Universität
Stuttgart

Fachbereich
Mathematik

Optimal regression rates for SVMs using Gaussian kernels

Mona Eberts, Ingo Steinwart

Preprint 2011/021

Universität Stuttgart

Fachbereich Mathematik

Optimal regression rates for SVMs using Gaussian kernels

Mona Eberts, Ingo Steinwart

Preprint 2011/021

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

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

ISSN **1613-8309**

LATEX-Style: Winfried Geis, Thomas Merkle

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) minimax 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 classifcations: Primary 62G08; secondary 62G05, 68Q32, 68T05.
Keywords: least squares regression, quantile estimation, support vector machines.

1 Introduction

Given i.i.d. observations $D := ((x_1, y_1), \dots, (x_n, y_n))$ 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 [14], which presents a lot of results in this direction. In contrast, the non-parametric quantile regression problem has attracted less attention, probably because for more advanced estimation procedures, a.k.a. 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 [17].

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 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 to considering 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, 28], $\lambda > 0$ is 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

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

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 [28] 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 [28, 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 [22], 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 [8, 16] and [32], respectively.

One of the main topics in recent theoretical investigations on SVMs have been learning rates. For example, the articles [9, 10, 25, 5, 20, 30] and the references therein establish rates for SVMs using the least squares loss, while SVMs using the pinball loss are investigated in [27, 29]. 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, 20, 30], the obtained learning rates are typically not optimal in a minimax sense. In addition, these three papers only consider some specific cases. For example, [5] only consider 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 [20, 30], 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, [20] 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 [30], 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 [30] shows that the inclusion assumptions made in this article are satisfied and that the resulting learning rates for SVMs are minimax 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

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

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 [31, 35]. Here we note that in both articles the authors actually consider binary classification, although a closer look reveals that at least the results of [35] 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 [35] establish rates of the form

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

Unfortunately, these rates are far from the known minimax 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 minimax 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 minimax 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 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. We begin with the input space $X \subset \mathbb{R}^d$, which is assumed to be a non-empty, open, and bounded set whose boundary ∂X has Lebesgue measure 0. Moreover, we only consider the case of bounded regression, that is, $Y := [-M, M]$ for some $M > 0$. We further assume that P is a probability measure on $X \times Y$ whose 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 bounded away from 0 and ∞. Recall that in this case the space $L_p(\mathrm{P}_X)$, $p \in (0, \infty)$, equals the space $L_p(\mu)$ and that the corresponding norms are equivalent. For the sake of simplicity, we thus restrict the formulations of our main results to the case, where P_X is the uniform distribution on X. Because of the described equivalence, however, it is straightforward to see that all results actually hold for the more general case in which P_X only has a Lebesgue density that is bounded away from 0 and ∞.

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

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

Since P lives on $X \times [-M, M]$, both the conditional mean and the conditional quantiles are $[-M, M]$-valued. It therefore suffices to consider $[-M, M]$-valued estimators of these quantities. 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

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

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

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

where again, L is one of the above loss functions.

The rest of this paper is organized as follows: The next section presents some upper bounds on the regularization 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 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) minimax 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

In this section, we present some approximation results that are essential to describe the infinite sample behaviour for *fixed* regularization parameter λ and kernel width γ. These results will turn out to be essential in the following sections, where we derive oracle inequalities and learning rates for SVMs with Gaussian kernels.

To formulate the approximation results, we need to introduce some function spaces that are later assumed to contain the target function. Let us begin by introducing some notations. We denote the Lebesgue spaces of order p with respect to the measure ν by $L_p(\nu)$ and for the Lebesgue measure μ on $X \subset \mathbb{R}^d$ we write $L_p(X) := L_p(\mu)$. Furthermore, B_E denotes the closed unit ball of a Banach space 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 recall the modulus of smoothness from, e.g. [11, p. 44], [12, 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 $f : X \to \mathbb{R}^d$ 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*

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

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

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

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

It is well-known, see e.g. [15, Equation (2.1)], that the modulus of smoothness with respect to $L_p\left(X\right)$ satisfies

$$\omega_{r,L_p(X)}\left(f,t\right) \leq \left(1+\frac{t}{s}\right)^r \omega_{r,L_p(X)}\left(f,s\right) \ , \tag{5}$$

for all $f \in L_p(X)$ and all $s > 0$. Moreover, the modulus of smoothness can be used to define the scale of Besov spaces. Namely, for $1 \le p, q \le \infty$, $\alpha > 0$, $r := \lfloor \alpha \rfloor + 1$, 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 $1 \le q < \infty$, 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. [11, pp. 54/55] and [12, p. 398]. In addition, for $q = \infty$, we often write $B^\alpha_{p,\infty}(\nu) = \mathrm{Lip}^*(\alpha, L_p(\nu))$ 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. [13, p. 25 and p. 44], that the Sobolev spaces $W^\alpha_p(\mathbb{R}^d)$ fall into the scale of Besov spaces, namely

$$W^\alpha_p(\mathbb{R}^d) \subset B^\alpha_{p,q}(\mathbb{R}^d) \tag{6}$$

for $\alpha \in \mathbb{N}$, $p \in (1, \infty)$, and $\max\{p, 2\} \le q \le \infty$. 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.

For our results, we need to extend functions $f : X \to \mathbb{R}$ to functions $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ such that the smoothness properties of f described by some Sobolev or Besov space are preserved by $\hat{f}$. Our main tool for this task is the following classical theorem.

Theorem 2.2 (Stein's Extension Theorem). *Let X be a bounded Lipschitz domain. Then there exists a linear operator $\mathfrak{E}$ mapping functions $f : X \to \mathbb{R}$ to functions $\mathfrak{E}f : \mathbb{R}^d \to \mathbb{R}$ with the following properties:*

(a) $\mathfrak{E}(f)_{|X} = f$, that is, $\mathfrak{E}$ is an extension operator.

(b) $\mathfrak{E}$ continuously maps $W^m_p(X)$ into $W^m_p(\mathbb{R}^d)$ for all $p \in [1, \infty]$ and all integers $m \ge 0$. That is, there exist constants $a_{m,p} \ge 0$, such that, for every $f \in W^m_p(X)$, we have

$$\|\mathfrak{E}f\|_{W^m_p(\mathbb{R}^d)} \le a_{m,p} \|f\|_{W^m_p(X)} . \tag{7}$$

(c) $\mathfrak{E}$ 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 $\alpha > 0$. That is, there exist constants $a_{\alpha,p,q} \ge 0$, such that, for every $f \in B^\alpha_{p,q}(X)$, we have

$$\|\mathfrak{E}f\|_{B^\alpha_{p,q}(\mathbb{R}^d)} \le a_{\alpha,p,q} \|f\|_{B^\alpha_{p,q}(X)} .$$

For more detailed conditions on X ensuring the existence of $\mathfrak{E}$, we refer to [26, p. 181] and [1, p. 83]. Property (c) follows from (a) and (b) by some interpolation argument since $B^\alpha_{p,q}$ can be described by the interpolation space $(W^{m_0}_p, W^{m_1}_p)_{\theta,q}$ of the real method, where $q \in [1, \infty]$, $p \in (1, \infty)$, $\theta \in (0, 1)$ and $m_0, m_1 \in \mathbb{N}_0$ satisfying $m_0 \ne m_1$ and $\alpha = m_0(1 - \theta) + m_1\theta$, see [34, pp. 65/66] for more details. In the following, we always assume that we do have such an extension operator $\mathfrak{E}$.

Throughout the paper we further assume that the boundary of X has zero Lebesgue measure, which for bounded Lipschitz domains is obviously satisfied. Note that if this assumption is satisfied, the uniform distribution on X can be identified with the uniform distribution on the interior and the closure of X, and hence we will not distinguish between them in terms of notation. Similarly, we typically view the uniform distribution on X as a probability measure defined on $\mathbb{R}^d$ rather than on X, that is, our notation does not distinguish between these two formally different measures, either.

To derive oracle inequalities for SVMs we have to estimate the regularization error

$$\min_{f\in H_\gamma} \lambda \left\|f\right\|_{H_\gamma}^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}^* \ .$$

The following two results construct a function f for which both the regularization term $\lambda \left\|f\right\|_{H_\gamma}^2$ and the excess risk $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}^*$ are small. To construct this function the next two theorems use, for $r \in \mathbb{N}$, and $\gamma > 0$, the function $K : \mathbb{R}^d \to \mathbb{R}$ defined by

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

where $K_\gamma\left(x\right) := \exp\left(-\gamma^{-2}\|x\|_2^2\right)$ for all $x \in \mathbb{R}^d$.

Now, the first result will be used to bound the excess risk.

Theorem 2.3. *Let $X \subset \mathbb{R}^d$ be a domain such that we have an extension operator $\mathfrak{E}$ of the form described in Theorem 2.2, P_X be the uniform distribution on X and $f \in L_\infty\left(X\right)$. Furthermore, let $\tilde{f}$ be defined by*

$$\tilde{f}\left(x\right) := \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} \mathfrak{E}f\left(x\right) \ , \qquad\qquad x \in \mathbb{R}^d \ . \tag{9}$$

Then, for $r \in \mathbb{N}$, $\gamma > 0$, and $q \in [1, \infty)$, we have $\mathfrak{E}f \in L_q(\mathrm{P}_X)$ and

$$\|K * \tilde{f} - f\|_{L_q(\mathrm{P}_X)}^q \leq C_{r,q}\, \omega_{r,L_q(\mathbb{R}^d)}^q \left(\mathfrak{E}f, \gamma/2\right) \ ,$$

where $C_{r,q}$ is a constant only depending on r, q and the volume $\mathrm{vol}(X)$ *of X.*

The second result will be used to bound the regularization term. In addition, it provides a very useful supremum bound.

Theorem 2.4. *Let $g \in L_2\left(\mathbb{R}^d\right)$, 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 $K * g \in H_\gamma$ with*

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

Moreover, if $g \in L_\infty\left(\mathbb{R}^d\right)$, then

$$\left|K * g\left(x\right)\right| \leq \left(\gamma\sqrt{\pi}\right)^{\frac{d}{2}} \left(2^r - 1\right) \left\|g\right\|_{L_\infty(\mathbb{R}^d)}$$

holds for all $x \in X$.

To illustrate the use of the two theorems above, we fix $g := \tilde{f} := (\gamma\sqrt{\pi})^{-\frac{d}{2}}\mathfrak{E}f^*$, where $f^* : X \to \mathbb{R}$ is a function satisfying $\mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}_{L,\mathrm{P}}^*$. Then it will turn out that Theorems 2.3 and 2.4 yield

$$\begin{aligned}
&\min_{f\in H_\gamma} \lambda \left\|f\right\|_{H_\gamma}^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}^* \\
&\leq \lambda \|K * \tilde{f}\|_{H_\gamma}^2 + \mathcal{R}_{L,\mathrm{P}}(K * \tilde{f}) - \mathcal{R}_{L,\mathrm{P}}^* \\
&\leq \lambda (2^r - 1)^2 (\gamma\sqrt{\pi})^{-d} \left\|\mathfrak{E}f^*\right\|_{L_2(\mathbb{R}^d)}^2 + c\|K * \tilde{f} - f^*\|_{L_2(\mathrm{P}_X)}^2 \\
&\leq \lambda (2^r - 1)^2 (\gamma\sqrt{\pi})^{-d} \left\|\mathfrak{E}f^*\right\|_{L_2(\mathbb{R}^d)}^2 + c\, C_{r,2}\, \omega_{r,L_2(\mathbb{R}^d)}^2 \left(\mathfrak{E}f^*, \gamma/2\right) \ ,
\end{aligned} \tag{10}$$

where the crucial intermediate estimate $\mathcal{R}_{L,\mathrm{P}}(K*\tilde{f})-\mathcal{R}^*_{L,\mathrm{P}} \leq c\|K*\tilde{f}-f^*\|^2_{L_2(\mathrm{P}_X)}$ will be discussed in Sections 3 and 4. Moreover, note that Theorem 2.4 also implies the estimate

$$|K * \tilde{f}| \leq (\gamma\sqrt{\pi})^{\frac{d}{2}}(2^r-1)\|\tilde{f}\|_{L_2(\mathbb{R}^d)} \leq (2^r-1)\|\mathfrak{E}f^*\|_{L_2(\mathbb{R}^d)}\,,$$

which will be important when applying concentration inequalities.

Besides the above bounds on the approximation properties of H_γ, we will also need to control the capacity of H_γ in terms of entropy numbers. To this end, the following definition recalls entropy numbers for the sake of completeness (cf. [6] or [28, Definition A.5.26] for more information).

Definition 2.5. *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*

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

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

For the empirical distribution D_X associated to the data set $D_X := (x_1,\ldots,x_n) \in X^n$, [28, Theorem 7.34] and [28, Corollary 7.31] immediately yield the following lemma regarding the capacity of H_γ.

Lemma 2.6. *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*

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

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

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 $L\left(y,t\right) = \left(y-t\right)^2$. It is well known that, for this loss, the function $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ defined by $f^*_{L,\mathrm{P}}\left(x\right) = \mathbb{E}_{\mathrm{P}}\left(Y|x\right)$, $x \in X$, is the only function for which the Bayes risk is attained. Furthermore, some simple and well-known transformations show

$$\mathcal{R}_{L,\mathrm{P}}\left(f\right) - \mathcal{R}^*_{L,\mathrm{P}} = \int_X \left|f - f^*_{L,\mathrm{P}}\right|^2 d\mathrm{P}_X = \left\|f - f^*_{L,\mathrm{P}}\right\|^2_{L_2(\mathrm{P}_X)}\,. \tag{11}$$

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

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

Theorem 3.1. *Let $X \subset B_{\ell_2^d}$ be a bounded domain with $\mu(\partial X) = 0$ such that we have an extension operator $\mathfrak{E}$ in the sense of Theorem 2.2. Furthermore, let $M > 0$, $Y := [-M,M]$, and P be a distribution on $X \times Y$ such that P_X is the uniform distribution on X. Assume that we have a fixed version $f^*_{L,\mathrm{P}}$ of the regression function such that $f^*_{L,\mathrm{P}}\left(x\right) = \mathbb{E}_{\mathrm{P}}\left(Y|x\right) \in [-M,M]$ for all $x \in X$. Assume further that, for $\alpha \geq 1$ and $r := \lfloor\alpha\rfloor + 1$, there exists a constant $c > 0$ such that, for all $t \in (0,1]$, we have*

$$\omega_{r,L_2(\mathbb{R}^d)}\left(\mathfrak{E}f^*_{L,\mathrm{P}},t\right) \leq c\,t^\alpha\,. \tag{12}$$

Then, for all $\varepsilon > 0$ and $p \in (0,1)$, there exists a constant $K > 0$ 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

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

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

With the help of this oracle inequality we can derive learning rates for the learning method (4).

Corollary 3.2. *Let $\varepsilon > 0$, $p \in (0,1)$ and $\rho \geq 1$ be fixed. Under the assumptions of Theorem 3.1 and with*

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

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

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n}) - \mathcal{R}^*_{L,\mathrm{P}} \leq C n^{-\frac{2\alpha}{2\alpha+2\alpha p+dp+(1-p)(1+\varepsilon)d}}$$

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.

It is rather easy to check that Theorem 3.1, Corollary 3.2, as well as the following Theorem 3.3 actually hold, if P_X is only Lebesgue absolutely continuous with bounded Lebesgue density. Indeed, the crucial result Theorem 2.3 also holds for such distributions, and the remaining arguments used to prove Theorem 3.1 and its consequences apply to all marginal distributions P_X.

To clarify the rates achieved in Corollary 3.2, we note that, for every $\xi > 0$, we can find $\varepsilon, p \in (0,1)$ sufficiently close to 0 such that the learning rate in Corollary 3.2 is at least as fast as

$$n^{-\frac{2\alpha}{2\alpha+d}+\xi}\,. \tag{13}$$

To achieve these rates, however, we need to set λ_n and γ_n as in Corollary 3.2, which in turn requires us 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. [28, 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

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

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

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

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 an δ_n-net of $(0,1]$ with $\epsilon_n \leq n^{-1}$ and $\delta_n \leq n^{-\frac{1}{2+d}}$. 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*

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

where $C_{\xi,\tau} > 0$ is a constant independent of n.

What is left to do is to relate Assumption (12) with the function spaces introduced in Section 2. This is the goal of the following two results.

Corollary 3.4. *Let $X \subset B_{\ell_2^d}$ be bounded domain with $\mu(\partial X) = 0$ such that we have an extension operator $\mathfrak{E}$ in the sense of Theorem 2.2. Furthermore, let $M > 0$, $Y := [-M, M]$, and P be a distribution on $X \times Y$ such that P_X is the uniform distribution on X. If, for some $\alpha \in \mathbb{N}$, we have $f^*_{L,\mathrm{P}} \in W_2^\alpha(X)$, then, for all $\xi > 0$, both the SVM considered in Corollary 3.2 and the TV-SVM considered in Theorem 3.3 learn with the rate*

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

Again, this result also holds for distributions P_X that have a bounded Lebesgue density. Moreover, for the uniform distribution P_X, or more generally, for distributions P_X having a Lebesgue density that is bounded away from 0 and infinity, it is well-known that the minmax rate for $\alpha > d/2$ and target f^* satisfying 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.

Similar to Corollary 3.4 we can show assumption (12) and essentially asymptotically optimal learning rates if the regression function is contained in a Besov space.

Corollary 3.5. *Let $X \subset B_{\ell_2^d}$ be a domain such that we have an extension operator $\mathfrak{E}$ in the sense of Theorem 2.2. Furthermore, let $M > 0$, $Y := [-M, M]$, and P be a distribution on $X \times Y$ such that P_X is the uniform distribution on X. If, for some $\alpha \geq 1$, we have $f^*_{L,\mathrm{P}} \in B^\alpha_{2,\infty}(X)$, then, for all $\xi > 0$, both the SVM considered in Corollary 3.2 and the TV-SVM considered in Theorem 3.3 learn with the rate*

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

Recall that we have $e_i(\,\mathrm{id} : B^\alpha_{2,\infty}(X) \to L_2(\mathrm{P}_X)) \sim i^{-\frac{\alpha}{d}}$, see e.g. [13, 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 [33, Theorem 2.2] that $n^{-\frac{2\alpha}{2\alpha+d}}$ is the optimal learning rate in a minimax sense for $\alpha > d$ (cf. [30, Theorem 13]). In other words, for $\alpha > d$, the learning rates obtained in Corollary 3.5 are again asymptotically optimal modulo ξ.

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. [9, 10, 25, 5, 20] and the references therein. In particular, optimal rates have been established in [5], if $f_P^* \in H$, and the eigenvalue behavior of the integral operator associated to H is known. Moreover, for $f_P^* \notin H$, the articles [20] and [30] 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 [20] actually considers a different exponent in the regularization term of (4). On the other hand, [30] shows that the rate $n^{-\beta/(\beta+p)}$ is often asymptotically optimal in a minmax sense. In particular, the latter is the case for $H = W_2^m(X)$, $f \in W_2^s(X)$, and $s \in (d/2, m]$, that is, when using a Sobolev space as the underlying RKHS H, then all target functions contained in a Sobolev of lower smoothness $s > d/2$ can be learned with the asymptotically optimal rate $n^{-\frac{2s}{2s+d}}$. Here we note that the condition $s > d/2$ ensures by Sobolev's embedding theorem that $W_2^s(X)$ consists of bounded functions, and hence $Y = [-M, M]$ does not impose an additional assumption on $f^*_{L,\mathrm{P}}$. If $s \in (0, d/2]$, then the results of [30] 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 [14].

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, [24] shows that for many analytic kernels the approximation error can only have polynomial decay 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 does not decay polynomially fast, see [24, Proposition 1.1.], and if $f^*_{L,\mathrm{P}} \in W^s_2(X)$, then, in general, the approximation error function only has a logarithmic decay. 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 [31], research has thus focused on the learning performance of SVMs with varying widths. The result that is probably the closest to ours is [35]. 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 $n^{-\frac{s}{s+2d+2}}$, again under the assumption $f^*_{L,\mathrm{P}} \in W^s_2(X)$ for some $s > 0$. Clearly, this is significantly worse than our rates. Furthermore, [36] treats the case, where X is isometrically embedded into a t-dimensional, connected and compact C^∞-submanifold of $\mathbb{R}^d$. In this case, it turns out that the resulting learning rate does not depend on the dimension d, but on the intrinsic dimension t of the data. Namely, the authors establish the rate $n^{-\frac{s}{8s+4t}}$ modulo a logarithmic factor, where $s \in (0,1]$ and $f^*_{L,\mathrm{P}} \in \mathrm{Lip}\,(s)$. Note that this rate is better than ours only if $t < \frac{d-14s}{8}$, that is, e.g. for $s = 1$, if $d > 8t + 14$.

Another direction of research that can be applied to Gaussian kernels with varying widths are multi-kernel regularization schemes, see [38, 21, 37] for some results in this direction. For example, [38] establishes learning rates of the form $n^{-\frac{2m-d}{4(4m-d)}+\xi}$ whenever $f^*_{L,\mathrm{P}} \in W^m_2(X)$ for some $m \in (d/2, d/2+2)$, 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

The goal of this section is to derive learning rates for SVMs for quantile regression. To this end, recall that the goal of quantile regression is to estimate the conditional τ-quantile, i.e. the set valued function

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

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

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

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 introduce some characteristics of the distribution P that make it possible 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, let Q be a distribution on $\mathbb{R}$ with support $\mathrm{supp}\,Q \subset [-1,1]$ and τ-quantile

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

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, i.e. $t^*_{\min} = t^*_{\max} =: t^*$ and $F_\tau^*(Q) = \{t^*\}$. The next definition describes 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 [29, 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 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 actually classifies the set of all distributions on $[-1,1]$.

Next, 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 for P_X-almost all $x \in X$, P_X is absolutely continuous with respect to the Lebesgue measure μ.

Furthermore, assume that the corresponding densities $f(\,\cdot\,,x) := \frac{d\mathrm{P}(\,\cdot\,|x)}{d\mu_{|[0,1]}}$ are uniformly bounded, that is, $f(y,x) \leq b$ 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 $f(\,\cdot\,,x)$ of $\mathrm{P}(\,\cdot\,|x)$ is bounded away from 0, i.e. $f(y,x) \geq \hat{b}$ 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 [29, 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 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 $q=1$.

Let us now 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 [29, Theorem 2.7] the following so-called self-calibration inequality

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_v(\mathrm{P}_X)} \leq 2^{1-\frac{1}{q}} q^{\frac{1}{q}} \|\kappa^{-1}\|^{\frac{1}{q}}_{L_p(\mathrm{P}_X)} (\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}})^{\frac{1}{q}} \ , \tag{17}$$

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

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq \|f - f^*_{\tau,\mathrm{P}}\|_{L_1(\mathrm{P}_X)} \tag{18}$$

for all $f : X \to [-1,1]$. For our purposes, this estimate can be substantially improved by the next theorem.

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*

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq q^{-1} \|b_{\mathrm{P}(\,\cdot\,|x)}\|_{L_p(\mathrm{P}_X)} \|f - f^*_{\tau,\mathrm{P}}\|^q_{L_u(\mathrm{P}_X)} \tag{19}$$

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

To see that (19) is indeed an improvement of (18) we consider $f_0 := K * \tilde{f}$ with K and $\tilde{f}$ as in (8) and (9). Assuming a standard bound on the modulus of continuity, see (21) below, we then obtain

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

from (18), while (19) yields

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

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 $q > 1$.

Theorem 4.8. *Let $X \subset B_{\ell_2^d}$ be a domain such that we have an extension operator $\mathfrak{E}$ in the sense of Theorem 2.2. Furthermore, let $Y := [-1,1]$, and P be a distribution on $X \times Y$ such that P_X is the uniform distribution on X. For $\tau \in (0,1)$ let $f^*_{\tau,\mathrm{P}} : X \to [-1,1]$ be the conditional τ-quantile function. Assume that there exist constants $\vartheta \in [0,1]$ and $V \geq 2^{2-\vartheta}$ such that the variance bound*

$$\mathbb{E}_{\mathrm{P}}(L_\tau \circ \widehat{f} - L_\tau \circ f^*_{\tau,\mathrm{P}})^2 \leq V \cdot \Big(\mathbb{E}_{\mathrm{P}}(L_\tau \circ \widehat{f} - L_\tau \circ f^*_{\tau,\mathrm{P}})\Big)^{\vartheta} \qquad (20)$$

is satisfied for all $f : X \to \mathbb{R}$ and that P has a τ-quantile of upper p-average type q with $p \in (1,\infty]$ and $q \in [1,\infty)$. Furthermore, assume that, for $\alpha \geq 1$ and $r := \lfloor \alpha \rfloor + 1$, there exists a constant $c > 0$ such that, for all $t \in (0,1]$, we have

$$\omega_{r,L_u(\mathbb{R}^d)}(\mathfrak{E} f^*_{\tau,\mathrm{P}}, t) \leq c t^{\alpha} \,. \qquad (21)$$

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 $1 - e^{-\rho}$.

Similarly to Theorem 3.1 and its corollaries, Theorem 4.8 and its consequences below actually hold, whenever P_X has a bounded Lebesgue density. 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 $\varepsilon > 0$, $\varsigma \in (0,1)$, and $\rho \geq 1$ be fixed. Under the assumptions of Theorem 4.8 and with*

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

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

$$\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C n^{-\frac{q\alpha}{q\alpha(2-\varsigma-\vartheta+\vartheta\varsigma)+q\alpha\varsigma+d\varsigma+(1-\varsigma)(1+\varepsilon)d}} \qquad (22)$$

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.

Analogously to Corollary 3.2, for every $\xi > 0$ we can find $\varepsilon, \varsigma \in (0,1)$ that are sufficiently close to 0 such that the learning rate in Corollary 4.9 is at least as fast as

$$n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d}+\xi} \,.$$

To achieve the learning rate (22), 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 same 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 $\epsilon_n \leq n^{-1}$ and $\delta_n \leq n^{-\frac{1}{1+d}}$. 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*

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

with a constant $C_{\xi,\tau} > 0$.

To apply Theorems 4.8 and 4.10 the variance bound (20) has to be fulfilled for the τ-pinball loss. But unfortunately, unlike for the least squares loss, (20) generally does not hold for some $\vartheta > 0$. However, if P has a lower quantile type, then the following result taken from [29, Theorem 2.8] establishes non-trivial variance bounds.

Theorem 4.11. *Let L_τ be the τ-pinball loss, $\tau \in (0,1)$, $F^*_{\tau,\mathrm{P}}$ consist of singletons and P be a distribution that has a τ-quantile of lower p-average type q with $p \in (0,\infty]$ 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*

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

Let us now combine this variance bound with the previous results. 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 realistic example $(p,q) = (\infty, 2)$, cf. Example 4.5.

Corollary 4.12. *Let P be a distribution that has a τ-quantile of lower and upper p-average type q for $q = 2$ and $p = \infty$. Under the assumptions of Theorems 4.8 and 4.11 we then obtain for the SVM considered in Corollary 4.9 that, for all $\xi > 0$ and $\rho \ge 1$,*

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

and

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

with constants $C_{\xi,\tau} > 0$ and $C'_{\xi,\tau} := 4\|\kappa^{-1}\|_{L_\infty(\mathrm{P}_X)} C_{\xi,\tau}$. In particular, these learning rates are obtained, if $f^*_{\tau,\mathrm{P}} \in W^\alpha_2(\mathrm{P}_X)$ or $f^*_{\tau,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_X)$. Moreover, the same learning rates can be obtained for the TV-SVM considered in Theorem 4.10.

Note that the convergence rates above equal the rates we achieved for the least squares SVMs in Section 3 (cf. Corollaries 3.4 and 3.5).

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

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \le \|f_0 - f^*_{\tau,\mathrm{P}}\|_{L_1(\mathrm{P}_X)} \le C_{r,1} c \gamma^\alpha, \tag{23}$$

in the corresponding proof, where $f_0 := K * \tilde{f}$ with K and $\tilde{f}$ as in (8) and (9). 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 (23) and (17) yield

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

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

In addition, we consider distributions P having a τ-quantile of upper p-average type q with $p = \infty$ and $q \ne 2$ in the following corollary, where we omit the obvious proof.

Corollary 4.13. *Let $p = \infty$. Under the assumptions of Theorem 4.8 and of Theorem 4.11 we obtain*

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

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 $\xi > 0$,

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

with a constant $C > 0$.

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 [32], 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. An approach similar to ours is used in [19] 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

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

is satisfied, then modulo some logarithmic factor, the rate $n^{-2/3}$ 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 (for the Gaussian case this assumption implies arbitrarily large values of α, see [28, Theorem 4.48]), and the technical hurdles are known to be significantly easier than for the general case. Nonetheless, it seems interesting that their rates can be essentially recovered by our results when setting $p = 1$, $q = 2$, and $\alpha = \infty$. Moreover, both articles also discuss algorithmic aspects of SVMs for quantile regression. Finally, [29] 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 [23], 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 [18] a partially linear quantile regression model is considered, where the parametric component learns with rate $n^{-\frac{1}{2}}$.

Finally, in [17, 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 regularization error. Let us begin with the proofs of these theorems. To this end, we need the convention $0^0 := 1$.

Proof of Lemma 2.3. First of all, we show $\mathfrak{C}f \in L_q(\mathrm{P}_X)$. Because of the assumption $f \in L_\infty(X)$, we have $f \in L_q(X)$ and $\mathfrak{C}f \in L_q(\mathbb{R}^d)$ for all $1 \leq q \leq \infty$. In addition,

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

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

$$\left\|K * \tilde{f} - f\right\|^q_{L_q(\mathrm{P}_X)} \leq C_{r,q}\, \omega^q_{r,L_q(\mathbb{R}^d)}\left(\mathfrak{C}f, \gamma/2\right) \ .$$

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 $x \in X$,

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

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

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

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

Next, Hölder's inequality yields, for $q > 1$,

$$\begin{aligned} &\left\| K * \tilde{f} - f \right\|_{L_q(\mathrm{P}_X)}^q \\ &\leq \int_{\mathbb{R}^d} \left(\left(\underbrace{\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\, dh}_{=1} \right)^{\frac{q-1}{q}} \left(\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\, |\triangle_h^r(\mathfrak{E}f, x)|^q \, dh \right)^{\frac{1}{q}} \right)^q d\,\mathrm{P}_X(x) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\, |\triangle_h^r(\mathfrak{E}f, x)|^q \, dh \, d\,\mathrm{P}_X(x) \\ &= \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h) \int_{\mathbb{R}^d} |\triangle_h^r(\mathfrak{E}f, x)|^q \, d\,\mathrm{P}_X(x)\, dh \\ &= \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\, \|\triangle_h^r(\mathfrak{E}f, \cdot)\|_{L_q(\mathrm{P}_X)}^q \, dh \\ &\leq \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}(h)\, \omega_{r, L_q(\mathrm{P}_X)}^q(\mathfrak{E}f, \|h\|_2)\, dh \ . \end{aligned} \tag{24}$$

Moreover, for $q = 1$, we have

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

Consequently, (24) holds for all $q \geq 1$. Furthermore, we have

$$\begin{aligned}
\omega^q_{r,L_q(\mathrm{P}_X)}\left(\mathfrak{E}f, t\right) &= \sup_{\|h\|_2 \leq t} \int_{\mathbb{R}^d} \Big| \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j}\, \mathfrak{E}f\left(x + jh\right) \Big|^q d\mathrm{P}_X\left(x\right) \\
&\leq \mu\left(X\right)^{-1} \sup_{\|h\|_2 \leq t} \int_{\mathbb{R}^d} \Big| \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j}\, \mathfrak{E}f\left(x + jh\right) \Big|^q d\mu\left(x\right) \\
&= \mu\left(X\right)^{-1} \omega^q_{r,L_q(\mathbb{R}^d)}\left(\mathfrak{E}f, t\right) \\
&\leq \mu\left(X\right)^{-1} \left(1 + \frac{2t}{\gamma}\right)^{rq} \omega^q_{r,L_q(\mathbb{R}^d)}\left(\mathfrak{E}f, \frac{\gamma}{2}\right)
\end{aligned}$$

for $t \geq 0$, where we used (5). Together with (24) this implies

$$\begin{aligned}
\left\| K * \tilde{f} - f \right\|^q_{L_q(\mathrm{P}_X)} &\leq \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}\left(h\right) \mu(X)^{-1} \left(1 + \frac{2\left\|h\right\|_2}{\gamma}\right)^{rq} \omega^q_{r,L_q(\mathbb{R}^d)}\left(\mathfrak{E}f, \frac{\gamma}{2}\right) dh \\
&= \mu(X)^{-1} \omega^q_{r,L_q(\mathbb{R}^d)}\left(\mathfrak{E}f, \frac{\gamma}{2}\right) \int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}\left(h\right) \left(1 + \frac{2\left\|h\right\|_2}{\gamma}\right)^{rq} dh\,. \qquad (25)
\end{aligned}$$

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

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

and Hölder's inequality yield

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}\left(h\right) \left(1 + \frac{2\left\|h\right\|_2}{\gamma}\right)^{rq} dh \\
&\leq \sum_{i=0}^{\lceil rq \rceil} \binom{\lceil rq \rceil}{i} \left(\frac{2}{\gamma}\right)^i \int_{\mathbb{R}^d} \left\|h\right\|_2^i \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}\left(h\right)\, dh \\
&\leq \sum_{i=0}^{\lceil rq \rceil} \binom{\lceil rq \rceil}{i} \left(\frac{2}{\gamma}\right)^i \left(\int_{\mathbb{R}^d} \left\|h\right\|_2^{2i} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}\left(h\right)\, dh\right)^{\frac{1}{2}}\,. \qquad (26)
\end{aligned}$$

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^d_{2i} to ℓ^d_2. This embedding constant leads to

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

$$
\begin{aligned}
&= d^{i-1}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{1}{2}} 2d\int_0^\infty t^{2i}\exp\left(-\frac{2t^2}{\gamma^2}\right)\,dt \\
&= 2d^{i}\left(\frac{2}{\gamma^2\pi}\right)^{\frac{1}{2}} \int_0^\infty t^{2i}\exp\left(-\frac{2t^2}{\gamma^2}\right)\,dt\ . \qquad (27)
\end{aligned}
$$

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

$$
\begin{aligned}
\int_0^\infty t^{2i}\exp\left(-\frac{2t^2}{\gamma^2}\right)\,dt &= \frac{1}{2}\frac{\gamma}{\sqrt{2}}\left(\frac{\gamma^2}{2}\right)^{i}\int_0^\infty u^{\left(i+\frac{1}{2}\right)-1}\exp\left(-u\right)\,du \\
&= \frac{1}{2}\frac{\gamma}{\sqrt{2}}\left(\frac{\gamma^2}{2}\right)^{i}\Gamma\left(i+\frac{1}{2}\right) \\
&= \frac{1}{2}\frac{\gamma}{\sqrt{2}}\left(\frac{\gamma^2}{2}\right)^{i}\Gamma\left(\frac{1}{2}\right)\prod_{j=1}^{i}\left(j-\frac{1}{2}\right) \\
&= \frac{1}{2}\frac{\gamma}{\sqrt{2}}\left(\frac{\gamma^2}{2}\right)^{i}\sqrt{\pi}\prod_{j=1}^{i}\left(j-\frac{1}{2}\right)\ . \qquad (28)
\end{aligned}
$$

Together, (27) and (28) 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 (26) 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, (25) implies

$$
\left\|K*\tilde{f}-f\right\|_{L_q(\mathrm{P}_X)}^{q} \le C_{r,q}\,\omega^q_{r,L_q(\mathbb{R}^d)}\left(\mathfrak{E}f,\frac{\gamma}{2}\right)
$$

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

Proof of Lemma 2.4. We define, for all $j\in\mathbb{N}$ and $x\in X$,

$$
g_j(x) := \left(\frac{2}{j\gamma\sqrt{\pi}}\right)^{\frac{d}{2}} K_{\frac{\gamma}{\sqrt{2}}}\left(\frac{x}{j}\right)\ . \qquad (29)
$$

By [28, Proposition 4.46] we obtain

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

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

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

Moreover, for the estimation of the norm we have

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

where we used [28, Proposition 4.46] in the first two steps. Finally, for all $x \in X$ and $g \in L_\infty\left(\mathbb{R}^d\right)$, Hölder's inequality implies

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

□

5.2 Proofs related to the least squares SVMs

To prove Theorem 3.1 we first deduce an oracle inequality for the least squares loss by specializing [28, Theorem 7.23].

Theorem 5.1. *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 $B_0 \ge 4M^2$ such that $\|L \circ f_0\|_\infty \le B_0$. Then, for all fixed $\rho \ge 1$, $\lambda > 0$, $\varepsilon > 0$ and $p \in (0, 1)$, the SVM using H_γ and L satisfies*

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

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 $L(y, \widehat{t}\,) \le L\,(y, t)$, i.e. it can be clipped at $M > 0$ (see [30, 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 [28, Definition 2.18]. See [28, Example 7.3] to verify that the least squares loss satisfies the supremum bound

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

and the variance bound

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

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 [28, Theorem 7.23] and Lemma 2.6 with $C_{\varepsilon,p} := C(\max\{c_{\varepsilon,p}, 4M^2\})^{2p}$, $c_{\varepsilon,p}$ as in Lemma 2.6 and a constant $C \geq 1$ which corresponds to the constant K of [28, 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 5.1.

Proof of Theorem 3.1. First of all, we want to apply Theorem 5.1 for $f_0 := K * \tilde{f}$ with

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

and

$$\tilde{f}(x) := \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} \mathfrak{E} f^*_{L,\mathrm{P}}(x)$$

for all $x \in \mathbb{R}^d$. The choice $f^*_{L,\mathrm{P}}(x) \in [-M, M]$ for all $x \in X$ implies $f^*_{L,\mathrm{P}} \in L_2(X)$ and the latter together with $X \subset B_{\ell_2^d}$ and (7) yields

$$\begin{aligned} \|\tilde{f}\|_{L_2(\mathbb{R}^d)} &= \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} \|\mathfrak{E} f^*_{L,\mathrm{P}}\|_{L_2(\mathbb{R}^d)} \\ &\leq \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} a_{0,2} \|f^*_{L,\mathrm{P}}\|_{L_2(X)} \\ &\leq \left(\frac{2}{\gamma\sqrt{\pi}}\right)^{\frac{d}{2}} a_{0,2} M \ , \end{aligned} \tag{30}$$

i.e. $\tilde{f} \in L_2(\mathbb{R}^d)$. Because of this and Theorem 2.4

$$f_0 = K * \tilde{f} \in H_\gamma$$

is satisfied. Since $f^*_{L,\mathrm{P}}(x) = \mathbb{E}_\mathrm{P}(Y|x) \in [-M, M]$ for all $x \in X$, we have $f^*_{L,\mathrm{P}} \in L_\infty(X)$ as well as $\mathfrak{E}(f^*_{L,\mathrm{P}}) \in L_\infty(\mathbb{R}^d)$. The latter yields $\tilde{f} \in L_\infty(\mathbb{R}^d)$ with

$$\begin{aligned} \|\tilde{f}\|_{L_\infty(\mathbb{R}^d)} &= \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} \| \ \mathfrak{E}(f^*_{L,\mathrm{P}})\|_{L_\infty(\mathbb{R}^d)} \\ &\leq a_{0,\infty} \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} \| \ f^*_{L,\mathrm{P}}\|_{L_\infty(X)} \\ &\leq a_{0,\infty} \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} M \ , \end{aligned}$$

where $a_{0,\infty}$ denotes the constant introduced in (7). With this and Theorem 2.4,

$$|K * \tilde{f}(x)| \leq \left(\gamma\sqrt{\pi}\right)^{\frac{d}{2}} (2^r - 1) \|\tilde{f}\|_{L_\infty(\mathbb{R}^d)} \leq a_{0,\infty} (2^r - 1) M$$

holds for all $x \in X$. Next, for all $(x, y) \in X \times Y$ and $a := \max\{a_{0,\infty}, 1\}$, we achieve

$$\begin{aligned} L(y, K * \tilde{f}(x)) &= (y - K * \tilde{f}(x))^2 \\ &= y^2 - 2y(K * \tilde{f}(x)) + (K * \tilde{f}(x))^2 \\ &\leq M^2 + 2a_{0,\infty} (2^r - 1) M^2 + a_{0,\infty}^2 (2^r - 1)^2 M^2 \\ &\leq 4^r a^2 M^2 \ . \end{aligned}$$

and

$$\|L \circ f_0\|_\infty = \sup_{(x,y) \in X \times Y} |L(y, f_0(x))| = \sup_{(x,y) \in X \times Y} \left| L\left(y, K * \tilde{f}(x)\right)\right| \leq 4^r a^2 M^2 =: B_0 \ .$$

Furthermore, (11) and Lemma 2.3 yield

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

where we used the assumption

$$\omega_{r,L_2(\mathbb{R}^d)}\left(\mathfrak{C}f^*_{L,\mathrm{P}}, \frac{\gamma}{2}\right) \le c\gamma^{\alpha}$$

for $\gamma \in (0,1]$, $\alpha \ge 1$, $r = \lfloor \alpha \rfloor + 1$ and a constant $c > 0$ in the last step. By Theorem 2.4 and (30) we know

$$\|f_0\|_{H_\gamma} = \|K * \tilde{f}\|_{H_\gamma} \le (2^r-1)\,\|\tilde{f}\|_{L_2(\mathbb{R}^d)} \le (2^r-1)\left(\frac{2}{\gamma\sqrt{\pi}}\right)^{\frac{d}{2}} a_{0,2}M\ .$$

Therefore, Theorem 5.1 and the above choice of f_0 yield, for all fixed $\rho \ge 1$, $\lambda > 0$, $\varepsilon > 0$ and $p \in (0,1)$, that the SVM using H_γ and L satisfies

$$\begin{aligned}&\lambda\left\|f_{\mathrm{D},\lambda,\gamma}\right\|^2_{H_\gamma} + \mathcal{R}_{L,\mathrm{P}}\left(\widehat{f}_{\mathrm{D},\lambda,\gamma}\right) - \mathcal{R}^*_{L,\mathrm{P}} \\ &\le 9\left(\lambda\left(2^r-1\right)^2\left(\frac{2}{\gamma\sqrt{\pi}}\right)^d a^2_{0,2}M^2 + C_{r,2}c^2\gamma^{2\alpha}\right) \\ &\quad + C_{\varepsilon,p}\frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p n} + \frac{\left(3456+15\cdot 4^r a^2\right)M^2(\ln(3)+1)\rho}{n} \\ &\le C_1\lambda\gamma^{-d} + 9\,C_r c^2\gamma^{2\alpha} + C_{\varepsilon,p}\frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p n} + \frac{C_2\rho}{n}\end{aligned}$$

with probability P^n not less than $1-e^{-\rho}$ and with constants $C_1 := 9\left(2^r-1\right)^2 2^d\pi^{-\frac{d}{2}}a^2_{0,2}M^2$, $C_2 := (\ln(3)+1)\left(3456+15\cdot 4^r a^2\right)M^2$, $a := \max\{a_{0,\infty},1\}$, $C_r := C_{r,2}$ only depending on r and $\mu(X)$ and $C_{\varepsilon,p}$ as in Theorem 5.1. □

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

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

with probability P^n not less than $1-e^{-\rho}$ and a constant $\widetilde{C} := \max\left\{C_1, 9\,C_r c^2, C_{\varepsilon,p}, C_2\rho\right\}$. In the next step [28, Lemma A.1.6.] shows that the sequences

$$\lambda_n = c_1 n^{-\frac{2\alpha+d}{2\alpha+2\alpha p+dp+(1-p)(1+\varepsilon)d}}$$

and

$$\gamma_n = c_2 n^{-\frac{1}{2\alpha+2\alpha p+dp+(1-p)(1+\varepsilon)d}}$$

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

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

where $c_3 > 0$ is a constant. With this, we finally obtain

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

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 5.2. *We fix finite 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 $0 < \epsilon_n < \hat{c}\, n^{-\frac{2\alpha+d}{2\alpha+2\alpha p+dp+(1-p)(1+\varepsilon)d}}$, a constant $\hat{c} > 0$ and $\delta_n > 0$. Then, for all $\varepsilon > 0$, $p \in (0,1)$, $d > 0$, $\alpha > 0$ and all $n \geq 1$, we have*

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

where $c > 0$ is a constant independent of n, Λ, ϵ_n, Γ, and δ_n.

Proof. Without loss of generality, we may assume that Λ and Γ are of the form $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$ and $\Gamma = \{\gamma_1, \ldots, \gamma_l\}$ with $\lambda_{i-1} < \lambda_i$ 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 $[0,1]^2$. [28, 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 $\epsilon_n \leq \frac{\hat{c}}{c_1}\lambda^*$. It is easy to see that

$$\lambda_i - \lambda_{i-1} \leq 2\epsilon_n \qquad \text{and} \qquad \gamma_j - \gamma_{j-1} \leq 2\delta_n \tag{31}$$

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 (31) 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 [28, Lemma A.1.6.], we obtain

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

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

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}_{L,\mathrm{P}}^* &\leq \frac{c_1}{2}\left(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p m} + \frac{\rho}{m}\right) \\ &\leq c_1\left(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\varepsilon)d}}{\lambda^p n} + \frac{\rho}{n}\right)\end{aligned} \tag{32}$$

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, [28, 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

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}})-\mathcal{R}^*_{L,\mathrm{P}}&<6\left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n}\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma})-\mathcal{R}^*_{L,\mathrm{P}}\right)+512M^2\frac{\rho_n}{n-m}\\&<6\left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n}\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma})-\mathcal{R}^*_{L,\mathrm{P}}\right)+2048M^2\frac{\rho_n}{n}\end{aligned}\qquad(33)$$

with probability P^{n-m} not less than $1-e^{-\rho}$. With (32), (33) and Lemma 5.2 we can conclude

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

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 $1-e^{-\rho}$. □

In the end, it remains to show that learning method (4) yields learning rates for regression functions contained in Sobolev or Besov spaces.

Proof of Corollary 3.4. The assumption $f^*_{L,\mathrm{P}}\in W_2^\alpha(\mathrm{P}_X)$ implies $f^*_{L,\mathrm{P}}\in W_2^\alpha(X)$. Then the extension operator $\mathfrak{E}$ in the sense of Theorem 2.2 yields $\mathfrak{E}f^*_{L,\mathrm{P}}\in W_2^\alpha\left(\mathbb{R}^d\right)$ and finally (6) implies $\mathfrak{E}f^*_{L,\mathrm{P}}\in B^\alpha_{2,\infty}(\mathbb{R}^d)=\mathrm{Lip}^*(\alpha,L_2(\mathbb{R}^d))$. By the definition of $\mathrm{Lip}^*(\alpha,L_2(\mathbb{R}^d))$ we obtain

$$\omega_{r,L_2(\mathbb{R}^d)}\left(\mathfrak{E}f^*_{L,\mathrm{P}},t\right)\leq c\,t^\alpha\;,\qquad t>0$$

for a suitable constant $c>0$. With this, all assumptions of Corollary 3.2 and of Theorem 3.3 are satisfied and hence we obtain the learning rate

$$n^{-\frac{2\alpha}{2\alpha+2\alpha p+dp+(1-p)(1+\varepsilon)d}}\;.$$

Finally, for every $\xi>0$ we can find $\varepsilon,p\in(0,1)$ sufficiently close to 0 such that the latter learning rate is at least as fast as

$$n^{-\frac{2\alpha}{2\alpha+d}+\xi}\;.$$

□

Proof of Corollary 3.5. For $\alpha\geq1$, $f^*_{L,\mathrm{P}}\in B^\alpha_{2,\infty}(\mathrm{P}_X)$ implies $f^*_{L,\mathrm{P}}\in B^\alpha_{2,\infty}(X)$, since P_X is the uniform distribution on X. With the help of the extension operator $\mathfrak{E}$ in the sense of Theorem 2.2 $f^*_{L,\mathrm{P}}\in B^\alpha_{2,\infty}(X)$ yields $\mathfrak{E}f^*_{L,\mathrm{P}}\in B^\alpha_{2,\infty}(\mathbb{R}^d)=\mathrm{Lip}^*(\alpha,L_2(\mathbb{R}^d))$. By the definition of $\mathrm{Lip}^*(\alpha,L_2(\mathbb{R}^d))$ we again obtain

$$\omega_{r,L_2(\mathbb{R}^d)}\left(\mathfrak{E}f^*_{L,\mathrm{P}},t\right)\leq c\,t^\alpha\;,\qquad t>0$$

for $\alpha\geq1$ and a suitable constant $c>0$. Now the assertion follows just as in the proof of Corollary 3.4. □

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

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

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 5.3. *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. [29, 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

$$\mathcal{C}_{L_\tau,Q}(t^*-t) - \mathcal{C}^*_{L_\tau,Q} = \int_0^t Q\left((t^*-s, t^*)\right) ds \leq \int_0^t b_Q s^{q-1} ds \leq \frac{b_Q}{q} t^q \tag{34}$$

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 (34). □

Proof of Theorem 4.7. With Lemma 5.3 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 [28, 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 2.6 we know that, for all $\varepsilon > 0$ and $0 < \varsigma < 1$, there exists a constant $c_{\varepsilon,\varsigma} \geq 0$ such that

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

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 (20) is satisfied for all $f : X \to \mathbb{R}$, we can apply [28, Theorem 7.23]. To this end, we choose $f_0 := K * \tilde{f}$, where $K : \mathbb{R}^d \to \mathbb{R}$ is defined by (8) and

$$\tilde{f}\,(x) := \left(\gamma\sqrt{\pi}\right)^{-\frac{d}{2}} \mathfrak{E} f^*_{\tau,\mathrm{P}}\,(x)$$

for all $x \in \mathbb{R}^d$. Because of $f^*_{\tau,\mathrm{P}}(x) \in [-1,1]$ for all $x \in X$, we know $\|f^*_{\tau,\mathrm{P}}\|_{L_2(X)} \leq 2^{\frac{d}{2}}$. This yields

$$\|\tilde{f}\|_{L_2(\mathbb{R}^d)} = (\gamma\sqrt{\pi})^{-\frac{d}{2}} \|\mathfrak{E} f^*_{\tau,\mathrm{P}}\|_{L_2(\mathbb{R}^d)} \leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} a_{0,2} \|f^*_{\tau,\mathrm{P}}\|_{L_2(X)} \leq 2^{\frac{d}{2}} (\gamma\sqrt{\pi})^{-\frac{d}{2}} a_{0,2}$$

and $\tilde{f} \in L_2(\mathbb{R}^d)$. Next, Theorem 2.4 implies $f_0 \in H_\gamma$ and

$$\|f_0\|_{H_\gamma} = (2^r - 1)\|\tilde{f}\|_{L_2(\mathbb{R}^d)} \leq (2^r - 1) 2^{\frac{d}{2}} (\gamma\sqrt{\pi})^{-\frac{d}{2}} a_{0,2}\ .$$

Because of $\|f^*_{\tau,\mathrm{P}}\|_{L_\infty(X)} \leq 1$ we have

$$\begin{aligned}\|\tilde{f}\|_{L_\infty(\mathbb{R}^d)} &= (\gamma\sqrt{\pi})^{-\frac{d}{2}} \|\mathfrak{E} f^*_{\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)} \\ &\leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} a_{0,\infty} \|f^*_{\tau,\mathrm{P}}\|_{L_\infty(X)} \\ &\leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} a_{0,\infty}\ ,\end{aligned}$$

i.e. $\tilde{f} \in L_\infty(\mathbb{R}^d)$. With this, Theorem 2.4 yields

$$|K * \tilde{f}(x)| \leq (\gamma\sqrt{\pi})^{\frac{d}{2}} (2^r - 1) \|\tilde{f}\|_{L_\infty(\mathbb{R}^d)} \leq a_{0,\infty}(2^r - 1) \tag{35}$$

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

$$\begin{aligned}L_\tau(y, K * \tilde{f}(x)) &\leq |y - K * \tilde{f}(x)| \\ &\leq 1 + (2^r - 1) a_{0,\infty} \\ &\leq 2^r a,\end{aligned}$$

where $a := \max\{a_{0,\infty}, 1\}$. With this, we obtain

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

where $B_0 = 2^r a \geq 2 = B$. 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 2.3 and derive

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

where we used (21). Finally, [28, Theorem 7.23] yields that, for all fixed $\rho > 0$ and $\lambda > 0$, the

SVM using H_γ and L_τ satisfies

$$\begin{aligned}
&\lambda \left\| f_{\mathrm{D},\lambda,\gamma} \right\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}\left(\widehat{f}_{\mathrm{D},\lambda,\gamma}\right) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\
&\le 9(\lambda \|f_0\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}}) \\
&\quad + c_1 \left(\frac{c_{\varepsilon,\varsigma}^{2\varsigma} \gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma n} \right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + 3\left(\frac{72V\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15B_0\rho}{n} \\
&\le 9\left(\lambda(2^r-1)^2 2^d (\gamma\sqrt{\pi})^{-d} a_{0,2}^2 + q^{-1} \|b_{\mathrm{P}(\,\cdot\,|x)}\|_{L_p(\mathrm{P}_X)} C_{r,u}^{\frac{q}{u}} c^q \gamma^{q\alpha}\right) \\
&\quad + c_1 \left(\frac{c_{\varepsilon,\varsigma}^{2\varsigma} \gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma n} \right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + 3\left(\frac{72V\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{15\cdot 2^r a\rho}{n} \\
&\le C\left(\lambda\gamma^{-d} + \gamma^{q\alpha} + \left(\frac{\gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma n} \right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + \left(\frac{\rho}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{\rho}{n} \right)
\end{aligned}$$

with probability P^n not less than $1-e^{-\rho}$ and a constant $C>0$ depending on r, d, $a_{0,2}$, $a_{0,\infty}$, q, p, $\|b_{\mathrm{P}(\,\cdot\,|x)}\|_{L_p(\mathrm{P}_X)}$, 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

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

where $c>0$ is a constant. In addition, we know by [28, Lemma A.1.6.] that the sequences

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

and

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

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

$$\lambda_n\gamma_n^{-d} + \gamma_n^{q\alpha} + \left(\frac{\gamma_n^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda_n^\varsigma n} \right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} \le c_3 n^{-\frac{q\alpha}{q\alpha(2-\varsigma-\vartheta+\vartheta\varsigma)+q\alpha\varsigma+d\varsigma+(1-\varsigma)(1+\varepsilon)d}} ,$$

where $c_3>0$ is a constant. With this, we finally obtain

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

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

To prove Theorem 4.10 we need the following lemma.

Lemma 5.4. *We fix finite 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 a constant* $\hat{c} > 0$, $0 < \epsilon_n < \hat{c}\, n^{-\frac{q\alpha+d}{q\alpha(2-\varsigma-\vartheta+\vartheta\varsigma)+q\alpha\varsigma+d\varsigma+(1-\varsigma)(1+\varepsilon)d}}$*, and* $\delta_n > 0$*. 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*

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

with 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 $[0,1]^2$. [28, 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 $\epsilon_n \leq \frac{\hat{c}}{c_1}\lambda^*$. Now the proof follows analogously to the proof of Lemma 5.2. □

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 $1 - |\Lambda_n \times \Gamma_n|\, e^{-\rho}$

$$
\begin{aligned}
\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} &\leq \frac{c_1}{2} \left(\lambda\gamma^{-d} + \gamma^{q\alpha} + \left(\frac{\gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma m} \right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + \left(\frac{\rho}{m} \right)^{\frac{1}{2-\vartheta}} + \frac{\rho}{m} \right) \\
&\leq c_1 \left(\lambda\gamma^{-d} + \gamma^{q\alpha} + \left(\frac{\gamma^{-(1-\varsigma)(1+\varepsilon)d}}{\lambda^\varsigma n} \right)^{\frac{1}{2-\varsigma-\vartheta+\vartheta\varsigma}} + \left(\frac{\rho}{n} \right)^{\frac{1}{2-\vartheta}} + \frac{\rho}{n} \right) \qquad (36)
\end{aligned}
$$

for all $(\lambda,\gamma) \in \Lambda_n \times \Gamma_n$ simultaneously. Here, $c_1 > 0$ is a constant. Furthermore, [28, 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

$$
\begin{aligned}
\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} &< 6 \left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \right) + 4 \left(\frac{8V\rho_n}{n-m} \right)^{\frac{1}{2-\vartheta}} \\
&< 6 \left(\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \right) + 4 \left(\frac{32V\rho_n}{n} \right)^{\frac{1}{2-\vartheta}} \qquad (37)
\end{aligned}
$$

with probability P^{n-m} not less than $1 - e^{-\rho}$. With (36), (37) and Lemma 5.4 we can conclude

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

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 $1-e^{-\rho}$. Finally, for every $\xi > 0$, we can find $\varepsilon, \varsigma \in (0,1)$ sufficiently close to 0 such that $n^{-\frac{q\alpha}{q\alpha(2-\varsigma-\vartheta+\vartheta\varsigma)+q\alpha\varsigma+d\varsigma+(1-\varsigma)(1+\varepsilon)d}}$ is at least as fast as

$$n^{-\frac{q\alpha}{q\alpha(2-\vartheta)+d}+\xi} \ .$$

□

Proof of Theorem 4.11. If $\vartheta := \min\{\frac{2}{q}, \frac{p}{p+1}\}$, we know by [29, 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 (20) 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 $\xi > 0$,

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

with a constant $C_\xi > 0$. Finally, the self calibration inequality (17) yields

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

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

Now, if $f^*_{\tau,\mathrm{P}} \in W^\alpha_2(\mathrm{P}_X)$ we also have $\mathfrak{E} f^*_{\tau,\mathrm{P}} \in W^\alpha_2\left(\mathbb{R}^d\right)$ by the proof of Corollary 3.4. Next, (6) implies $\mathfrak{E} f^*_{\tau,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathbb{R}^d) = \mathrm{Lip}^*(\alpha, L_2(\mathbb{R}^d))$ and therefore we obtain

$$\omega_{r,L_2(\mathbb{R}^d)}\left(\mathfrak{E} f^*_{\tau,\mathrm{P}}, t\right) \leq c\, t^\alpha \ , \qquad\qquad t > 0 \ ,$$

for a suitable constant $c > 0$. This yields the assertions by the first part of the corollary. To prove the assertion for $f^*_{\tau,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_X)$, we proceed in the same way using the proof of Corollary 3.5. □

References

[1] R. Adams and J. Fournier. *Sobolev Spaces.* Academic Press, 2^{nd} edition, 2003.

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

[3] H. Berens and R. DeVore. Quantitative Korovin theorems for positive linear operators on L_p-spaces. *AMS*, Volume 245, 1978.

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

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

[6] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators.* Cambridge University Press, 1990.

[7] P. Chaudhuri. Global Nonparametric Estimation of Conditional Quantile Functions and Their Derivatives. *J. Multivariate Anal.*, 39:246–269, 1991.

[8] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge, 2000.

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

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

[11] R. DeVore and G. Lorentz. *Constructive Approximation.* Springer-Verlag Berlin Heidelberg, 1993.

[12] R. DeVore and V. Popov. Interpolation of Besov Spaces. *AMS*, 305, 1988.

[13] D. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential 0perators.* Cambridge University Press, 1996.

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

[15] H. Johnen and K. Scherer. On the equivalence of the K-functional and moduli of continuity and some applications. In *Lecture Notes in Math.*, volume 571, pages 119–140. Springer-Verlag Berlin, 1976.

[16] S. S. Keerthi and S. K. Shevade. SMO algorithm for least squares SVM formulations. *Neural Computation*, 15:487–507, 2003.

[17] R. Koenker. *Quantile Regression.* Cambridge University Press, 1 edition, 2005.

[18] S. Lee. Efficient Semiparametric Estimation of a Partially Linear Quantile Regression Model. *Econometric Theory*, 19:1–31, 2003.

[19] Y. Li, Y. Liu, and J. Zhu. Quantile Regression in Reproducing Kernel Hilbert Spaces. *J. Amer. Statist. Assoc.*, 102(477):255–268, 2007.

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

[21] C. Micchelli, M. Pontil, Q. Wu, and D.-X. Zhou. Error bounds for learning the kernel. 2005.

[22] B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[23] X. Shen. On the Method of Penalization. *Statist. Sinica*, pages 337–357, 1998.

[24] S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Anal. Appl.*, Volume 1, 2003.

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

[26] E. Stein. *Singular Integrals and Differentiability Properties of Functions.* Princeton Univ. Press, 1970.

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

[28] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer-Verlag, New York, 2008.

[29] I. Steinwart and A. Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17:211–225, 2011.

[30] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[31] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35:575–607, 2007.

[32] I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola. Nonparametric Quantile Estimation. *J. Mach. Learn. Res. 7*, pages 1231–1264, 2006.

[33] V. Temlyakov. Optimal estimators in learning theory. *Banach Center Publications, Inst. Math. Polish Academy of Sciences*, 72:341–366, 2006.

[34] H. Triebel. *Theory of Function Spaces III.* Birkhäuser Verlag, 2006.

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

[36] G.-B. Ye and D.-X. Zhou. Learning and approximation by Gaussians on Riemannian manifolds. *Adv. Comput. Math.*, Volume 29, 2008.

[37] Y. Ying and C. Campbell. Generalization bounds for learning the kernel. In S. Dasgupta and A. Klivans, editors, *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[38] Y. Ying and D.-X. Zhou. Learnability of Gaussians with flexible variances. *J. Mach. Learn. Res. 8*, 2007.

Mona Eberts
Pfaffenwaldring 57
70569 Stuttgart
Germany
E-Mail: eberts@mathematik.uni-stuttgart.de
WWW: http://www.mathematik.uni-stuttgart.de/visitenkarte.jsp?pid=729

Ingo Steinwart
Pfaffenwaldring 57
70569 Stuttgart
Germany
E-Mail: ingo.steinwart@mathematik.uni-stuttgart.de
WWW: http://www.isa.uni-stuttgart.de/Steinwart/

Erschienene Preprints ab Nummer 2007/001

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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