625}
	5 000 (10 000)	100	25.57	11.60	0.6254 ±0.0074	4.00 ±0.00	1250	{1250}
	10 000 (50 000)	10	89.55	83.44	0.5649 ±0.0047	4.00 ±0.00	2500	{2500}
	25 000 (50 000)	10	905.89	243.32	0.4746 ±0.0062	4.00 ±0.00	6250	{6250}
	50 000 (50 000)	3	2160	529.96	0.3990 ±0.0032	4.00 ±0.00	12 500	{12500}
	100 000 (50 000)	3	4191	826.40	0.3224 ±0.0016	4.00 ±0.00	25 000	{25000}
	250 000 (50 000)	3	29 909	1677	0.2251 ±0.0012	4.00 ±0.00	62 500	{62500}
	500 000 (50 000)	0 (3)	NA	NA	NA ±NA	4.00 ±0.00	125 000	{125000}
RC-SVM # of ws = 5	1 000 (10 000)	100	0.69	0.30	0.7674 ±0.0142	5.00 ±0.00	200	{200}
	2 500 (10 000)	100	3.59	2.97	0.6798 ±0.0092	5.00 ±0.00	500	{500}
	5 000 (10 000)	100	18.79	14.25	0.6404 ±0.0082	5.00 ±0.00	1000	{1000}
	10 000 (50 000)	10	72.89	112.97	0.5790 ±0.0070	5.00 ±0.00	2000	{2000}
	25 000 (50 000)	10	659.31	297.07	0.4886 ±0.0048	5.00 ±0.00	5000	{5000}
	50 000 (50 000)	3	1620	470.55	0.4216 ±0.0024	5.00 ±0.00	10 000	{10000}
	100 000 (50 000)	3	5441	1298	0.3420 ±0.0040	5.00 ±0.00	20 000	{20000}
	250 000 (50 000)	3	20 828	1614	0.2395 ±0.0015	5.00 ±0.00	50 000	{50000}
	500 000 (50 000)	3	81 660	2794	0.1931 ±0.0007	5.00 ±0.00	100 000	{100000}
RC-SVM # of ws = 6	1 000 (10 000)	100	0.73	0.29	0.7776 ±0.0138	6.00 ±0.00	167	[166 , 167]
	2 500 (10 000)	100	3.49	3.27	0.6911 ±0.0101	6.00 ±0.00	417	[416 , 417]
	5 000 (10 000)	100	15.61	16.57	0.6519 ±0.0089	6.00 ±0.00	833	[833 , 834]
	10 000 (50 000)	10	50.90	72.44	0.5937 ±0.0060	6.00 ±0.00	1667	[1666 , 1667]
	25 000 (50 000)	10	230.76	259.58	0.5035 ±0.0055	6.00 ±0.00	4167	[4166 , 4167]
	50 000 (50 000)	3	1406	425.29	0.4355 ±0.0028	6.00 ±0.00	8333	[8333 , 8334]
	100 000 (50 000)	3	5139	1203	0.3579 ±0.0047	6.00 ±0.00	16 667	[16666 , 16667]
	250 000 (50 000)	3	17 099	1577	0.2542 ±0.0027	6.00 ±0.00	41 667	[41666 , 41667]
	500 000 (50 000)	3	66 335	2534	0.2013 ±0.0004	6.00 ±0.00	83 333	[83333 , 83334]
RC-SVM # of ws = 7	1 000 (10 000)	100	0.72	0.30	0.7843 ±0.0146	7.00 ±0.00	143	[142 , 143]
	2 500 (10 000)	100	3.76	6.22	0.6991 ±0.0090	7.00 ±0.00	357	[357 , 358]
	5 000 (10 000)	100	9.51	10.49	0.6608 ±0.0072	7.00 ±0.00	714	[714 , 715]
	10 000 (50 000)	10	51.30	75.81	0.6045 ±0.0078	7.00 ±0.00	1429	[1428 , 1429]
	25 000 (50 000)	10	258.70	255.08	0.5163 ±0.0057	7.00 ±0.00	3571	[3571 , 3572]
	50 000 (50 000)	3	1087	440.26	0.4463 ±0.0026	7.00 ±0.00	7143	[7142 , 7143]
	100 000 (50 000)	3	3174	861.05	0.3765 ±0.0033	7.00 ±0.00	14 286	[14285 , 14286]
	250 000 (50 000)	3	15 472	1480	0.2638 ±0.0012	7.00 ±0.00	35 714	[35714 , 35715]
	500 000 (50 000)	3	56 601	2628	0.2094 ±0.0013	7.00 ±0.00	71 429	[71428 , 71429]

	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
RC-SVM # of ws = 10	1 000 (10 000)	100	0.99	0.32	0.8134 ±0.0133	10.00 ±0.00	100	{100}
	2 500 (10 000)	100	2.73	5.29	0.7188 ±0.0090	10.00 ±0.00	250	{250}
	5 000 (10 000)	100	6.61	15.24	0.6825 ±0.0074	10.00 ±0.00	500	{500}
	10 000 (50 000)	10	31.94	69.23	0.6278 ±0.0079	10.00 ±0.00	1000	{1000}
	25 000 (50 000)	10	259.18	304.44	0.5444 ±0.0041	10.00 ±0.00	2500	{2500}
	50 000 (50 000)	3	867.90	645.72	0.4904 ±0.0044	10.00 ±0.00	5000	{5000}
	100 000 (50 000)	3	2065	1204	0.4120 ±0.0076	10.00 ±0.00	10 000	{10000}
	250 000 (50 000)	3	9534	1398	0.3039 ±0.0017	10.00 ±0.00	25 000	{25000}
	500 000 (50 000)	3	43 146	2689	0.2300 ±0.0008	10.00 ±0.00	50 000	{50000}
RC-SVM # of ws = 20	1 000 (10 000)	100	1.67	0.33	0.8646 ±0.0082	20.00 ±0.00	50	{50}
	2 500 (10 000)	100	3.16	4.43	0.7574 ±0.0076	20.00 ±0.00	125	{125}
	5 000 (10 000)	100	3.99	13.46	0.7172 ±0.0069	20.00 ±0.00	250	{250}
	10 000 (50 000)	10	22.55	170.42	0.6770 ±0.0068	20.00 ±0.00	500	{500}
	25 000 (50 000)	10	87.06	236.67	0.5958 ±0.0054	20.00 ±0.00	1250	{1250}
	50 000 (50 000)	3	150.59	269.68	0.5470 ±0.0029	20.00 ±0.00	2500	{2500}
	100 000 (50 000)	3	885.28	685.50	0.4760 ±0.0028	20.00 ±0.00	5000	{5000}
	250 000 (50 000)	3	4559	1634	0.3752 ±0.0057	20.00 ±0.00	12 500	{12500}
	500 000 (50 000)	3	19 033	2802	0.2971 ±0.0008	20.00 ±0.00	25 000	{25000}
RC-SVM # of ws = 50	1 000 (10 000)	100	3.75	0.38	0.9280 ±0.0046	50.00 ±0.00	20	{20}
	2 500 (10 000)	100	6.58	5.52	0.8220 ±0.0073	50.00 ±0.00	50	{50}
	5 000 (10 000)	100	6.04	15.28	0.7679 ±0.0081	50.00 ±0.00	100	{100}
	10 000 (50 000)	10	11.50	124.14	0.7245 ±0.0058	50.00 ±0.00	200	{200}
	25 000 (50 000)	10	19.95	277.44	0.6551 ±0.0056	50.00 ±0.00	500	{500}
	50 000 (50 000)	3	54.19	221.31	0.6101 ±0.0032	50.00 ±0.00	1000	{1000}
	100 000 (50 000)	3	319.38	1018	0.5600 ±0.0023	50.00 ±0.00	2000	{2000}
	250 000 (50 000)	3	2680	1659	0.4688 ±0.0037	50.00 ±0.00	5000	{5000}
	500 000 (50 000)	3	6984	2352	0.4011 ±0.0022	50.00 ±0.00	10 000	{10000}
RC-SVM # of ws = 100	1 000 (10 000)	100	7.14	0.42	0.9562 ±0.0035	100.00 ±0.00	10	{10}
	2 500 (10 000)	100	12.05	5.85	0.8752 ±0.0049	100.00 ±0.00	25	{25}
	5 000 (10 000)	100	10.70	16.60	0.8239 ±0.0061	100.00 ±0.00	50	{50}
	10 000 (50 000)	10	12.05	92.09	0.7593 ±0.0069	100.00 ±0.00	100	{100}
	25 000 (50 000)	10	21.60	276.46	0.6926 ±0.0037	100.00 ±0.00	250	{250}
	50 000 (50 000)	3	32.39	206.27	0.6493 ±0.0033	100.00 ±0.00	500	{500}
	100 000 (50 000)	3	141.39	798.77	0.6082 ±0.0022	100.00 ±0.00	1000	{1000}
	250 000 (50 000)	3	1325	1970	0.5306 ±0.0029	100.00 ±0.00	2500	{2500}
	500 000 (50 000)	3	3053	2315	0.4670 ±0.0025	100.00 ±0.00	5000	{5000}
RC-SVM # of ws = 150	1 000 (10 000)	100	11.98	0.63	0.9650 ±0.0023	150.00 ±0.00	7	[6 , 7]
	2 500 (10 000)	100	15.21	4.66	0.9086 ±0.0040	150.00 ±0.00	17	[16 , 17]
	5 000 (10 000)	100	16.94	14.18	0.8610 ±0.0053	150.00 ±0.00	33	[33 , 34]
	10 000 (50 000)	10	18.06	142.93	0.7936 ±0.0068	150.00 ±0.00	67	[66 , 67]
	25 000 (50 000)	10	16.47	183.81	0.7141 ±0.0039	150.00 ±0.00	167	[166 , 167]
	50 000 (50 000)	3	26.25	187.73	0.6742 ±0.0008	150.00 ±0.00	333	[333 , 334]
	100 000 (50 000)	3	100.66	755.96	0.6337 ±0.0020	150.00 ±0.00	667	[666 , 667]
	250 000 (50 000)	3	537.02	1455	0.5658 ±0.0004	150.00 ±0.00	1667	[1666 , 1667]
	500 000 (50 000)	3	1859	2336	0.5081 ±0.0019	150.00 ±0.00	3333	[3333 , 3334]

Table 4: RC-SVM results relating to the COVTYPE data sets

	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
LS-SVM	1 000 (10 000)	100	1.21	0.25	0.1916 ±0.0055	1.00		
	2 500 (10 000)	100	8.25	0.57	0.1842 ±0.0043	1.00		
	5 000 (10 000)	100	33.98	1.03	0.1672 ±0.0032	1.00		
	10 000 (50 000)	10	129.26	11.76	0.1596 ±0.0032	1.00		
	25 000 (50 000)	10	791.14	52.75	0.1453 ±0.0020	1.00		
	50 000 (50 000)	3	3029	169.78	0.1410 ±0.0022	1.00		
	100 000 (50 000)	3	11 078	201.54	0.1302 ±0.0008	1.00		

	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
VP-SVM radius = 1	1 000 (10 000)	100	1.81	0.46	0.2586 ±0.0090	21.62 ±1.23	15	[1 , 469]
	2 500 (10 000)	100	2.90	0.90	0.2243 ±0.0058	25.03 ±1.58	28	[1 , 1289]
	5 000 (10 000)	100	5.19	1.34	0.1891 ±0.0036	26.50 ±1.63	62	[2 , 2520]
	10 000 (50 000)	10	12.27	9.21	0.1724 ±0.0033	28.50 ±1.51	84	[1 , 5018]
	25 000 (50 000)	10	43.56	21.51	0.1522 ±0.0024	34.10 ±1.52	209	[1 , 6015]
	50 000 (50 000)	3	245.63	169.75	0.1462 ±0.0027	39.67 ±2.31	275	[3 , 18273]
	100 000 (50 000)	3	932.30	347.92	0.1352 ±0.0024	41.00 ±1.00	535	[6 , 30230]
	250 000 (50 000)	3	9404	869.18	0.1277 ±0.0014	44.67 ±0.58	1120	[7 , 112012]
	400 000 (50 000)	3	17 110	1437	0.1152 ±0.0014	44.67 ±1.15	1804	[10 , 116673]
VP-SVM radius = 2	1 000 (10 000)	100	1.02	0.40	0.2056 ±0.0070	4.32 ±0.72	86	[5 , 956]
	2 500 (10 000)	100	2.69	0.81	0.1953 ±0.0050	4.84 ±0.72	200	[8 , 2263]
	5 000 (10 000)	100	10.51	1.25	0.1743 ±0.0040	5.06 ±0.81	304	[3 , 4670]
	10 000 (50 000)	10	45.12	8.41	0.1636 ±0.0022	4.70 ±0.48	583	[63 , 8697]
	25 000 (50 000)	10	264.42	34.60	0.1466 ±0.0038	5.70 ±0.67	1360	[51 , 19740]
	50 000 (50 000)	3	1268	186.33	0.1445 ±0.0008	6.33 ±1.15	1663	[180 , 42173]
	100 000 (50 000)	3	3670	425.33	0.1279 ±0.0016	6.33 ±0.58	7167	[359 , 65686]
	250 000 (50 000)	0 (3)	NA	NA	NA ±NA	7.33 ±2.08	34 091	[547 , 210793]
	400 000 (50 000)	0 (3)	NA	NA	NA ±NA	7.33 ±1.53	54 545	[738 , 336202]
VP-SVM radius = 3	1 000 (10 000)	100	1.18	0.27	0.1932 ±0.0061	1.41 ±0.57	957	[36 , 1000]
	2 500 (10 000)	100	5.28	0.59	0.1875 ±0.0055	1.60 ±0.53	2371	[55 , 2500]
	5 000 (10 000)	100	24.31	1.08	0.1687 ±0.0035	1.36 ±0.48	4795	[188 , 5000]
	10 000 (50 000)	10	116.00	8.76	0.1592 ±0.0032	1.10 ±0.32	10 000	[615 , 10000]
	25 000 (50 000)	10	580.23	41.56	0.1473 ±0.0029	1.80 ±0.42	21 051	[752 , 25000]
	50 000 (50 000)	3	3130	186.34	0.1429 ±0.0015	1.67 ±0.58	42 736	[1861 , 50000]
	100 000 (50 000)	3	8886	371.73	0.1303 ±0.0029	2.33 ±0.58	12 300	[2023 , 97977]
	250 000 (50 000)	0 (3)	NA	NA	NA ±NA	2.33 ±0.58	107 143	[2796 , 247204]
	400 000 (50 000)	0 (3)	NA	NA	NA ±NA	2.00 ±0.00	200 000	[4773 , 395227]
VP-SVM radius = 4	1 000 (10 000)	100	1.15	0.23	0.1913 ±0.0056	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	6.45	0.57	0.1842 ±0.0043	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	28.18	1.06	0.1672 ±0.0032	1.00 ±0.00	5000	{5000}
	10 000 (50 000)	10	119.95	8.58	0.1596 ±0.0032	1.00 ±0.00	10 000	{10000}
	25 000 (50 000)	10	807.92	72.42	0.1452 ±0.0020	1.10 ±0.32	25 000	[6292 , 25000]
	50 000 (50 000)	3	3539	197.21	0.1410 ±0.0022	1.00 ±0.00	50 000	{50000}
	100 000 (50 000)	3	12 679	389.43	0.1302 ±0.0008	1.00 ±0.00	100 000	{100000}
	250 000 (50 000)	0 (3)	NA	NA	NA ±NA	1.00 ±0.00	250 000	{250000}
	400 000 (50 000)	0 (3)	NA	NA	NA ±NA	1.00 ±0.00	400 000	{400000}

	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
RC-SVM # of ws = 1	1 000 (10 000)	100	1.10	0.23	0.1912 ±0.0054	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	7.11	0.53	0.1842 ±0.0042	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	32.00	0.87	0.1672 ±0.0032	1.00 ±0.00	5000	{5000}
	10 000 (50 000)	10	119.75	11.36	0.1595 ±0.0033	1.00 ±0.00	10 000	{10000}
	25 000 (50 000)	10	737.28	24.31	0.1453 ±0.0020	1.00 ±0.00	25 000	{25000}
	50 000 (50 000)	3	2688	39.19	0.1410 ±0.0022	1.00 ±0.00	50 000	{50000}
	100 000 (50 000)	3	10 227	163.67	0.1302 ±0.0008	1.00 ±0.00	100 000	{100000}
	250 000 (50 000)	0	NA	NA	NA ±NA	NA ±NA	NA	{NA}
	400 000 (50 000)	0	NA	NA	NA ±NA	NA ±NA	NA	{NA}
RC-SVM # of ws = 2	1 000 (10 000)	100	1.06	0.38	0.1964 ±0.0049	2.00 ±0.00	500	{500}
	2 500 (10 000)	100	3.88	0.59	0.1881 ±0.0037	2.00 ±0.00	1250	{1250}
	5 000 (10 000)	100	14.59	0.94	0.1721 ±0.0030	2.00 ±0.00	2500	{2500}
	10 000 (50 000)	10	55.61	11.33	0.1644 ±0.0026	2.00 ±0.00	5000	{5000}
	25 000 (50 000)	10	369.81	18.29	0.1465 ±0.0022	2.00 ±0.00	12 500	{12500}
	50 000 (50 000)	3	1422	39.07	0.1439 ±0.0017	2.00 ±0.00	25 000	{25000}
	100 000 (50 000)	3	5465	184.18	0.1325 ±0.0003	2.00 ±0.00	50 000	{50000}
	250 000 (50 000)	0 (3)	NA	NA	NA ±NA	2.00 ±0.00	125 000	{125000}
	400 000 (50 000)	0 (3)	NA	NA	NA ±NA	2.00 ±0.00	200 000	{200000}

	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
RC-SVM # of ws = 3	1 000 (10 000)	100	0.87	0.39	0.2036 ±0.0050	3.00 ±0.00	333	[333 , 334]
	2 500 (10 000)	100	2.65	0.63	0.1925 ±0.0035	3.00 ±0.00	833	[833 , 834]
	5 000 (10 000)	100	9.58	0.99	0.1758 ±0.0027	3.00 ±0.00	1667	[1666 , 1667]
	10 000 (50 000)	10	34.21	11.14	0.1660 ±0.0025	3.00 ±0.00	3333	[3333 , 3334]
	25 000 (50 000)	10	240.40	20.92	0.1506 ±0.0017	3.00 ±0.00	8333	[8333 , 8334]
	50 000 (50 000)	3	972.97	33.79	0.1487 ±0.0003	3.00 ±0.00	16 667	[16666 , 16667]
	100 000 (50 000)	3	3732	178.60	0.1353 ±0.0038	3.00 ±0.00	33 333	[33333 , 33334]
	250 000 (50 000)	3	21 952	697.16	0.1251 ±0.0021	3.00 ±0.00	83 333	[83333 , 83334]
	400 000 (50 000)	0 (3)	NA	NA	NA ±NA	3.00 ±0.00	133 333	[133333 , 133334]
RC-SVM # of ws = 4	1 000 (10 000)	100	0.82	0.40	0.2087 ±0.0050	4.00 ±0.00	250	{250}
	2 500 (10 000)	100	1.92	0.57	0.1967 ±0.0032	4.00 ±0.00	625	{625}
	5 000 (10 000)	100	6.82	0.93	0.1795 ±0.0028	4.00 ±0.00	1250	{1250}
	10 000 (50 000)	10	23.97	9.69	0.1716 ±0.0020	4.00 ±0.00	2500	{2500}
	25 000 (50 000)	10	172.79	21.55	0.1525 ±0.0015	4.00 ±0.00	6250	{6250}
	50 000 (50 000)	3	723.49	36.48	0.1497 ±0.0005	4.00 ±0.00	12 500	{12500}
	100 000 (50 000)	3	2882	189.13	0.1373 ±0.0009	4.00 ±0.00	25 000	{25000}
	250 000 (50 000)	3	18 055	807.67	0.1272 ±0.0010	4.00 ±0.00	62 500	{62500}
	400 000 (50 000)	1 (3)	44 182	1193	0.1225 ±0.0000	4.00 ±0.00	100 000	{100000}
RC-SVM # of ws = 5	1 000 (10 000)	100	0.84	0.42	0.2137 ±0.0044	5.00 ±0.00	200	{200}
	2 500 (10 000)	100	1.65	0.50	0.2001 ±0.0030	5.00 ±0.00	500	{500}
	5 000 (10 000)	100	5.58	1.17	0.1819 ±0.0027	5.00 ±0.00	1000	{1000}
	10 000 (50 000)	10	20.48	11.74	0.1738 ±0.0026	5.00 ±0.00	2000	{2000}
	25 000 (50 000)	10	132.96	22.39	0.1556 ±0.0015	5.00 ±0.00	5000	{5000}
	50 000 (50 000)	3	582.94	41.03	0.1510 ±0.0005	5.00 ±0.00	10 000	{10000}
	100 000 (50 000)	3	2317	174.35	0.1397 ±0.0012	5.00 ±0.00	20 000	{20000}
	250 000 (50 000)	3	13 655	684.71	0.1290 ±0.0013	5.00 ±0.00	50 000	{50000}
	400 000 (50 000)	3	35 724	1360	0.1233 ±0.0003	5.00 ±0.00	80 000	{80000}
RC-SVM # of ws = 6	1 000 (10 000)	100	0.86	0.42	0.2176 ±0.0050	6.00 ±0.00	167	[166 , 167]
	2 500 (10 000)	100	1.55	0.50	0.2036 ±0.0037	6.00 ±0.00	417	[416 , 417]
	5 000 (10 000)	100	5.29	1.26	0.1849 ±0.0022	6.00 ±0.00	833	[833 , 834]
	10 000 (50 000)	10	19.15	12.14	0.1748 ±0.0010	6.00 ±0.00	1667	[1666 , 1667]
	25 000 (50 000)	10	106.73	22.69	0.1580 ±0.0016	6.00 ±0.00	4167	[4166 , 4167]
	50 000 (50 000)	3	478.62	41.69	0.1540 ±0.0013	6.00 ±0.00	8333	[8333 , 8334]
	100 000 (50 000)	3	1946	180.19	0.1396 ±0.0018	6.00 ±0.00	16 667	[16666 , 16667]
	250 000 (50 000)	3	12 080	690.60	0.1315 ±0.0012	6.00 ±0.00	41 667	[41666 , 41667]
	400 000 (50 000)	3	31 533	1298	0.1251 ±0.0006	6.00 ±0.00	66 667	[66666 , 66667]
RC-SVM # of ws = 10	1 000 (10 000)	100	1.02	0.45	0.2341 ±0.0048	10.00 ±0.00	100	{100}
	2 500 (10 000)	100	1.45	0.51	0.2136 ±0.0030	10.00 ±0.00	250	{250}
	5 000 (10 000)	100	3.42	1.30	0.1928 ±0.0026	10.00 ±0.00	500	{500}
	10 000 (50 000)	10	12.03	13.35	0.1831 ±0.0028	10.00 ±0.00	1000	{1000}
	25 000 (50 000)	10	58.63	22.87	0.1650 ±0.0013	10.00 ±0.00	2500	{2500}
	50 000 (50 000)	3	264.45	47.21	0.1613 ±0.0008	10.00 ±0.00	5000	{5000}
	100 000 (50 000)	3	1166	190.86	0.1450 ±0.0022	10.00 ±0.00	10 000	{10000}
	250 000 (50 000)	3	7119	643.38	0.1374 ±0.0012	10.00 ±0.00	25 000	{25000}
	400 000 (50 000)	3	18 495	1317	0.1306 ±0.0014	10.00 ±0.00	40 000	{40000}
RC-SVM # of ws = 20	1 000 (10 000)	100	1.61	0.50	0.2736 ±0.0071	20.00 ±0.00	50	{50}
	2 500 (10 000)	100	1.76	0.55	0.2365 ±0.0038	20.00 ±0.00	125	{125}
	5 000 (10 000)	100	2.84	1.33	0.2067 ±0.0024	20.00 ±0.00	250	{250}
	10 000 (50 000)	10	6.79	14.07	0.1946 ±0.0026	20.00 ±0.00	500	{500}
	25 000 (50 000)	10	31.20	23.20	0.1730 ±0.0021	20.00 ±0.00	1250	{1250}
	50 000 (50 000)	3	115.38	46.53	0.1699 ±0.0003	20.00 ±0.00	2500	{2500}
	100 000 (50 000)	3	524.48	201.80	0.1548 ±0.0009	20.00 ±0.00	5000	{5000}
	250 000 (50 000)	3	3688	616.18	0.1418 ±0.0006	20.00 ±0.00	12 500	{12500}
	400 000 (50 000)	3	9783	1119	0.1374 ±0.0003	20.00 ±0.00	20 000	{20000}
RC-SVM # of ws = 50	1 000 (10 000)	100	3.35	0.51	0.4617 ±0.0284	50.00 ±0.00	20	{20}
	2 500 (10 000)	100	3.31	0.63	0.2958 ±0.0079	50.00 ±0.00	50	{50}
	5 000 (10 000)	100	4.79	1.56	0.2351 ±0.0034	50.00 ±0.00	100	{100}
	10 000 (50 000)	10	5.77	15.00	0.2125 ±0.0014	50.00 ±0.00	200	{200}
	25 000 (50 000)	10	25.22	23.75	0.1867 ±0.0013	50.00 ±0.00	500	{500}
	50 000 (50 000)	3	52.65	43.21	0.1812 ±0.0004	50.00 ±0.00	1000	{1000}
	100 000 (50 000)	3	184.55	170.12	0.1666 ±0.0002	50.00 ±0.00	2000	{2000}
	250 000 (50 000)	3	1324	588.68	0.1547 ±0.0003	50.00 ±0.00	5000	{5000}
	400 000 (50 000)	3	3825	1008	0.1491 ±0.0009	50.00 ±0.00	8000	{8000}

Table 5: Experimental results relating to the COD-RNA data sets

	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
LS-SVM	1 000 (10 000)	100	1.13	0.41	0.1582 ±0.0052	1.00		
	2 500 (10 000)	100	4.08	0.80	0.1099 ±0.0028	1.00		
	5 000 (10 000)	100	18.37	1.31	0.1010 ±0.0016	1.00		
	10 000 (50 000)	10	82.65	9.29	0.0728 ±0.0006	1.00		
	25 000 (50 000)	10	526.48	24.21	0.0542 ±0.0004	1.00		
	50 000 (50 000)	3	2146	46.89	0.0448 ±0.0003	1.00		
	100 000 (50 000)	3	8907	246.99	0.0365 ±0.0002	1.00		
	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
VP-SVM radius = 1	1 000 (10 000)	100	4.22	0.58	0.2496 ±0.0074	58.07 ±1.99	10	[1 , 83]
	2 500 (10 000)	100	5.79	0.95	0.1489 ±0.0043	76.80 ±1.93	17	[1 , 177]
	5 000 (10 000)	100	7.72	1.40	0.1218 ±0.0029	93.44 ±2.24	21	[1 , 336]
	10 000 (50 000)	10	11.40	9.93	0.0834 ±0.0023	115.70 ±2.63	29	[1 , 618]
	25 000 (50 000)	10	24.17	30.30	0.0614 ±0.0014	154.10 ±3.35	43	[1 , 1835]
	50 000 (50 000)	3	86.99	201.41	0.0455 ±0.0012	162.33 ±4.04	92	[2 , 2808]
	100 000 (50 000)	3	264.22	356.52	0.0365 ±0.0019	199.00 ±7.81	127	[1 , 5386]
VP-SVM radius = 2	1 000 (10 000)	100	1.48	0.47	0.2038 ±0.0061	18.08 ±0.69	67	[12 , 110]
	2 500 (10 000)	100	1.92	0.89	0.1185 ±0.0034	18.93 ±0.74	169	[23 , 276]
	5 000 (10 000)	100	3.42	1.42	0.0977 ±0.0024	19.41 ±0.57	339	[16 , 539]
	10 000 (50 000)	10	7.24	14.77	0.0660 ±0.0015	20.40 ±0.52	348	[41 , 1010]
	25 000 (50 000)	10	28.04	34.35	0.0513 ±0.0010	22.20 ±0.92	664	[64 , 2516]
	50 000 (50 000)	3	196.07	235.05	0.0408 ±0.0002	23.33 ±0.58	1236	[171 , 4336]
	100 000 (50 000)	3	688.80	410.05	0.0343 ±0.0006	24.33 ±0.58	2141	[306 , 9959]
VP-SVM radius = 3	1 000 (10 000)	100	0.97	0.47	0.1857 ±0.0050	9.98 ±0.14	103	[84 , 152]
	2 500 (10 000)	100	1.34	0.90	0.1109 ±0.0036	10.00 ±0.00	252	[226 , 276]
	5 000 (10 000)	100	3.38	1.93	0.0966 ±0.0023	10.00 ±0.00	498	[467 , 539]
	10 000 (50 000)	10	8.67	14.17	0.0660 ±0.0008	10.00 ±0.00	1000	[962 , 1059]
	25 000 (50 000)	10	64.97	74.29	0.0507 ±0.0005	10.00 ±0.00	2494	[2447 , 2566]
	50 000 (50 000)	3	310.74	248.32	0.0409 ±0.0004	10.00 ±0.00	5002	[4902 , 5147]
	100 000 (50 000)	3	1127	421.73	0.0342 ±0.0004	10.00 ±0.00	9988	[9912 , 10070]
VP-SVM radius = 4	1 000 (10 000)	100	1.12	0.39	0.1582 ±0.0052	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	3.94	0.81	0.1099 ±0.0028	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	21.67	1.36	0.1010 ±0.0016	1.00 ±0.00	5000	{5000}
	10 000 (50 000)	10	92.45	11.62	0.0728 ±0.0006	1.00 ±0.00	10 000	{10000}
	25 000 (50 000)	10	1052	110.79	0.0542 ±0.0004	1.00 ±0.00	25 000	{25000}
	50 000 (50 000)	3	3621	201.31	0.0448 ±0.0003	1.00 ±0.00	50 000	{50000}
	100 000 (50 000)	3	12 439	333.57	0.0365 ±0.0002	1.00 ±0.00	100 000	{100000}

	data set sizes	runs	train time	test time	test error	# of ws	ws size: median	ws size: range
RC-SVM # of ws = 1	1 000 (10 000)	100	1.08	0.38	0.1582 ±0.0052	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	4.01	0.85	0.1099 ±0.0028	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	17.47	1.29	0.1010 ±0.0016	1.00 ±0.00	5000	{5000}
	10 000 (50 000)	10	79.67	8.62	0.0728 ±0.0006	1.00 ±0.00	10 000	{10000}
	25 000 (50 000)	10	500.55	22.20	0.0542 ±0.0004	1.00 ±0.00	25 000	{25000}
	50 000 (50 000)	3	2388	116.48	0.0448 ±0.0003	1.00 ±0.00	50 000	{50000}
	100 000 (50 000)	3	8592	231.08	0.0365 ±0.0002	1.00 ±0.00	100 000	{100000}
RC-SVM # of ws = 5	1 000 (10 000)	100	0.66	0.39	0.2250 ±0.0075	5.00 ±0.00	200	{200}
	2 500 (10 000)	100	1.26	0.84	0.1698 ±0.0050	5.00 ±0.00	500	{500}
	5 000 (10 000)	100	3.73	1.45	0.1480 ±0.0031	5.00 ±0.00	1000	{1000}
	10 000 (50 000)	10	12.55	11.46	0.1093 ±0.0015	5.00 ±0.00	2000	{2000}
	25 000 (50 000)	10	88.31	27.18	0.0805 ±0.0009	5.00 ±0.00	5000	{5000}
	50 000 (50 000)	3	612.93	226.31	0.0624 ±0.0003	5.00 ±0.00	10 000	{10000}
	100 000 (50 000)	3	1623	301.96	0.0488 ±0.0001	5.00 ±0.00	20 000	{20000}
RC-SVM # of ws = 10	1 000 (10 000)	100	0.94	0.36	0.2648 ±0.0059	10.00 ±0.00	100	{100}
	2 500 (10 000)	100	1.40	0.90	0.2064 ±0.0049	10.00 ±0.00	250	{250}
	5 000 (10 000)	100	3.20	1.39	0.1806 ±0.0039	10.00 ±0.00	500	{500}
	10 000 (50 000)	10	6.99	10.92	0.1372 ±0.0027	10.00 ±0.00	1000	{1000}
	25 000 (50 000)	10	40.29	29.37	0.0992 ±0.0007	10.00 ±0.00	2500	{2500}
	50 000 (50 000)	3	199.33	168.55	0.0763 ±0.0006	10.00 ±0.00	5000	{5000}
	100 000 (50 000)	3	892.77	499.74	0.0590 ±0.0004	10.00 ±0.00	10 000	{10000}
RC-SVM # of ws = 20	1 000 (10 000)	100	1.61	0.37	0.3066 ±0.0038	20.00 ±0.00	50	{50}
	2 500 (10 000)	100	1.95	0.86	0.2461 ±0.0039	20.00 ±0.00	125	{125}
	5 000 (10 000)	100	2.95	1.35	0.2156 ±0.0043	20.00 ±0.00	250	{250}
	10 000 (50 000)	10	4.67	10.56	0.1704 ±0.0031	20.00 ±0.00	500	{500}
	25 000 (50 000)	10	22.59	29.35	0.1275 ±0.0008	20.00 ±0.00	1250	{1250}
	50 000 (50 000)	3	92.33	168.15	0.0964 ±0.0011	20.00 ±0.00	2500	{2500}
	100 000 (50 000)	3	1081	613.04	0.0731 ±0.0004	20.00 ±0.00	5000	{5000}
RC-SVM # of ws = 50	1 000 (10 000)	100	4.14	0.43	0.3478 ±0.0048	50.00 ±0.00	20	{20}
	2 500 (10 000)	100	3.74	0.84	0.2966 ±0.0026	50.00 ±0.00	50	{50}
	5 000 (10 000)	100	4.97	1.31	0.2694 ±0.0036	50.00 ±0.00	100	{100}
	10 000 (50 000)	10	5.46	10.90	0.2220 ±0.0029	50.00 ±0.00	200	{200}
	25 000 (50 000)	10	18.23	25.04	0.1715 ±0.0031	50.00 ±0.00	500	{500}
	50 000 (50 000)	3	39.18	143.68	0.1366 ±0.0014	50.00 ±0.00	1000	{1000}
	100 000 (50 000)	3	133.67	264.38	0.1036 ±0.0028	50.00 ±0.00	2000	{2000}
RC-SVM # of ws = 100	1 000 (10 000)	100	6.86	0.51	0.3502 ±0.0040	100.00 ±0.00	10	{10}
	2 500 (10 000)	100	7.37	0.92	0.3268 ±0.0031	100.00 ±0.00	25	{25}
	5 000 (10 000)	100	7.58	1.31	0.3123 ±0.0022	100.00 ±0.00	50	{50}
	10 000 (50 000)	10	8.58	8.93	0.2678 ±0.0032	100.00 ±0.00	100	{100}
	25 000 (50 000)	10	15.91	21.95	0.2077 ±0.0021	100.00 ±0.00	250	{250}
	50 000 (50 000)	3	27.94	132.07	0.1679 ±0.0023	100.00 ±0.00	500	{500}
	100 000 (50 000)	3	79.05	227.69	0.1341 ±0.0017	100.00 ±0.00	1000	{1000}
RC-SVM # of ws = 150	1 000 (10 000)	100	10.52	0.62	0.3468 ±0.0026	150.00 ±0.00	7	[6 , 7]
	2 500 (10 000)	100	10.43	1.00	0.3328 ±0.0024	150.00 ±0.00	17	[16 , 17]
	5 000 (10 000)	100	9.34	1.23	0.3395 ±0.0082	150.00 ±0.00	33	[33 , 34]
	10 000 (50 000)	10	11.38	9.10	0.2943 ±0.0014	150.00 ±0.00	67	[66 , 67]
	25 000 (50 000)	10	17.25	19.51	0.2336 ±0.0041	150.00 ±0.00	167	[166 , 167]
	50 000 (50 000)	3	25.66	122.27	0.1927 ±0.0005	150.00 ±0.00	333	[333 , 334]
	100 000 (50 000)	3	73.12	188.29	0.1538 ±0.0012	150.00 ±0.00	667	[666 , 667]
RC-SVM # of ws = 200	1 000 (10 000)	100	13.70	0.64	0.3447 ±0.0021	200.00 ±0.00	5	{5}
	2 500 (10 000)	100	14.45	1.11	0.3346 ±0.0020	200.00 ±0.00	12	[12 , 13]
	5 000 (10 000)	100	12.32	1.22	0.3534 ±0.0092	200.00 ±0.00	25	{25}
	10 000 (50 000)	10	13.19	8.93	0.3091 ±0.0024	200.00 ±0.00	50	{50}
	25 000 (50 000)	10	19.80	18.41	0.2512 ±0.0030	200.00 ±0.00	125	{125}
	50 000 (50 000)	3	26.89	87.32	0.2095 ±0.0033	200.00 ±0.00	250	{250}
	100 000 (50 000)	3	70.20	185.72	0.1664 ±0.0021	200.00 ±0.00	500	{500}

Table 6: Experimental results relating to the IJCNN1 data sets

	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
LS-SVM	1 000 (10 000)	100	0.33	0.07	0.0284 ±0.0003	0.0541 ±0.0033	1.00		
	2 500 (10 000)	100	1.74	0.07	0.0275 ±0.0002	0.0461 ±0.0019	1.00		
	5 000 (10 000)	100	7.14	0.14	0.0269 ±0.0002	0.0396 ±0.0019	1.00		
	10 000 (10 000)	100	27.90	0.22	0.0265 ±0.0001	0.0323 ±0.0014	1.00		
	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
VP-SVM $r = 0.1$	1 000 (10 000)	100	1.84	0.28	0.0287 ±0.0005	0.0561 ±0.0041	15.70 ±2.43	64	[27 , 109]
	2 500 (10 000)	100	1.90	0.34	0.0267 ±0.0002	0.0368 ±0.0027	15.77 ±2.34	159	[59 , 268]
	5 000 (10 000)	100	2.01	0.40	0.0260 ±0.0001	0.0248 ±0.0018	16.04 ±2.45	312	[125 , 532]
	10 000 (10 000)	100	3.05	0.55	0.0257 ±0.0001	0.0197 ±0.0010	15.55 ±2.43	643	[245 , 1054]
VP-SVM $r = 0.25$	1 000 (10 000)	100	0.65	0.16	0.0277 ±0.0004	0.0465 ±0.0037	6.17 ±0.88	162	[66 , 251]
	2 500 (10 000)	100	0.75	0.18	0.0265 ±0.0002	0.0338 ±0.0024	6.12 ±0.87	408	[152 , 645]
	5 000 (10 000)	100	1.58	0.24	0.0260 ±0.0001	0.0256 ±0.0016	6.17 ±0.79	810	[305 , 1291]
	10 000 (10 000)	100	4.93	0.38	0.0258 ±0.0001	0.0213 ±0.0013	6.27 ±0.83	1595	[600 , 2550]
VP-SVM $r = 0.5$	1 000 (10 000)	100	0.37	0.11	0.0275 ±0.0003	0.0447 ±0.0032	3.49 ±0.50	287	[121 , 502]
	2 500 (10 000)	100	0.77	0.14	0.0266 ±0.0002	0.0355 ±0.0025	3.44 ±0.50	727	[329 , 1202]
	5 000 (10 000)	100	2.31	0.19	0.0262 ±0.0001	0.0290 ±0.0021	3.41 ±0.49	1466	[627 , 2511]
	10 000 (10 000)	100	8.88	0.31	0.0259 ±0.0001	0.0242 ±0.0019	3.51 ±0.50	2849	[1253 , 5003]
VP-SVM $r = 1$	1 000 (10 000)	100	0.27	0.08	0.0282 ±0.0003	0.0517 ±0.0030	2.00 ±0.00	500	[266 , 734]
	2 500 (10 000)	100	0.97	0.11	0.0272 ±0.0002	0.0426 ±0.0024	2.00 ±0.00	1250	[658 , 1842]
	5 000 (10 000)	100	3.68	0.15	0.0268 ±0.0002	0.0382 ±0.0021	2.00 ±0.00	2500	[1224 , 3776]
	10 000 (10 000)	100	14.93	0.44	0.0263 ±0.0001	0.0309 ±0.0020	2.00 ±0.00	5000	[2501 , 7499]
VP-SVM $r = 2$	1 000 (10 000)	100	0.32	0.07	0.0284 ±0.0004	0.0541 ±0.0033	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	1.75	0.08	0.0275 ±0.0002	0.0461 ±0.0020	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	7.16	0.15	0.0269 ±0.0002	0.0396 ±0.0019	1.00 ±0.00	5000	{5000}
	10 000 (10 000)	100	27.91	0.24	0.0265 ±0.0001	0.0323 ±0.0014	1.00 ±0.00	10 000	{10000}
	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
RC-SVM # ws = 1	1 000 (10 000)	100	0.32	0.07	0.0284 ±0.0003	0.0540 ±0.0031	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	1.76	0.08	0.0275 ±0.0002	0.0461 ±0.0020	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	7.16	0.15	0.0269 ±0.0002	0.0396 ±0.0019	1.00 ±0.00	5000	{5000}
	10 000 (10 000)	100	28.05	0.24	0.0265 ±0.0001	0.0323 ±0.0014	1.00 ±0.00	10 000	{10000}
RC-SVM # ws = 2	1 000 (10 000)	100	0.26	0.09	0.0287 ±0.0005	0.0571 ±0.0042	2.00 ±0.00	500	{500}
	2 500 (10 000)	100	0.94	0.11	0.0276 ±0.0002	0.0475 ±0.0018	2.00 ±0.00	1250	{1250}
	5 000 (10 000)	100	3.46	0.14	0.0273 ±0.0001	0.0444 ±0.0014	2.00 ±0.00	2500	{2500}
	10 000 (10 000)	100	14.37	0.28	0.0268 ±0.0001	0.0376 ±0.0016	2.00 ±0.00	5000	{5000}
RC-SVM # ws = 3	1 000 (10 000)	100	0.31	0.11	0.0288 ±0.0004	0.0583 ±0.0037	3.00 ±0.00	333	[333 , 334]
	2 500 (10 000)	100	0.76	0.14	0.0277 ±0.0002	0.0488 ±0.0022	3.00 ±0.00	833	[833 , 834]
	5 000 (10 000)	100	2.32	0.17	0.0274 ±0.0001	0.0456 ±0.0014	3.00 ±0.00	1667	[1666 , 1667]
	10 000 (10 000)	100	9.58	0.27	0.0271 ±0.0001	0.0417 ±0.0015	3.00 ±0.00	3333	[3333 , 3334]
RC-SVM # ws = 4	1 000 (10 000)	100	0.43	0.12	0.0289 ±0.0004	0.0588 ±0.0037	4.00 ±0.00	250	{250}
	2 500 (10 000)	100	0.69	0.16	0.0278 ±0.0003	0.0500 ±0.0025	4.00 ±0.00	625	{625}
	5 000 (10 000)	100	1.90	0.20	0.0275 ±0.0002	0.0469 ±0.0016	4.00 ±0.00	1250	{1250}
	10 000 (10 000)	100	6.95	0.27	0.0272 ±0.0001	0.0432 ±0.0010	4.00 ±0.00	2500	{2500}
RC-SVM # ws = 5	1 000 (10 000)	100	0.52	0.15	0.0287 ±0.0004	0.0577 ±0.0033	5.00 ±0.00	200	{200}
	2 500 (10 000)	100	0.67	0.18	0.0280 ±0.0002	0.0515 ±0.0022	5.00 ±0.00	500	{500}
	5 000 (10 000)	100	1.65	0.23	0.0276 ±0.0002	0.0480 ±0.0020	5.00 ±0.00	1000	{1000}
	10 000 (10 000)	100	5.52	0.30	0.0273 ±0.0001	0.0438 ±0.0012	5.00 ±0.00	2000	{2000}
RC-SVM # ws = 6	1 000 (10 000)	100	0.64	0.16	0.0288 ±0.0004	0.0586 ±0.0031	6.00 ±0.00	167	[166 , 167]
	2 500 (10 000)	100	0.69	0.19	0.0281 ±0.0003	0.0525 ±0.0028	6.00 ±0.00	417	[416 , 417]
	5 000 (10 000)	100	1.52	0.25	0.0277 ±0.0002	0.0490 ±0.0022	6.00 ±0.00	833	[833 , 834]
	10 000 (10 000)	100	4.67	0.32	0.0273 ±0.0001	0.0446 ±0.0011	6.00 ±0.00	1667	[1666 , 1667]
RC-SVM # ws = 10	1 000 (10 000)	100	1.05	0.22	0.0288 ±0.0003	0.0584 ±0.0029	10.00 ±0.00	100	{100}
	2 500 (10 000)	100	1.08	0.27	0.0282 ±0.0003	0.0534 ±0.0026	10.00 ±0.00	250	{250}
	5 000 (10 000)	100	1.39	0.34	0.0280 ±0.0002	0.0522 ±0.0023	10.00 ±0.00	500	{500}
	10 000 (10 000)	100	3.45	0.44	0.0275 ±0.0001	0.0471 ±0.0012	10.00 ±0.00	1000	{1000}
RC-SVM # ws = 20	1 000 (10 000)	100	2.28	0.34	0.0293 ±0.0004	0.0626 ±0.0027	20.00 ±0.00	50	{50}
	2 500 (10 000)	100	2.06	0.43	0.0282 ±0.0002	0.0531 ±0.0022	20.00 ±0.00	125	{125}
	5 000 (10 000)	100	2.38	0.51	0.0283 ±0.0002	0.0544 ±0.0017	20.00 ±0.00	250	{250}
	10 000 (10 000)	100	3.14	0.64	0.0280 ±0.0001	0.0520 ±0.0014	20.00 ±0.00	500	{500}

Table 7: Experimental results relating to the artificial data of Type I

	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
LS-SVM	1 000 (10 000)	100	0.60	0.12	0.0178 ±0.0006	0.0641 ±0.0047	1.00		
	2 500 (10 000)	100	2.29	0.21	0.0168 ±0.0004	0.0559 ±0.0033	1.00		
	5 000 (10 000)	100	8.37	0.25	0.0165 ±0.0003	0.0531 ±0.0026	1.00		
	10 000 (10 000)	100	31.61	0.44	0.0163 ±0.0003	0.0511 ±0.0033	1.00		
	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
VP-SVM $r = 0.1$	1 000 (10 000)	100	1.55	0.29	0.0157 ±0.0004	0.0427 ±0.0040	16.08 ±2.50	62	[22 , 114]
	2 500 (10 000)	100	1.66	0.36	0.0146 ±0.0002	0.0288 ±0.0028	15.79 ±2.40	158	[51 , 292]
	5 000 (10 000)	100	1.88	0.43	0.0142 ±0.0001	0.0208 ±0.0016	15.95 ±2.53	313	[105 , 549]
	10 000 (10 000)	100	3.09	0.59	0.0140 ±0.0001	0.0163 ±0.0013	15.84 ±2.50	631	[231 , 1043]
VP-SVM $r = 0.25$	1 000 (10 000)	100	0.58	0.17	0.0156 ±0.0004	0.0426 ±0.0045	6.21 ±0.83	161	[60 , 267]
	2 500 (10 000)	100	0.77	0.22	0.0147 ±0.0002	0.0299 ±0.0030	6.29 ±0.81	397	[133 , 676]
	5 000 (10 000)	100	1.64	0.30	0.0144 ±0.0001	0.0247 ±0.0023	6.22 ±0.82	804	[285 , 1299]
	10 000 (10 000)	100	4.99	0.44	0.0141 ±0.0001	0.0194 ±0.0016	6.38 ±0.80	1567	[597 , 2535]
VP-SVM $r = 0.5$	1 000 (10 000)	100	0.36	0.12	0.0159 ±0.0004	0.0465 ±0.0050	3.47 ±0.50	288	[122 , 482]
	2 500 (10 000)	100	0.81	0.18	0.0151 ±0.0002	0.0370 ±0.0031	3.52 ±0.50	710	[297 , 1248]
	5 000 (10 000)	100	2.52	0.28	0.0150 ±0.0002	0.0346 ±0.0032	3.47 ±0.50	1441	[604 , 2508]
	10 000 (10 000)	100	9.36	0.43	0.0147 ±0.0002	0.0301 ±0.0034	3.58 ±0.50	2793	[1249 , 4962]
VP-SVM $r = 1$	1 000 (10 000)	100	0.33	0.09	0.0169 ±0.0005	0.0562 ±0.0047	2.00 ±0.00	500	[265 , 735]
	2 500 (10 000)	100	1.11	0.14	0.0160 ±0.0004	0.0477 ±0.0041	2.00 ±0.00	1250	[613 , 1887]
	5 000 (10 000)	100	4.11	0.23	0.0158 ±0.0004	0.0452 ±0.0044	2.00 ±0.00	2500	[1232 , 3768]
	10 000 (10 000)	100	16.26	0.42	0.0154 ±0.0005	0.0402 ±0.0064	2.00 ±0.00	5000	[2533 , 7467]
VP-SVM $r = 2$	1 000 (10 000)	100	0.37	0.08	0.0178 ±0.0006	0.0641 ±0.0045	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	2.03	0.12	0.0168 ±0.0003	0.0558 ±0.0033	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	8.12	0.20	0.0165 ±0.0003	0.0531 ±0.0026	1.00 ±0.00	5000	{5000}
	10 000 (10 000)	100	31.12	0.36	0.0163 ±0.0003	0.0511 ±0.0033	1.00 ±0.00	10 000	{10000}
	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
RC-SVM # ws = 1	1 000 (10 000)	100	0.37	0.07	0.0178 ±0.0006	0.0640 ±0.0046	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	2.02	0.13	0.0168 ±0.0003	0.0558 ±0.0033	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	8.14	0.19	0.0165 ±0.0003	0.0531 ±0.0026	1.00 ±0.00	5000	{5000}
	10 000 (10 000)	100	31.45	0.36	0.0163 ±0.0003	0.0511 ±0.0033	1.00 ±0.00	10 000	{10000}
RC-SVM # ws = 2	1 000 (10 000)	100	0.30	0.11	0.0172 ±0.0005	0.0595 ±0.0046	2.00 ±0.00	500	{500}
	2 500 (10 000)	100	1.10	0.15	0.0169 ±0.0004	0.0567 ±0.0037	2.00 ±0.00	1250	{1250}
	5 000 (10 000)	100	4.06	0.25	0.0165 ±0.0003	0.0528 ±0.0029	2.00 ±0.00	2500	{2500}
	10 000 (10 000)	100	16.33	0.41	0.0163 ±0.0002	0.0516 ±0.0020	2.00 ±0.00	5000	{5000}
RC-SVM # ws = 3	1 000 (10 000)	100	0.32	0.14	0.0168 ±0.0004	0.0555 ±0.0040	3.00 ±0.00	333	[333 , 334]
	2 500 (10 000)	100	0.88	0.17	0.0169 ±0.0004	0.0566 ±0.0033	3.00 ±0.00	833	[833 , 834]
	5 000 (10 000)	100	2.71	0.28	0.0165 ±0.0002	0.0534 ±0.0022	3.00 ±0.00	1667	[1666 , 1667]
	10 000 (10 000)	100	10.99	0.45	0.0163 ±0.0002	0.0515 ±0.0018	3.00 ±0.00	3333	[3333 , 3334]
RC-SVM # ws = 4	1 000 (10 000)	100	0.39	0.17	0.0166 ±0.0004	0.0544 ±0.0039	4.00 ±0.00	250	{250}
	2 500 (10 000)	100	0.81	0.21	0.0166 ±0.0003	0.0537 ±0.0032	4.00 ±0.00	625	{625}
	5 000 (10 000)	100	2.23	0.31	0.0165 ±0.0003	0.0532 ±0.0027	4.00 ±0.00	1250	{1250}
	10 000 (10 000)	100	8.16	0.49	0.0163 ±0.0002	0.0516 ±0.0020	4.00 ±0.00	2500	{2500}
RC-SVM # ws = 5	1 000 (10 000)	100	0.46	0.19	0.0167 ±0.0004	0.0553 ±0.0041	5.00 ±0.00	200	{200}
	2 500 (10 000)	100	0.78	0.25	0.0165 ±0.0004	0.0532 ±0.0035	5.00 ±0.00	500	{500}
	5 000 (10 000)	100	1.94	0.32	0.0164 ±0.0003	0.0524 ±0.0029	5.00 ±0.00	1000	{1000}
	10 000 (10 000)	100	6.51	0.51	0.0164 ±0.0002	0.0521 ±0.0022	5.00 ±0.00	2000	{2000}
RC-SVM # ws = 6	1 000 (10 000)	100	0.53	0.21	0.0168 ±0.0004	0.0556 ±0.0031	6.00 ±0.00	167	[166 , 167]
	2 500 (10 000)	100	0.76	0.28	0.0164 ±0.0003	0.0520 ±0.0029	6.00 ±0.00	417	[416 , 417]
	5 000 (10 000)	100	1.77	0.37	0.0164 ±0.0003	0.0518 ±0.0027	6.00 ±0.00	833	[833 , 834]
	10 000 (10 000)	100	5.49	0.54	0.0165 ±0.0002	0.0530 ±0.0019	6.00 ±0.00	1667	[1666 , 1667]
RC-SVM # ws = 10	1 000 (10 000)	100	0.85	0.27	0.0171 ±0.0003	0.0585 ±0.0027	10.00 ±0.00	100	{100}
	2 500 (10 000)	100	0.98	0.40	0.0160 ±0.0003	0.0487 ±0.0027	10.00 ±0.00	250	{250}
	5 000 (10 000)	100	1.56	0.52	0.0160 ±0.0002	0.0478 ±0.0023	10.00 ±0.00	500	{500}
	10 000 (10 000)	100	3.91	0.63	0.0164 ±0.0002	0.0524 ±0.0017	10.00 ±0.00	1000	{1000}
RC-SVM # ws = 20	1 000 (10 000)	100	1.83	0.37	0.0196 ±0.0007	0.0774 ±0.0042	20.00 ±0.00	50	{50}
	2 500 (10 000)	100	1.97	0.56	0.0163 ±0.0002	0.0512 ±0.0020	20.00 ±0.00	125	{125}
	5 000 (10 000)	100	2.20	0.79	0.0156 ±0.0002	0.0446 ±0.0018	20.00 ±0.00	250	{250}
	10 000 (10 000)	100	3.19	0.99	0.0160 ±0.0002	0.0488 ±0.0018	20.00 ±0.00	500	{500}

Table 8: Experimental results relating to the artificial data of Type II

	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
LS-SVM	1 000 (10 000)	100	0.36	0.05	0.0588 ±0.0009	0.0799 ±0.0056	1.00		
	2 500 (10 000)	100	1.82	0.07	0.0572 ±0.0005	0.0677 ±0.0039	1.00		
	5 000 (10 000)	100	7.47	0.11	0.0580 ±0.0003	0.0745 ±0.0019	1.00		
	10 000 (10 000)	100	29.56	0.24	0.0559 ±0.0003	0.0563 ±0.0028	1.00		
	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
VP-SVM $r = 0.1$	1 000 (10 000)	100	1.52	0.29	0.0580 ±0.0009	0.0731 ±0.0057	15.55 ±2.36	64	[26 , 109]
	2 500 (10 000)	100	1.61	0.35	0.0553 ±0.0004	0.0493 ±0.0038	15.99 ±2.39	156	[56 , 279]
	5 000 (10 000)	100	1.61	0.42	0.0539 ±0.0002	0.0347 ±0.0024	15.80 ±2.45	316	[116 , 531]
	10 000 (10 000)	100	3.14	0.58	0.0534 ±0.0001	0.0257 ±0.0016	15.73 ±2.58	636	[236 , 1068]
VP-SVM $r = 0.25$	1 000 (10 000)	100	0.53	0.16	0.0563 ±0.0007	0.0600 ±0.0057	6.18 ±0.87	162	[56 , 259]
	2 500 (10 000)	100	0.77	0.19	0.0547 ±0.0003	0.0429 ±0.0034	6.26 ±0.81	399	[150 , 658]
	5 000 (10 000)	100	1.65	0.26	0.0537 ±0.0001	0.0315 ±0.0024	6.27 ±0.81	797	[321 , 1270]
	10 000 (10 000)	100	5.11	0.39	0.0534 ±0.0001	0.0244 ±0.0021	6.34 ±0.79	1577	[597 , 2534]
VP-SVM $r = 0.5$	1 000 (10 000)	100	0.34	0.11	0.0564 ±0.0007	0.0620 ±0.0057	3.48 ±0.50	287	[126 , 501]
	2 500 (10 000)	100	0.81	0.14	0.0551 ±0.0004	0.0478 ±0.0047	3.49 ±0.50	716	[297 , 1244]
	5 000 (10 000)	100	2.44	0.22	0.0543 ±0.0004	0.0409 ±0.0048	3.50 ±0.50	1429	[641 , 2517]
	10 000 (10 000)	100	9.40	0.30	0.0538 ±0.0004	0.0322 ±0.0054	3.54 ±0.50	2825	[1209 , 4993]
VP-SVM $r = 1$	1 000 (10 000)	100	0.28	0.08	0.0579 ±0.0009	0.0738 ±0.0062	2.00 ±0.00	500	[248 , 752]
	2 500 (10 000)	100	1.04	0.11	0.0565 ±0.0007	0.0617 ±0.0059	2.00 ±0.00	1250	[625 , 1875]
	5 000 (10 000)	100	3.92	0.15	0.0564 ±0.0005	0.0622 ±0.0044	2.00 ±0.00	2500	[1281 , 3719]
	10 000 (10 000)	100	15.93	0.23	0.0555 ±0.0004	0.0520 ±0.0035	2.00 ±0.00	5000	[2404 , 7596]
VP-SVM $r = 2$	1 000 (10 000)	100	0.35	0.06	0.0588 ±0.0009	0.0799 ±0.0057	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	1.84	0.08	0.0572 ±0.0005	0.0677 ±0.0039	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	7.51	0.12	0.0580 ±0.0003	0.0745 ±0.0019	1.00 ±0.00	5000	{5000}
	10 000 (10 000)	100	29.75	0.25	0.0559 ±0.0003	0.0563 ±0.0028	1.00 ±0.00	10 000	{10000}
	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
RC-SVM # ws = 1	1 000 (10 000)	100	0.35	0.07	0.0588 ±0.0009	0.0801 ±0.0058	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	1.85	0.08	0.0572 ±0.0005	0.0678 ±0.0039	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	7.73	0.12	0.0580 ±0.0003	0.0745 ±0.0019	1.00 ±0.00	5000	{5000}
	10 000 (10 000)	100	30.24	0.27	0.0559 ±0.0003	0.0563 ±0.0028	1.00 ±0.00	10 000	{10000}
RC-SVM # ws = 2	1 000 (10 000)	100	0.30	0.08	0.0596 ±0.0009	0.0855 ±0.0052	2.00 ±0.00	500	{500}
	2 500 (10 000)	100	1.01	0.11	0.0579 ±0.0005	0.0736 ±0.0033	2.00 ±0.00	1250	{1250}
	5 000 (10 000)	100	3.81	0.15	0.0573 ±0.0005	0.0699 ±0.0038	2.00 ±0.00	2500	{2500}
	10 000 (10 000)	100	15.22	0.24	0.0562 ±0.0004	0.0598 ±0.0035	2.00 ±0.00	5000	{5000}
RC-SVM # ws = 3	1 000 (10 000)	100	0.32	0.12	0.0591 ±0.0010	0.0825 ±0.0067	3.00 ±0.00	333	[333 , 334]
	2 500 (10 000)	100	0.82	0.14	0.0585 ±0.0006	0.0776 ±0.0038	3.00 ±0.00	833	[833 , 834]
	5 000 (10 000)	100	2.52	0.18	0.0574 ±0.0004	0.0710 ±0.0031	3.00 ±0.00	1667	[1666 , 1667]
	10 000 (10 000)	100	10.05	0.24	0.0564 ±0.0004	0.0622 ±0.0032	3.00 ±0.00	3333	[3333 , 3334]
RC-SVM # ws = 4	1 000 (10 000)	100	0.37	0.15	0.0589 ±0.0011	0.0814 ±0.0069	4.00 ±0.00	250	{250}
	2 500 (10 000)	100	0.77	0.16	0.0587 ±0.0006	0.0793 ±0.0038	4.00 ±0.00	625	{625}
	5 000 (10 000)	100	2.07	0.20	0.0579 ±0.0004	0.0747 ±0.0025	4.00 ±0.00	1250	{1250}
	10 000 (10 000)	100	7.41	0.27	0.0563 ±0.0004	0.0616 ±0.0036	4.00 ±0.00	2500	{2500}
RC-SVM # ws = 5	1 000 (10 000)	100	0.46	0.17	0.0588 ±0.0010	0.0811 ±0.0064	5.00 ±0.00	200	{200}
	2 500 (10 000)	100	0.73	0.19	0.0586 ±0.0007	0.0787 ±0.0045	5.00 ±0.00	500	{500}
	5 000 (10 000)	100	1.87	0.24	0.0583 ±0.0004	0.0775 ±0.0027	5.00 ±0.00	1000	{1000}
	10 000 (10 000)	100	5.88	0.30	0.0563 ±0.0003	0.0617 ±0.0029	5.00 ±0.00	2000	{2000}
RC-SVM # ws = 6	1 000 (10 000)	100	0.52	0.19	0.0586 ±0.0010	0.0796 ±0.0070	6.00 ±0.00	167	[166 , 167]
	2 500 (10 000)	100	0.73	0.21	0.0586 ±0.0006	0.0788 ±0.0044	6.00 ±0.00	417	[416 , 417]
	5 000 (10 000)	100	1.67	0.26	0.0586 ±0.0005	0.0793 ±0.0033	6.00 ±0.00	833	[833 , 834]
	10 000 (10 000)	100	5.01	0.33	0.0565 ±0.0003	0.0637 ±0.0026	6.00 ±0.00	1667	[1666 , 1667]
RC-SVM # ws = 10	1 000 (10 000)	100	0.89	0.25	0.0593 ±0.0011	0.0845 ±0.0069	10.00 ±0.00	100	{100}
	2 500 (10 000)	100	0.99	0.33	0.0579 ±0.0006	0.0744 ±0.0043	10.00 ±0.00	250	{250}
	5 000 (10 000)	100	1.51	0.36	0.0590 ±0.0005	0.0824 ±0.0033	10.00 ±0.00	500	{500}
	10 000 (10 000)	100	3.59	0.44	0.0574 ±0.0003	0.0705 ±0.0024	10.00 ±0.00	1000	{1000}
RC-SVM # ws = 20	1 000 (10 000)	100	1.90	0.38	0.0671 ±0.0019	0.1238 ±0.0082	20.00 ±0.00	50	{50}
	2 500 (10 000)	100	2.01	0.53	0.0578 ±0.0007	0.0742 ±0.0052	20.00 ±0.00	125	{125}
	5 000 (10 000)	100	2.19	0.64	0.0583 ±0.0005	0.0779 ±0.0031	20.00 ±0.00	250	{250}
	10 000 (10 000)	100	3.20	0.67	0.0582 ±0.0003	0.0770 ±0.0021	20.00 ±0.00	500	{500}

Table 9: Experimental results relating to the artificial data of Type III

	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
LS-SVM	1 000 (10 000)	100	0.61	0.07	0.0155 ±0.0002	0.0848 ±0.0012	1.00		
	2 500 (10 000)	100	3.28	0.16	0.0151 ±0.0003	0.0821 ±0.0016	1.00		
	5 000 (10 000)	100	11.89	0.63	0.0132 ±0.0007	0.0694 ±0.0049	1.00		
	10 000 (10 000)	100	45.94	0.58	0.0137 ±0.0007	0.0730 ±0.0046	1.00		
	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
VP-SVM $r = 0.25$	1 000 (10 000)	100	4.27	0.62	0.0172 ±0.0005	0.0946 ±0.0024	44.23 ±2.19	23	[4 , 54]
	2 500 (10 000)	100	4.78	0.85	0.0142 ±0.0002	0.0775 ±0.0015	46.09 ±2.58	54	[13 , 108]
	5 000 (10 000)	100	5.03	1.16	0.0129 ±0.0002	0.0682 ±0.0010	46.32 ±2.61	108	[19 , 244]
	10 000 (10 000)	100	5.72	1.76	0.0119 ±0.0001	0.0607 ±0.0007	47.52 ±2.72	210	[43 , 416]
VP-SVM $r = 0.5$	1 000 (10 000)	100	1.45	0.30	0.0158 ±0.0003	0.0867 ±0.0019	14.50 ±1.25	69	[18 , 155]
	2 500 (10 000)	100	1.62	0.49	0.0139 ±0.0002	0.0752 ±0.0013	15.22 ±1.47	164	[40 , 323]
	5 000 (10 000)	100	2.17	0.78	0.0129 ±0.0002	0.0683 ±0.0011	15.00 ±1.45	333	[78 , 635]
	10 000 (10 000)	100	4.52	1.24	0.0120 ±0.0001	0.0611 ±0.0009	15.72 ±1.62	636	[144 , 1261]
VP-SVM $r = 1$	1 000 (10 000)	100	0.49	0.18	0.0152 ±0.0003	0.0833 ±0.0017	5.16 ±0.53	194	[69 , 478]
	2 500 (10 000)	100	0.99	0.32	0.0140 ±0.0003	0.0753 ±0.0019	5.40 ±0.51	463	[156 , 1260]
	5 000 (10 000)	100	2.77	0.58	0.0132 ±0.0003	0.0697 ±0.0020	5.48 ±0.54	912	[264 , 2437]
	10 000 (10 000)	100	10.03	0.80	0.0124 ±0.0002	0.0638 ±0.0013	5.56 ±0.64	1799	[529 , 4868]
VP-SVM $r = 2$	1 000 (10 000)	100	0.43	0.10	0.0155 ±0.0002	0.0847 ±0.0014	1.72 ±0.45	581	[283 , 1000]
	2 500 (10 000)	100	2.29	0.15	0.0149 ±0.0004	0.0808 ±0.0024	1.55 ±0.50	1613	[696 , 2500]
	5 000 (10 000)	100	7.96	0.42	0.0138 ±0.0007	0.0739 ±0.0049	1.67 ±0.47	2994	[1345 , 5000]
	10 000 (10 000)	100	31.92	0.82	0.0126 ±0.0009	0.0656 ±0.0065	1.64 ±0.48	6098	[2702 , 10000]
VP-SVM $r = 3$	1 000 (10 000)	100	0.54	0.07	0.0155 ±0.0002	0.0847 ±0.0012	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	3.09	0.12	0.0151 ±0.0003	0.0822 ±0.0016	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	11.64	0.53	0.0132 ±0.0007	0.0694 ±0.0049	1.00 ±0.00	5000	{5000}
	10 000 (10 000)	100	45.40	0.54	0.0137 ±0.0007	0.0730 ±0.0046	1.00 ±0.00	10 000	{10000}
	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
RC-SVM # ws = 1	1 000 (10 000)	100	0.53	0.06	0.0155 ±0.0002	0.0847 ±0.0012	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	3.08	0.12	0.0151 ±0.0003	0.0822 ±0.0016	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	11.67	0.52	0.0132 ±0.0007	0.0694 ±0.0049	1.00 ±0.00	5000	{5000}
	10 000 (10 000)	100	46.22	0.55	0.0137 ±0.0007	0.0730 ±0.0046	1.00 ±0.00	10 000	{10000}
RC-SVM # ws = 2	1 000 (10 000)	100	0.37	0.14	0.0153 ±0.0002	0.0837 ±0.0015	2.00 ±0.00	500	{500}
	2 500 (10 000)	100	1.61	0.13	0.0152 ±0.0002	0.0830 ±0.0012	2.00 ±0.00	1250	{1250}
	5 000 (10 000)	100	6.12	0.22	0.0151 ±0.0002	0.0825 ±0.0014	2.00 ±0.00	2500	{2500}
	10 000 (10 000)	100	23.51	0.97	0.0130 ±0.0006	0.0680 ±0.0046	2.00 ±0.00	5000	{5000}
RC-SVM # ws = 5	1 000 (10 000)	100	0.46	0.22	0.0152 ±0.0002	0.0830 ±0.0014	5.00 ±0.00	200	{200}
	2 500 (10 000)	100	0.96	0.30	0.0150 ±0.0002	0.0815 ±0.0009	5.00 ±0.00	500	{500}
	5 000 (10 000)	100	2.70	0.28	0.0153 ±0.0001	0.0833 ±0.0008	5.00 ±0.00	1000	{1000}
	10 000 (10 000)	100	9.95	0.36	0.0152 ±0.0001	0.0826 ±0.0007	5.00 ±0.00	2000	{2000}
RC-SVM # ws = 10	1 000 (10 000)	100	0.88	0.29	0.0154 ±0.0002	0.0844 ±0.0011	10.00 ±0.00	100	{100}
	2 500 (10 000)	100	1.02	0.46	0.0148 ±0.0001	0.0808 ±0.0007	10.00 ±0.00	250	{250}
	5 000 (10 000)	100	1.96	0.58	0.0150 ±0.0001	0.0818 ±0.0008	10.00 ±0.00	500	{500}
	10 000 (10 000)	100	5.67	0.53	0.0152 ±0.0001	0.0827 ±0.0005	10.00 ±0.00	1000	{1000}
RC-SVM # ws = 15	1 000 (10 000)	100	1.50	0.34	0.0157 ±0.0002	0.0864 ±0.0014	15.00 ±0.00	67	[66 , 67]
	2 500 (10 000)	100	1.62	0.53	0.0149 ±0.0001	0.0814 ±0.0007	15.00 ±0.00	167	[166 , 167]
	5 000 (10 000)	100	2.10	0.84	0.0149 ±0.0001	0.0812 ±0.0006	15.00 ±0.00	333	[333 , 334]
	10 000 (10 000)	100	4.80	0.87	0.0150 ±0.0001	0.0819 ±0.0005	15.00 ±0.00	667	[666 , 667]
RC-SVM # ws = 20	1 000 (10 000)	100	1.88	0.39	0.0162 ±0.0003	0.0891 ±0.0017	20.00 ±0.00	50	{50}
	2 500 (10 000)	100	1.99	0.59	0.0151 ±0.0001	0.0822 ±0.0006	20.00 ±0.00	125	{125}
	5 000 (10 000)	100	2.27	0.89	0.0149 ±0.0001	0.0814 ±0.0006	20.00 ±0.00	250	{250}
	10 000 (10 000)	100	4.24	1.14	0.0149 ±0.0001	0.0812 ±0.0005	20.00 ±0.00	500	{500}
RC-SVM # ws = 40	1 000 (10 000)	100	3.96	0.59	0.0194 ±0.0007	0.1053 ±0.0033	40.00 ±0.00	25	{25}
	2 500 (10 000)	100	4.06	0.82	0.0157 ±0.0002	0.0859 ±0.0010	40.00 ±0.00	62	[62 , 63]
	5 000 (10 000)	100	4.29	1.17	0.0152 ±0.0001	0.0829 ±0.0005	40.00 ±0.00	125	{125}
	10 000 (10 000)	100	4.93	1.75	0.0148 ±0.0001	0.0808 ±0.0004	40.00 ±0.00	250	{250}
RC-SVM # ws = 50	1 000 (10 000)	100	4.96	0.69	0.0222 ±0.0011	0.1181 ±0.0045	50.00 ±0.00	20	{20}
	2 500 (10 000)	100	5.09	0.93	0.0161 ±0.0002	0.0883 ±0.0011	50.00 ±0.00	50	{50}
	5 000 (10 000)	100	5.30	1.29	0.0153 ±0.0001	0.0839 ±0.0005	50.00 ±0.00	100	{100}
	10 000 (10 000)	100	6.43	1.94	0.0149 ±0.0001	0.0810 ±0.0004	50.00 ±0.00	200	{200}

Table 10: Experimental results relating to the artificial data of Type IV

	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
LS-SVM	1 000 (10 000)	100	0.53	0.05	0.0649 ±0.0004	0.0370 ±0.0055	1.00		
	2 500 (10 000)	100	3.26	0.09	0.0647 ±0.0002	0.0330 ±0.0032	1.00		
	5 000 (10 000)	100	12.51	0.15	0.0640 ±0.0001	0.0240 ±0.0015	1.00		
	10 000 (10 000)	100	49.29	0.28	0.0640 ±0.0001	0.0223 ±0.0015	1.00		

	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
VP-SVM $r = 0.25$	1 000 (10 000)	100	4.85	0.66	0.0797 ±0.0021	0.1258 ±0.0076	44.56 ±2.07	22	[6 , 54]
	2 500 (10 000)	100	4.76	0.77	0.0719 ±0.0009	0.0882 ±0.0043	46.18 ±2.55	54	[14 , 116]
	5 000 (10 000)	100	4.88	0.95	0.0682 ±0.0005	0.0663 ±0.0033	47.45 ±2.61	105	[20 , 214]
	10 000 (10 000)	100	5.77	1.14	0.0657 ±0.0003	0.0460 ±0.0027	48.05 ±2.78	208	[38 , 407]
VP-SVM $r = 0.5$	1 000 (10 000)	100	1.65	0.30	0.0703 ±0.0011	0.0820 ±0.0068	14.98 ±1.44	67	[16 , 129]
	2 500 (10 000)	100	1.64	0.37	0.0670 ±0.0006	0.0565 ±0.0045	15.17 ±1.52	165	[40 , 334]
	5 000 (10 000)	100	2.26	0.45	0.0652 ±0.0003	0.0404 ±0.0035	15.69 ±1.47	319	[74 , 644]
	10 000 (10 000)	100	4.81	0.62	0.0645 ±0.0002	0.0307 ±0.0023	15.83 ±1.49	632	[167 , 1343]
VP-SVM $r = 1$	1 000 (10 000)	100	0.56	0.16	0.0670 ±0.0008	0.0578 ±0.0064	5.33 ±0.47	188	[68 , 464]
	2 500 (10 000)	100	1.05	0.19	0.0656 ±0.0004	0.0431 ±0.0042	5.33 ±0.47	469	[171 , 1162]
	5 000 (10 000)	100	3.01	0.26	0.0643 ±0.0002	0.0290 ±0.0025	5.42 ±0.52	923	[297 , 2445]
	10 000 (10 000)	100	10.81	0.41	0.0641 ±0.0001	0.0243 ±0.0022	5.54 ±0.61	1805	[555 , 4840]
VP-SVM $r = 2$	1 000 (10 000)	100	0.49	0.13	0.0653 ±0.0006	0.0427 ±0.0061	1.65 ±0.48	606	[266 , 1000]
	2 500 (10 000)	100	2.21	0.11	0.0649 ±0.0003	0.0357 ±0.0042	1.70 ±0.48	1471	[603 , 2500]
	5 000 (10 000)	100	9.32	0.18	0.0641 ±0.0001	0.0248 ±0.0018	1.54 ±0.50	3247	[1348 , 5000]
	10 000 (10 000)	100	35.16	0.31	0.0640 ±0.0001	0.0227 ±0.0018	1.62 ±0.49	6173	[2608 , 10000]
VP-SVM $r = 3$	1 000 (10 000)	100	0.54	0.07	0.0649 ±0.0004	0.0371 ±0.0055	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	3.25	0.09	0.0647 ±0.0002	0.0331 ±0.0032	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	12.54	0.16	0.0640 ±0.0001	0.0240 ±0.0015	1.00 ±0.00	5000	{5000}
	10 000 (10 000)	100	49.35	0.29	0.0640 ±0.0001	0.0223 ±0.0015	1.00 ±0.00	10 000	{10000}

	data set sizes	runs	train time	test time	test error	L_2-error	# of ws	ws size: median	ws size: range
RC-SVM # ws = 1	1 000 (10 000)	100	0.53	0.07	0.0649 ±0.0004	0.0373 ±0.0056	1.00 ±0.00	1000	{1000}
	2 500 (10 000)	100	3.27	0.09	0.0647 ±0.0002	0.0331 ±0.0033	1.00 ±0.00	2500	{2500}
	5 000 (10 000)	100	12.55	0.16	0.0640 ±0.0001	0.0240 ±0.0015	1.00 ±0.00	5000	{5000}
	10 000 (10 000)	100	50.04	0.29	0.0640 ±0.0001	0.0223 ±0.0015	1.00 ±0.00	10 000	{10000}
RC-SVM # ws = 2	1 000 (10 000)	100	0.38	0.09	0.0652 ±0.0005	0.0409 ±0.0056	2.00 ±0.00	500	{500}
	2 500 (10 000)	100	1.71	0.12	0.0649 ±0.0003	0.0352 ±0.0037	2.00 ±0.00	1250	{1250}
	5 000 (10 000)	100	6.39	0.18	0.0640 ±0.0001	0.0242 ±0.0019	2.00 ±0.00	2500	{2500}
	10 000 (10 000)	100	25.47	0.30	0.0640 ±0.0001	0.0221 ±0.0014	2.00 ±0.00	5000	{5000}
RC-SVM # ws = 5	1 000 (10 000)	100	0.48	0.17	0.0655 ±0.0005	0.0440 ±0.0056	5.00 ±0.00	200	{200}
	2 500 (10 000)	100	1.02	0.19	0.0654 ±0.0004	0.0426 ±0.0044	5.00 ±0.00	500	{500}
	5 000 (10 000)	100	2.82	0.24	0.0643 ±0.0002	0.0295 ±0.0026	5.00 ±0.00	1000	{1000}
	10 000 (10 000)	100	10.63	0.37	0.0641 ±0.0001	0.0248 ±0.0017	5.00 ±0.00	2000	{2000}
RC-SVM # ws = 10	1 000 (10 000)	100	0.85	0.25	0.0657 ±0.0005	0.0471 ±0.0054	10.00 ±0.00	100	{100}
	2 500 (10 000)	100	1.05	0.33	0.0656 ±0.0004	0.0448 ±0.0041	10.00 ±0.00	250	{250}
	5 000 (10 000)	100	2.04	0.35	0.0646 ±0.0002	0.0335 ±0.0032	10.00 ±0.00	500	{500}
	10 000 (10 000)	100	6.34	0.51	0.0644 ±0.0001	0.0299 ±0.0019	10.00 ±0.00	1000	{1000}
RC-SVM # ws = 15	1 000 (10 000)	100	1.47	0.32	0.0664 ±0.0006	0.0537 ±0.0059	15.00 ±0.00	67	[66 , 67]
	2 500 (10 000)	100	1.63	0.43	0.0656 ±0.0004	0.0449 ±0.0040	15.00 ±0.00	167	[166 , 167]
	5 000 (10 000)	100	2.18	0.49	0.0647 ±0.0002	0.0350 ±0.0030	15.00 ±0.00	333	[333 , 334]
	10 000 (10 000)	100	5.23	0.58	0.0646 ±0.0001	0.0335 ±0.0022	15.00 ±0.00	667	[666 , 667]
RC-SVM # ws = 20	1 000 (10 000)	100	1.92	0.38	0.0671 ±0.0007	0.0601 ±0.0054	20.00 ±0.00	50	{50}
	2 500 (10 000)	100	2.00	0.52	0.0659 ±0.0004	0.0473 ±0.0037	20.00 ±0.00	125	{125}
	5 000 (10 000)	100	2.32	0.61	0.0647 ±0.0002	0.0356 ±0.0029	20.00 ±0.00	250	{250}
	10 000 (10 000)	100	4.75	0.69	0.0647 ±0.0002	0.0353 ±0.0021	20.00 ±0.00	500	{500}
RC-SVM # ws = 40	1 000 (10 000)	100	3.89	0.60	0.0707 ±0.0011	0.0850 ±0.0065	40.00 ±0.00	25	{25}
	2 500 (10 000)	100	4.05	0.81	0.0669 ±0.0005	0.0577 ±0.0039	40.00 ±0.00	62	[62 , 63]
	5 000 (10 000)	100	4.47	1.00	0.0650 ±0.0002	0.0388 ±0.0030	40.00 ±0.00	125	{125}
	10 000 (10 000)	100	5.66	1.24	0.0649 ±0.0002	0.0379 ±0.0022	40.00 ±0.00	250	{250}
RC-SVM # ws = 50	1 000 (10 000)	100	4.70	0.69	0.0726 ±0.0010	0.0960 ±0.0051	50.00 ±0.00	20	{20}
	2 500 (10 000)	100	5.06	0.90	0.0675 ±0.0005	0.0627 ±0.0038	50.00 ±0.00	50	{50}
	5 000 (10 000)	100	5.33	1.15	0.0651 ±0.0002	0.0410 ±0.0027	50.00 ±0.00	100	{100}
	10 000 (10 000)	100	6.57	1.49	0.0649 ±0.0002	0.0383 ±0.0020	50.00 ±0.00	200	{200}

Table 11: Experimental results relating to the artificial data of Type V

Mona Eberts
Universität Stuttgart
Fachbereich Mathematik
Pfaffenwaldring 57
70569 Stuttgart
Germany
E-Mail: Mona.Eberts@mathematik.uni-stuttgart.de
WWW: www.mathematik.uni-stuttgart.de/fachbereich/mitarbeiter/visitenkarte.jsp?pid=729

Ingo Steinwart
Universität Stuttgart
Fachbereich Mathematik
Pfaffenwaldring 57
70569 Stuttgart
Germany
E-Mail: Ingo.Steinwart@mathematik.uni-stuttgart.de
WWW: www.isa.uni-stuttgart.de/Steinwart

Erschienene Preprints ab Nummer 2007/2007-001

Komplette Liste: `http://www.mathematik.uni-stuttgart.de/preprints`

2014-002 *Eberts, M.; Steinwart, I.:* Optimal Learning Rates for Localized SVMs

2014-001 *Giesselmann, J.:* A relative entropy approach to convergence of a low order approximation to a nonlinear elasticity model with viscosity and capillarity

2013-016 *Steinwart, I.:* Fully Adaptive Density-Based Clustering

2013-015 *Steinwart, I.:* Some Remarks on the Statistical Analysis of SVMs and Related Methods

2013-014 *Rohde, C.; Zeiler, C.:* A Relaxation Riemann Solver for Compressible Two-Phase Flow with Phase Transition and Surface Tension

2013-013 *Moroianu, A.; Semmelmann, U.:* Generalized Killling spinors on Einstein manifolds

2013-012 *Moroianu, A.; Semmelmann, U.:* Generalized Killing Spinors on Spheres

2013-011 *Kohls, K; Rösch, A.; Siebert, K.G.:* Convergence of Adaptive Finite Elements for Control Constrained Optimal Control Problems

2013-010 *Corli, A.; Rohde, C.; Schleper, V.:* Parabolic Approximations of Diffusive-Dispersive Equations

2013-009 *Nava-Yazdani, E.; Polthier, K.:* De Casteljau's Algorithm on Manifolds

2013-008 *Bächle, A.; Margolis, L.:* Rational conjugacy of torsion units in integral group rings of non-solvable groups

2013-007 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups over composition algebras

2013-006 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups, semifields, and translation planes

2013-005 *Eck, C.; Kutter, M.; Sändig, A.-M.; Rohde, C.:* A Two Scale Model for Liquid Phase Epitaxy with Elasticity: An Iterative Procedure

2013-004 *Griesemer, M.; Wellig, D.:* The Strong-Coupling Polaron in Electromagnetic Fields

2013-003 *Kabil, B.; Rohde, C.:* The Influence of Surface Tension and Configurational Forces on the Stability of Liquid-Vapor Interfaces

2013-002 *Devroye, L.; Ferrario, P.G.; Györfi, L.; Walk, H.:* Strong universal consistent estimate of the minimum mean squared error

2013-001 *Kohls, K.; Rösch, A.; Siebert, K.G.:* A Posteriori Error Analysis of Optimal Control Problems with Control Constraints

2012-018 *Kimmerle, W.; Konovalov, A.:* On the Prime Graph of the Unit Group of Integral Group Rings of Finite Groups II

2012-017 *Stroppel, B.; Stroppel, M.:* Desargues, Doily, Dualities, and Exceptional Isomorphisms

2012-016 *Moroianu, A.; Pilca, M.; Semmelmann, U.:* Homogeneous almost quaternion-Hermitian manifolds

2012-015 *Steinke, G.F.; Stroppel, M.J.:* Simple groups acting two-transitively on the set of generators of a finite elation Laguerre plane

2012-014 *Steinke, G.F.; Stroppel, M.J.:* Finite elation Laguerre planes admitting a two-transitive group on their set of generators

2012-013 *Diaz Ramos, J.C.; Dominguez Vázquez, M.; Kollross, A.:* Polar actions on complex hyperbolic spaces

2012-012 *Moroianu; A.; Semmelmann, U.:* Weakly complex homogeneous spaces

2012-011 *Moroianu; A.; Semmelmann, U.:* Invariant four-forms and symmetric pairs

2012-010 *Hamilton, M.J.D.:* The closure of the symplectic cone of elliptic surfaces

2012-009 *Hamilton, M.J.D.:* Iterated fibre sums of algebraic Lefschetz fibrations

2012-008 *Hamilton, M.J.D.:* The minimal genus problem for elliptic surfaces

2012-007 *Ferrario, P.:* Partitioning estimation of local variance based on nearest neighbors under censoring

2012-006 *Stroppel, M.:* Buttons, Holes and Loops of String: Lacing the Doily

2012-005 *Hantsch, F.:* Existence of Minimizers in Restricted Hartree-Fock Theory

2012-004 *Grundhöfer, T.; Stroppel, M.; Van Maldeghem, H.:* Unitals admitting all translations

2012-003 *Hamilton, M.J.D.:* Representing homology classes by symplectic surfaces

2012-002 *Hamilton, M.J.D.:* On certain exotic 4-manifolds of Akhmedov and Park

2012-001 *Jentsch, T.:* Parallel submanifolds of the real 2-Grassmannian

2011-028 *Spreer, J.:* Combinatorial 3-manifolds with cyclic automorphism group

2011-027 *Griesemer, M.; Hantsch, F.; Wellig, D.:* On the Magnetic Pekar Functional and the Existence of Bipolarons

2011-026 *Müller, S.:* Bootstrapping for Bandwidth Selection in Functional Data Regression

2011-025 *Felber, T.; Jones, D.; Kohler, M.; Walk, H.:* Weakly universally consistent static forecasting of stationary and ergodic time series via local averaging and least squares estimates

2011-024 *Jones, D.; Kohler, M.; Walk, H.:* Weakly universally consistent forecasting of stationary and ergodic time series

2011-023 *Györfi, L.; Walk, H.:* Strongly consistent nonparametric tests of conditional independence

2011-022 *Ferrario, P.G.; Walk, H.:* Nonparametric partitioning estimation of residual and local variance based on first and second nearest neighbors

2011-021 *Eberts, M.; Steinwart, I.:* Optimal regression rates for SVMs using Gaussian kernels

2011-020 *Frank, R.L.; Geisinger, L.:* Refined Semiclassical Asymptotics for Fractional Powers of the Laplace Operator

2011-019 *Frank, R.L.; Geisinger, L.:* Two-term spectral asymptotics for the Dirichlet Laplacian on a bounded domain

2011-018 *Hänel, A.; Schulz, C.; Wirth, J.:* Embedded eigenvalues for the elastic strip with cracks

2011-017 *Wirth, J.:* Thermo-elasticity for anisotropic media in higher dimensions

2011-016 *Höllig, K.; Hörner, J.:* Programming Multigrid Methods with B-Splines

2011-015 *Ferrario, P.:* Nonparametric Local Averaging Estimation of the Local Variance Function

2011-014 *Müller, S.; Dippon, J.:* k-NN Kernel Estimate for Nonparametric Functional Regression in Time Series Analysis

2011-013 *Knarr, N.; Stroppel, M.:* Unitals over composition algebras

2011-012 *Knarr, N.; Stroppel, M.:* Baer involutions and polarities in Moufang planes of characteristic two

2011-011 *Knarr, N.; Stroppel, M.:* Polarities and planar collineations of Moufang planes

2011-010 *Jentsch, T.; Moroianu, A.; Semmelmann, U.:* Extrinsic hyperspheres in manifolds with special holonomy

2011-009 *Wirth, J.:* Asymptotic Behaviour of Solutions to Hyperbolic Partial Differential Equations

2011-008 *Stroppel, M.:* Orthogonal polar spaces and unitals

2011-007 *Nagl, M.:* Charakterisierung der Symmetrischen Gruppen durch ihre komplexe Gruppenalgebra

2011-006 *Solanes, G.; Teufel, E.:* Horo-tightness and total (absolute) curvatures in hyperbolic spaces

2011-005 *Ginoux, N.; Semmelmann, U.:* Imaginary Kählerian Killing spinors I

2011-004 *Scherer, C.W.; Köse, I.E.:* Control Synthesis using Dynamic D-Scales: Part II — Gain-Scheduled Control

2011-003 *Scherer, C.W.; Köse, I.E.:* Control Synthesis using Dynamic D-Scales: Part I — Robust Control

2011-002 *Alexandrov, B.; Semmelmann, U.:* Deformations of nearly parallel G_2-structures

2011-001 *Geisinger, L.; Weidl, T.:* Sharp spectral estimates in domains of infinite volume

2010-018 *Kimmerle, W.; Konovalov, A.:* On integral-like units of modular group rings

2010-017 *Gauduchon, P.; Moroianu, A.; Semmelmann, U.:* Almost complex structures on quaternion-Kähler manifolds and inner symmetric spaces

2010-016 *Moroianu, A.; Semmelmann,U.:* Clifford structures on Riemannian manifolds

2010-015 *Grafarend, E.W.; Kühnel, W.:* A minimal atlas for the rotation group $SO(3)$

2010-014 *Weidl, T.:* Semiclassical Spectral Bounds and Beyond

2010-013 *Stroppel, M.:* Early explicit examples of non-desarguesian plane geometries

2010-012 *Effenberger, F.:* Stacked polytopes and tight triangulations of manifolds

2010-011 *Györfi, L.; Walk, H.:* Empirical portfolio selection strategies with proportional transaction costs

2010-010 *Kohler, M.; Krzyżak, A.; Walk, H.:* Estimation of the essential supremum of a regression function

2010-009 *Geisinger, L.; Laptev, A.; Weidl, T.:* Geometrical Versions of improved Berezin-Li-Yau Inequalities

2010-008 *Poppitz, S.; Stroppel, M.:* Polarities of Schellhammer Planes

2010-007 *Grundhöfer, T.; Krinn, B.; Stroppel, M.:* Non-existence of isomorphisms between certain unitals

2010-006 *Höllig, K.; Hörner, J.; Hoffacker, A.:* Finite Element Analysis with B-Splines: Weighted and Isogeometric Methods

2010-005 *Kaltenbacher, B.; Walk, H.:* On convergence of local averaging regression function estimates for the regularization of inverse problems

2010-004 *Kühnel, W.; Solanes, G.:* Tight surfaces with boundary

2010-003 *Kohler, M; Walk, H.:* On optimal exercising of American options in discrete time for stationary and ergodic data

2010-002 *Gulde, M.; Stroppel, M.:* Stabilizers of Subspaces under Similitudes of the Klein Quadric, and Automorphisms of Heisenberg Algebras

2010-001 *Leitner, F.:* Examples of almost Einstein structures on products and in cohomogeneity one

2009-008 *Griesemer, M.; Zenk, H.:* On the atomic photoeffect in non-relativistic QED

2009-007 *Griesemer, M.; Moeller, J.S.:* Bounds on the minimal energy of translation invariant n-polaron systems

2009-006 *Demirel, S.; Harrell II, E.M.:* On semiclassical and universal inequalities for eigenvalues of quantum graphs

2009-005 *Bächle, A, Kimmerle, W.:* Torsion subgroups in integral group rings of finite groups

2009-004 *Geisinger, L.; Weidl, T.:* Universal bounds for traces of the Dirichlet Laplace operator

2009-003 *Walk, H.:* Strong laws of large numbers and nonparametric estimation

2009-002 *Leitner, F.:* The collapsing sphere product of Poincaré-Einstein spaces

2009-001 *Brehm, U.; Kühnel, W.:* Lattice triangulations of E^3 and of the 3-torus

2008-006 *Kohler, M.; Krzyżak, A.; Walk, H.:* Upper bounds for Bermudan options on Markovian data using nonparametric regression and a reduced number of nested Monte Carlo steps

2008-005 *Kaltenbacher, B.; Schöpfer, F.; Schuster, T.:* Iterative methods for nonlinear ill-posed problems in Banach spaces: convergence and applications to parameter identification problems

2008-004 *Leitner, F.:* Conformally closed Poincaré-Einstein metrics with intersecting scale singularities

2008-003 *Effenberger, F.; Kühnel, W.:* Hamiltonian submanifolds of regular polytope

2008-002 *Hertweck, M.; Höfert, C.R.; Kimmerle, W.:* Finite groups of units and their composition factors in the integral group rings of the groups $PSL(2,q)$

2008-001 *Kovarik, H.; Vugalter, S.; Weidl, T.:* Two dimensional Berezin-Li-Yau inequalities with a correction term

2007-006 *Weidl, T.:* Improved Berezin-Li-Yau inequalities with a remainder term

2007-005 *Frank, R.L.; Loss, M.; Weidl, T.:* Polya's conjecture in the presence of a constant magnetic field

2007-004 *Ekholm, T.; Frank, R.L.; Kovarik, H.:* Eigenvalue estimates for Schrödinger operators on metric trees

2007-003 *Lesky, P.H.; Racke, R.:* Elastic and electro-magnetic waves in infinite waveguides

2007-002 *Teufel, E.:* Spherical transforms and Radon transforms in Moebius geometry

2007-001 *Meister, A.:* Deconvolution from Fourier-oscillating error densities under decay and smoothness restrictions