Electronic Journal of Statistics
Vol. 7 (2013) 1–42
ISSN: 1935-7524
DOI: 10.1214/12-EJS760

Optimal regression rates for SVMs using Gaussian kernels

Mona Eberts and Ingo Steinwart

Institute for Stochastics and Applications
University of Stuttgart
Pfaffenwaldring 57
D-70569 Stuttgart
e-mail: mona.eberts@mathematik.uni-stuttgart.de
ingo.steinwart@mathematik.uni-stuttgart.de

Abstract: Support vector machines (SVMs) using Gaussian kernels are one of the standard and state-of-the-art learning algorithms. In this work, we establish new oracle inequalities for such SVMs when applied to either least squares or conditional quantile regression. With the help of these oracle inequalities we then derive learning rates that are (essentially) minmax optimal under standard smoothness assumptions on the target function. We further utilize the oracle inequalities to show that these learning rates can be adaptively achieved by a simple data-dependent parameter selection method that splits the data set into a training and a validation set.

AMS 2000 subject classifications: Primary 62G08; secondary 62G05, 68Q32, 68T05.
Keywords and phrases: Least squares regression, quantile estimation, support vector machines.

Received October 2011.

Contents

1. Introduction

Given i.i.d. observations $D := ((x_1, y_1), \ldots, (x_n, y_n))$ of input/output observations drawn from an unknown distribution P on $X \times Y$, where $Y \subset \mathbb{R}$, the goal of non-parametric regression is to find a function $f_D : X \to \mathbb{R}$ that captures important characteristics of the conditional distribution $\mathrm{P}(Y|x)$, $x \in X$. For example,

in non-parametric least squares regression, an f_D is sought that approximates the conditional mean $\mathbb{E}(Y|x)$, while in quantile regression the goal is to find an estimate f_D of the quantiles of $\mathrm{P}(Y|x)$, $x \in X$. Non-parametric least squares regression is one of the classical non-parametric problems, which has been extensively studied for decades. We refer to the book [16], which presents many results in this direction. In contrast, the non-parametric quantile regression problem has attracted less attention, probably because for more advanced estimation procedures, or learning algorithms, the problem is often less tractable, both mathematically and algorithmically. Nonetheless, also for this problem important contributions have been made, which, besides other questions regarding quantile regression, are summarized in the recent book [19].

A typical way to assess the quality of a found estimator f_D in these regression problems is the distance of f_D to the target function. To be more precise, if f^* denotes the conditional function of interest, that is, either the conditional mean or a conditional quantile, and P is the marginal distribution of P on X, then, for some $p \in (0, \infty)$, the norm

$$\|f_D - f^*\|^p_{L_p(\mathrm{P}_X)}\,, \tag{1}$$

is often used to describe how well f_D approximates f^*. Here we note, that taking the p-th power of the norm is, of course, not dictated by mathematics but more by historically grown habits for the least squares loss. Recall that, for least squares regression, one usually considers $p = 2$ due to the very nature of the least squares loss, while for quantile regression various values for p have actually been considered. In both cases, we say the learning algorithm that produces the estimates f_D is consistent, if the norm in (1) converges to 0 in probability for $n \to \infty$. Likewise, learning rates describe the corresponding convergence rates, either in probability or in expectation.

One of the learning algorithms that have recently attracted many theoretical investigations are support vector machines (SVMs), or more precisely, kernel-based regularized empirical risk minimizers. Reasons for this grown interest include their state-of-the-art empirical performance in applications, their relatively simple implementation and application, and last-but-not-least, their flexibility. To describe this flexibility, which is key for considering the two regression scenarios simultaneously, let us briefly recall that SVMs solve an optimization problem of the form

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

where H is a reproducing kernel Hilbert space (RKHS) with reproducing kernel k, see e.g. [2, 4, 30], $\lambda > 0$ is a user-specified regularization parameter, $L : Y \times \mathbb{R} \to [0, \infty)$ is a loss function, and $\mathcal{R}_{L,\mathrm{D}}(f)$ denotes the empirical error or risk of a function $f : X \to \mathbb{R}$, that is

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

It is well-known that the optimization problem above has a unique solution whenever the loss L is convex in its second argument. In addition, under mild assumptions on the richness of H and the way the regularization parameter λ is chosen, the corresponding SVM is L-risk consistent. We refer to [30] for detailed descriptions of these and other results. Now, the above mentioned flexibility of SVMs is made possible by their two main ingredients, namely the RKHS H and the loss function L. To be more precise, the loss function can be used to model the learning target, see [30, Chapter 3], while the RKHS can be used to adapt to the nature of the input domain X. For example, when using the standard least squares loss in the optimization problem (2), the SVM estimates the conditional mean, and for the so-called pinball loss, see Section 4 for a definition, the SVM estimates conditional quantiles. On the other hand, RKHSs can be defined on arbitrary input domains X, so that, besides standard $\mathbb{R}^d$-valued data, various other types of data can be dealt with. Moreover, due to the so-called kernel-trick [24], the choice of H has little to no algorithmic consequences for solving the SVM optimization problem. The latter is not true for the choice of L, where each different L demands a different optimization algorithm. However, for standard loss functions including the least-squares loss and the pinball loss, these optimization problems, which reduce to convex quadratic optimization problems, have been well-understood. For solvers, we exemplarily refer to [9, 18] and [35], respectively.

One of the main topics in recent theoretical investigations on SVMs have been learning rates. For example, the articles [10, 11, 27, 5, 22, 32] and the references therein establish rates for SVMs using the least squares loss, while SVMs using the pinball loss are investigated in [29, 31]. We discuss the findings of these articles and compare them to our results in more detail at the end of Sections 3 and 4 after we have presented our main results. Here, we only note that besides a very few articles, namely [5, 22, 32], the obtained learning rates are typically not optimal in a minmax sense. In addition, these three papers only consider some specific cases. For example, [5] only considers the case, when the target function, in this case the conditional mean, is contained in the used RKHS H. On the other hand, H is assumed to be generic in this article, that is, no specific family of kernels is considered. The latter generality is also adopted in [22, 32], where the authors establish optimal rates in the more realistic case in which H does not contain the target function. Unfortunately, however, these articles require additional assumptions on the interplay between H and the marginal distribution P_X. Namely, [22] assumes that the eigenfunctions of the integral operator associated to the kernel k of H are (almost) uniformly bounded. This assumption, however, cannot be easily guaranteed, neither in practice nor in theory. This issue is partially addressed in [32], where the eigenfunction assumption is replaced by a weaker assumption in terms of inclusions of certain interpolation spaces of H and $L_2(\mathrm{P}_X)$. While in practice, these inclusions can not be checked either, there are, at least, certain combinations of H and $L_2(\mathrm{P}_X)$ in which they are satisfied. For example, if $X \subset \mathbb{R}^d$ is a bounded domain satisfying some standard regularity assumptions and H is a Sobolev space $W_2^m(X)$ of sufficient smoothness m, that is $m > d/2$, then [32] shows that the inclusion assumptions

made in this article are satisfied and that the resulting learning rates for SVMs are minmax optimal. While this result is interesting from a theoretical point of view, in practice Sobolev spaces of large order m are rarely used for SVMs, probably because of computational issues.

The discussion so far may already indicate the fact that most articles, including the three establishing optimal rates, only consider the case, where H is *fixed* during the training process. This scenario, however, is rather unrealistic, since in most applications, H is chosen in a data-dependent way. For example, for input domains $X \subset \mathbb{R}^d$, the standard way of using SVMs is to equip them with Gaussian RBF kernels k_γ defined by

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

and to determine the free width parameter $\gamma > 0$ in a data-dependent way, e.g., by cross-validation. Despite the dominance of this approach, however, only a very few articles analyze the learning behaviour of SVMs with Gaussian kernels. To be more concrete, the currently best learning rates have been established in [33, 40]. Here we note that in both articles the authors actually consider binary classification, although a closer look reveals that at least the results of [40] can also be applied to least squares regression. Indeed, if the conditional mean is assumed to be contained in the Sobolev space $W_2^s(X)$ for some $s > 0$, then [40] establishes rates of the form

$$n^{-\frac{s}{s+2d+2}}.$$

Unfortunately, these rates are far from the known minmax rates $n^{-\frac{2s}{2s+d}}$ of this setting, and up to now, it has been unknown, whether SVMs with Gaussian kernels can actually achieve these minmax rates, as their good empirical performance may suggest, or whether they can only learn with sub-optimal rates like classical kernel rules with Gaussian kernels do. The first goal of this paper is to answer this question. More precisely, we show that SVMs with least squares loss and Gaussian kernels can learn with rate

$$n^{-\frac{2s}{2s+d}+\xi} \tag{3}$$

for all $\xi > 0$. In other words, we establish learning rates that are arbitrarily close to the minmax rates. Moreover, we show that these rates can be achieved by a simple but completely data-driven procedure that splits the data set D into a training and a validation set. Our second goal is to show that these rates as well as the adaptivity to the unknown smoothness s is preserved when considering quantile regression, instead. More precisely, we show under mild additional assumptions on the conditional distributions that the conditional quantile functions f^* are approximated by SVM decision functions in the L_2-norm (1) with rate (3). Moreover, it turns out that splitting D into a training and validation set again leads to a learning procedure that is fully adaptive to the unknown smoothness s.

In the remainder of this section we introduce some assumptions and notations used throughout the paper. Except a passage in Section 3, where we discuss the case of unbounded noise, we mainly consider the case of bounded regression. Thus, we begin with the probability measure P on $\mathbb{R}^d \times Y$ where $Y := [-M, M]$ for some $M > 0$. We further assume that $X := \operatorname{supp} \mathrm{P}_X \subset B_{\ell_2^d}$ and that the marginal distribution P_X on X is absolutely continuous with respect to the Lebesgue measure μ on X. In addition, the corresponding density of P_X is assumed to be contained in $L_q(X)$ for some $q \in [1, \infty]$, where $L_q(\nu)$ is the Lebesgue spaces of order q w.r.t. the measure ν and for the Lebesgue measure μ on $X \subset \mathbb{R}^d$ we define $L_q(X) := L_q(\mu)$.

Throughout the paper we further assume that the boundary of X has zero Lebesgue measure. Note that if this assumption is satisfied, the distribution on X can be identified with the distribution on the interior and the closure of X, since it has a Lebesgue density on X. Hence, we will not distinguish between these distributions in terms of notation. Similarly, we often view the distribution on X as a probability measure defined on $\mathbb{R}^d$ rather than on X.

Since we consider both least squares regression and quantile regression, it is helpful to consider some concepts in a generic way. To this end, we say that a function $L : Y \times \mathbb{R} \to [0, \infty)$ is a loss function, if it is measurable. In the following, L will be either the least squares loss or the pinball loss introduced in Section 4. Moreover, for a measurable $f : X \to \mathbb{R}$, the L-risk is defined by

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

and the Bayes L-risk is the smallest possible L-risk, that is

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

Since P lives on $X \times [-M, M]$, both the conditional mean and the conditional quantiles are $[-M, M]$-valued on X. It therefore suffices to consider estimators of these quantities that are $[-M, M]$-valued on X. To make this precise, we denote the clipped value of some $t \in \mathbb{R}$ by $\widehat{t}$, that is

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

It is easy to check that the risks of both the least squares loss and the pinball loss satisfy

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) \leq \mathcal{R}_{L,\mathrm{P}}(f) \ ,$$

for all $f : X \to \mathbb{R}$. In other words, clipping the decision functions at $\pm M$ does not increase the L-risk, and hence we will always consider clipped versions of the SVM decision functions. Finally, since we do not consider SVMs with a fixed

kernel, a notation that is slightly more detailed than (2) is helpful. Namely, if H_γ is the RKHS of the Gaussian RBF kernel k_γ, then we write

$$f_{\mathrm{D},\lambda,\gamma} = \arg\min_{f\in H_\gamma} \lambda \|f\|_{H_\gamma}^2 + \mathcal{R}_{L,\mathrm{D}}(f) \ , \tag{4}$$

where again, L is one of the above loss functions. Note that projection (4) has already been used in the literature, see e.g. [8, 33, 39].

The rest of this paper is organized as follows: The next section presents some upper bounds on the approximation error of SVMs using Gaussian kernels. These bounds are then used to derive new oracle inequalities for the least squares loss and for the pinball loss in Sections 3 and 4, respectively. In these sections we also present and discuss the learning rates that result from these oracle inequalities. In particular, it turns out that the rates are (essentially) minmax optimal if the target function is contained in some Sobolev or Besov spaces. Section 5 finally presents, besides some technical lemmata, the proofs of our results.

2. Estimates on the approximation error

The main goal of this work is to derive new oracle inequalities and learning rates for SVMs with Gaussian kernels using the least squares loss (cf. Section 3) and the pinball loss (cf. Section 4), respectively. To this end, we need to describe the infinite sample behaviour for *fixed* regularization parameter λ and kernel width γ, i.e. we need to find an upper bound for the *approximation error function* $A_\gamma\colon [0,\infty) \to [0,\infty)$ defined by

$$A_\gamma(\lambda) := \inf_{f\in H_\gamma} \lambda\|f\|_{H_\gamma}^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \,, \tag{5}$$

where the infimum is actually attained by a unique element $f_{\mathrm{P},\lambda,\gamma} \in H_\gamma$, see [30, Lemma 5.1 and Theorem 5.2]. In this section we thus estimate the approximation error of some function contained in the RKHS H_γ.

Let us begin by introducing some function spaces that are later assumed to contain the target function. To this end, we first present some notations. We denote the closed unit ball of a Banach space E by B_E. In particular, for the d-dimensional Euclidean space ℓ_2^d, we write $B_{\ell_2^d}$. For $s \in \mathbb{R}$, $\lfloor s \rfloor$ is the greatest integer smaller or equal s and $\lceil s \rceil$ is the smallest integer greater or equal s.

Let us now introduce a first type of subspaces of $L_p(\nu)$, namely Sobolev spaces (cf. [1, Section 3] and [38, Sections 2 and 3]). To this end, let $\partial^{(\alpha)}$ be the α-th weak derivative for a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $|\alpha| = \sum_{i=1}^d \alpha_i$. Then, for an integer $m \geq 0$, $1 \leq p \leq \infty$, and a measure ν, the Sobolev space of order m w.r.t. ν is defined by

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

i.e. it is the space of all functions in $L_p(\nu)$, whose weak derivatives up to order m exist and are contained in $L_p(\nu)$. The Sobolev space is equipped with the Sobolev norm

$$\|f\|_{W_p^m(\nu)} := \left(\sum_{|\alpha| \leq m} \left\| \partial^{(\alpha)} f \right\|_{L_p(\nu)}^p \right)^{\frac{1}{p}} ,$$

(cf. [1, page 60]). Moreover, we write $W_p^0(\nu) = L_p(\nu)$ and, for the Lebesgue measure μ on $X \subset \mathbb{R}^d$, we define $W_p^m(X) := W_p^m(\mu)$.

In order to bound the excess risk in (5), it turns out that we need a finer scale of smoothness given by the Besov space scale. To recall these function spaces, we first have to introduce another device to measure the smoothness of functions, namely the modulus of smoothness (see e.g. [12, p. 44], [13, p. 398], and [3, p. 360]).

Definition 2.1. Let $X \subset \mathbb{R}^d$ be a subset with non-empty interior, ν be an arbitrary measure on X, and $f : X \to \mathbb{R}$ be a function with $f \in L_p(\nu)$ for some $p \in (0, \infty]$. For $r \in \mathbb{N}$, the r-th modulus of smoothness of f is defined by

$$\omega_{r, L_p(\nu)}(f, t) = \sup_{\|h\|_2 \leq t} \|\triangle_h^r(f, \cdot)\|_{L_p(\nu)} , \qquad t \geq 0 ,$$

where $\|\cdot\|_2$ denotes the Euclidean norm and the r-th difference $\triangle_h^r(f, \cdot)$ is defined by

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

for $h = (h_1, \ldots, h_d) \in [0, \infty)^d$ and $X_{r,h} := \{x \in X : x + sh \in X \text{ f.a. } s \in [0, r]\}$.

To illustrate the idea of the modulus of smoothness, let us consider the case $d = 1$ and $r = 1$. Then, we obtain

$$h^{-1} \triangle_h^1(f, x) = \frac{f(x + h) - f(x)}{h} \xrightarrow{h \to 0} f'(x) ,$$

if the derivative f' of f exists in x. Consequently, $h^{-1} \triangle_h^1(f, x)$ equals the secant's slope and is bounded, if f is differentiable at x. Analogously, $h^{-r} \triangle_h^r(f, x)$ is bounded, if, for example, second order derivatives exist.

It is well-known, see e.g. [17, Equation (2.1)], that the modulus of smoothness with respect to $L_p(\mathbb{R}^d)$ satisfies

$$\omega_{r, L_p(\mathbb{R}^d)}(f, t) \leq \left(1 + \frac{t}{s}\right)^r \omega_{r, L_p(\mathbb{R}^d)}(f, s) , \tag{6}$$

for all $f \in L_p(\mathbb{R}^d)$ and all $s > 0$. As already mentioned, the modulus of smoothness can be used to define the scale of Besov spaces (see e.g. [1, Section 7] and

[38, Sections 2 and 3]), which besides Sobolev spaces will later be assumed to contain the target function and thus indicate the smoothness of the latter. Namely, for $1 \leq p, q \leq \infty$, $\alpha > 0$, $r := \lfloor \alpha \rfloor + 1$, and an arbitrary measure ν, the Besov space $B^{\alpha}_{p,q}(\nu)$ is

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

where, for $1 \leq q < \infty$, the seminorm $|\cdot|_{B^{\alpha}_{p,q}(\nu)}$ is defined by

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

and, for $q = \infty$, it is defined by

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

In both cases, the norm of $B^{\alpha}_{p,q}(\nu)$ can be defined by $\|f\|_{B^{\alpha}_{p,q}(\nu)} := \|f\|_{L_p(\nu)} + |f|_{B^{\alpha}_{p,q}(\nu)}$, see e.g. [12, pp. 54/55] and [13, p. 398]. In addition, for $q = \infty$, we often write $\mathrm{Lip}^*(\alpha, L_p(\nu)) := B^{\alpha}_{p,\infty}(\nu)$ and call $\mathrm{Lip}^*(\alpha, L_p(\nu))$ the generalized Lipschitz space of order α. Finally, if ν is the Lebesgue measure on X, we write $B^{\alpha}_{p,q}(X) := B^{\alpha}_{p,q}(\nu)$.

It is well-known, see e.g. [15, p. 25 and p. 44], that the Sobolev spaces $W^{\alpha}_p(\mathbb{R}^d)$ fall into the scale of Besov spaces, namely

$$W^{\alpha}_p(\mathbb{R}^d) \subset B^{\alpha}_{p,q}(\mathbb{R}^d) \tag{7}$$

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.

As alluded at the beginning of this section, our main goal is to derive new oracle inequalities for SVMs and for this purpose we need to estimate the approximation error (5). In order to cope with this task, we already introduced all necessary concepts in the previous part of this section. It remains to show how they can be applied. Now, to bound (5), it suffices to find a function $f_0 \in H_\gamma$ such that both the regularization term $\lambda \|f_0\|^2_{H_\gamma}$ and the excess risk $\mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}$ are small. To construct this function f_0 we define, for $r \in \mathbb{N}$ and $\gamma > 0$, the function $K : \mathbb{R}^d \to \mathbb{R}$ by

$$K(x) := \sum_{j=1}^{r} \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2 \pi} \right)^{\frac{d}{2}} K_{\frac{j\gamma}{\sqrt{2}}}(x) , \tag{8}$$

where $K_\gamma(x) := \exp\left(-\gamma^{-2} \|x\|_2^2\right)$ for all $x \in \mathbb{R}^d$. Let us now assume that there exists a function $f^*_{L,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ such that $f^*_{L,\mathrm{P}}$ is a Bayes decision function, i.e. $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$, and such that $f^*_{L,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$. Then we define f_0 by convolving K with this Bayes decision function, that is

$$f_0(x) := K * f^*_{L,\mathrm{P}}(x) := \int_{\mathbb{R}^d} K(x-t) f^*_{L,\mathrm{P}}(t)\, dt , \qquad x \in \mathbb{R}^d . \tag{9}$$

Now, to show that f_0 is indeed a suitable function to bound (5), we first need to ensure that f_0 is contained in H_γ. In addition, we need to derive bounds for both, the H_γ-norm and the excess risk of f_0. Theorem 2.2 concentrates on the latter with the help of the modulus of smoothness, while Theorem 2.3 estimates the regularization term.

Theorem 2.2. *Let us fix some $q \in [1,\infty)$. Furthermore, assume that P_X is a distribution on $\mathbb{R}^d$ that has a Lebesgue density $g \in L_p(\mathbb{R}^d)$ for some $p \in [1,\infty]$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be such that $f \in L_q\left(\mathbb{R}^d\right) \cap L_\infty\left(\mathbb{R}^d\right)$. Then, for $r \in \mathbb{N}$, $\gamma > 0$, and $s \geq 1$ with $1 = \frac{1}{s} + \frac{1}{p}$, we have*

$$\|K * f - f\|^q_{L_q(\mathrm{P}_X)} \leq C_{r,q} \|g\|_{L_p(\mathbb{R}^d)} \, \omega^q_{r,L_{qs}(\mathbb{R}^d)}\left(f, \gamma/2\right) \ ,$$

where $C_{r,q}$ is a constant only depending on r and q.

The next result will be used to bound the regularization term and to prove that the convolution of a function from $L_2(\mathbb{R}^d)$ with K is contained in the RKHS H_γ. In addition, it provides a very useful supremum bound.

Theorem 2.3. *Let $f \in L_2\left(\mathbb{R}^d\right)$, H_γ be the RKHS of the Gaussian RBF kernel k_γ over $X \subset \mathbb{R}^d$ with $\gamma > 0$ and $K : \mathbb{R}^d \to \mathbb{R}$ be defined by (8) for a fixed $r \in \mathbb{N}$. Then we have $K * f \in H_\gamma$ with*

$$\|K * f\|_{H_\gamma} \leq \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} \left(2^r - 1\right) \|f\|_{L_2(\mathbb{R}^d)} \ .$$

Moreover, if $f \in L_\infty\left(\mathbb{R}^d\right)$, we have

$$|K * f\,(x)| \leq \left(2^r - 1\right) \|f\|_{L_\infty(\mathbb{R}^d)} \ , \qquad x \in X \ .$$

To illustrate the theorems above, let us now consider f_0 defined by (9), where $f^*_{L,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ is a Bayes decision function contained in $L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$. For the sake of simplicity, we fix $q = 2$ and $p, s \geq 1$ with $\frac{1}{p} + \frac{1}{s} = 1$, i.e. the Lebesgue density g of P_X has to be contained in $L_p(X)$. Then is turns out that together the two theorems yield

$$\begin{aligned}
&\min_{f \in H_\gamma} \lambda \|f\|^2_{H_\gamma} + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq \lambda \|f_0\|^2_{H_\gamma} + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}} \\
&= \lambda \|K * f^*_{L,\mathrm{P}}\|^2_{H_\gamma} + \mathcal{R}_{L,\mathrm{P}}(K * f^*_{L,\mathrm{P}}) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq \lambda (\gamma\sqrt{\pi})^{-d} (2^r - 1)^2 \left\|f^*_{L,\mathrm{P}}\right\|^2_{L_2(\mathbb{R}^d)} + c \|K * f^*_{L,\mathrm{P}} - f^*_{L,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)} \\
&\leq \lambda (\gamma\sqrt{\pi})^{-d} (2^r - 1)^2 \left\|f^*_{L,\mathrm{P}}\right\|^2_{L_2(\mathbb{R}^d)} + c\, C_{r,2} \|g\|_{L_p(\mathbb{R}^d)} \, \omega^2_{r,L_{2s}(\mathbb{R}^d)}\left(f^*_{L,\mathrm{P}}, \gamma/2\right) , \quad (10)
\end{aligned}$$

where the crucial intermediate estimate

$$\mathcal{R}_{L,\mathrm{P}}(K * f^*_{L,\mathrm{P}}) - \mathcal{R}^*_{L,\mathrm{P}} \leq c \|K * f^*_{L,\mathrm{P}} - f^*_{L,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)}$$

will be discussed in Sections 3 and 4, respectively. Now, to further bound (10), we have to estimate the modulus of smoothness. To this end, recall that for $f^*_{L,\mathrm{P}} \in B^\alpha_{2s,\infty}(\mathbb{R}^d)$, or $f^*_{L,\mathrm{P}} \in W^\alpha_{2s}(\mathbb{R}^d)$ by (7), we have

$$\omega_{r,L_{2s}(\mathbb{R}^d)}\left(f^*_{L,\mathrm{P}}, t\right) \leq c_s\, t^\alpha\,, \qquad\qquad t > 0\,,$$

where $r := \lfloor \alpha \rfloor + 1$ and $c_s > 0$ is a constant. Using this inequality the upper bound of the approximation error only depends on the kernel width γ, the regularization parameter λ, the smoothness parameter α of the target function and some positive constants, i.e.

$$\min_{f \in H_\gamma} \lambda \|f\|^2_{H_\gamma} + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \leq c_1 \lambda \gamma^{-d} + c_2 \gamma^{2\alpha}\,. \tag{11}$$

At first sight, it seems surprising that starting from (10) the parameters s and p do not appear in (11) any more, but it has to be pointed out that they affect the constant $c_2 > 0$. Moreover, note that Theorem 2.3 also implies the estimate

$$\|f_0\|_\infty = \|K * f^*_{L,\mathrm{P}}\|_\infty \leq (2^r - 1)\|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}\,,$$

which will be important when applying concentration inequalities to prove the new oracle inequalities of Sections 3 and 4.

Remark 2.4. To bound the approximation error in (5), we assumed that there exists a Bayes decision function $f^*_{L,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ such that $f^*_{L,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$. This assumption could be significantly weakened if functions $f : X \to \mathbb{R}$ could be extended to functions $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ such that $\hat{f}$ inherits the smoothness properties of f described by some Sobolev or Besov space. Fortunately, Stein's Extension Theorem (cf. [28, p. 181]) guarantees the existence of such an extension operator with the desired features, whenever $X \subset \mathbb{R}^d$ is a bounded Lipschitz domain. To be more precise, in this case there exists a linear operator $\mathfrak{E}$ mapping functions $f : X \to \mathbb{R}$ to functions $\mathfrak{E}f : \mathbb{R}^d \to \mathbb{R}$ such that

i) $\mathfrak{E}f_{|X} = f$,
ii) $\mathfrak{E}$ continuously maps $W^m_p(X)$ into $W^m_p(\mathbb{R}^d)$ for all $p \in [1,\infty]$ and all integers $m \geq 0$, and
iii) $\mathfrak{E}$ continuously maps $B^\alpha_{p,q}(X)$ into $B^\alpha_{p,q}(\mathbb{R}^d)$ for all $p \in (1,\infty)$, $q \in (0,\infty]$ and all $\alpha > 0$.

For more details we refer to [28, p. 181], [1, p. 83], and [37, pp. 65/66].

Now, in addition to the general assumptions made in Section 1, let $X \subset \mathbb{R}^d$ be a *bounded* domain such that the extension operator $\mathfrak{E}$ exists and $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ be a Bayes decision function such that $f^*_{L,\mathrm{P}} \in L_\infty(X)$. Using Stein's extension operator we then obtain a function $\mathfrak{E}f^*_{L,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ with $\mathfrak{E}f^*_{L,\mathrm{P}} \in L_q(\mathbb{R}^d)$ for *all* $1 \leq q \leq \infty$. With this and with the choice $f := \mathfrak{E}f^*_{L,\mathrm{P}}$, Theorems 2.2 and 2.3 can be applied and thus estimation (10) of the approximation error holds for $\mathfrak{E}f^*_{L,\mathrm{P}}$. In [14] these considerations are carried out precisely.

3. Learning rates for least squares SVMs

In this section, we consider the non-parametric least squares regression problem based on 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, for this loss, the function $f^*_{L,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ defined by $f^*_{L,\mathrm{P}}(x) = \mathbb{E}_{\mathrm{P}}(Y|x)$, $x \in \mathbb{R}^d$, is the only function for which the Bayes risk is attained. Furthermore, some simple and well-known transformations show

$$\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} = \left\| f - f^*_{L,\mathrm{P}} \right\|^2_{L_2(\mathrm{P}_X)} . \tag{12}$$

In other words, the motivating estimate (10) is satisfied for $c = 1$.

In the following, we present our main results of this section including the essentially optimal rates for LS-SVMs using Gaussian kernels.

Theorem 3.1. *Let $Y := [-M, M]$ for $M > 0$, and P be a distribution on $\mathbb{R}^d \times Y$ such that $X := \operatorname{supp} \mathrm{P}_X \subset B_{\ell_2^d}$ is a bounded domain with $\mu(\partial X) = 0$. Furthermore, let P_X have a Lebesgue density $g \in L_q(\mathbb{R}^d)$ for some $q \geq 1$. 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 $f^*_{L,\mathrm{P}} \in B^\alpha_{2s,\infty}(\mathbb{R}^d)$ for $\alpha \geq 1$ and $s \geq 1$ with $\frac{1}{q} + \frac{1}{s} = 1$. Then, for all $\varepsilon > 0$ and $p \in (0,1)$, there exists a constant $C > 0$ such that for all $n \geq 1$, $\rho \geq 1$, $\gamma \in (0,1]$, and $\lambda > 0$, the SVM using the RKHS H_γ and the least squares loss L satisfies*

$$\lambda \left\| f_{\mathrm{D},\lambda,\gamma} \right\|^2_{H_\gamma} + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \leq C \left(\lambda \gamma^{-d} + c^2 \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p n} + \frac{\rho}{n} \right)$$

with probability P^n not less than $1 - e^{-\rho}$.

For the proof of Theorem 2.2 it is essential that P_X is absolutely continuous w.r.t. the Lebesgue measure. Apart from that the remaining arguments used to prove Theorem 3.1 and its consequences below apply to all marginal distributions P_X.

Note that Theorem 3.1 in particular holds for Bayes decision functions $f^*_{L,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ with $f^*_{L,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $f^*_{L,\mathrm{P}} \in W^\alpha_{2s}(\mathbb{R}^d)$ for $\alpha \in \mathbb{N}$ and $s \geq 1$ by (7).

With the help of Theorem 3.1 we can immediately derive learning rates for the learning method (4).

Corollary 3.2. *Let $\rho \geq 1$ be fixed. Under the assumptions of Theorem 3.1 and with*

$$\lambda_n = c_1 n^{-1} ,$$
$$\gamma_n = c_2 n^{-\frac{1}{2\alpha+d}} ,$$

we have, for all $n \geq 1$ and $\xi > 0$,

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n}) - \mathcal{R}^*_{L,\mathrm{P}} \leq C \rho \, n^{-\frac{2\alpha}{2\alpha+d}+\xi} \tag{13}$$

with probability P^n not less than $1 - e^{-\rho}$. Here, $c_1 > 0$ and $c_2 > 0$ are user-specified constants and $C > 0$ is a constant independent of n and ρ.

Note that the choice of λ_n in the Corollary above is actually *independent* of the unknown smoothness parameter α. Moreover, the kernel width depends on α in the same way, it does in classical kernel methods for density estimation and regression.

Consequently, to achieve rate (13), we need to know α. Since in practice we usually do not know this value nor its existence, we now show that a standard training/validation approach, see e.g. [30, Chapters 6.5, 7.4, 8.2], achieves the same rates adaptively, i.e. without knowing α. To this end, let $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be sequences of finite subsets $\Lambda_n, \Gamma_n \subset (0,1]$. For a data set $D := ((x_1, y_1), \ldots, (x_n, y_n))$, we define

$$D_1 := ((x_1, y_1), \ldots, (x_m, y_m))$$
$$D_2 := ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n))$$

where $m := \lfloor \frac{n}{2} \rfloor + 1$ and $n \geq 4$. We will use D_1 as a training set by computing the SVM decision functions

$$f_{\mathrm{D}_1,\lambda,\gamma} := \arg\min_{f \in H_\gamma} \lambda \|f\|_{H_\gamma}^2 + \mathcal{R}_{L,\mathrm{D}_1}(f), \qquad (\lambda,\gamma) \in \Lambda_n \times \Gamma_n \tag{14}$$

and use D_2 to determine (λ, γ) by choosing a $(\lambda_{\mathrm{D}_2}, \gamma_{\mathrm{D}_2}) \in \Lambda_n \times \Gamma_n$ such that

$$\mathcal{R}_{L,\mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}\right) = \min_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L,\mathrm{D}_2}\left(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}\right). \tag{15}$$

In the following, we call this training/validation approach TV-SVM. For suitably chosen candidate sets Λ_n and Γ_n that only depend on n and d, the next theorem establishes the rates (13) for TV-SVMs.

Theorem 3.3. *Under the assumptions of Theorem 3.1 we fix sequences $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ of finite subsets $\Lambda_n, \Gamma_n \subset (0,1]$ such that Λ_n is an ϵ_n-net of $(0,1]$ and Γ_n is a δ_n-net of $(0,1]$ with $\epsilon_n \leq n^{-1}$, $\delta_n \leq n^{-\frac{1}{2+d}}$, $1 \in \Lambda_n$, and $1 \in \Gamma_n$. Furthermore, assume that the cardinalities $|\Lambda_n|$ and $|\Gamma_n|$ grow polynomially in n. Then, for all $\xi > 0$ and $\rho \geq 1$, the TV-SVM producing the decision functions $f_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}$ satisfies*

$$\mathrm{P}^n\left(\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}) - \mathcal{R}^*_{L,\mathrm{P}} \leq C\rho\, n^{-\frac{2\alpha}{2\alpha+d}+\xi}\right) \geq 1 - e^{-\rho} \tag{16}$$

where $C > 0$ is a constant independent of n and ρ.

Strictly speaking, it suffices to consider $\lambda_n = cn^{-1}$ and a δ_n-net Γ_n of $(0,1]$ in the theorem above, since we have already seen in Corollary 3.2 that the optimal λ_n is of this form for $f^*_{L,\mathrm{P}} \in B^\alpha_{2s,\infty}(\mathbb{R}^d)$. However, since we do not know, whether this is also true under other distributional assumptions, we decided to formulate the training/validation approach over both λ and γ as a safety measure.

Remark 3.4. The learning rates obtained so far in particular hold, if P_X has a Lebesgue density that is bounded away from 0 and ∞. It is well-known that in this case the minmax rate for $\alpha > d/2$ and target functions $f^*_{L,\mathrm{P}} \in W_2^\alpha(X)$

is $n^{-\frac{2\alpha}{2\alpha+d}}$. Modulo ξ, our rate is therefore asymptotically optimal in a minmax sense. In addition, recall that for the latter distributions P_X the entropy numbers (cf. Definition 5.1) of the embedding $\mathrm{id} : B^{\alpha}_{2,\infty}(X) \to L_2(\mathrm{P}_X)$ behave like $i^{-\frac{\alpha}{d}}$, see e.g. [15, p. 151]. Therefore and since $B^{\alpha}_{2,\infty}(X)$ is continuously embedded into the space $\ell_\infty(X)$ of all bounded functions on X, we obtain by [36, Theorem 2.2] that $n^{-\frac{2\alpha}{2\alpha+d}}$ is the optimal learning rate in a minmax sense for $\alpha > d$ (cf. [32, Theorem 13]). In other words, for $\alpha > d$, the learning rates (13) and (16) obtained for $f^*_{L,\mathrm{P}} \in B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ are again asymptotically optimal modulo ξ.

Remark 3.5. Recall the extension operator $\mathfrak{E}$ and the assumptions made in Remark 2.4. Then the results of Theorem 3.1, Corollary 3.2, and Theorem 3.3 also hold, if $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ is a Bayes decision function with $f^*_{L,\mathrm{P}} \in L_\infty(X)$ and $f^*_{L,\mathrm{P}} \in B^{\alpha}_{2s,\infty}(X)$ for some $\alpha \geq 1$ and $s \geq 1$. Indeed, these assumptions on $f^*_{L,\mathrm{P}}$ are sufficient, since $f^*_{L,\mathrm{P}} \in L_\infty(X)$ implies $\mathfrak{E}f^*_{L,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$. Analogously, $f^*_{L,\mathrm{P}} \in B^{\alpha}_{2s,\infty}(X)$ implies $\mathfrak{E}f^*_{L,\mathrm{P}} \in B^{\alpha}_{2s,\infty}(\mathbb{R}^d)$, and hence, Theorem 3.1, Corollary 3.2, and Theorem 3.3 can be applied for $\mathfrak{E}f^*_{L,\mathrm{P}}$. Note that we can also assume $f^*_{L,\mathrm{P}} \in W^{\alpha}_{2s}(X)$, $\alpha \in \mathbb{N}$, since this yields $\mathfrak{E}f^*_{L,\mathrm{P}} \in W^{\alpha}_{2s}(\mathbb{R}^d) \subset B^{\alpha}_{2s,\infty}(\mathbb{R}^d)$ by (7). In both cases, the resulting rates obviously coincide with (13), and in addition, it is not hard to see, that these rates are, again, optimal in a minmax sense. Again, we refer to [14] for a detailed version of the results of this section using the extension operator $\mathfrak{E}$.

So far we only considered the case of bounded regression but in practice unbounded noise is relevant as well. In the following, we briefly examine a regression problem with exponentially decaying Y-tails. That is, for $\varepsilon := y - f^*_{L,\mathrm{P}}(x)$ we assume that there are constants $c \geq 1$ and $l > 0$ such that

$$\mathrm{P}\left(\left\{(x,y) \in X \times Y : |\varepsilon| \leq c\rho^l\right\}\right) \geq 1 - e^{-\rho} \tag{17}$$

for all $\rho > 1$. In other words, the probability of having large noise is very small. Additionally, we assume that the Bayes decision function $f^*_{L,\mathrm{P}}$ is P_X-almost surely bounded on X. The next theorem establishes learning rates for least squares SVMs in the spirit of Corollary 3.2 under these new assumptions.

Theorem 3.6. *Let $Y \subset \mathbb{R}$ and P be a distribution on $\mathbb{R}^d \times Y$ such that $X := \operatorname{supp} \mathrm{P}_X \subset B_{\ell_2^d}$ is a bounded domain with $\mu(\partial X) = 0$. Furthermore, let P_X have a Lebesgue density $g \in L_q(\mathbb{R}^d)$ for $q \geq 1$. Assume that $f^*_{L,\mathrm{P}}(x) \in [-1,1]$ for P_X-almost all $x \in X$ and that $f^*_{L,\mathrm{P}} \in L_2(\mathbb{R}^d)$ as well as $f^*_{L,\mathrm{P}} \in B^{\alpha}_{2s,\infty}(\mathbb{R}^d)$ for some $\alpha \geq 1$ and $s \geq 1$ defined by $\frac{1}{s} + \frac{1}{q} = 1$. Finally, assume that (17) holds. We define*

$$\lambda_n = c_1 n^{-1} \ ,$$
$$\gamma_n = c_2 n^{-\frac{1}{2\alpha+d}} \ ,$$

where $c_1 > 0$ and $c_2 > 0$ are user-specified constants independent of n. Now, for some fixed $\hat{\rho} \geq 1$ and $n \geq 3$, we define $\rho := \hat{\rho} + \ln n$ and $M_n := 2c\rho^l$. Let us consider the SVM that uses λ_n and γ_n and that clips its decision function

$f_{\mathrm{D},\lambda_n,\gamma_n}$ at M_n after training. Then, for all $\xi > 0$, there exists a constant $C > 0$ independent of n and ρ such that for all $\bar{\rho} > 1$ we have

$$\lambda \|f_{\mathrm{D},\lambda_n,\gamma_n}\|_{H_\gamma}^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n}) - \mathcal{R}_{L,\mathrm{P}}^* \leq C\bar{\rho}\hat{\rho}^{2l+\xi} n^{-\frac{2\alpha}{2\alpha+d}+\xi} \tag{18}$$

with probability P^n *not less than* $1 - (e^{-\bar{\rho}} + e^{-\hat{\rho}})$.

Note that the tail assumption (17) does not change the learning rates achieved in Corollary 3.2. Moreover, the learning rate (18) is essentially optimal for $f_{L,\mathrm{P}}^* \in B_{2s,\infty}^{\alpha}(\mathbb{R}^d)$ and $f_{L,\mathrm{P}}^* \in W_{2s}^{\alpha}(\mathbb{R}^d)$, respectively. Finally, these rates can again be achieved by the TV-SVM approach considered in Theorem 3.3, if we assume an upper bound on the unknown parameter l. The following two examples illustrate that such an assumption may not be to unrealistic.

Example 3.7. Let us consider the case of Gaussian noise, i.e. $\varepsilon \sim \mathcal{N}(0,1)$. The choice $m = \sqrt{2}\rho^{\frac{1}{2}}$ yields

$$\begin{aligned}
\mathrm{P}\left(\{(x,y) \in X \times Y : |\varepsilon| \leq m\}\right) &= 1 - \frac{2}{\sqrt{2\pi}} \int_m^\infty e^{-\frac{t^2}{2}} dt \\
&\geq 1 - \frac{2}{\sqrt{2\pi}} \sqrt{\frac{\pi}{2}} e^{-\frac{m^2}{2}} \\
&= 1 - e^{-\rho} \,,
\end{aligned}$$

where we used

$$\int_m^\infty e^{-\frac{t^2}{2}} dt \leq \int_m^\infty e^{-\frac{m^2}{2}} e^{-\frac{(t-m)^2}{2}} dt = e^{-\frac{m^2}{2}} \int_0^\infty e^{-\frac{t^2}{2}} dt = \sqrt{\frac{\pi}{2}} e^{-\frac{m^2}{2}} \,.$$

Hence, assumption (17) is satisfied for $l = \frac{1}{2}$ and $c = \sqrt{2}$. Obviously, a similar result holds for the general case $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Example 3.8. Assume that the distribution of the noise ε has the density

$$h(t) = \frac{1}{2} e^{-|t|} \,, \qquad\qquad t \in \mathbb{R} \,.$$

Then we have

$$\mathrm{P}\left(\{(x,y) \in X \times Y : |\varepsilon| \leq \rho\}\right) = \int_0^\rho e^{-t} dt = 1 - e^{-\rho} \,,$$

i.e. assumption (17) holds for $l = 1$ and $c = 1$.

Finally, it seems fair to say that for distributions that have fatter Y-tails than those considered by assumption (17) the simple clipping approach of Theorem 3.6 will lead to slower rates.

Let us now compare our results with previously obtained learning rates for SVMs. To begin recall that there have already been made several investigations on learning rates for SVMs using the least squares loss, see e.g. [5, 10, 11, 27, 22]

and the references therein. In particular, optimal rates have been established in [5], if $f^*_{L,\mathrm{P}} \in H$, and the eigenvalue behavior of the integral operator associated to H is known. Moreover, for $f^*_{L,\mathrm{P}} \notin H$, the articles [22] and [32] establish learning rates of the form $n^{-\beta/(\beta+p)}$, where β is a parameter describing the approximation properties of H and p is a parameter describing the eigenvalue decay. In both cases, however, additional assumptions on the interplay between H and $L_2(\mathrm{P}_X)$ are required, and [22] actually considers a different exponent in the regularization term of (4). On the other hand, [32] shows that the rate $n^{-\beta/(\beta+p)}$ is often asymptotically optimal in a minmax sense. In particular, the latter is the case for $H = W_2^m(X)$, $f^*_{L,\mathrm{P}} \in W_2^\alpha(X)$, and $\alpha \in (d/2, m]$, that is, when using a Sobolev space as underlying RKHS H, then all target functions contained in a Sobolev space of lower smoothness $\alpha > d/2$ can be learned with the asymptotically optimal rate $n^{-\frac{2\alpha}{2\alpha+d}}$. Here we note that the condition $\alpha > d/2$ ensures by Sobolev's embedding theorem that $W_2^\alpha(X)$ consists of bounded functions, and hence $Y = [-M, M]$ does not impose an additional assumption on $f^*_{L,\mathrm{P}}$. If $\alpha \in (0, d/2]$, then the results of [32] still yield the above mentioned rates, but we no longer know whether they are optimal in a minmax sense, since $Y = [-M, M]$ does impose an additional assumption. In addition, note that for Sobolev spaces this result, modulo an extra log factor, has already been proved by [16, Section 3.2]. Besides, similar results to those of [32] have been recently achieved in [34] for the case of multiple kernel learning. Note that one of the reasons for these learning rates is the fact that the approximation error decays polynomially fast if (and only if) the regression function $f^*_{L,\mathrm{P}}$ is contained in some interpolation space $[L_2(X), W_2^m(X)]_{\beta,\infty}$, $m \in \mathbb{Z}_+$, $\beta \in (0,1)$, see [26] for more details.

These results suggest that by using a fixed C^∞-kernel such as the Gaussian RBF kernel, one could actually learn the entire scale of Sobolev spaces with the above mentioned rates. Unfortunately, however, there are good reasons to believe that this is not the case. Indeed, [26] shows that for many analytic kernels the approximation error function $A_H(\lambda)$ defined by (5) can *only* have polynomial decay for $\lambda \to 0$ if $f^*_{L,\mathrm{P}}$ is analytic, too. In particular, for Gaussian kernels with *fixed* width γ and $f^*_{L,\mathrm{P}} \notin C^\infty$, the approximation error has only logarithmic decay, see [26, Proposition 1.1.]. Since it seems rather unlikely that these poor approximation properties can be balanced by superior bounds on the estimation error, the above-mentioned results indicate that Gaussian kernels with *fixed* width may have a poor performance. This conjecture is justified by many empirical experience gained throughout the last decade. Beginning with [33], research has thus focused on the learning performance of SVMs with varying widths. In this direction the result that is probably the closest to ours is [40]. Although these authors actually consider binary classification using convex loss functions including the least squares loss, it is relatively straightforward to translate their findings to our least squares regression scenario. The resulting learning rate is $n^{-\frac{\alpha}{\alpha+2d+2}}$, again under the assumption $f^*_{L,\mathrm{P}} \in W_2^\alpha(X)$ for some $\alpha > 0$. Clearly, this is significantly worse than our rates.

In [41] multi-kernel regularization schemes are treated, where X is isometrically embedded into a t-dimensional, connected and compact C^∞-submanifold

of $\mathbb{R}^d$. In [41, Section 5] it is pointed out that the regularization scheme with one fixed Gaussian kernel has weak approximation ability since the regularization error only decays polynomially if the regression function is analytic. Hence, the multi-kernel regularization scheme using Gaussian kernels with flexible variances is advantageous for many applications. In addition, it turns out that the resulting learning rate for this multi-kernel regularization scheme does not depend on the dimension d, but on the intrinsic dimension t of the data. Namely, the authors establish the rate $n^{-\frac{\alpha}{8\alpha+4t}}$ modulo a logarithmic factor, where $\alpha \in (0,1]$ and $f^*_{L,\mathrm{P}} \in \mathrm{Lip}\,(\alpha)$. Note that this rate is better than ours only if $t < \frac{d-14\alpha}{8}$, that is, e.g. for $\alpha = 1$, if $d > 8t + 14$. Other results on multi-kernel regularization schemes applied to Gaussian kernels with varying widths can be found in [23, 41, 42, 43]. For example, [43] treats the multi-kernel least squares regularized regression problem and, for $f^*_{L,\mathrm{P}} \in W_2^\alpha(X)$, establishes learning rates of the form $n^{-\frac{\alpha}{2(4\alpha+d)}}$ modulo a logarithmic factor for $\alpha \leq 2$ and bounded X as well as rates of the form $n^{-\frac{2\alpha-d}{4(4\alpha-d)}+\xi}$ whenever $f^*_{L,\mathrm{P}} \in W_2^\alpha\,(X)$ for some $\alpha \in (d/2, d/2+2)$, respectively, where again $\xi > 0$ can be chosen to be arbitrarily close to 0.

Again all these rates are far from being optimal, so that it seems fair to conclude that our results represent a significant advance. Furthermore, we can conclude that, in terms of asymptotical minmax rates, multi-kernel approaches applied to Gaussian RBFs *cannot* provide any significant improvement over a simple training/validation approach for determining the kernel width and the regularization parameter, since the latter already leads to rates that are optimal modulo an arbitrarily small ξ in the exponent.

4. Learning rates for SVMs for quantile regression

In the previous section we used the approximation results of Section 2 to derive essentially optimal learning rates for least squares SVMs. In this section we focus on quantile regression with the goal to derive learning rates for SVMs that are comparable with the rates achieved for least squares SVMs. Recall that the goal of quantile regression is to estimate the conditional τ-quantile, i.e. the set valued function

$$F^*_{\tau,\mathrm{P}}(x) := \{t \in \mathbb{R} : \mathrm{P}\,(Y \leq t|x) \geq \tau \text{ and } \mathrm{P}\,(Y \geq t|x) \geq 1-\tau\}\,,$$

where $\tau \in (0,1)$ is a fixed constant. Throughout this section, we assume $Y := [-1,1]$ and that $F^*_{\tau,\mathrm{P}}$ consists of singletons, i.e. there exists an $f^*_{\tau,\mathrm{P}} : X \to [-1,1]$, such that $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$ for P_X-almost all $x \in X$. In the following, $f^*_{\tau,\mathrm{P}}$ is called the conditional τ-quantile function. To estimate the latter one can use the so-called τ-pinball loss $L_\tau : Y \times \mathbb{R} \to [0,\infty)$ represented by

$$\psi(r) := \begin{cases} -(1-\tau)r, & \text{if } r < 0 \\ \tau r, & \text{if } r \geq 0\,, \end{cases}$$

where $r := y - t$ and $L_\tau(y,t) = \psi(r)$. Recall that the conditional τ-quantile function is, modulo P_X-zero sets, the only function that minimizes the L_τ-risk, that is $\mathcal{R}^*_{L_\tau,\mathrm{P}} = \mathcal{R}_{L_\tau,\mathrm{P}}(f^*_{\tau,\mathrm{P}})$.

To derive meaningful learning rates for SVMs for quantile regression, we need to compare the excess L_τ-risk of some estimator f_D to the distance

$$\|f_D - f^*_{\tau,\mathrm{P}}\|_{L_v(\mathrm{P}_X)} .$$

For that purpose, we have to introduce some characteristics of the distribution P. For the sake of simplicity, we use Q as template for the conditional distribution $\mathrm{P}(\,\cdot\,|x)$. Hence, let Q be a distribution on $\mathbb{R}$ with support $\operatorname{supp} Q \subset [-1,1]$ and τ-quantile

$$F^*_\tau(Q) := \{t \in \mathbb{R} : Q\left((-\infty,t]\right) \geq \tau \text{ and } Q\left([t,\infty)\right) \geq 1-\tau\} .$$

Recall that $F^*_\tau(Q)$ is a bounded and closed interval, i.e. $F^*_\tau(Q) = [t^*_{\min}, t^*_{\max}]$ with $t^*_{\min} := \min F^*_\tau(Q)$ and $t^*_{\max} := \max F^*_\tau(Q)$. Since we assumed that $F^*_{\tau,\mathrm{P}}$ consists of singletons, we also assume that $F^*_\tau(Q)$ consists of singletons for the sake of uniformity, i.e. $t^*_{\min} = t^*_{\max} =: t^*$ and $F^*_\tau(Q) = \{t^*\}$. We start with a definition describing the concentration of Q around the τ-quantile t^*.

Definition 4.1. A distribution Q with $\operatorname{supp} Q \subset [-1,1]$ is said to have a τ-quantile t^* of lower type $q \in (1,\infty)$, if there exist constants $\alpha_Q \in (0,2]$ and $b_Q > 0$ such that

$$Q\left((t^*-s,t^*)\right) \geq b_Q s^{q-1}$$
$$Q\left((t^*,t^*+s)\right) \geq b_Q s^{q-1}$$

for all $s \in [0,\alpha_Q]$. Moreover, Q has a τ-quantile of type $q = 1$, if $Q(\{t^*\}) > 0$. In this case we define $\alpha_Q := 2$ and $b_Q := \min\{\tau - Q((-\infty,t^*)), Q((-\infty,t^*]) - \tau\}$, where we note that this implies $b_Q > 0$. For $q \geq 1$, we finally write $\kappa_Q := b_Q \alpha_Q^{q-1}$.

Definition 4.1 has already been introduced in [31, Section 2], where more details including examples that go beyond the ones we discuss below can be found.

Since we are interested in distributions P on $X \times \mathbb{R}$ and not only in distributions Q on $\mathbb{R}$, we extend Definition 4.1 to such P.

Definition 4.2. Let $p \in (0,\infty]$, $q \in [1,\infty)$, and P be a distribution on $X \times \mathbb{R}$ with $\operatorname{supp} \mathrm{P}(\,\cdot\,|x) \subset [-1,1]$ for P_X-almost all $x \in X$. Then P is said to have a τ-quantile of lower p-average type q, if $\mathrm{P}(\,\cdot\,|x)$ has a τ-quantile of lower type q for P_X-almost all $x \in X$, and the function $\kappa : X \to [0,\infty]$ defined, for P_X-almost all $x \in X$, by

$$\kappa(x) := \kappa_{\mathrm{P}(\cdot|x)} ,$$

where $\kappa_{\mathrm{P}(\cdot|x)} = b_{\mathrm{P}(\cdot|x)} \alpha^{q-1}_{\mathrm{P}(\cdot|x)}$ is defined in Definition 4.1, satisfies $\kappa^{-1} \in L_p(\mathrm{P}_X)$.

Definition 4.1 describes the concentration around t^* by lower bounds. Analogously, the next definition measures the concentration of Q around t^* by upper bounds.

Definition 4.3. A distribution Q on $[-1,1]$ is said to have a τ-quantile t^* of upper type $q \in [1,\infty)$, if there exists a constant $b_Q > 0$ such that

$$Q\left((t^* - s, t^*)\right) \leq b_Q s^{q-1}$$
$$Q\left((t^*, t^* + s)\right) \leq b_Q s^{q-1}$$

for all $s \in [0,2]$.

By setting $q = 1$ and $b_Q = 1$, we see that Q always has a τ-quantile of upper type q. On the other hand, for $q > 1$ Definition 4.3 divides the set of all distributions on $[-1,1]$ into various classes.

Finally, based on Definition 4.3 we define quantiles of upper p-average type q analogously to the quantiles of lower p-average type q.

Definition 4.4. Let $p \in (1,\infty]$, $q \in [1,\infty)$, and P be a distribution on $X \times [-1,1]$. Then P is said to have a τ-quantile of upper p-average type q, if $\mathrm{P}(\,\cdot\,|x)$ has a τ-quantile of upper type q for P_X-almost all $x \in X$, and the function $\varphi : X \to [0,\infty]$ defined, for P_X-almost all $x \in X$, by $\varphi(x) := b_{\mathrm{P}(\,\cdot\,|x)}$, satisfies $\varphi \in L_p(\mathrm{P}_X)$.

Let us now present some examples to illustrate the notion of quantiles of upper and lower p-average type q.

Example 4.5. Let ν be a distribution on $[-1,1]$ having a bounded Lebesgue density h, i.e. $h(y) \leq b$ for some $b \in (0,\infty)$ and Lebesgue-almost all $y \in [-1,1]$. Then a simple integration yields that ν has a τ-quantile of upper type $q = 2$ for all $\tau \in (0,1)$. Here, we set $b_\nu := b$.

In addition, we assume that P is a distribution on $X \times [-1,1]$ with $X \subset \mathbb{R}^d$ and such that P_X is absolutely continuous with respect to the Lebesgue measure μ. Furthermore, assume that the corresponding conditional densities $h(\,\cdot\,,x) := \frac{d\mathrm{P}(\,\cdot\,|x)}{d\mu_{|[-1,1]}}$ are uniformly bounded, that is, $h(y,x) \leq b$ for Lebesgue-almost all $y \in [-1,1]$. Then, for $p = \infty$, P has a τ-quantile of upper p-average type $q = 2$ with $\varphi(x) := b$.

If we further assume that, for P_X-almost all $x \in X$, the density $h(\,\cdot\,,x)$ of $\mathrm{P}(\,\cdot\,|x)$ is bounded away from 0, i.e. $h(y,x) \geq \hat{b}$ for some $0 < \hat{b} \leq b$ for Lebesgue-almost all $y \in [-1,1]$, then, for $p = \infty$, P also has a τ-quantile of lower p-average type $q = 2$ with $\kappa(x) := 2\hat{b}$.

Example 4.6. Let δ_{t^*} be the Dirac measure at $t^* \in (0,1)$, ν be a distribution on $[-1,1]$ with $\nu(\{t^*\}) = 0$ and $Q := \alpha\nu + (1-\alpha)\delta_{t^*}$ for some $\alpha \in [0,1)$. By [31, Example 2.4] we know that, for $\tau \in (\alpha\nu((-\infty,t^*)), \alpha\nu((-\infty,t^*)) + 1 - \alpha)$, $\{t^*\}$ is a τ-quantile of lower type $q = 1$ with $\kappa_Q := \min\{\tau - \alpha\nu((-\infty,t^*)), \alpha\nu((-\infty,t^*)) + 1 - \alpha - \tau\}$.

Now assume that P is a distribution on $X \times [-1,1]$ such that each conditional distribution $\mathrm{P}(\,\cdot\,|x)$ is of the above form Q, where t^* may depend on x but ν and

α do not. Then, for $p = \infty$, P has a τ-quantile of lower p-average type $q = 1$. Moreover, for $p = \infty$, P also has a τ-quantile of upper p-average type $q = 1$.

Using the property of the distribution P having quantiles of lower/upper p-average type q, we return to our initial goal of comparing the excess L_τ-risk of some estimator f_D to the distance $\|f_D - f^*_{\tau,\mathrm{P}}\|_{L_v(\mathrm{P}_X)}$. To this end, we first recall from [31, Theorem 2.7] the following so-called self-calibration inequality

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_v(\mathrm{P}_X)} \leq 2^{1-\frac{1}{q}} q^{\frac{1}{q}} \|\kappa^{-1}\|^{\frac{1}{q}}_{L_p(\mathrm{P}_X)} (\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}})^{\frac{1}{q}} \,, \tag{19}$$

which holds for $p \in (0,\infty]$, $q \in [1,\infty)$, $v := \frac{pq}{p+1}$, and all $f : X \to [-1,1]$, whenever P is a distribution that has a τ-quantile of lower p-average type q. Initially, our statistical analysis will provide oracle inequalities for the excess L_τ-risk, and hence self-calibration inequalities provide a natural mean to translate such oracle inequalities into bounds on the distance $\|f_D - f^*_{\tau,\mathrm{P}}\|_{L_v(\mathrm{P}_X)}$. Interestingly, however, if we want to use the approximation results from Section 2, we also need inverse self-calibration inequalities. In this respect, we first note that the Lipschitz continuity of L_τ immediately yields

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq \|f - f^*_{\tau,\mathrm{P}}\|_{L_1(\mathrm{P}_X)} \tag{20}$$

for all $f : X \to [-1,1]$. For our purposes, this estimate can be substantially improved by the next theorem for distributions having quantiles of upper p-average type q.

Theorem 4.7. *Let* P *be a distribution on* $X \times [-1,1]$ *that has a* τ*-quantile of upper* p*-average type* q *with* $p \in (1,\infty]$ *and* $q \in [1,\infty)$*. In addition, assume that, for all* $x \in X$*, we have* $\mathrm{P}(\{f^*_{\tau,\mathrm{P}}(x)\}|x) = 0$*. Then we have*

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq q^{-1} \|b_{\mathrm{P}(\cdot\,|x)}\|_{L_p(\mathrm{P}_X)} \|f - f^*_{\tau,\mathrm{P}}\|^q_{L_u(\mathrm{P}_X)} \tag{21}$$

for all $f : X \to [-1,1]$*, where* $u := \frac{pq}{p-1}$.

To see that (21) is indeed an improvement of (20) we consider $f_0 := K * f^*_{\tau,\mathrm{P}}$ with K as in (8) and $f^*_{\tau,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ such that it is a Bayes decision function with $f^*_{\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$. Assuming $f^*_{\tau,\mathrm{P}} \in B^\alpha_{1,\infty}(\mathbb{R}^d)$ or $f^*_{\tau,\mathrm{P}} \in B^\alpha_{u,\infty}(\mathbb{R}^d)$ with u as in Theorem 4.7, we obtain by Theorem 2.2 and by the definition of Besov spaces

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq \|f_0 - f^*_{\tau,\mathrm{P}}\|_{L_1(\mathrm{P}_X)} \leq c_1\, \omega_{r,L_1(\mathbb{R}^d)}\left(f^*_{\tau,\mathrm{P}}, \gamma/2\right) \leq c_2 \gamma^\alpha$$

from (20), while (21) yields

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq c_3 \|f_0 - f^*_{\tau,\mathrm{P}}\|^q_{L_u(\mathrm{P}_X)} \leq c_4 \left(\omega^u_{r,L_u(\mathbb{R}^d)}\left(f^*_{\tau,\mathrm{P}}, \gamma/2\right)\right)^{\frac{q}{u}} \leq c_5 \gamma^{q\alpha},$$

for suitable positive constants $c_1, \ldots, c_5$. Since $\gamma \in (0,1]$, it is obvious that the second estimate is tighter than the first one whenever $q > 1$. Taking advantage of Theorem 4.7 and the improved estimate of the excess L_τ-risk, we achieve a new oracle inequality for SVMs for quantile regression similar to the one of Theorem 3.1 for LS-SVMs.

Theorem 4.8. *Let $Y := [-1,1]$, and P be a distribution on $\mathbb{R}^d \times Y$ that has a τ-quantile of upper p-average type q for some $p \in (1,\infty]$ and $q \in [1,\infty)$. Assume that $X := \operatorname{supp} \mathrm{P}_X \subset B_{\ell_2^d}$ is a domain and that P_X has a Lebesgue density $g \in L_w(\mathbb{R}^d)$ for some $w \geq 1$. For $u := \frac{pq}{p-1}$, suppose that $f^*_{\tau,\mathrm{P}} \in L_u(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ if $u \in [1,2)$ and $f^*_{\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ if $u \geq 2$. In addition, we assume that $f^*_{\tau,\mathrm{P}} \in B^\alpha_{us,\infty}(\mathbb{R}^d)$ for some $\alpha \geq 1$ and $s \geq 1$ defined by $\frac{1}{s} + \frac{1}{w} = 1$. Finally, suppose that there exist constants $\vartheta \in [0,1]$ and $V \geq 2^{2-\vartheta}$ such that the variance bound*

$$\mathbb{E}_{\mathrm{P}}(L_\tau \circ \widehat{f} - L_\tau \circ f^*_{\tau,\mathrm{P}})^2 \leq V \cdot \Big(\mathbb{E}_{\mathrm{P}}(L_\tau \circ \widehat{f} - L_\tau \circ f^*_{\tau,\mathrm{P}})\Big)^\vartheta \tag{22}$$

is satisfied for all $f : \mathbb{R}^d \to \mathbb{R}$. Then, for all $\varepsilon > 0$ and $\varsigma \in (0,1)$, there exists a constant $C > 0$ such that for all $n \geq 1$, $\rho \geq 1$, $\gamma \in (0,1]$, and $\lambda > 0$, the SVM using the RKHS H_γ and the pinball loss L_τ satisfies

$$\begin{aligned}
&\lambda \|f_{\mathrm{D},\lambda,\gamma}\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\leq C\left(\lambda\gamma^{-d} + \gamma^{q\alpha} + \left(\frac{\gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma n}\right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + \left(\frac{\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{\rho}{n}\right)
\end{aligned}$$

with probability P^n not less than $1 - e^{-\rho}$.

To prove Theorem 4.8 we have to use the approximation result of Theorem 2.2. Similarly to Theorem 3.1 and its corollaries, it is for this reason essential in Theorem 4.8 that P_X is absolutely continuous w.r.t. the Lebesgue measure and that the associated density is contained in $L_w(X)$. The remaining arguments used to prove Theorem 4.8 and its consequences, on the contrary, hold for all marginal distributions P_X. Our next goal is to illustrate these consequences. We begin with a general form of the learning rates that result from Theorem 4.8.

Corollary 4.9. *Let $\rho \geq 1$ be fixed, and $\xi > 0$. Under the assumptions of Theorem 4.8 and with*

$$\begin{aligned}
\lambda_n &= c_1 n^{-\frac{q\alpha+d}{q\alpha(2-\vartheta)+d}}\,, \\
\gamma_n &= c_2 n^{-\frac{1}{q\alpha(2-\vartheta)+d}}\,,
\end{aligned}$$

we have, for all $n \geq 1$,

$$\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C\rho\, n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d}+\xi} \tag{23}$$

with probability P^n not less than $1 - e^{-\rho}$. Here, $c_1 > 0$ and $c_2 > 0$ are user-specified constants and $C > 0$ is a constant independent of n and ρ.

To achieve the learning rate (23), λ_n and γ_n have to be set as in Corollary 4.9. To this end, we again have to know α and ϑ, which is usually not the case in practice. Nevertheless, we derive the same learning rates without knowing neither α nor ϑ by the standard training/validation approach of Section 3.

Theorem 4.10. *Under the assumptions of Theorem 4.8 we fix sequences* $\Lambda := (\Lambda_n)$ *and* $\Gamma := (\Gamma_n)$ *of finite subsets* $\Lambda_n, \Gamma_n \subset (0,1]$ *such that* Λ_n *is an* ϵ_n*-net of* $(0,1]$ *and* Γ_n *is an* δ_n*-net of* $(0,1]$ *with* $\epsilon_n \leq n^{-1}$, $\delta_n \leq n^{-\frac{1}{1+d}}$, $1 \in \Lambda_n$, *and* $1 \in \Gamma_n$. *Furthermore, assume that the cardinalities* $|\Lambda_n|$ *and* $|\Gamma_n|$ *grow polynomially in* n. *Then, for all* $\xi > 0$ *and* $\rho \geq 1$, *the TV-SVM using* L_τ *satisfies*

$$\mathrm{P}^n \left(\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C\rho\, n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d}+\xi} \right) \geq 1 - e^{-\rho}$$

with a constant $C > 0$ *independent of* n *and* ρ

To apply Theorems 4.8 and 4.10, we need the variance bound (22) for the τ-pinball loss. But unfortunately, unlike for the least squares loss, (22) generally does not hold for some $\vartheta > 0$. However, if P has a lower quantile type, then the following result taken from [31, Theorem 2.8] establishes non-trivial variance bounds.

Theorem 4.11. *Let* P *be a distribution that has a* τ*-quantile of lower* p*-average type* q *for some* $p \in (0,\infty]$ *and* $q \in [1,\infty)$. *Then, for* $\vartheta := \min\{\frac{2}{q}, \frac{p}{p+1}\}$, $V := 2^{2-\vartheta} q^\vartheta \|\kappa^{-1}\|^\vartheta_{L_p(\mathrm{P}_X)}$, *and all* $f : X \to \mathbb{R}$, *we have*

$$\mathbb{E}_\mathrm{P}(L_\tau \circ \widehat{f} - L_\tau \circ f^*_{\tau,\mathrm{P}})^2 \leq V \cdot \left(\mathbb{E}_\mathrm{P}(L_\tau \circ \widehat{f} - L_\tau \circ f^*_{\tau,\mathrm{P}}) \right)^\vartheta .$$

Let us now combine this variance bound with the previous results such that learning rates can be achieved with simplified assumptions. For the sake of simplicity, we restrict our considerations to distributions P that have both a τ-quantile of lower and upper p-average type q. Let us begin with the probably most interesting example $(p,q) = (\infty, 2)$, cf. Example 4.5.

Corollary 4.12. *Let* $Y := [-1,1]$, *and* P *be a distribution on* $\mathbb{R}^d \times Y$ *that has a* τ*-quantile of lower and upper* p*-average type* q *for* $q = 2$ *and* $p = \infty$. *Assume that* $X := \operatorname{supp} \mathrm{P}_X \subset B_{\ell_2^d}$ *is a domain and that* P_X *has a Lebesgue density* $g \in L_w(\mathbb{R}^d)$ *for some* $w \geq 1$. *Suppose that* $f^*_{\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ *and* $f^*_{\tau,\mathrm{P}} \in B^\alpha_{2s,\infty}(\mathbb{R}^d)$ *for some* $\alpha \geq 1$ *and* $s \geq 1$ *defined by* $\frac{1}{s} + \frac{1}{w} = 1$. *Then we obtain for the SVM considered in Corollary 4.9 that, for all* $\xi > 0$ *and* $\rho \geq 1$,

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

and

$$\mathrm{P}^n \left(\|\widehat{f}_{\mathrm{D},\lambda,\gamma} - f^*_{\tau,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)} \leq C'\rho\, n^{-\frac{2\alpha}{2\alpha+d}+\xi} \right) \geq 1 - e^{-\rho} ,$$

with constants $C > 0$ *and* $C' := 4\|\kappa^{-1}\|_{L_\infty(\mathrm{P}_X)} C$ *independent of* n *and* ρ. *Moreover, the same learning rates can be obtained for the TV-SVM considered in Theorem 4.10.*

In particular, by (7) Corollary 4.12 also holds for a conditional τ-quantile function $f^*_{\tau,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ such that $f^*_{\tau,\mathrm{P}}|_X : X \to [-1,1]$, $f^*_{\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $f^*_{\tau,\mathrm{P}} \in W^\alpha_{2s}(\mathbb{R}^d)$ for $\alpha \in \mathbb{N}$ and $s \geq 1$.

Note that the convergence rates above equal the rates we achieved for the least squares SVMs in Section 3 (cf. Remark 3.4).

Let us now again quickly discuss the influence of the assumed *upper* quantile type. To this end, assume that we are not using a possibly non-trivial upper quantile type. Then, as discussed in front of Theorem 4.8, we can only use the estimate

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq \|f_0 - f^*_{\tau,\mathrm{P}}\|_{L_1(\mathrm{P}_X)} \leq c\gamma^\alpha, \tag{24}$$

in the corresponding proof, where $f_0 := K * f^*_{\tau,\mathrm{P}}$. Assuming that P has a τ-quantile of *lower* p-average type q with $p = \infty$ and $q = 2$, i.e. $v = 2$ and $\vartheta = 1$, then (24) and (19) yield

$$\|\widehat{f}_{\mathrm{D},\lambda,\gamma} - f^*_{\tau,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)} \leq Cn^{-\frac{\alpha}{\alpha+d}+\xi}$$

for all $\xi > 0$. Clearly, this rate is significantly worse than that of Corollary 4.12.

For this reason and for the sake of completeness, we consider distributions P having a τ-quantile of upper p-average type q with $p = \infty$ and $q \neq 2$ in the following corollary, where we omit the obvious proof.

Corollary 4.13. *Let* P *be a distribution having a τ-quantile of lower and upper p-average type q with $p = \infty$ and $q \in [1,2) \cup (2,\infty)$. Under the assumptions of Theorem 4.8 and of Theorem 4.11 we obtain*

$$\vartheta = \begin{cases} 1\,, & \text{if } q < 2\,, \\ \frac{2}{q}\,, & \text{if } q > 2\,. \end{cases}$$

Then, for the SVM considered in Corollary 4.9 as well as for the TV-SVM considered in Theorem 4.10, we obtain, for all $\xi > 0$ and $\rho \geq 1$,

$$\|\widehat{f}_{\mathrm{D},\lambda,\gamma} - f^*_{\tau,\mathrm{P}}\|^q_{L_q(\mathrm{P}_X)} \leq \begin{cases} C\rho\, n^{-\frac{q\alpha}{q\alpha+d}+\xi}\,, & \text{if } q < 2\,, \\ C\rho\, n^{-\frac{q\alpha}{2\alpha(q-1)+d}+\xi}\,, & \text{if } q > 2\,. \end{cases}$$

with probability P^n *not less than* $1 - e^{-\rho}$*, where $C > 0$ is a constant independent of n and ρ.*

Remark 4.14. Again, recall the extension operator $\mathfrak{E}$ and the assumptions made in Remark 2.4. Let $X \subset B_{\ell_2^d}$ be a domain such that we have such an extension operator $\mathfrak{E}$, and let $f^*_{\tau,\mathrm{P}} : X \to [-1,1]$ be the conditional τ-quantile function such that $f^*_{\tau,\mathrm{P}} \in B^\alpha_{us,\infty}(X)$ for some $\alpha \geq 1$, $s \geq 1$, and $u = \frac{pq}{p-1}$. Then applying $\mathfrak{E} f^*_{\tau,\mathrm{P}}$ we achieve the results of Theorem 4.8, Corollary 4.9, Theorem 4.10, Corollary 4.12, and Corollary 4.13 for the modified assumptions analogously to Remark 3.5. Moreover, the same holds for $f^*_{\tau,\mathrm{P}} \in W^\alpha_{us}(X)$, $\alpha \in \mathbb{N}$, instead of $f^*_{\tau,\mathrm{P}} \in B^\alpha_{us,\infty}(X)$ since $f^*_{\tau,\mathrm{P}} \in W^\alpha_{us}(X)$ implies $\mathfrak{E} f^*_{\tau,\mathrm{P}} \in W^\alpha_{us}(\mathbb{R}^d) \subset B^\alpha_{us,\infty}(\mathbb{R}^d)$ by (7).

Like learning rates for least squares regression, learning rates for quantile regression have already been obtained in the literature, although it seems fair to say that the latter regression problem has attracted less attention. Let us begin the discussion of such rates with the case of SVMs. Probably the first result in this direction is [35], where a learning rate of $n^{-\frac{1}{2}}$ for the excess risk is shown under some assumptions including that $f^*_{\tau,\mathrm{P}}$ is contained in the RKHS used by the SVM. In addition, algorithmic aspects of SVMs for quantile regression are discussed. An approach similar to ours is used in [21] to estimate the distance of the SVM estimator to $f^*_{\tau,\mathrm{P}}$. There, the authors show for example, that if $f^*_{\tau,\mathrm{P}}$ is contained in some known H_γ and the following calibration inequality

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_1(\mathrm{P}_X)} \leq c\sqrt{\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}} \tag{25}$$

is satisfied, then modulo some logarithmic factor, the rate $n^{-1/3}$ can be achieved for $\|f_{D,\lambda_n,\gamma} - f^*_{\tau,\mathrm{P}}\|_{L_1(\mathrm{P}_X)}$. Unfortunately, assuming that $f^*_{\tau,\mathrm{P}}$ is contained in the used RKHS is rather restrictive as discussed in the previous section. In addition, it is well-known that establishing rates under such an assumption is rather simple compared to the general case. Nonetheless, it seems interesting that their rates can be essentially recovered by combining Theorem 4.8 with Theorem 4.11 and (19) for $p = 1$, $q = 2$, and $\alpha = \infty$, since in this case (19) reduces to (25). Moreover, for the example discussed after [21, Corollary 1] our general results actually achieve a rate of the form $n^{-2/3+\xi}$, whereas the authors only obtain a rate of the form $n^{-1/3}$. Furthermore, [31] achieves our rate $n^{-\frac{2\alpha}{2\alpha+d}}$ if $H = W_2^\alpha(X)$ for some $\alpha > \frac{d}{2}$, P has a τ-quantile of lower p-average type q with $p = \infty$ and $q = 2$, and, again, $f^*_{\tau,\mathrm{P}} \in H$.

The Sobolev setting is also treated in [25], where the author considers a penalized estimate with hypothesis space $W_p^\alpha[a,b]$. In particular, he obtains the same learning rate as we do for $d = 1$. In [20] a partially linear quantile regression model is considered, where the parametric component learns with rate $n^{-\frac{1}{2}}$.

Finally, [19, Chapter 7] presents learning rates for a polynomial model and locally polynomial quantile regression estimators. Here, the rate $n^{-\frac{2\alpha}{2\alpha+d}} \ln n$ is achieved, where α describes the order of smoothness. In fact, the author refers to [7], where a similar rate is also achieved for arbitrary L_p-norms with $1 \leq p < \infty$.

5. Proofs

5.1. Proofs of Section 2

In Section 2 we presented two theorems that estimate parts of the approximation error. Let us begin with the proofs of these theorems. To this end, we need the convention $0^0 := 1$.

Proof of Theorem 2.2. First of all, we show $f \in L_q(\mathrm{P}_X)$. Because of the assumption $f \in L_q\left(\mathbb{R}^d\right) \cap L_\infty\left(\mathbb{R}^d\right)$, we have $f \in L_u\left(\mathbb{R}^d\right)$ and thus $f \in L_u\left(X\right)$

for every $u \in [q,\infty]$ and $q \in [1,\infty]$. In addition,

$$\|f\|_{L_q(\mathrm{P}_X)} = \left(\int_{\mathbb{R}^d} |f(x)|^q \, d\mathrm{P}_X(x)\right)^{\frac{1}{q}} = \left(\int_X |f(x)|^q \, d\mathrm{P}_X(x)\right)^{\frac{1}{q}} \leq \|f\|_{L_\infty(X)} < \infty$$

holds, i.e. $f \in L_q(\mathrm{P}_X)$ for all $q \in [1,\infty)$. It remains to show

$$\|K * f - f\|^q_{L_q(\mathrm{P}_X)} \leq C_{r,q} \|g\|_{L_p(\mathbb{R}^d)} \, \omega^q_{r,L_{qs}(\mathbb{R}^d)}(f, \gamma/2) \ .$$

To this end, we use the translation invariance of the Lebesgue measure and $K_\gamma(u) = K_\gamma(-u)$ $(u \in \mathbb{R}^d)$ to obtain, for $x \in X$,

$$\begin{aligned}
K * f(x) &= \int_{\mathbb{R}^d} \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{j\gamma}{\sqrt{2}}}(x-t) \, f(t) \, dt \\
&= \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} \int_{\mathbb{R}^d} K_{\frac{\gamma}{\sqrt{2}}}\left(\frac{x-t}{j}\right) f(t) \, dt \\
&= \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} \int_{\mathbb{R}^d} K_{\frac{\gamma}{\sqrt{2}}}(h) \, f(x+jh) \, j^d \, dh \\
&= \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h) \left(\sum_{j=1}^r \binom{r}{j} (-1)^{1-j} f(x+jh)\right) dh \ .
\end{aligned}$$

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

$$\begin{aligned}
&\|K * f - f\|^q_{L_q(\mathrm{P}_X)} \\
&= \int_X |K * f(x) - f(x)|^q \, d\mathrm{P}_X(x) \\
&= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h) \left(\sum_{j=1}^r \binom{r}{j} (-1)^{1-j} f(x+jh)\right) dh - f(x) \right|^q d\mathrm{P}_X(x) \\
&= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h) \left(\left(\sum_{j=1}^r \binom{r}{j} (-1)^{2r+1-j} f(x+jh)\right) - f(x)\right) dh \right|^q d\mathrm{P}_X(x) \\
&= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h) \left(\sum_{j=0}^r \binom{r}{j} (-1)^{2r+1-j} f(x+jh)\right) dh \right|^q d\mathrm{P}_X(x) \\
&= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} (-1)^{r+1} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h) \, \triangle^r_h(f,x) \, dh \right|^q d\mathrm{P}_X(x) \ .
\end{aligned}$$

Next, Hölder's inequality and $\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{d/2} K_{\frac{\gamma}{\sqrt{2}}}(h) \, dh = 1$ yield, for $q > 1$,

$$\|K * f - f\|^q_{L_q(\mathrm{P}_X)}$$

$$
\begin{aligned}
&\leq \int_{\mathbb{R}^d}\left(\left(\int_{\mathbb{R}^d}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)dh\right)^{\frac{q-1}{q}}\left(\int_{\mathbb{R}^d}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\left|\triangle_h^r(f,x)\right|^q dh\right)^{\frac{1}{q}}\right)^q d\mathrm{P}_X(x)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\left|\triangle_h^r(f,x)\right|^q dh\, d\mathrm{P}_X(x)\\
&= \int_{\mathbb{R}^d}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\int_{\mathbb{R}^d}\left|\triangle_h^r(f,x)\right|^q d\mathrm{P}_X(x)\, dh\\
&= \int_{\mathbb{R}^d}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\left\|\triangle_h^r(f,\cdot)\right\|_{L_q(\mathrm{P}_X)}^q dh\\
&\leq \int_{\mathbb{R}^d}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\,\omega_{r,L_q(\mathrm{P}_X)}^q(f,\|h\|_2)\, dh\,. \qquad (26)
\end{aligned}
$$

Moreover, for $q=1$, we have

$$
\begin{aligned}
\|K*f-f\|_{L_1(\mathrm{P}_X)} &= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}(-1)^{r+1}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\,\triangle_h^r(f,x)\, dh\right| d\mathrm{P}_X(x)\\
&\leq \int_{\mathbb{R}^d}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\int_{\mathbb{R}^d}\left|\triangle_h^r(f,x)\right| d\mathrm{P}_X(x)\, dh\\
&\leq \int_{\mathbb{R}^d}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\,\omega_{r,L_1(\mathrm{P}_X)}(f,\|h\|_2)\, dh\,.
\end{aligned}
$$

Consequently, (26) holds for all $q\geq 1$. Furthermore, we have

$$
\begin{aligned}
\omega_{r,L_q(\mathrm{P}_X)}^q(f,t) &= \sup_{\|h\|_2\leq t}\int_{\mathbb{R}^d}\left|\triangle_h^r(f,x)\right|^q d\mathrm{P}_X(x)\\
&= \sup_{\|h\|_2\leq t}\int_{\mathbb{R}^d}\left|\triangle_h^r(f,x)\right|^q g(x)d\mu(x)\\
&= \sup_{\|h\|_2\leq t}\int_{\mathbb{R}^d}\left|\triangle_h^r(f,x)\,(g(x))^{\frac{1}{q}}\right|^q d\mu(x)\\
&= \sup_{\|h\|_2\leq t}\left\|\triangle_h^r(f,\cdot)\,g^{\frac{1}{q}}\right\|_{L_q(\mathbb{R}^d)}^q\\
&\leq \sup_{\|h\|_2\leq t}\left(\left\|\triangle_h^r(f,\cdot)\right\|_{L_{qs}(\mathbb{R}^d)}\left\|g^{\frac{1}{q}}\right\|_{L_{qp}(\mathbb{R}^d)}\right)^q\\
&= \|g\|_{L_p(\mathbb{R}^d)}\,\omega_{r,L_{qs}(\mathbb{R}^d)}^q(f,t)\\
&\leq \|g\|_{L_p(\mathbb{R}^d)}\left(1+\frac{2t}{\gamma}\right)^{rq}\omega_{r,L_{qs}(\mathbb{R}^d)}^q\left(f,\frac{\gamma}{2}\right)
\end{aligned}
$$

for $t\geq 0$, where we used (6). Together with (26) this implies

$$\|K*f-f\|_{L_q(\mathrm{P}_X)}^q$$

$$\leq \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h) \, \|g\|_{L_p(\mathbb{R}^d)} \left(1 + \frac{2\,\|h\|_2}{\gamma}\right)^{rq} \omega^q_{r,L_{qs}(\mathbb{R}^d)}\left(f, \frac{\gamma}{2}\right) dh$$

$$= \|g\|_{L_p(\mathbb{R}^d)} \, \omega^q_{r,L_{qs}(\mathbb{R}^d)}\Big(f, \frac{\gamma}{2}\Big) \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h) \left(1 + \frac{2\,\|h\|_2}{\gamma}\right)^{rq} dh\,. \quad (27)$$

Because $\left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(\cdot)$ is the density of a probability measure on $\mathbb{R}^d$, the estimate

$$\left(1 + \frac{2\,\|h\|_2}{\gamma}\right)^{rq} \leq \left(1 + \frac{2\,\|h\|_2}{\gamma}\right)^{\lceil rq \rceil} \leq \sum_{i=0}^{\lceil rq \rceil} \binom{\lceil rq \rceil}{i} \left(\frac{2}{\gamma}\,\|h\|_2\right)^i$$

and Hölder's inequality yield

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h) \left(1 + \frac{2\,\|h\|_2}{\gamma}\right)^{rq} dh \\
&\leq \sum_{i=0}^{\lceil rq \rceil} \binom{\lceil rq \rceil}{i} \left(\frac{2}{\gamma}\right)^i \int_{\mathbb{R}^d} \|h\|_2^i \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\, dh \\
&\leq \sum_{i=0}^{\lceil rq \rceil} \binom{\lceil rq \rceil}{i} \left(\frac{2}{\gamma}\right)^i \left(\int_{\mathbb{R}^d} \|h\|_2^{2i} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\, dh\right)^{\frac{1}{2}} . \qquad (28)
\end{aligned}$$

Since, for $s \geq 0$ and an integer $i \geq 0$, the function $s \mapsto s^i$ is convex, we have for every integer $i \geq 0$ the transformation

$$\left(\sum_{j=1}^d h_j^2\right)^i = d^i \left(\sum_{j=1}^d \frac{1}{d} h_j^2\right)^i \leq d^i \sum_{j=1}^d \frac{1}{d} \left(h_j^2\right)^i = d^{i-1} \sum_{j=1}^d h_j^{2i}\,.$$

Note that $d^{\frac{i-1}{2i}}$ is just the embedding constant of ℓ_{2i}^d to ℓ_2^d. This embedding constant leads to

$$\begin{aligned}
&\int_{\mathbb{R}^d} \|h\|_2^{2i} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\, dh \\
&= \int_{\mathbb{R}^d} \|h\|_2^{2i} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\,\|h\|_2^2}{\gamma^2}\right) dh \\
&\leq d^{i-1} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} \sum_{j=1}^d \int_{\mathbb{R}^d} h_j^{2i} \prod_{l=1}^d \exp\left(-\frac{2h_l^2}{\gamma^2}\right) d(h_1, \ldots, h_d) \\
&= d^{i-1} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} \sum_{j=1}^d \left(\frac{\gamma^2\pi}{2}\right)^{\frac{d-1}{2}} \int_{\mathbb{R}} h_j^{2i} \exp\left(-\frac{2h_j^2}{\gamma^2}\right) dh_j \\
&= d^{i-1} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{1}{2}} 2d \int_0^\infty t^{2i} \exp\left(-\frac{2t^2}{\gamma^2}\right) dt
\end{aligned}$$

$$= 2d^i \left(\frac{2}{\gamma^2\pi}\right)^{\frac{1}{2}} \int_0^\infty t^{2i} \exp\left(-\frac{2t^2}{\gamma^2}\right) dt \ . \tag{29}$$

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

$$\begin{aligned}
\int_0^\infty t^{2i} \exp\left(-\frac{2t^2}{\gamma^2}\right) dt &= \frac{1}{2}\frac{\gamma}{\sqrt{2}} \left(\frac{\gamma^2}{2}\right)^i \int_0^\infty u^{\left(i+\frac{1}{2}\right)-1} \exp\left(-u\right) du \\
&= \frac{1}{2}\frac{\gamma}{\sqrt{2}} \left(\frac{\gamma^2}{2}\right)^i \Gamma\left(i+\frac{1}{2}\right) \\
&= \frac{1}{2}\frac{\gamma}{\sqrt{2}} \left(\frac{\gamma^2}{2}\right)^i \Gamma\left(\frac{1}{2}\right) \prod_{j=1}^{i}\left(j-\frac{1}{2}\right) \\
&= \frac{1}{2}\frac{\gamma}{\sqrt{2}} \left(\frac{\gamma^2}{2}\right)^i \sqrt{\pi} \prod_{j=1}^{i}\left(j-\frac{1}{2}\right) \ .
\end{aligned} \tag{30}$$

Together, (29) and (30) lead to

$$\int_{\mathbb{R}^d} \|h\|_2^{2i} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\, dh \le d^i \left(\frac{\gamma^2}{2}\right)^i \prod_{j=1}^{i}\left(j-\frac{1}{2}\right) \ ,$$

and with (28) we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h) \left(1+\frac{2\,\|h\|_2}{\gamma}\right)^{rq} dh \\
&\le \sum_{i=0}^{\lceil rq\rceil} \binom{\lceil rq\rceil}{i} \left(\frac{2}{\gamma}\right)^i \left(d^i \left(\frac{\gamma^2}{2}\right)^i \prod_{j=1}^{i}\left(j-\frac{1}{2}\right)\right)^{\frac{1}{2}} \\
&= \sum_{i=0}^{\lceil rq\rceil} \binom{\lceil rq\rceil}{i} (2d)^{\frac{i}{2}} \prod_{j=1}^{i}\left(j-\frac{1}{2}\right)^{\frac{1}{2}} \ ,
\end{aligned}$$

where the empty product is defined to equal one. Finally, (27) implies

$$\|K*f-f\|_{L_q(\mathrm{P}_X)}^q \le C_{r,q}\, \|g\|_{L_p(\mathbb{R}^d)}\, \omega_{r,L_{qs}(\mathbb{R}^d)}^q \left(f, \frac{\gamma}{2}\right)$$

for $C_{r,q} := \sum_{i=0}^{\lceil rq\rceil} \binom{\lceil rq\rceil}{i} (2d)^{\frac{i}{2}} \prod_{j=1}^{i}\left(j-\frac{1}{2}\right)^{\frac{1}{2}}$. □

Proof of Lemma 2.3. We define, for all $j \in \mathbb{N}$ and $x \in X$,

$$g_j(x) := \left(\frac{2}{j\gamma\sqrt{\pi}}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}\left(\frac{x}{j}\right) \ . \tag{31}$$

By [30, Proposition 4.46] we obtain

$$g_j * f \in H_{j\gamma}(X) \subset H_\gamma(X)$$

for all $j \in \mathbb{N}$. Due to the properties of the convolution, we finally obtain

$$K * f = \sum_{j=1}^{r} \binom{r}{j} (-1)^{1-j} j^{-\frac{d}{2}} \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} (g_j * f) \in H_\gamma(X) \ .$$

Moreover, for the estimation of the norm we have

$$\begin{aligned}
\|K * f\|_{H_\gamma} &\leq \sum_{j=1}^{r} j^{\frac{d}{2}} \left\| \binom{r}{j} (-1)^{1-j} \left(\frac{2}{j^2\gamma^2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|\cdot\|_2^2}{j^2\gamma^2}\right) * f \right\|_{H_{j\gamma}} \\
&\leq \sum_{j=1}^{r} j^{\frac{d}{2}} \binom{r}{j} \left(j\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} \|f\|_{L_2(\mathbb{R}^d)} \\
&= \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} \left(2^r - 1\right) \|f\|_{L_2(\mathbb{R}^d)} \ ,
\end{aligned}$$

where we used [30, Proposition 4.46] in the first two steps. Finally, for all $x \in X$ and $f \in L_\infty\left(\mathbb{R}^d\right)$, Hölder's inequality implies

$$\begin{aligned}
|K * f(x)| &\leq \sup_{\hat{x}\in X} |K * f(\hat{x})| \\
&\leq \sup_{\hat{x}\in X} \int_{\mathbb{R}^d} |K(\hat{x} - t) f(t)| \, dt \\
&\leq \|f\|_{L_\infty(\mathbb{R}^d)} \sum_{j=1}^{r} \binom{r}{j} \sup_{\hat{x}\in X} \int_{\mathbb{R}^d} \left(\frac{2}{j^2\gamma^2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|\hat{x} - t\|_2^2}{(j\gamma)^2}\right) dt \\
&= \left(2^r - 1\right) \|f\|_{L_\infty(\mathbb{R}^d)} \ .
\end{aligned}$$

□

5.2. Proofs related to the least squares SVMs

To be able to prove the new oracle inequality of Theorem 3.1 we need to control the capacity of H_γ in terms of entropy numbers. For the sake of completeness, we start by recalling entropy numbers (cf. [6] or [30, Definition A.5.26] for more information).

Definition 5.1. Let $S : E \to F$ be a bounded, linear operator between the normed spaces E and F and $i \geq 1$ be an integer. Then the i-th (dyadic) entropy number of S is defined by

$$e_i(S) := \inf\left\{ \varepsilon > 0 : \exists t_1, \ldots, t_{2^{i-1}} \in SB_E \text{ such that } SB_E \subset \bigcup_{j=1}^{2^{i-1}} (t_j + \varepsilon B_F) \right\}$$

where the convention $\inf \emptyset := \infty$ is used.

Immediately, [30, Theorem 7.34] and [30, Corollary 7.31] yield the following lemma regarding the capacity of H_γ, where D_X is the empirical distribution associated to the data set $D_X := (x_1, \ldots, x_n) \in X^n$.

Lemma 5.2. *Let P_X be a distribution on $X \subset B_{\ell_2^d}$, k_γ be the Gaussian RBF kernel over X with width $\gamma \in (0,1]$ and H_γ be the associated RKHS. Then, for all $\varepsilon > 0$ and $0 < p < 1$, there exists a constant $c_{\varepsilon,p} \geq 0$ such that*

$$\mathbb{E}_{D_X \sim \mathrm{P}_X^n} e_i \left(\mathrm{id} : H_\gamma \to L_2(\mathrm{D}_X) \right) \leq c_{\varepsilon,p} \gamma^{-\frac{(1-p)(1+\varepsilon)d}{2p}} i^{-\frac{1}{2p}}$$

for all $i \geq 1$ and $n \geq 1$.

With the help of the above lemma we are now able to deduce an oracle inequality for the least squares loss by specializing [30, Theorem 7.23], which will later be used to prove Theorem 3.1.

Theorem 5.3. *Let $X \subset B_{\ell_2^d}$, $Y := [-M, M] \subset \mathbb{R}$ be a closed subset with $M > 0$ and P be a distribution on $X \times Y$. Furthermore, let $L : Y \times \mathbb{R} \to [0, \infty)$ be the least squares loss, k_γ be the Gaussian RBF kernel over X with width $\gamma \in (0,1]$ and H_γ be the associated RKHS. Fix an $f_0 \in H_\gamma$ and a constant $B_0 \geq 4M^2$ such that $\|L \circ f_0\|_\infty \leq B_0$. Then, for all fixed $\rho \geq 1$, $\lambda > 0$, $\varepsilon > 0$ and $p \in (0,1)$, the SVM using H_γ and L satisfies*

$$\begin{aligned}
&\lambda \|f_{\mathrm{D},\lambda,\gamma}\|_{H_\gamma}^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}_{L,\mathrm{P}}^* \\
&\leq 9\Big(\lambda \|f_0\|_{H_\gamma}^2 + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}_{L,\mathrm{P}}^*\Big) + C_{\varepsilon,p} \frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p n} + \frac{\big(3456M^2 + 15B_0\big)(1+\ln 3)\rho}{n}
\end{aligned}$$

with probability P^n not less than $1 - e^{-\rho}$, where $C_{\varepsilon,p}$ is a constant only depending on ε, p and M.

Proof. First of all, note that, for all $t \in \mathbb{R}$ and $y \in [-M, M]$, the least squares loss satisfies $L(y, \widehat{t}\,) \leq L(y,t)$, i.e. it can be clipped at $M > 0$ (see [32, Section 1]). Furthermore, the least squares loss is locally Lipschitz continuous with the local Lipschitz constant $|L|_{a,1} = 2(a + M)$ for $a > 0$ in the sense of [30, Definition 2.18]. See [30, Example 7.3] to verify that the least squares loss satisfies the supremum bound

$$L(y,t) = (y - t)^2 \leq 4M^2$$

and the variance bound

$$\mathbb{E}_{\mathrm{P}}(L \circ \widehat{f} - L \circ f_{L,\mathrm{P}}^*)^2 \leq 16M^2 \, \mathbb{E}_{\mathrm{P}}(L \circ \widehat{f} - L \circ f_{L,\mathrm{P}}^*)$$

for all $y \in Y$, $t \in [-M, M]$ and $f \in H_\gamma$ with constants $B := 4M^2$, $V := 16M^2$ and $\vartheta := 1$. Consequently, the assertion follows from [30, Theorem 7.23] and Lemma 5.2 with $C_{\varepsilon,p} := C(\max\{c_{\varepsilon,p}, 4M^2\})^{2p}$, $c_{\varepsilon,p}$ as in Lemma 5.2 and a constant $C \geq 1$ which corresponds to the constant K of [30, Theorem 7.23]. Finally, a variable transformation adjusts P^n not to be less than $1 - e^{-\rho}$. □

Now, we can prove the oracle inequality introduced in Theorem 3.1 on the basis of Theorem 5.3.

Proof of Theorem 3.1. First of all, we want to apply Theorem 5.3 for $f_0 := K * f^*_{L,\mathrm{P}}$, where $f^*_{L,\mathrm{P}}$ is a Bayes decision function w.r.t. L and P and

$$K(x) := \sum_{j=1}^{r} \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|x\|_2^2}{j^2\gamma^2}\right)$$

for all $x \in \mathbb{R}^d$. The assumption $f^*_{L,\mathrm{P}} \in L_2\left(\mathbb{R}^d\right)$ and Theorem 2.3 immediately yield

$$f_0 = K * f^*_{L,\mathrm{P}} \in H_\gamma \ .$$

Furthermore, because of $f^*_{L,\mathrm{P}} \in L_\infty\left(\mathbb{R}^d\right)$ and Theorem 2.3, the estimate

$$|K * f^*_{L,\mathrm{P}}(x)| \leq (2^r - 1)\, \|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}$$

holds for all $x \in X$. This implies, for all $(x, y) \in X \times Y$,

$$\begin{aligned} L(y, K * f^*_{L,\mathrm{P}}(x)) &= (y - K * f^*_{L,\mathrm{P}}(x))^2 \\ &= y^2 - 2y(K * f^*_{L,\mathrm{P}}(x)) + (K * f^*_{L,\mathrm{P}}(x))^2 \\ &\leq M^2 + 2M\,(2^r - 1)\, \|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)} + (2^r - 1)^2\, \|f^*_{L,\mathrm{P}}\|^2_{L_\infty(\mathbb{R}^d)} \\ &\leq 4^r \max\{M, \|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}\}^2 \end{aligned}$$

and

$$\begin{aligned} \|L \circ f_0\|_\infty &= \sup_{(x,y)\in X\times Y} |L(y, f_0(x))| \\ &= \sup_{(x,y)\in X\times Y} \left|L\left(y, K * f^*_{L,\mathrm{P}}(x)\right)\right| \\ &\leq 4^r \max\{M, \|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}\}^2 =: B_0 \ . \end{aligned}$$

Furthermore, (12) and Theorem 2.2 yield

$$\begin{aligned} \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}} &= \mathcal{R}_{L,\mathrm{P}}(K * f^*_{L,\mathrm{P}}) - \mathcal{R}^*_{L,\mathrm{P}} \\ &= \left\|K * f^*_{L,\mathrm{P}} - f^*_{L,\mathrm{P}}\right\|^2_{L_2(\mathrm{P}_X)} \\ &\leq C_{r,2}\, \|g\|_{L_q(\mathbb{R}^d)}\, \omega^2_{r,L_{2s}(\mathbb{R}^d)}\left(f^*_{L,\mathrm{P}}, \frac{\gamma}{2}\right) \\ &\leq C_{r,2}\, \|g\|_{L_q(\mathbb{R}^d)}\, c^2\gamma^{2\alpha} \ , \end{aligned}$$

where we used

$$\omega_{r,L_{2s}(\mathbb{R}^d)}\left(f^*_{L,\mathrm{P}}, \frac{\gamma}{2}\right) \leq c\gamma^\alpha$$

for $\gamma > 0$, $\alpha \geq 1$, $r = \lfloor \alpha \rfloor + 1$, and a constant $c > 0$ in the last step, which in turn immediately results from the assumption $f^*_{L,\mathrm{P}} \in B^{\alpha}_{2s,\infty}(\mathbb{R}^d)$. By Theorem 2.3 we know

$$\|f_0\|_{H_\gamma} = \|K * f^*_{L,\mathrm{P}}\|_{H_\gamma} \leq \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} \left(2^r - 1\right) \|f^*_{L,\mathrm{P}}\|_{L_2(\mathbb{R}^d)} \,.$$

Therefore, Theorem 5.3 and the above choice of f_0 yield, for all fixed $\rho \geq 1$, $\lambda > 0$, $\varepsilon > 0$, and $p \in (0,1)$, that the SVM using H_γ and L satisfies

$$\begin{aligned}
&\lambda \|f_{\mathrm{D},\lambda,\gamma}\|^2_{H_\gamma} + \mathcal{R}_{L,\mathrm{P}}\left(\widehat{f}_{\mathrm{D},\lambda,\gamma}\right) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq 9\left(\lambda\left(\gamma\sqrt{\pi}\right)^{-d}\left(2^r-1\right)^2 \|f^*_{L,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)} + C_{r,2}\,\|g\|_{L_q(\mathbb{R}^d)}\, c^2\gamma^{2\alpha}\right) \\
&\quad + C_{\varepsilon,p}\frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p n} + \frac{\left(3456M^2 + 15\cdot 4^r \max\{M, \|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}\}^2\right)(1+\ln 3)\rho}{n} \\
&\leq C_1\lambda\gamma^{-d} + 9\,C_r c^2\gamma^{2\alpha} + C_{\varepsilon,p}\frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p n} + \frac{C_2\rho}{n} \qquad (32)
\end{aligned}$$

with probability P^n not less than $1 - e^{-\rho}$ and with constants $C_{\varepsilon,p}$ as in Theorem 5.3, $C_r := C_{r,2}\,\|g\|_{L_q(\mathbb{R}^d)}$, $C_1 := 9\left(2^r-1\right)^2 \pi^{-\frac{d}{2}} \|f^*_{L,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)}$, and $C_2 := (\ln(3)+1)\left(3456M^2 + 15\cdot 4^r \max\{M, \|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}\}^2\right)$. □

Remark 5.4. Consider the case $M \geq 1$ in the proof of Theorem 3.1. Then we have

$$C_2 \leq M^2(\ln(3)+1)\left(3456 + 15\cdot 4^r \max\{1, \|f^*_{L,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}\}^2\right)$$

as well as

$$C_{\varepsilon,p} = (\max\{c_{\varepsilon,p}, 4M^2\})^{2p} \cdot \max\{3\tilde{K}, 2B\}$$

by Theorem 5.3 and [30, Theorem 7.23], where $c_{\varepsilon,p}$ is independent of M, $B = 4M^2$, and

$$\tilde{K} = \max\{c_1|L|^p_{M,1}V^{\frac{1-p}{2}}, c_2|L|^{2p}_{M,1}B^{1-p}\} = \max\{4c_1M, 4^{1+p}c_2M^2\} = \hat{c}M^2$$

with Lipschitz constant $|L|_{M,1} = 4M$ and $V = 16M^2$. Since

$$\begin{aligned}
C_{\varepsilon,p} &= (\max\{c_{\varepsilon,p}, 4M^2\})^{2p} \cdot \max\{3\hat{c}M^2, 8M^2\} \\
&\leq (\max\{c_{\varepsilon,p}, 4\})^{2p} \max\{3\hat{c}, 8\} M^{2+4p} \,,
\end{aligned}$$

we can rewrite (32) such that

$$\begin{aligned}
&\lambda \|f_{\mathrm{D},\lambda,\gamma}\|^2_{H_\gamma} + \mathcal{R}_{L,\mathrm{P}}\left(\widehat{f}_{\mathrm{D},\lambda,\gamma}\right) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq CM^{2+4p}\left(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p n} + \frac{\rho}{n}\right) \,,
\end{aligned}$$

where the positive constant C is independent of M, λ, γ, ρ, and n.

With the help of the oracle inequality achieved in Theorem 3.1 the learning rate stated in Corollary 3.2 can be shown in a few steps.

Proof of Corollary 3.2. In a first step, Theorem 3.1 can be applied which yields

$$\begin{aligned}
&\lambda_n \left\| f_{\mathrm{D},\lambda_n,\gamma_n} \right\|_{H_{\gamma_n}}^2 + \mathcal{R}_{L,\mathrm{P}} \left(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n} \right) - \mathcal{R}_{L,\mathrm{P}}^* \\
&\leq C_1 \lambda_n \gamma_n^{-d} + 9\, C_r c^2 \gamma_n^{2\alpha} + C_{\varepsilon,p} \frac{\gamma_n^{-(1-p)(1+\varepsilon)d}}{\lambda_n^p n} + \frac{C_2 \rho}{n} \\
&\leq \widetilde{C} \rho \left(\lambda_n \gamma_n^{-d} + \gamma_n^{2\alpha} + \gamma_n^{-(1-p)(1+\varepsilon)d} \lambda_n^{-p} n^{-1} + n^{-1} \right)
\end{aligned}$$

for all $\varepsilon > 0$ and $p \in (0,1)$ with probability P^n not less than $1 - e^{-\rho}$ and a constant $\widetilde{C} := \max\left\{ C_1, 9\, C_r c^2, C_{\varepsilon,p}, C_2 \right\}$. In a next step the sequences

$$\lambda_n = c_1 n^{-1}$$

and

$$\gamma_n = c_2 n^{-\frac{1}{2\alpha+d}}$$

with arbitrary constants $c_1 > 0$ and $c_2 > 0$ yield

$$\lambda_n \gamma_n^{-d} + \gamma_n^{2\alpha} + \gamma_n^{-(1-p)(1+\varepsilon)d} \lambda_n^{-p} n^{-1} \leq c_3 n^{-\frac{2\alpha}{2\alpha+d} + \frac{2\alpha p + \varepsilon(1-p)d}{2\alpha+d}} \leq c_3 n^{-\frac{2\alpha}{2\alpha+d}+\xi} \,,$$

where $c_3 > 0$ is a constant and $\xi \geq \frac{2\alpha p + \varepsilon(1-p)d}{2\alpha+d}$. With this, we finally obtain

$$\begin{aligned}
\lambda_n \left\| f_{\mathrm{D},\lambda_n,\gamma_n} \right\|_{H_{\gamma_n}}^2 + \mathcal{R}_{L,\mathrm{P}} \left(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n} \right) - \mathcal{R}_{L,\mathrm{P}}^* &\leq \widetilde{C} \rho \left(c_3 n^{-\frac{2\alpha}{2\alpha+d}+\xi} + n^{-1} \right) \\
&\leq C \rho\, n^{-\frac{2\alpha}{2\alpha+d}+\xi}
\end{aligned}$$

with the constant $C := \widetilde{C}\,(c_3 + 1)$. □

Next, we want to prove Theorem 3.3. To this end, we need the following technical lemma.

Lemma 5.5. *We fix finite subsets $\Lambda_n, \Gamma_n \subset (0,1]$ such that Λ_n is an ϵ_n-net of $(0,1]$ and Γ_n is an δ_n-net of $(0,1]$ with $0 < \epsilon_n \leq n^{-1}$, $\delta_n > 0$, $1 \in \Lambda_n$, and $1 \in \Gamma_n$. Then, for all $\varepsilon > 0$, $p \in (0,1)$, $d > 0$, $\alpha > 0$ and all $n \geq 1$, we have*

$$\inf_{(\lambda,\gamma) \in \Lambda \times \Gamma} \left(\lambda \gamma^{-d} + \gamma^{2\alpha} + n^{-1} \lambda^{-p} \gamma^{-(1-p)(1+\varepsilon)d} \right) \leq c \left(n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \delta_n^{2\alpha} \right) \,,$$

where $\xi > 0$ depends on ε and p, and $c > 0$ is a constant independent of n, Λ, ϵ_n, Γ, and δ_n.

Proof. Without loss of generality, we may assume that Λ_n and Γ_n are of the form $\Lambda_n = \{\lambda_1, \ldots, \lambda_m\}$ and $\Gamma_n = \{\gamma_1, \ldots, \gamma_l\}$ with $\lambda_m = 1$ and $\gamma_l = 1$ as well as $\lambda_{i-1} < \lambda_i$ and $\gamma_{j-1} < \gamma_j$ for all $i = 2, \ldots, m$ and $j = 2, \ldots, l$. Furthermore, we fix a minimizer (λ^*, γ^*) of the function $(\lambda, \gamma) \to \lambda \gamma^{-d} + \gamma^{2\alpha} + n^{-1} \lambda^{-p} \gamma^{-(1-p)(1+\varepsilon)d}$

defined on $[0,1]^2$. [30, Lemma A.1.6.] shows that $\lambda^* = c_1 n^{-\frac{2\alpha+d}{2\alpha+2\alpha p+dp+(1-p)(1+\varepsilon)d}}$ with a constant $c_1 > 0$. This implies $\epsilon_n \leq \frac{1}{c_1}\lambda^*$. With $\lambda_0 := 0$ and $\gamma_0 := 0$ it is easy to see that

$$\lambda_i - \lambda_{i-1} \leq 2\epsilon_n \qquad \text{and} \qquad \gamma_j - \gamma_{j-1} \leq 2\delta_n \tag{33}$$

hold for all $i = 1,\ldots,m$ and $j = 1,\ldots,l$. Furthermore, there exist indices $i \in \{1,\ldots,m\}$ and $j \in \{1,\ldots,l\}$ such that $\lambda_{i-1} \leq \lambda^* \leq \lambda_i$ and $\gamma_{j-1} \leq \gamma^* \leq \gamma_j$. Together with (33) this yields $\lambda^* \leq \lambda_i \leq \lambda^* + 2\epsilon_n$ and $\gamma^* \leq \gamma_j \leq \gamma^* + 2\delta_n$. Using this result and [30, Lemma A.1.6.], we obtain

$$\begin{aligned}
&\inf_{(\lambda,\gamma)\in\Lambda\times\Gamma} \left(\lambda\gamma^{-d} + \gamma^{2\alpha} + n^{-1}\lambda^{-p}\gamma^{-(1-p)(1+\varepsilon)d}\right)\\
&\leq \lambda_i\gamma_j^{-d} + \gamma_j^{2\alpha} + n^{-1}\lambda_i^{-p}\gamma_j^{-(1-p)(1+\varepsilon)d}\\
&\leq (\lambda^* + 2\epsilon_n)\,(\gamma^*)^{-d} + (\gamma^* + 2\delta_n)^{2\alpha} + n^{-1}\,(\lambda^*)^{-p}\,(\gamma^*)^{-(1-p)(1+\varepsilon)d}\\
&\leq (1 + \frac{2}{c_1})\lambda^*\,(\gamma^*)^{-d} + (\gamma^* + 2\delta_n)^{2\alpha} + n^{-1}\,(\lambda^*)^{-p}\,(\gamma^*)^{-(1-p)(1+\varepsilon)d}\\
&\leq c_2\left(\lambda^*\,(\gamma^*)^{-d} + (\gamma^*)^{2\alpha} + n^{-1}\,(\lambda^*)^{-p}\,(\gamma^*)^{-(1-p)(1+\varepsilon)d} + \delta_n^{2\alpha}\right)\\
&= c_2 \min_{\lambda,\gamma\in[0,1]}\left(\lambda\gamma^{-d} + \gamma^{2\alpha} + n^{-1}\lambda^{-p}\gamma^{-(1-p)(1+\varepsilon)d}\right) + c_2\delta_n^{2\alpha}\\
&\leq c_2\,c_3\,n^{-\frac{2\alpha}{2\alpha+2\alpha p+dp+(1-p)(1+\varepsilon)d}} + c_2\delta_n^{2\alpha}\\
&\leq c\left(n^{-\frac{2\alpha}{2\alpha+2\alpha p+dp+(1-p)(1+\varepsilon)d}} + \delta_n^{2\alpha}\right)\\
&\leq c\left(n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \delta_n^{2\alpha}\right)
\end{aligned}$$

with constants $c_2 > 0$, $c_3 > 0$ and $c := \max\{c_2\,c_3, c_2\}$ independent of n, Λ, ϵ_n, Γ, and δ_n. □

Proof of Theorem 3.3. Let m be defined by $m := \lfloor \frac{n}{2} \rfloor + 1$, i.e. $m \geq \frac{n}{2}$. Then Theorem 3.1 yields with probability P^m not less than $1 - |\Lambda_n \times \Gamma_n|\, e^{-\rho}$

$$\begin{aligned}
\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} &\leq \frac{c_1}{2}\left(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p m} + \frac{\rho}{m}\right)\\
&\leq c_1\left(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p n} + \frac{\rho}{n}\right)
\end{aligned} \tag{34}$$

for all $(\lambda,\gamma) \in \Lambda_n \times \Gamma_n$ simultaneously. Here, $c_1 > 0$ is a constant independent of n, ρ, λ, and γ. Furthermore, [30, Theorem 7.2], $n - m \geq \frac{n}{2} - 1 \geq \frac{n}{4}$, and $\rho_n := \rho + \ln(1 + |\Lambda_n \times \Gamma_n|)$ yield

$$\begin{aligned}
&\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}) - \mathcal{R}^*_{L,\mathrm{P}}\\
&< 6\left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}}\right) + 512M^2\frac{\rho_n}{n-m}
\end{aligned}$$

$$< 6 \left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \right) + 2048M^2\frac{\rho_n}{n} \tag{35}$$

with probability P^{n-m} not less than $1-e^{-\rho}$. With (34), (35), and Lemma 5.5 we can conclude

$$\begin{aligned}
&\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}) - \mathcal{R}^*_{L,\mathrm{P}} \\
&< 6 \left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \right) + 2048M^2\frac{\rho_n}{n} \\
&\le 6c_1 \left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \left(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p n} \right) + \frac{\rho}{n} \right) + 2048M^2\frac{\rho_n}{n} \\
&\le 6c_1 \left(c\left(n^{-\frac{2\alpha}{2\alpha+d}+\xi} + \delta_n^{2\alpha} \right) + \frac{\rho}{n} \right) + 2048M^2\frac{\rho_n}{n} \\
&\le \left(6c_1c + 6c_1\rho + 2048M^2\rho_n \right) n^{-\frac{2\alpha}{2\alpha+d}+\xi} + 6c_1c\delta_n^{2\alpha} \\
&\le \left(12c_1c + 6c_1\rho + 2048M^2\rho_n \right) n^{-\frac{2\alpha}{2\alpha+d}+\xi}
\end{aligned}$$

with probability P^n not less than $1-(1+|\Lambda_n\times\Gamma_n|)\,e^{-\rho}$. With a variable transformation P^n can be adjusted such that it is not less than $1-e^{-\rho}$. □

For the least squares loss it finally remains to prove learning rates in the case of unbounded noise.

Proof of Theorem 3.6. By (17), we obtain

$$\begin{aligned}
\mathrm{P}^n\left(\left\{ D\in(X\times Y)^n : \max_{i\in\{1,\ldots,n\}} \{|\varepsilon_i|\} \le c\rho^l \right\} \right) &\ge 1 - \sum_{i=1}^n \mathrm{P}\left(|\varepsilon_i| > c\rho^l \right) \\
&\ge 1 - e^{-(\rho-\ln n)} \,.
\end{aligned}$$

Thus, we have

$$\mathrm{P}^n\left(\left\{ D\in(X\times Y)^n : \max_{i\in\{1,\ldots,n\}} \{|\varepsilon_i|\} \le c\left(\hat{\rho}+\ln n\right)^l \right\} \right) \ge 1 - e^{-\hat{\rho}} \,,$$

i.e. with probability P^n not less than $1-e^{-\hat{\rho}}$ we have $|y_i| \le M_n$ for all $i\in\{1,\ldots,n\}$. Therefore, the usual LS-SVM with belatedly clipped decision function at M_n is with probability P^n not less than $1-e^{-\hat{\rho}}$ clipped regularized empirical risk minimization (CR-ERM) in the sense of [30, Definition 7.18]. Since in the proof of [30, Theorem 7.20] the CR-ERM property is used exactly once, namely at the very beginning of the proof, while the rest of the proof only considers clipped decision functions independently of whether they are CR-ERMs or not, the oracle inequality of [30, Theorem 7.20] holds for $\hat{Y} := [-M_n, M_n]$ modulo a set of probability P^n not less than $1-e^{-\hat{\rho}}$. Analogously to Theorem 3.1 we then obtain that

$$\lambda \left\| f_{\mathrm{D},\lambda,\gamma} \right\|^2_{H_\gamma} + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \le C M_n^{2+4p} \left(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\epsilon)d}}{\lambda^p n} + \frac{\rho}{n} \right)$$

holds with probability P^n not less than $1 - e^{-\bar{\rho}} - e^{-\hat{\rho}}$, where $\bar{\rho} \geq 1$, $\epsilon > 0$, and $p \in (0,1)$. Here we used Remark 5.4 to determine how the arising constants depend on M_n. Together with Corollary 3.2 and $(a+b)^c \leq (2ab)^c$ for $a, b \geq 1$ and $c > 0$ this yields

$$\begin{aligned}
&\lambda \|f_{\mathrm{D},\lambda,\gamma}\|_{H_\gamma}^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}_{L,\mathrm{P}}^* \\
&\leq \hat{C}\bar{\rho} M_n^{2+4p} n^{-\frac{2\alpha}{2\alpha+d}+\xi} \\
&= \hat{C}\bar{\rho} \left(2c\left(\hat{\rho} + \ln n\right)^l\right)^{2+4p} n^{-\frac{2\alpha}{2\alpha+d}+\xi} \\
&\leq \hat{C}\bar{\rho}\, 2^{6(1+l)} c^6 \hat{\rho}^{2l(1+2p)} n^{-\frac{2\alpha}{2\alpha+d}+\xi} \left(\ln n\right)^{2l(1+2p)} \\
&\leq C\bar{\rho}\hat{\rho}^{2l+\xi'} n^{-\frac{2\alpha}{2\alpha+d}+\hat{\xi}}
\end{aligned}$$

for $\xi, \hat{\xi}, \xi' > 0$, and $n \geq 3$ with probability P^n not less than $1 - e^{-\bar{\rho}} - e^{-\hat{\rho}}$. □

5.3. Proofs related to SVMs for quantile regression

Let Q be a distribution on $\mathbb{R}$ with $\operatorname{supp} Q \subset [-1,1]$ and, for $\tau \in (0,1)$, L_τ be the τ-pinball loss. We define the inner L_τ-risk by

$$\mathcal{C}_{L_\tau,Q}(t) := \int_Y L_\tau(y,t)\, dQ(y), \qquad t \in \mathbb{R},$$

and the minimal inner L_τ-risk by $\mathcal{C}_{L_\tau,Q}^* := \inf_{t\in\mathbb{R}} \mathcal{C}_{L_\tau,Q}(t)$. With this definition we first present an estimate of the inner L_τ-risk in the following lemma and afterwards we can prove Theorem 4.7 that estimates the excess risk.

Lemma 5.6. *Let Q be a distribution on $\mathbb{R}$ with $\operatorname{supp} Q \subset [-1,1]$ that has a τ-quantile of upper type $q > 1$. For $\tau \in (0,1)$, let $F_{\tau,Q}^*$ consist of singletons, i.e. there exists an $t^* \in \mathbb{R}$ with $F_{\tau,Q}^* = \{t^*\}$. Furthermore, let $Q(\{t^*\}) = 0$. Then*

$$\mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}_{L_\tau,Q}^* \leq \frac{b_Q}{q}|t - t^*|^q$$

holds for all $t \in \mathbb{R}$.

Proof. [31, Proposition 4.1] yields

$$\mathcal{C}_{L_\tau,Q}(t^* + t) - \mathcal{C}_{L_\tau,Q}^* = \int_0^t Q\left((t^*, t^* + s)\right) ds \leq \int_0^t b_Q s^{q-1} ds \leq \frac{b_Q}{q} t^q$$

and

$$\mathcal{C}_{L_\tau,Q}(t^* - t) - \mathcal{C}_{L_\tau,Q}^* = \int_0^t Q\left((t^* - s, t^*)\right) ds \leq \int_0^t b_Q s^{q-1} ds \leq \frac{b_Q}{q} t^q \tag{36}$$

for all $t \geq 0$. With this, we have, for $t \geq t^*$,

$$\mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}_{L_\tau,Q}^* = \mathcal{C}_{L_\tau,Q}(t^* + (t - t^*)) - \mathcal{C}_{L_\tau,Q}^* \leq \frac{b_Q}{q}(t - t^*)^q = \frac{b_Q}{q}|t - t^*|^q .$$

The case $t < t^*$ follows analogously with (36). □

Proof of Theorem 4.7. With Lemma 5.6 and the choice $Q := \mathrm{P}(\cdot|x)$ for all $x \in X$, we obtain

$$\begin{aligned}
&\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&= \int_X \int_Y L_\tau(y, f(x))\, d\mathrm{P}(y|x) d\mathrm{P}_X(x) - \int_X \int_Y L_\tau(y, f^*_{\tau,\mathrm{P}}(x))\, d\mathrm{P}(y|x) d\mathrm{P}_X(x) \\
&= \int_X \mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}(f(x)) - \mathcal{C}^*_{L_\tau,\mathrm{P}(\cdot|x)}\, d\mathrm{P}_X(x) \\
&\leq \int_X \frac{b_{\mathrm{P}(\cdot|x)}}{q} |f(x) - f^*_{\tau,\mathrm{P}}(x)|^q\, d\mathrm{P}_X(x) \\
&= q^{-1} \|b_{\mathrm{P}(\cdot|x)}\|_{L_p(\mathrm{P}_X)} \|f - f^*_{\tau,\mathrm{P}}\|^q_{L_u(\mathrm{P}_X)}
\end{aligned}$$

for every $f : X \to [-1,1]$. □

Proof of Theorem 4.8. By [30, Section 9.3 and Lemma 2.23] we know that, for all $\tau \in (0,1)$, the τ-pinball loss L_τ is Lipschitz continuous and can be clipped at $M = 1$ for $Y := [-1,1]$. Furthermore, for all $\tau \in (0,1)$, the supremum bound is satisfied for the τ-pinball loss, since

$$L_\tau(y,t) = \max\{\tau, 1-\tau\}|y-t| \leq 2 =: B$$

holds for all $y \in Y$ and all $t \in [-1,1]$. By Lemma 5.2 we know that, for all $\varepsilon > 0$ and $0 < \varsigma < 1$, there exists a constant $c_{\varepsilon,\varsigma} \geq 0$ such that

$$\mathbb{E}_{D_X \sim \mathrm{P}^n_X} e_i\left(\mathrm{id} : H_\gamma \to L_2(\mathrm{D}_X)\right) \leq c_{\varepsilon,\varsigma} \gamma^{-\frac{(1-\varsigma)(1+\varepsilon)d}{2\varsigma}} i^{-\frac{1}{2\varsigma}}$$

for all $i \geq 1$ and $n \geq 1$.

Since we assume that there exist constants $\vartheta \in [0,1]$ and $V \geq B^{2-\vartheta} = 2^{2-\vartheta}$ such that the variance bound (22) is satisfied for all $f : \mathbb{R}^d \to \mathbb{R}$, we can apply [30, Theorem 7.23]. To this end, we choose $f_0 := K * f^*_{\tau,\mathrm{P}}$, where $K : \mathbb{R}^d \to \mathbb{R}$ is defined by (8). Theorem 2.3 and the assumption $f^*_{\tau,\mathrm{P}} \in L_2(\mathbb{R}^d)$ then imply $f_0 \in H_\gamma$ and

$$\|f_0\|_{H_\gamma} \leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} (2^r - 1) \|f^*_{\tau,\mathrm{P}}\|_{L_2(\mathbb{R}^d)}\,.$$

Next, the assumption $f^*_{\tau,\mathrm{P}} \in L_\infty(\mathbb{R}^d)$ together with Theorem 2.3 yields

$$|K * f^*_{\tau,\mathrm{P}}(x)| \leq (2^r - 1) \|f^*_{\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)} \tag{37}$$

for all $x \in X$. Furthermore, for all $(x,y) \in X \times Y$, the latter implies

$$\begin{aligned}
L_\tau(y, K * f^*_{\tau,\mathrm{P}}(x)) &\leq |y - K * f^*_{\tau,\mathrm{P}}(x)| \\
&\leq 1 + (2^r - 1) \|f^*_{\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)} \\
&\leq 2^r a\,,
\end{aligned}$$

where $a := \max\{1, \|f^*_{\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}\}$. With this, we obtain

$$\|L_\tau \circ f_0\|_\infty = \sup_{(x,y)\in X\times Y} |L_\tau(y, K * f^*_{\tau,\mathrm{P}}(x))| \le 2^r a =: B_0\,,$$

where $B_0 = 2^r a \ge 2 = B$. In addition, we have to estimate the excess risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$. To this end, we apply Theorem 4.7 and Theorem 2.2 and derive

$$\begin{aligned}
\mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}} &\le q^{-1}\|b_{\mathrm{P}(\,\cdot\,|x)}\|_{L_p(\mathrm{P}_X)}\|f_0 - f^*_{\tau,\mathrm{P}}\|^q_{L_u(\mathrm{P}_X)} \\
&= q^{-1}\|b_{\mathrm{P}(\,\cdot\,|x)}\|_{L_p(\mathrm{P}_X)}\|K * f^*_{\tau,\mathrm{P}} - f^*_{\tau,\mathrm{P}}\|^q_{L_u(\mathrm{P}_X)} \\
&\le q^{-1}\|b_{\mathrm{P}(\,\cdot\,|x)}\|_{L_p(\mathrm{P}_X)}\left(C_{r,u}\|g\|_{L_w(\mathbb{R}^d)}\omega^u_{r,L_{us}(\mathbb{R}^d)}(f^*_{\tau,\mathrm{P}}, \frac{\gamma}{2})\right)^{\frac{q}{u}} \\
&\le q^{-1}\|b_{\mathrm{P}(\,\cdot\,|x)}\|_{L_p(\mathrm{P}_X)}C^{\frac{q}{u}}_{r,u}\|g\|^{\frac{q}{u}}_{L_w(\mathbb{R}^d)}c^q\gamma^{q\alpha}\,,
\end{aligned}$$

where we used $f^*_{\tau,\mathrm{P}} \in B^\alpha_{us,\infty}(\mathbb{R}^d)$. Finally, [30, Theorem 7.23] yields that, for all fixed $\rho > 0$ and $\lambda > 0$, the SVM using H_γ and L_τ satisfies

$$\begin{aligned}
&\lambda\,\|f_{\mathrm{D},\lambda,\gamma}\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}\left(\widehat{f}_{\mathrm{D},\lambda,\gamma}\right) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\le 9(\lambda\|f_0\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}}) \\
&\quad + c_1\left(\frac{c^{2\varsigma}_{\varepsilon,\varsigma}\gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma n}\right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + 3\left(\frac{72V\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\rho}{n} \\
&\le 9\left(\lambda(2^r-1)^2(\gamma\sqrt{\pi})^{-d}\|f^*_{\tau,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)} + \frac{\|b_{\mathrm{P}(\,\cdot\,|x)}\|_{L_p(\mathrm{P}_X)}C^{\frac{q}{u}}_{r,u}\|g\|^{\frac{q}{u}}_{L_w(\mathbb{R}^d)}c^q}{q}\gamma^{q\alpha}\right) \\
&\quad + c_1\left(\frac{c^{2\varsigma}_{\varepsilon,\varsigma}\gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma n}\right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + 3\left(\frac{72V\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15\cdot 2^r a\rho}{n} \\
&\le C\left(\lambda\gamma^{-d} + \gamma^{q\alpha} + \left(\frac{\gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma n}\right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + \left(\frac{\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{\rho}{n}\right)
\end{aligned}$$

with probability P^n not less than $1 - e^{-\rho}$ and a constant $C > 0$ depending on $\|f^*_{\tau,\mathrm{P}}\|_{L_2(\mathbb{R}^d)}$, $\|f^*_{\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}$, r, d, q, p, $\|b_{\mathrm{P}(\,\cdot\,|x)}\|_{L_p(\mathrm{P}_X)}$, $\|g\|_{L_w(\mathbb{R}^d)}$, ε, ς, ϑ, and V. □

With the help of the just proven oracle inequality we now derive the learning rates of Corollary 4.9.

Proof of Corollary 4.9. Theorem 4.8 yields

$$\begin{aligned}
&\lambda_n\,\|f_{\mathrm{D},\lambda_n,\gamma_n}\|^2_{H_{\gamma_n}} + \mathcal{R}_{L,\mathrm{P}}\left(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n}\right) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\le c\left(\lambda_n\gamma_n^{-d} + \gamma_n^{q\alpha} + \left(\frac{\gamma_n^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda_n^\varsigma n}\right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + \left(\frac{\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{\rho}{n}\right)\,,
\end{aligned}$$

where $c > 0$ is a constant. In addition, the sequences

$$\lambda_n = c_1 n^{-\frac{q\alpha+d}{q\alpha(2-\vartheta)+d}}$$

and

$$\gamma_n = c_2 n^{-\frac{1}{q\alpha(2-\vartheta)+d}}$$

with arbitrary constants $c_1 > 0$ and $c_2 > 0$ yield

$$\begin{aligned}
&\lambda_n \gamma_n^{-d} + \gamma_n^{q\alpha} + \left(\frac{\gamma_n^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda_n^{\varsigma} n} \right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} \\
&\leq c_3 n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d} + \frac{q\alpha\vartheta\rho+\varepsilon d(1-\rho)}{(q\alpha(2-\vartheta)+d)(2-\rho-\vartheta+\vartheta\rho)}} \\
&\leq c_3 n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d}+\xi} \,,
\end{aligned}$$

where $c_3 > 0$ is a constant and $\xi \geq \frac{q\alpha\vartheta\rho+\varepsilon d(1-\rho)}{(q\alpha(2-\vartheta)+d)(2-\rho-\vartheta+\vartheta\rho)}$. With this, we finally obtain

$$\begin{aligned}
&\lambda_n \left\| f_{\mathrm{D},\lambda_n,\gamma_n} \right\|_{H_{\gamma_n}}^2 + \mathcal{R}_{L,\mathrm{P}} \left(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n} \right) - \mathcal{R}_{L,\mathrm{P}}^* \\
&\leq c \left(c_3 n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d}+\xi} + \left(\frac{\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{\rho}{n} \right) \\
&\leq C\rho\, n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d}+\xi}
\end{aligned}$$

with probability P^n not less than $1 - e^{-\rho}$ and with the constant $C := c(c_3 + 2)$. □

To prove Theorem 4.10 we need the following lemma.

Lemma 5.7. *We fix finite subsets $\Lambda_n, \Gamma_n \subset (0,1]$ such that Λ_n is an ϵ_n-net of $(0,1]$ and Γ_n is an δ_n-net of $(0,1]$ with $0 < \epsilon_n < n^{-1}$, $\delta_n > 0$, $1 \in \Lambda_n$, and $1 \in \Gamma_n$. Then, for all $\varepsilon > 0$, $\varsigma \in (0,1)$, $\vartheta \in [0,1]$, $q \in [1,\infty)$, $d > 0$, $\alpha > 0$, and all $n \geq 1$, we have*

$$\begin{aligned}
&\inf_{(\lambda,\gamma)\in\Lambda\times\Gamma} \left(\lambda\gamma^{-d} + \gamma^{q\alpha} + \left(\lambda^{-\varsigma} n^{-1} \gamma^{-(1-\varsigma)(1+\varepsilon)d} \right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} \right) \\
&\leq c \left(n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d}+\xi} + \delta_n^{q\alpha} \right)
\end{aligned}$$

with $\xi > 0$ depending on ε and ς, and a constant $c > 0$ independent of n, Λ, ϵ_n, Γ, and δ_n.

Proof. Let (λ^*, γ^*) be the minimizer of the function

$$(\lambda,\gamma) \to \lambda\gamma^{-d} + \gamma^{q\alpha} + \left(\lambda^{-\varsigma} n^{-1} \gamma^{-(1-\varsigma)(1+\varepsilon)d} \right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}}$$

defined on $[0,1]^2$. [30, Lemma A.1.6.] shows that

$$\lambda^* = c_1 n^{-\frac{q\alpha+d}{q\alpha(2-\varsigma-\vartheta+\vartheta\varsigma)+q\alpha\varsigma+d\varsigma+(1-\varsigma)(1+\varepsilon)d}}$$

with a constant $c_1 > 0$. This implies $\epsilon_n \leq \frac{1}{c_1}\lambda^*$. Now the proof follows analogously to the proof of Lemma 5.5. □

Proof of Theorem 4.10. Let m be defined by $m := \lfloor \frac{n}{2} \rfloor + 1$, i.e. $m \geq \frac{n}{2}$. Therefore, Theorem 4.8 yields with probability P^m not less than $1 - |\Lambda_n \times \Gamma_n|\, e^{-\rho}$

$$\begin{aligned}
&\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\leq \frac{c_1}{2}\left(\lambda\gamma^{-d} + \gamma^{q\alpha} + \left(\frac{\gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma m}\right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + \left(\frac{\rho}{m}\right)^{\frac{1}{2-\vartheta}} + \frac{\rho}{m}\right) \\
&\leq c_1\left(\lambda\gamma^{-d} + \gamma^{q\alpha} + \left(\frac{\gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma n}\right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + \left(\frac{\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{\rho}{n}\right) \qquad (38)
\end{aligned}$$

for all $(\lambda,\gamma) \in \Lambda_n \times \Gamma_n$ simultaneously. Here, $c_1 > 0$ is a constant. Furthermore, [30, Theorem 7.2], $n - m \geq \frac{n}{2} - 1 \geq \frac{n}{4}$, and $\rho_n := \rho + \ln(1 + |\Lambda_n \times \Gamma_n|)$ yield

$$\begin{aligned}
&\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&< 6\left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}\right) + 4\left(\frac{8V\rho_n}{n-m}\right)^{\frac{1}{2-\vartheta}} \\
&< 6\left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}\right) + 4\left(\frac{32V\rho_n}{n}\right)^{\frac{1}{2-\vartheta}} \qquad (39)
\end{aligned}$$

with probability P^{n-m} not less than $1 - e^{-\rho}$. With (38), (39), and Lemma 5.7 we can conclude

$$\begin{aligned}
&\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&< 6\left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}\right) + 4\left(\frac{32V\rho_n}{n}\right)^{\frac{1}{2-\vartheta}} \\
&\leq 6c_1 \inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n}\left(\lambda\gamma^{-d} + \gamma^{q\alpha} + \left(\frac{\gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma n}\right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}}\right) \\
&\quad + 6c_1\left(\frac{\rho}{n}\right)^{\frac{1}{2-\vartheta}} + 6c_1\frac{\rho}{n} + 4\left(\frac{32V\rho_n}{n}\right)^{\frac{1}{2-\vartheta}} \\
&\leq 6c_1\left(c\left(n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d}+\xi} + \delta_n^{q\alpha}\right) + \left(\frac{\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{\rho}{n}\right) + 4\left(\frac{32V\rho_n}{n}\right)^{\frac{1}{2-\vartheta}} \\
&\leq \left(6c_1(2c + \rho^{\frac{1}{2-\vartheta}} + \rho) + 4\,(32V\rho_n)^{\frac{1}{2-\vartheta}}\right) n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d}+\xi}
\end{aligned}$$

with probability P^n not less than $1-(1+|\Lambda_n \times \Gamma_n|)\, e^{-\rho}$. With a variable transformation P^n can be adjusted such that it is not less than $1 - e^{-\rho}$. □

Proof of Theorem 4.11. If $\vartheta := \min\{\frac{2}{q}, \frac{p}{p+1}\}$, we know by [31, Theorem 2.8] that, for all $f : X \to [-1,1]$, there exists an $f^*_{\tau,\mathrm{P}} : X \to [-1,1]$ with $f^*_{\tau,\mathrm{P}}(x) \in F^*_{\tau,\mathrm{P}}(x)$ for P_X-almost all $x \in X$ such that the variance bound (22) is satisfied with $V = 2^{2-\vartheta} q^\vartheta \|\nu^{-1}\|^\vartheta_{L_p(\mathrm{P}_x)}$. Since $F^*_{\tau,\mathrm{P}}$ consists of singletons, the variance bound is fulfilled for all $f : X \to [-1,1]$ with $f^*_{\tau,\mathrm{P}}$. □

Proof of Corollary 4.12. For $q = 2$ and $p = \infty$, Theorem 4.11 and Corollary 4.9 immediately yield $\vartheta = 1$, $V = 4\|\kappa^{-1}\|_{L_\infty(\mathrm{P}_X)}$, and, for every $\xi > 0$,

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

with a constant $C > 0$. Finally, the self calibration inequality (19) yields

$$\begin{aligned} \|\widehat{f}_{\mathrm{D},\lambda,\gamma} - f^*_{\tau,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)} &\le 4\|\kappa^{-1}\|_{L_\infty(\mathrm{P}_X)} \left(\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \right) \\ &\le C'\rho\, n^{-\frac{2\alpha}{2\alpha+d}+\xi} \end{aligned}$$

for all $\xi > 0$ and $C' := 4\|\kappa^{-1}\|_{L_\infty(\mathrm{P}_X)} C$. □

Acknowledgements

We thank the two anonymous reviewers and the associate editor for their helpful comments.

References

[1] Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*, 2nd ed. Academic Press, New York. MR2424078
[2] Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437
[3] Berens, H. and Vore, R. D. (1978). Quantitative Korovkin Theorems for Positive Linear Operators on L_p- Spaces. *Trans. Amer. Math. Soc.* **245** 349–361. MR0511414
[4] Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, Boston.
[5] Caponnetto, A. and De Vito, E. (2007). Optimal rates for regularized least squares algorithm. *Found. Comput. Math.* **7** 331–368. MR2335249
[6] Carl, B. and Stephani, I. (1990). *Entropy, Compactness, and the Approximation of Operators. Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge. MR1098497
[7] Chaudhuri, P. (1991). Global nonparametric estimation of conditional quantile functions and their derivatives. *J. Multivariate Anal.* **39** 246–269. MR1147121
[8] Chen, D.-R., Wu, Q., Ying, Y. and Zhou, D.-X. (2004). Support vector machine soft margin classifiers: Error analysis. *J. Mach. Learn. Res.* **5** 1143–1175. MR2248013

[9] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge.
[10] Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc.* **39** 1–49. MR1864085
[11] De Vito, E., Caponnetto, A. and Rosasco, L. (2005). Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.* **5** 59–85. MR2125691
[12] DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation. Grundlehren Der Mathematischen Wissenschaften*. Springer-Verlag, Berlin. MR1261635
[13] DeVore, R. A. and Popov, V. A. (1988). Interpolation of Besov Spaces. *Trans. Amer. Math. Soc.* **305** 397–414. MR0920166
[14] Eberts, M. and Steinwart, I. (2011). Optimal learning rates for least squares SVMs using Gaussian kernels. In *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira and K. Q. Weinberger, eds.) 1539–1547.
[15] Edmunds, D. E. and Triebel, H. (1996). *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge. MR1410258
[16] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.
[17] Johnen, H. and Scherer, K. (1976). On the equivalence of the K-functional and moduli of continuity and some applications. In *Lecture Notes in Math.*, **571** 119–140. Springer-Verlag, Berlin. MR0487423
[18] Keerthi, S. S. and Shevade, S. K. (2003). SMO Algorithm for Least Squares SVM Formulations. *Neural Computation* **15** 487–507.
[19] Koenker, R. (2005). *Quantile Regression*, 1st ed. *Econometric Society Monographs*. Cambridge University Press. MR2268657
[20] Lee, S. S. (2003). Efficient Semiparametric estimation of a partially linear quantile regression model. *Econometric Theory* **19** 1–31. MR1965840
[21] Li, Y., Liu, Y. and Zhu, J. (2007). Quantile Regression in Reproducing Kernel Hilbert Spaces. *J. Amer. Statist. Assoc.* **102** 255–268. MR2293307
[22] Mendelson, S. and Neeman, J. (2010). Regularization in Kernel Learning. *Ann. Statist.* **38** 526–565. MR2590050
[23] Micchelli, C. A., Pontil, M., Wu, Q. and Zhou, D. X. (2005). Error bounds for learning the kernel. http://eprints.pascal-network.org/archive/00001014/.
[24] Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.
[25] Shen, X. (1998). On the Method of Penalization. *Statist. Sinica* **8** 337–357. MR1624410
[26] Smale, S. and Zhou, D. X. (2003). Estimating the approximation error in learning theory. *Anal. Appl.* **1** 17–41. MR1959283
[27] Smale, S. and Zhou, D. X. (2007). Learning theory estimates via integral operators and their approximations. *Constr. Approx.* **26** 153–172. MR2327597

[28] Stein, E. M. (1970). *Singular integrals and differentiability properties of functions*. Princeton University Press, Princeton, NJ. MR0290095
[29] Steinwart, I. and Christmann, A. (2008a). How SVMs can estimate quantiles and the median. In *Advances in Neural Information Processing Systems 20* (J. C. Platt, D. Koller, Y. Singer and S. Roweis, eds.) 305–312. MIT Press, Cambridge, MA.
[30] Steinwart, I. and Christmann, A. (2008b). *Support Vector Machines*. Springer, New York. MR2450103
[31] Steinwart, I. and Christmann, A. (2011). Estimating Conditional Quantiles with the Help of the Pinball Loss. *Bernoulli* **17** 211–225. MR2797989
[32] Steinwart, I., Hush, D. and Scovel, C. (2009). Optimal Rates for Regularized Least Squares Regression. In *Proceedings of the 22nd Annual Conference on Learning Theory* (S. Dasgupta and A. Klivans, eds.) 79–93. MR2277920
[33] Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.* **35** 575–607. MR2336860
[34] Suzuki, T. (2011). Unifying Framework for Fast Learning Rate of Non-Sparse Multiple Kernel Learning. In *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira and K. Q. Weinberger, eds.) 1575–1583.
[35] Takeuchi, I., Le, Q. V., Sears, T. D. and Smola, A. J. (2006). Nonparametric Quantile Estimation. *J. Mach. Learn. Res.* **7** 1231–1264. MR2274404
[36] Temlyakov, V. (2006). Optimal estimators in learning theory. *Banach Center Publications, Inst. Math. Polish Academy of Sciences* **72** 341–366. MR2325756
[37] Triebel, H. (2006). *Theory of function spaces III*. Birkhäuser, Basel [u.a.]. MR2250142
[38] Triebel, H. (2010). *Theory of Function Spaces*, Repr. of the 1983 ed. Birkhäuser, Basel.
[39] Wu, Q., Ying, Y. and Zhou, D.-X. (2006). Learning Rates of Least-Square Regularized Regression. *Found. Comput. Math.* **6** 171–192. MR2228738
[40] Xiang, D. H. and Zhou, D. X. (2009). Classification with Gaussians and Convex Loss. *J. Mach. Learn. Res.* **10** 1447–1468. MR2534867
[41] Ye, G.-B. and Zhou, D.-X. (2008). Learning and approximation by Gaussians on Riemannian manifolds. *Adv. Comput. Math.* **29** 291–310. MR2438346
[42] Ying, Y. and Campbell, C. (2009). Generalization bounds for learning the kernel. In *Proceedings of the 22nd Annual Conference on Learning Theory* (S. Dasgupta and A. Klivans, eds.).
[43] Ying, Y. and Zhou, D.-X. (2007). Learnability of Gaussians with Flexible Variances. *J. Mach. Learn. Res.* **8** 249–276. MR2320669