Bernoulli **17**(1), 2011, 211–225
DOI: 10.3150/10-BEJ267

Estimating conditional quantiles with the help of the pinball loss

INGO STEINWART[1] and ANDREAS CHRISTMANN[2]

[1] *University of Stuttgart, Department of Mathematics, D-70569 Stuttgart, Germany.*
E-mail: ingo.steinwart@mathematik.uni-stuttgart.de
[2] *University of Bayreuth, Department of Mathematics, D-95440 Bayreuth.*
E-mail: Andreas.Christmann@uni-bayreuth.de

The so-called pinball loss for estimating conditional quantiles is a well-known tool in both statistics and machine learning. So far, however, only little work has been done to quantify the efficiency of this tool for nonparametric approaches. We fill this gap by establishing inequalities that describe how close approximate pinball risk minimizers are to the corresponding conditional quantile. These inequalities, which hold under mild assumptions on the data-generating distribution, are then used to establish so-called variance bounds, which recently turned out to play an important role in the statistical analysis of (regularized) empirical risk minimization approaches. Finally, we use both types of inequalities to establish an oracle inequality for support vector machines that use the pinball loss. The resulting learning rates are min–max optimal under some standard regularity assumptions on the conditional quantile.

Keywords: nonparametric regression; quantile estimation; support vector machines

1. Introduction

Let P be a distribution on $X \times \mathbb{R}$, where X is an arbitrary set equipped with a σ-algebra. The goal of quantile regression is to estimate the conditional quantile, that is, the set-valued function

$$F^*_{\tau,\mathrm{P}}(x) := \{t \in \mathbb{R} : \mathrm{P}((-\infty, t]|x) \ge \tau \text{ and } \mathrm{P}([t, \infty)|x) \ge 1 - \tau\}, \qquad x \in X,$$

where $\tau \in (0, 1)$ is a fixed constant specifying the desired quantile level and $\mathrm{P}(\cdot|x)$, $x \in X$, is the regular conditional probability of P. Throughout this paper, we assume that $\mathrm{P}(\cdot|x)$ has its support in $[-1, 1]$ for P_X-almost all $x \in X$, where P_X denotes the marginal distribution of P on X. (By a simple scaling argument, all our results can be generalized to distributions living on $X \times [-M, M]$ for some $M > 0$. The uniform boundedness of the conditionals $\mathrm{P}(\cdot|x)$ is, however, crucial.) Let us additionally assume for a moment that $F^*_{\tau,\mathrm{P}}(x)$ consists of singletons, that is, there exists an $f^*_{\tau,\mathrm{P}} : X \to \mathbb{R}$, called the conditional τ-quantile function, such that $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$ for P_X-almost all $x \in X$. (Most of our main results do not require this assumption, but here, in the introduction, it makes the exposition more transparent.) Then one approach to estimate the conditional τ-quantile function is based on the so-called *τ-pinball loss* $L : Y \times \mathbb{R} \to [0, \infty)$, which is defined by

$$L(y, t) := \begin{cases} (1 - \tau)(t - y), & \text{if } y < t, \\ \tau(y - t), & \text{if } y \ge t. \end{cases}$$

1350-7265 © 2011 ISI/BS

With the help of this loss function we define the L-risk of a function $f : X \to \mathbb{R}$ by

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

Recall that $f^*_{\tau,\mathrm{P}}$ is up to P_X-zero sets the *only* function satisfying $\mathcal{R}_{L,\mathrm{P}}(f^*_{\tau,\mathrm{P}}) = \inf \mathcal{R}_{L,\mathrm{P}}(f) =: \mathcal{R}^*_{L,\mathrm{P}}$, where the infimum is taken over *all* measurable functions $f : X \to \mathbb{R}$. Based on this observation, several estimators minimizing a (modified) empirical L-risk were proposed (see [13] for a survey on both parametric and nonparametric methods) for situations where P is unknown, but i.i.d. samples $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times \mathbb{R})^n$ drawn from P are given.

Empirical methods estimating quantile functions with the help of the pinball loss typically obtain functions f_D for which $\mathcal{R}_{L,\mathrm{P}}(f_D)$ is close to $\mathcal{R}^*_{L,\mathrm{P}}$ with high probability. In general, however, this only implies that f_D is close to $f^*_{\tau,\mathrm{P}}$ in a very weak sense (see [21], Remark 3.18) but recently, [23], Theorem 2.5, established *self-calibration inequalities* of the form

$$\| f - f^*_{\tau,\mathrm{P}} \|_{L_r(\mathrm{P}_X)} \le c_{\mathrm{P}} \sqrt{\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}}, \tag{1}$$

which hold under mild assumptions on P described by the parameter $r \in (0, 1]$. The first goal of this paper is to generalize and to improve these inequalities. Moreover, we will use these new self-calibration inequalities to establish *variance bounds* for the pinball risk, which in turn are known to improve the statistical analysis of empirical risk minimization (ERM) approaches.

The second goal of this paper is to apply the self-calibration inequalities and the variance bounds to support vector machines (SVMs) for quantile regression. Recall that [12,20,26] proposed an SVM that finds a solution $f_{D,\lambda} \in H$ of

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

where $\lambda > 0$ is a regularization parameter, H is a reproducing kernel Hilbert space (RKHS) over X, and $\mathcal{R}_{L,\mathrm{D}}(f)$ denotes the empirical risk of f, that is, $\mathcal{R}_{L,\mathrm{D}}(f) := \frac{1}{n}\sum_{i=1}^n L(y_i, f(x_i))$. In [9] robustness properties and consistency for *all* distributions P on $X \times \mathbb{R}$ were established for this SVM, while [12,26] worked out how to solve this optimization problem with standard techniques from machine learning. Moreover, [26] also provided an exhaustive empirical study, which shows the excellent performance of this SVM. We have recently established an oracle inequality for these SVMs in [23], which was based on (1) and the resulting variance bounds. In this paper, we improve this oracle inequality with the help of the new self-calibration inequalities and variance bounds. It turns out that the resulting learning rates are substantially faster than those of [23]. Finally, we briefly discuss an adaptive parameter selection strategy.

The rest of this paper is organized as follows. In Section 2, we present both our new self-calibration inequality and the new variance bound. We also introduce the assumptions on P that lead to these inequalities and discuss how these inequalities improve our former results in [23]. In Section 3, we use these new inequalities to establish an oracle inequality for the SVM approach above. In addition, we discuss the resulting learning rates and how these can be achieved in an adaptive way. Finally, all proofs are contained in Section 4.

2. Main results

In order to formulate the main results of this section, we need to introduce some assumptions on the data-generating distribution P. To this end, let Q be a distribution on $\mathbb{R}$ and $\operatorname{supp} \mathrm{Q}$ be its support. For $\tau \in (0, 1)$, the τ*-quantile* of Q is the set

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

It is well known that $F_\tau^*(\mathrm{Q})$ is a bounded and closed interval. We write

$$t_{\min}^*(\mathrm{Q}) := \min F_\tau^*(\mathrm{Q}) \quad \text{and} \quad t_{\max}^*(\mathrm{Q}) := \max F_\tau^*(\mathrm{Q}),$$

which implies $F_\tau^*(\mathrm{Q}) = [t_{\min}^*(\mathrm{Q}), t_{\max}^*(\mathrm{Q})]$. Moreover, it is easy to check that the interior of $F_\tau^*(\mathrm{Q})$ is a Q-zero set, that is, $\mathrm{Q}((t_{\min}^*(\mathrm{Q}), t_{\max}^*(\mathrm{Q}))) = 0$. To avoid notational overload, we usually omit the argument Q if the considered distribution is clearly determined from the context.

Definition 2.1 (Quantiles of type q). *A distribution* Q *with* $\operatorname{supp} \mathrm{Q} \subset [-1, 1]$ *is said to have a* τ*-quantile of type* $q \in (1, \infty)$ *if there exist constants* $\alpha_{\mathrm{Q}} \in (0, 2]$ *and* $b_{\mathrm{Q}} > 0$ *such that*

$$\mathrm{Q}\big((t_{\min}^* - s, t_{\min}^*)\big) \geq b_{\mathrm{Q}} s^{q-1}, \tag{3}$$

$$\mathrm{Q}\big((t_{\max}^*, t_{\max}^* + s)\big) \geq b_{\mathrm{Q}} s^{q-1} \tag{4}$$

for all $s \in [0, \alpha_{\mathrm{Q}}]$. *Moreover,* Q *has a* τ*-quantile of type* $q = 1$, *if* $\mathrm{Q}(\{t_{\min}^*\}) > 0$ *and* $\mathrm{Q}(\{t_{\max}^*\}) > 0$. *In this case, we define* $\alpha_{\mathrm{Q}} := 2$ *and*

$$b_{\mathrm{Q}} := \begin{cases} \min\{\mathrm{Q}(\{t_{\min}^*\}), \mathrm{Q}(\{t_{\max}^*\})\}, & \textit{if } t_{\min}^* \neq t_{\max}^*, \\ \min\{\tau - \mathrm{Q}((-\infty, t_{\min}^*)), \mathrm{Q}((-\infty, t_{\max}^*]) - \tau\}, & \textit{if } t_{\min}^* = t_{\max}^*, \end{cases}$$

where we note that $b_{\mathrm{Q}} > 0$ *in both cases. For all* $q \geq 1$, *we finally write* $\gamma_{\mathrm{Q}} := b_{\mathrm{Q}} \alpha_{\mathrm{Q}}^{q-1}$.

Since τ-quantiles of type q are the central concept of this work, let us illustrate this notion by a few examples. We begin with an example for which all quantiles are of type 2.

Example 2.2. Let ν be a distribution with $\operatorname{supp} \nu \subset [-1, 1]$, μ be a distribution with $\operatorname{supp} \mu \subset [-1, 1]$ that has a density h with respect to the Lebesgue measure and $\mathrm{Q} := \alpha\nu + (1 - \alpha)\mu$ for some $\alpha \in [0, 1)$. If h is bounded away from 0, that is, $h(y) \geq b$ for some $b > 0$ and Lebesgue-almost all $y \in [-1, 1]$, then Q has a τ-quantile of type $q = 2$ for all $\tau \in (0, 1)$ as simple integration shows. In this case, we set $b_{\mathrm{Q}} := (1 - \alpha)b$ and $\alpha_{\mathrm{Q}} := \min\{1 + t_{\min}^*, 1 - t_{\max}^*\}$.

Example 2.3. Again, let ν be a distribution with $\operatorname{supp} \nu \subset [-1, 1]$, μ be a distribution with $\operatorname{supp} \mu \subset [-1, 1]$ that has a Lebesgue density h, and $\mathrm{Q} := \alpha\nu + (1 - \alpha)\mu$ for some $\alpha \in [0, 1)$. If, for a fixed $\tau \in (0, 1)$, there exist constants $b > 0$ and $p > -1$ such that

$$h(y) \geq b\big(t_{\min}^*(\mathrm{Q}) - y\big)^p, \qquad y \in [-1, t_{\min}^*(\mathrm{Q})],$$

$$h(y) \geq b\big(y - t_{\max}^*(\mathrm{Q})\big)^p, \qquad y \in [t_{\max}^*(\mathrm{Q}), 1].$$

Lebesgue-almost surely, then simple integration shows that Q has a τ-quantile of type $q = 2 + p$ and we may set $b_Q := (1-\alpha)b/(1+p)$ and $\alpha_Q := \min\{1 + t^*_{\min}(Q), 1 - t^*_{\max}(Q)\}$.

Example 2.4. Let ν be a distribution with $\operatorname{supp}\nu \subset [-1,1]$ and $Q := \alpha\nu + (1-\alpha)\delta_{t^*}$ for some $\alpha \in [0,1)$, where δ_{t^*} denotes the Dirac measure at $t^* \in (0,1)$. If $\nu(\{t^*\}) = 0$, we then have $Q((-\infty, t^*)) = \alpha\nu((-\infty, t^*))$ and $Q((-\infty, t^*]) = \alpha\nu((-\infty, t^*)) + 1 - \alpha$, and hence $\{t^*\}$ is a τ-quantile of type $q = 1$ for all τ satisfying $\alpha\nu((-\infty, t^*)) < \tau < \alpha\nu((-\infty, t^*)) + 1 - \alpha$.

Example 2.5. Let ν be a distribution with $\operatorname{supp}\nu \subset [-1,1]$ and $Q := (1-\alpha-\beta)\nu + \alpha\delta_{t_{\min}} + \beta\delta_{t_{\max}}$ for some $\alpha, \beta \in (0,1]$ with $\alpha + \beta \le 1$. If $\nu([t_{\min}, t_{\max}]) = 0$, we have $Q((-\infty, t_{\min}]) = (1-\alpha-\beta)\nu((-\infty, t_{\min}]) + \alpha$ and $Q([t_{\max}, \infty)) = (1-\alpha-\beta)(1 - \nu((-\infty, t_{\min}])) + \beta$. Consequently, $[t_{\min}, t_{\max}]$ is the $\tau := (1-\alpha-\beta)\nu((-\infty, t^*]) + \alpha$ quantile of Q and this quantile is of type $q = 1$.

As outlined in the introduction, we are not interested in a single distribution Q on $\mathbb{R}$ but in distributions P on $X \times \mathbb{R}$. The following definition extends the previous definition to such P.

Definition 2.6 (Quantiles of p-average type q). *Let $p \in (0,\infty]$, $q \in [1,\infty)$, and P be a distribution on $X \times \mathbb{R}$ with $\operatorname{supp} P(\cdot|x) \subset [-1,1]$ for P_X-almost all $x \in X$. Then P is said to have a τ-quantile of p-average type q, if $P(\cdot|x)$ has a τ-quantile of type q for P_X-almost all $x \in X$, and the function $\gamma : X \to [0,\infty]$ defined, for P_X-almost all $x \in X$, by*

$$\gamma(x) := \gamma_{P(\cdot|x)},$$

where $\gamma_{P(\cdot|x)} = b_{P(\cdot|x)}\alpha^{q-1}_{P(\cdot|x)}$ is defined in Definition 2.1, satisfies $\gamma^{-1} \in L_p(P_X)$.

To establish the announced self-calibration inequality, we finally need the distance

$$\operatorname{dist}(t, A) := \inf_{s \in A} |t - s|$$

between an element $t \in \mathbb{R}$ and an $A \subset \mathbb{R}$. Moreover, $\operatorname{dist}(f, F^*_{\tau,P})$ denotes the function $x \mapsto \operatorname{dist}(f(x), F^*_{\tau,P}(x))$. With these preparations the self-calibration inequality reads as follows.

Theorem 2.7. *Let L be the τ-pinball loss, $p \in (0,\infty]$ and $q \in [1,\infty)$ be real numbers, and $r := \frac{pq}{p+1}$. Moreover, let P be a distribution that has a τ-quantile of p-average type $q \in [1,\infty)$. Then, for all $f : X \to [-1,1]$, we have*

$$\|\operatorname{dist}(f, F^*_{\tau,P})\|_{L_r(P_X)} \le 2^{1-1/q} q^{1/q} \|\gamma^{-1}\|^{1/q}_{L_p(P_X)} \bigl(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\bigr)^{1/q}.$$

Let us briefly compare the self-calibration inequality above with the one established in [23]. To this end, we can solely focus on the case $q = 2$, since this was the only case considered in [23]. For the same reason, we can restrict our considerations to distributions P that have a unique conditional τ-quantile $f^*_{\tau,P}(x)$ for P_X-almost all $x \in X$. Then Theorem 2.7 yields

$$\|f - f^*_{\tau,P}\|_{L_r(P_X)} \le 2\|\gamma^{-1}\|^{1/2}_{L_p(P_X)} \bigl(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\bigr)^{1/2}$$

for $r := \frac{2p}{p+1}$. On the other hand, it was shown in [23], Theorem 2.5, that

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_{r/2}(\mathrm{P}_X)} \le \sqrt{2}\|\gamma^{-1}\|^{1/2}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\big)^{1/2}$$

under the *additional* assumption that the conditional widths $\alpha_{\mathrm{P}(\cdot|x)}$ considered in Definition 2.1 are *independent* of x. Consequently, our new self-calibration inequality is more general and, modulo the constant $\sqrt{2}$, also sharper.

It is well known that self-calibration inequalities for Lipschitz continuous losses lead to variance bounds, which in turn are important for the statistical analysis of ERM approaches; see [1,2,14–17,28]. For the pinball loss, we obtain the following variance bound.

Theorem 2.8. *Let L be the τ-pinball loss, $p \in (0, \infty]$ and $q \in [1, \infty)$ be real numbers, and*

$$\vartheta := \min\left\{\frac{2}{q}, \frac{p}{p+1}\right\}.$$

Let P be a distribution that has a τ-quantile of p-average type q. Then, 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

$$\mathbb{E}_{\mathrm{P}}(L \circ f - L \circ f^*_{\tau,\mathrm{P}})^2 \le 2^{2-\vartheta} q^{\vartheta}\|\gamma^{-1}\|^{\vartheta}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\big)^{\vartheta},$$

where we used the shorthand $L \circ f$ for the function $(x, y) \mapsto L(y, f(x))$.

Again, it is straightforward to show that the variance bound above is both more general and stronger than the variance bound established in [23], Theorem 2.6.

3. An application to support vector machines

The goal of this section is to establish an oracle inequality for the SVM defined in (2). The use of this oracle inequality is then illustrated by some learning rates we derive from it.

Let us begin by recalling some RKHS theory (see, e.g., [24], Chapter 4, for a more detailed account). To this end, let $k : X \times X \to \mathbb{R}$ be a measurable kernel, that is, a measurable function that is symmetric and positive definite. Then the associated RKHS H consists of measurable functions. Let us additionally assume that k is bounded with $\|k\|_\infty := \sup_{x\in X}\sqrt{k(x, x)} \le 1$, which in turn implies that H consists of bounded functions and $\|f\|_\infty \le \|f\|_H$ for all $f \in H$.

Suppose now that we have a distribution P on $X \times Y$. To describe the approximation error of SVMs we use the *approximation error function*

$$A(\lambda) := \inf_{f\in H} \lambda\|f\|^2_H + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}, \qquad \lambda > 0,$$

where L is the τ-pinball loss. Recall that [24], Lemma 5.15 and Theorem 5.31, showed that $\lim_{\lambda\to 0} A(\lambda) = 0$, if the RKHS H is dense in $L_1(\mathrm{P}_X)$ and the speed of this convergence describes how well H approximates the Bayes L-risk $\mathcal{R}^*_{L,\mathrm{P}}$. In particular, [24], Corollary 5.18, shows that

$A(\lambda) \le c\lambda$ for some constant $c > 0$ and all $\lambda > 0$ if and only if there exists an $f \in H$ such that $f(x) \in F^*_{\tau,\mathrm{P}}(x)$ for P_X-almost all $x \in X$.

We further need the integral operator $T_k : L_2(\mathrm{P}_X) \to L_2(\mathrm{P}_X)$ defined by

$$T_k f(\cdot) := \int_X k(x, \cdot) f(x) \, d\mathrm{P}_X(x), \qquad f \in L_2(\mathrm{P}_X).$$

It is well known that T_k is self-adjoint and nuclear; see, for example, [24], Theorem 4.27. Consequently, it has at most countably many eigenvalues (including geometric multiplicities), which are all non-negative and summable. Let us order these eigenvalues $\lambda_i(T_k)$. Moreover, if we only have finitely many eigenvalues, we extend this finite sequence by zeros. As a result, we can always deal with a decreasing, non-negative sequence $\lambda_1(T_k) \ge \lambda_2(T_k) \ge \cdots$, which satisfies $\sum_{i=1}^{\infty} \lambda_i(T_k) < \infty$. The finiteness of this sum can already be used to establish oracle inequalities; see [24], Theorem 7.22. But in the following we assume that the eigenvalues converge even faster to zero, since (a) this case is satisfied for many RKHSs and (b) it leads to better oracle inequalities. To be more precise, we assume that there exist constants $a \ge 1$ and $\varrho \in (0, 1)$ such that

$$\lambda_i(T_k) \le a i^{-1/\varrho}, \qquad i \ge 1. \tag{5}$$

Recall that (5) was first used in [6] to establish an oracle inequality for SVMs using the hinge loss, while [7,18,25] consider (5) for SVMs using the least-squares loss. Furthermore, one can show (see [22]) that (5) is equivalent (modulo a constant only depending on ϱ) to

$$e_i\big(\mathrm{id} : H \to L_2(\mathrm{P}_X)\big) \le \sqrt{a} i^{-1/(2\varrho)}, \qquad i \ge 1, \tag{6}$$

where $e_i(\mathrm{id} : H \to L_2(\mathrm{P}_X))$ denotes the ith (dyadic) entropy number [8] of the inclusion map from H into $L_2(\mathrm{P}_X)$. In addition, [22] shows that (6) implies a bound on expectations of random entropy numbers, which in turn are used in [24], Chapter 7.4, to establish general oracle inequalities for SVMs. On the other hand, (6) has been extensively studied in the literature. For example, for m-times differentiable kernels on Euclidean balls X of $\mathbb{R}^d$, it is known that (6) holds for $\varrho := \frac{d}{2m}$. We refer to [10], Chapter 5, and [24], Theorem 6.26, for a precise statement. Analogously, if $m > d/2$ is some integer, then the Sobolev space $H := W^m(X)$ is an RKHS that satisfies (6) for $\varrho := \frac{d}{2m}$, and this estimate is also asymptotically sharp; see [5,11].

We finally need the clipping operation defined by

$$\widehat{t} := \max\{-1, \min\{1, t\}\}$$

for all $t \in \mathbb{R}$. We can now state the following oracle inequality for SVMs using the pinball loss.

Theorem 3.1. *Let L be the τ-pinball loss 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 $x \in X$. Assume that there exists a function $f^*_{\tau,\mathrm{P}} : X \to \mathbb{R}$ with $f^*_{\tau,\mathrm{P}}(x) \in F^*_{\tau,\mathrm{P}}(x)$ for P_X-almost all $x \in X$ and constants $V \ge 2^{2-\vartheta}$ and $\vartheta \in [0, 1]$ such that*

$$\mathbb{E}_{\mathrm{P}}(L \circ f - L \circ f^*_{\tau,\mathrm{P}})^2 \le V\big(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\big)^{\vartheta} \tag{7}$$

for all $f : X \to [-1, 1]$. *Moreover, let* H *be a separable RKHS over* X *with a bounded measurable kernel satisfying* $\|k\|_\infty \le 1$. *In addition, assume that* (5) *is satisfied for some* $a \ge 1$ *and* $\varrho \in (0, 1)$. *Then there exists a constant* K *depending only on* ϱ, V, *and* ϑ *such that, for all* $\varsigma \ge 1$, $n \ge 1$ *and* $\lambda > 0$, *we have with probability* P^n *not less than* $1 - 3e^{-\varsigma}$ *that*

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{D,\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} \le 9A(\lambda) + 30\sqrt{\frac{A(\lambda)}{\lambda}\frac{\varsigma}{n}} + K\left(\frac{a^\varrho}{\lambda^\varrho n}\right)^{1/(2-\varrho-\vartheta+\vartheta\varrho)} + 3\left(\frac{72V\varsigma}{n}\right)^{1/(2-\vartheta)}.$$

Let us now discuss the learning rates obtained from this oracle inequality. To this end, we assume in the following that there exist constants $c > 0$ and $\beta \in (0, 1]$ such that

$$A(\lambda) \le c\lambda^\beta, \qquad \lambda > 0. \tag{8}$$

Recall from [24], Corollary 5.18, that, for $\beta = 1$, this assumption holds if and only if there exists a τ-quantile function $f^*_{\tau,\mathrm{P}}$ with $f^*_{\tau,\mathrm{P}} \in H$. Moreover, for $\beta < 1$, there is a tight relationship between (8) and the behavior of the approximation error of the balls $\lambda^{-1}B_H$; see [24], Theorem 5.25. In addition, one can show (see [24], Chapter 5.6) that if $f^*_{\tau,\mathrm{P}}$ is contained in the real interpolation space $(L_1(\mathrm{P}_X), H)_{\vartheta,\infty}$, see [4], then (8) is satisfied for $\beta := \vartheta/(2-\vartheta)$. For example, if $H := W^m(X)$ is a Sobolev space over a Euclidean ball $X \subset \mathbb{R}^d$ of order $m > d/2$ and P_X has a Lebesgue density that is bounded away from 0 and ∞, then $f^*_{\tau,\mathrm{P}} \in W^s(X)$ for some $s \in (d/2, m]$ implies (8) for $\beta := s/(2m - s)$.

Now assume that (8) holds. We further assume that λ is determined by $\lambda_n = n^{-\gamma/\beta}$, where

$$\gamma := \min\left\{\frac{\beta}{\beta(2-\vartheta+\varrho\vartheta-\varrho)+\varrho}, \frac{2\beta}{\beta+1}\right\}. \tag{9}$$

Then Theorem 3.1 shows that $\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{D,\lambda_n})$ converges to $\mathcal{R}^*_{L,\mathrm{P}}$ with rate $n^{-\gamma}$; see [24], Lemma A.1.7, for calculating the value of γ. Note that this choice of λ yields the best learning rates from Theorem 3.1. Unfortunately, however, this choice requires knowledge of the usually unknown parameters β, ϑ and ϱ. To address this issue, let us consider the following scheme that is close to approaches taken in practice (see [19] for a similar technique that has a fast implementation based on regularization paths).

Definition 3.2. *Let* H *be an RKHS over* X *and* $\Lambda := (\Lambda_n)$ *be a sequence of finite subsets* $\Lambda_n \subset (0, 1]$. *Given a data set* $D := ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times \mathbb{R})^n$, *we define*

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

where $m := \lfloor n/2 \rfloor + 1$ *and* $n \ge 3$. *Then we use* D_1 *to compute the SVM decision functions*

$$f_{D_1,\lambda} := \arg\min_{f \in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}_1}(f), \qquad \lambda \in \Lambda_n,$$

and D_2 to determine λ by choosing a $\lambda_{D_2} \in \Lambda_n$ such that

$$\mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{D_1,\lambda_{D_2}}) = \min_{\lambda\in\Lambda_n} \mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{D_1,\lambda}).$$

In the following, we call this learning method, which produces the decision functions $\widehat{f}_{D_1,\lambda_{D_2}}$, a training validation SVM with respect to Λ.

Training validation SVMs have been extensively studied in [24], Chapter 7.4. In particular, [24], Theorem 7.24, gives the following result that shows that the learning rate $n^{-\gamma}$ can be achieved without knowing of the existence of the parameters β, ϑ and ϱ or their particular values.

Theorem 3.3. *Let (Λ_n) be a sequence of n^{-2}-nets Λ_n of $(0,1]$ such that the cardinality $|\Lambda_n|$ of Λ_n grows polynomially in n. Furthermore, consider the situation of Theorem 3.1 and assume that (8) is satisfied for some $\beta \in (0,1]$. Then the training validation SVM with respect to $\Lambda := (\Lambda_n)$ learns with rate $n^{-\gamma}$, where γ is defined by (9).*

Let us now consider how these learning rates in terms of risks translate into rates for

$$\|\widehat{f}_{D,\lambda_n} - f^*_{\tau,\mathrm{P}}\|_{L_r(\mathrm{P}_X)} \to 0. \tag{10}$$

To this end, we assume that P has a τ-quantile of p-average type q, where we additionally assume for the sake of simplicity that $r := \frac{pq}{p+1} \le 2$. Note that the latter is satisfied for all p if $q \le 2$, that is, if all conditional distributions are concentrated around the quantile at least as much as the uniform distribution; see the discussion following Definition 2.1. We further assume that the conditional quantiles $F^*_{\tau,\mathrm{P}}(x)$ are singletons for P_X-almost all $x \in X$. Then Theorem 2.8 provides a variance bound of the form (7) for $\vartheta := p/(p+1)$, and hence γ defined in (9) becomes

$$\gamma = \min\left\{\frac{\beta(p+1)}{\beta(2+p-\varrho)+\varrho(p+1)}, \frac{2\beta}{\beta+1}\right\}.$$

By Theorem 2.7 we consequently see that (10) converges with rate $n^{-\gamma/q}$, where $r := pq/(p+1)$. To illustrate this learning rate, let us assume that we have picked an RKHS H with $f^*_{\tau,\mathrm{P}} \in H$. Then we have $\beta = 1$, and hence it is easy to check that the latter learning rate reduces to

$$n^{-(p+1)/(q(2+p+\varrho p))}.$$

For the sake of simplicity, let us further assume that the conditional distributions do not change too much in the sense that $p = \infty$. Then we have $r = q$, and hence

$$\int_X |\widehat{f}_{D,\lambda_n} - f^*_{\tau,\mathrm{P}}|^q \, d\mathrm{P}_X \tag{11}$$

converges to zero with rate $n^{-1/(1+\varrho)}$. The latter shows that the value of q does not change the learning rate for (11), but only the exponent in (11). Now note that by our assumption on P and the definition of the clipping operation we have

$$\|\widehat{f}_{D,\lambda_n} - f^*_{\tau,\mathrm{P}}\|_\infty \le 2,$$

and consequently small values of q emphasize the discrepancy of $\widehat{f}_{D,\lambda_n}$ to $f^*_{\tau,\mathrm{P}}$ more than large values of q do. In this sense, a stronger average concentration around the quantile is helpful for the learning process.

Let us now have a closer look at the special case $q = 2$, which is probably the most interesting case for applications. Then we have the learning rate $n^{-1/(2(1+\varrho))}$ for

$$\|\widehat{f}_{D,\lambda_n} - f^*_{\tau,\mathrm{P}}\|_{L_2(\mathrm{P}_X)}.$$

Now recall that the conditional median equals the conditional mean for *symmetric* conditional distributions $\mathrm{P}(\cdot|x)$. Moreover, if H is a Sobolev space $W^m(X)$, where $m > d/2$ denotes the smoothness index and X is a Euclidean ball in $\mathbb{R}^d$, then H consists of continuous functions, and [11] shows that H satisfies (5) for $\varrho := d/(2m)$. Consequently, we see that in this case the latter convergence rate is optimal in a min–max sense [27,29] if P_X is the uniform distribution. Finally, recall that in the case $\beta = 1$, $q = 2$ and $p = \infty$ discussed so far, the results derived in [23] only yield a learning rate of $n^{-1/(3(1+\varrho))}$ for

$$\|\widehat{f}_{D,\lambda_n} - f^*_{\tau,\mathrm{P}}\|_{L_1(\mathrm{P}_X)}.$$

In other words, the earlier rates from [23] are not only worse by a factor of $3/2$ in the exponent but also are stated in terms of the weaker $L_1(\mathrm{P}_X)$-norm. In addition, [23] only considered the case $q = 2$, and hence we see that our new results are also more general.

4. Proofs

Since the proofs of Theorems 2.7 and 2.8 use some notation developed in [21] and [24], Chapter 3, let us begin by recalling these. To this end, let L be the τ-pinball loss for some fixed $\tau \in (0, 1)$ and Q be a distribution on $\mathbb{R}$ with $\operatorname{supp} \mathrm{Q} \subset [-1, 1]$. Then [21,24] defined the *inner L-risks* by

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

and the *minimal inner L-risk* was denoted by $\mathcal{C}^*_{L,\mathrm{Q}} := \inf_{t\in\mathbb{R}} \mathcal{C}_{L,\mathrm{Q}}(t)$. Moreover, we write $\mathcal{M}_{L,\mathrm{Q}}(0^+) = \{t \in \mathbb{R} : \mathcal{C}_{L,\mathrm{Q}}(t) = \mathcal{C}^*_{L,\mathrm{Q}}\}$ for the set of exact minimizers.

Our first goal is to compute the excess inner risks and the set of exact minimizers for the pinball loss. To this end recall that (see [3], Theorem 23.8), given a distribution Q on $\mathbb{R}$ and a measurable function $g : X \to [0, \infty)$, we have

$$\int_{\mathbb{R}} g\, \mathrm{dQ} = \int_0^\infty \mathrm{Q}(\{g \ge s\})\, \mathrm{d}s. \tag{12}$$

With these preparations we can now show the following generalization of [24], Proposition 3.9.

Proposition 4.1. *Let L be the τ-pinball loss and Q be a distribution on $\mathbb{R}$ with $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$. Then there exist $q_+, q_- \in [0,1]$ with $q_+ + q_- = \mathrm{Q}([t^*_{\min}, t^*_{\max}])$, and, for all $t \geq 0$, we have*

$$\mathcal{C}_{L,\mathrm{Q}}(t^*_{\max} + t) - \mathcal{C}^*_{L,\mathrm{Q}} = tq_+ + \int_0^t \mathrm{Q}\big((t^*_{\max}, t^*_{\max} + s)\big)\,\mathrm{d}s, \tag{13}$$

$$\mathcal{C}_{L,\mathrm{Q}}(t^*_{\min} - t) - \mathcal{C}^*_{L,\mathrm{Q}} = tq_- + \int_0^t \mathrm{Q}\big((t^*_{\min} - s, t^*_{\min})\big)\,\mathrm{d}s. \tag{14}$$

Moreover, if $t^*_{\min} \neq t^*_{\max}$, then we have $q_- = \mathrm{Q}(\{t^*_{\min}\})$ and $q_+ = \mathrm{Q}(\{t^*_{\max}\})$. Finally, $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ equals the τ-quantile, that is, $\mathcal{M}_{L,\mathrm{Q}}(0^+) = [t^*_{\min}, t^*_{\max}]$.

Proof. Obviously, we have $\mathrm{Q}((-\infty, t^*_{\max}]) + \mathrm{Q}([t^*_{\max}, \infty)) = 1 + \mathrm{Q}(\{t^*_{\max}\})$, and hence we obtain $\tau \leq \mathrm{Q}((-\infty, t^*_{\max}]) \leq \tau + \mathrm{Q}(\{t^*_{\max}\})$. In other words, there exists a $q_+ \in [0,1]$ satisfying $0 \leq q_+ \leq \mathrm{Q}(\{t^*_{\max}\})$ and

$$\mathrm{Q}((-\infty, t^*_{\max}]) = \tau + q_+. \tag{15}$$

Let us consider the distribution $\tilde{\mathrm{Q}}$ defined by $\tilde{\mathrm{Q}}(A) := \mathrm{Q}(t^*_{\max} + A)$ for all measurable $A \subset \mathbb{R}$. Then it is not hard to see that $t^*_{\max}(\tilde{\mathrm{Q}}) = 0$. Moreover, we obviously have $\mathcal{C}_{L,\mathrm{Q}}(t^*_{\max} + t) = \mathcal{C}_{L,\tilde{\mathrm{Q}}}(t)$ for all $t \in \mathbb{R}$. Let us now compute the inner risks of L with respect to $\tilde{\mathrm{Q}}$. To this end, we fix a $t \geq 0$. Then we have

$$\int_{y<t} (y-t)\,\mathrm{d}\tilde{\mathrm{Q}}(y) = \int_{y<0} y\,\mathrm{d}\tilde{\mathrm{Q}}(y) - t\tilde{\mathrm{Q}}((-\infty, t)) + \int_{0 \leq y < t} y\,\mathrm{d}\tilde{\mathrm{Q}}(y)$$

and

$$\int_{y \geq t} (y-t)\,\mathrm{d}\tilde{\mathrm{Q}}(y) = \int_{y \geq 0} y\,\mathrm{d}\tilde{\mathrm{Q}}(y) - t\tilde{\mathrm{Q}}([t, \infty)) - \int_{0 \leq y < t} y\,\mathrm{d}\tilde{\mathrm{Q}}(y)$$

and hence we obtain

$$\begin{aligned}\mathcal{C}_{L,\tilde{\mathrm{Q}}}(t) &= (\tau - 1)\int_{y<t} (y-t)\,\mathrm{d}\tilde{\mathrm{Q}}(y) + \tau \int_{y \geq t} (y-t)\,\mathrm{d}\tilde{\mathrm{Q}}(y) \\ &= \mathcal{C}_{L,\tilde{\mathrm{Q}}}(0) - \tau t + t\tilde{\mathrm{Q}}((-\infty, 0)) + t\tilde{\mathrm{Q}}([0,t)) - \int_{0 \leq y < t} y\,\mathrm{d}\tilde{\mathrm{Q}}(y).\end{aligned}$$

Moreover, using (12) we find

$$t\tilde{\mathrm{Q}}([0,t)) - \int_{0 \leq y < t} y\,\mathrm{d}\tilde{\mathrm{Q}}(y) = \int_0^t \tilde{\mathrm{Q}}([0,t))\,\mathrm{d}s - \int_0^t \tilde{\mathrm{Q}}([s,t))\,\mathrm{d}s = t\tilde{\mathrm{Q}}(\{0\}) + \int_0^t \tilde{\mathrm{Q}}((0,s))\,\mathrm{d}s,$$

and since (15) implies $\tilde{\mathrm{Q}}((-\infty, 0)) + \tilde{\mathrm{Q}}(\{0\}) = \tilde{\mathrm{Q}}((-\infty, 0]) = \tau + q_+$, we thus obtain

$$\mathcal{C}_{L,\mathrm{Q}}(t^*_{\max} + t) = \mathcal{C}_{L,\mathrm{Q}}(t)^*_{\max} + tq_+ + \int_0^t \mathrm{Q}\big((t^*_{\max}, t^*_{\max} + s)\big)\,\mathrm{d}s. \tag{16}$$

By considering the pinball loss with parameter $1-\tau$ and the distribution $\bar{\mathrm{Q}}$ defined by $\bar{\mathrm{Q}}(A) := \mathrm{Q}(-t^*_{\min} - A)$, $A \subset \mathbb{R}$ measurable, we further see that (16) implies

$$\mathcal{C}_{L,\mathrm{Q}}(t^*_{\min} - t) = \mathcal{C}_{L,\mathrm{Q}}(t)^*_{\min} + tq_- + \int_0^t \mathrm{Q}\big((t^*_{\min} - s, t^*_{\min})\big)\,\mathrm{d}s, \qquad t \geq 0, \tag{17}$$

where q_- satisfies $0 \leq q_- \leq \mathrm{Q}(\{t^*_{\min}\})$ and $\mathrm{Q}([t^*_{\min}, \infty)) = 1 - \tau + q_-$. By (15) we then find $q_+ + q_- = \mathrm{Q}([t^*_{\min}, t^*_{\max}])$. Moreover, if $t^*_{\min} \neq t^*_{\max}$, the fact $\mathrm{Q}((t^*_{\min}, t^*_{\max})) = 0$ yields

$$q_+ + q_- = \mathrm{Q}([t^*_{\min}, t^*_{\max}]) = \mathrm{Q}(\{t^*_{\min}\}) + \mathrm{Q}(\{t^*_{\max}\}).$$

Using the earlier established $q_+ \leq \mathrm{Q}(\{t^*_{\max}\})$ and $q_- \leq \mathrm{Q}(\{t^*_{\min}\})$, we then find both $q_- = \mathrm{Q}(\{t^*_{\min}\})$ and $q_+ = \mathrm{Q}(\{t^*_{\max}\})$.

To prove (13) and (14), we first consider the case $t^*_{\min} = t^*_{\max}$. Then (16) and (17) yield $\mathcal{C}_{L,\mathrm{Q}}(t)^*_{\min} = \mathcal{C}_{L,\mathrm{Q}}(t)^*_{\max} \leq \mathcal{C}_{L,\mathrm{Q}}(t)$, $t \in \mathbb{R}$. This implies $\mathcal{C}_{L,\mathrm{Q}}(t)^*_{\min} = \mathcal{C}_{L,\mathrm{Q}}(t)^*_{\max} = \mathcal{C}^*_{L,\mathrm{Q}}$, and hence we conclude that (16) and (17) are equivalent to (13) and (14), respectively. Moreover, in the case $t^*_{\min} \neq t^*_{\max}$, we have $\mathrm{Q}((t^*_{\min}(\mathrm{Q}), t^*_{\max}(\mathrm{Q}))) = 0$, which in turn implies $\mathrm{Q}((-\infty, t^*_{\min}]) = \tau$ and $\mathrm{Q}([t^*_{\max}, \infty)) = 1 - \tau$. For $t \in (t^*_{\min}, t^*_{\max}]$, we consequently find

$$\begin{aligned} \mathcal{C}_{L,\mathrm{Q}}(t) &= (\tau - 1)\int_{y<t} (y-t)\,\mathrm{dQ}(y) + \tau \int_{y \geq t} (y-t)\,\mathrm{dQ}(y) \\ &= (\tau - 1)\int_{y<t^*_{\max}} y\,\mathrm{dQ}(y) + \tau \int_{y \geq t^*_{\max}} y\,\mathrm{dQ}(y), \end{aligned} \tag{18}$$

where we used $\mathrm{Q}((-\infty, t)) = \mathrm{Q}((-\infty, t^*_{\min}]) = \tau$ and $\mathrm{Q}([t, \infty)) = \mathrm{Q}([t^*_{\max}, \infty)) = 1 - \tau$. Since the right-hand side of (18) is independent of t, we thus conclude $\mathcal{C}_{L,\mathrm{Q}}(t) = \mathcal{C}_{L,\mathrm{Q}}(t)^*_{\max}$ for all $t \in (t^*_{\min}, t^*_{\max}]$. Analogously, we find $\mathcal{C}_{L,\mathrm{Q}}(t) = \mathcal{C}_{L,\mathrm{Q}}(t)^*_{\min}$ for all $t \in [t^*_{\min}, t^*_{\max})$, and hence we can, again, conclude $\mathcal{C}_{L,\mathrm{Q}}(t)^*_{\min} = \mathcal{C}_{L,\mathrm{Q}}(t)^*_{\max} \leq \mathcal{C}_{L,\mathrm{Q}}(t)$ for all $t \in \mathbb{R}$. As in the case $t^*_{\min} = t^*_{\max}$, the latter implies that (16) and (17) are equivalent to (13) and (14), respectively.

For the proof of $\mathcal{M}_{L,\mathrm{Q}}(0^+) = [t^*_{\min}, t^*_{\max}]$, we first note that the previous discussion has already shown $\mathcal{M}_{L,\mathrm{Q}}(0^+) \supset [t^*_{\min}, t^*_{\max}]$. Let us assume that $\mathcal{M}_{L,\mathrm{Q}}(0^+) \not\subset [t^*_{\min}, t^*_{\max}]$. By a symmetry argument, we then may assume without loss of generality that there exists a $t \in \mathcal{M}_{L,\mathrm{Q}}(0^+)$ with $t > t^*_{\max}$. From (13) we then conclude that $q_+ = 0$ and $\mathrm{Q}((t^*_{\max}, t)) = 0$. Now, $q_+ = 0$ together with (15) shows $\mathrm{Q}((-\infty, t^*_{\max}]) = \tau$, which in turn implies $\mathrm{Q}((-\infty, t]) \geq \tau$. Moreover, $\mathrm{Q}((t^*_{\max}, t)) = 0$ yields

$$\mathrm{Q}([t, \infty)) = \mathrm{Q}([t^*_{\max}, \infty)) - \mathrm{Q}(\{t^*_{\max}\}) = 1 - \mathrm{Q}((-\infty, t^*_{\max}]) = 1 - \tau.$$

In other words, t is a τ-quantile, which contradicts $t > t^*_{\max}$. □

For the proof of Theorem 2.7 we further need the *self-calibration loss* of L that is defined by

$$\breve{L}(\mathrm{Q}, t) := \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}}(0^+)), \qquad t \in \mathbb{R}, \tag{19}$$

where Q is a distribution with $\operatorname{supp} \mathrm{Q} \subset [-1,1]$. Let us define the *self-calibration function* by

$$\delta_{\max,\breve{L},L}(\varepsilon,\mathrm{Q}) := \inf_{t\in\mathbb{R}:\breve{L}(\mathrm{Q},t)\geq\varepsilon} \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}, \qquad \varepsilon \geq 0.$$

Note that if, for $t \in \mathbb{R}$, we write $\varepsilon := \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}}(0^+))$, then we have $\breve{L}(\mathrm{Q},t) \geq \varepsilon$, and hence the definition of the self-calibration function yields

$$\delta_{\max,\breve{L},L}(\operatorname{dist}(t,\mathcal{M}_{L,\mathrm{Q}}(0^+)),\mathrm{Q}) \leq \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}, \qquad t \in \mathbb{R}. \tag{20}$$

In other words, the self-calibration function measures how well an ε-approximate L-risk minimizer t approximates the set of exact L-risk minimizers.

Our next goal is to estimate the self-calibration function for the pinball loss. To this end we need the following simple technical lemma.

Lemma 4.2. *For $\alpha \in [0,2]$ and $q \in [1,\infty)$ consider the function $\delta : [0,2] \to [0,\infty)$ defined by*

$$\delta(\varepsilon) := \begin{cases} \varepsilon^q, & \text{if } \varepsilon \in [0,\alpha], \\ q\alpha^{q-1}\varepsilon - \alpha^q(q-1), & \text{if } \varepsilon \in [\alpha,2]. \end{cases}$$

Then, for all $\varepsilon \in [0,2]$, we have

$$\delta(\varepsilon) \geq \left(\frac{\alpha}{2}\right)^{q-1} \varepsilon^q.$$

Proof. Since $\alpha \leq 2$ and $q \geq 1$ we easily see by the definition of δ that the assertion is true for $\varepsilon \in [0,\alpha]$. Now consider the function $h : [\alpha,2] \to \mathbb{R}$ defined by

$$h(\varepsilon) := q\alpha^{q-1}\varepsilon - \alpha^q(q-1) - \left(\frac{\alpha}{2}\right)^{q-1}\varepsilon^q, \qquad \varepsilon \in [\alpha,2].$$

It suffices to show that $h(\varepsilon) \geq 0$ for all $\varepsilon \in [\alpha,2]$. To show the latter we first check that

$$h'(\varepsilon) = q\alpha^{q-1} - q\left(\frac{\alpha}{2}\right)^{q-1}\varepsilon^{q-1}, \qquad \varepsilon \in [\alpha,2]$$

and hence we have $h'(\varepsilon) \geq 0$ for all $\varepsilon \in [\alpha,2]$. Now we obtain the assertion from this, $\alpha \in [0,2]$ and

$$h(\alpha) = \alpha^q - \left(\frac{\alpha}{2}\right)^{q-1}\alpha^q = \alpha^q\left(1 - \left(\frac{\alpha}{2}\right)^{q-1}\right) \geq 0. \qquad \square$$

Lemma 4.3. *Let L be the τ-pinball loss and Q be a distribution on $\mathbb{R}$ with $\operatorname{supp}\mathrm{Q} \subset [-1,1]$ that has a τ-quantile of type $q \in [1,\infty)$. Moreover, let $\alpha_\mathrm{Q} \in (0,2]$ and $b_\mathrm{Q} > 0$ denote the corresponding constants. Then, for all $\varepsilon \in [0,2]$, we have*

$$\delta_{\max,\breve{L},L}(\varepsilon,\mathrm{Q}) \geq q^{-1}b_\mathrm{Q}\left(\frac{\alpha_\mathrm{Q}}{2}\right)^{q-1}\varepsilon^q = q^{-1}2^{1-q}\gamma_\mathrm{Q}\varepsilon^q.$$

Proof. Since L is convex, the map $t \mapsto \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}$ is convex, and thus it is decreasing on $(-\infty, t^*_{\min}]$ and increasing on $[t^*_{\max}, \infty)$. Using $\mathcal{M}_{L,\mathrm{Q}}(0^+) = [t^*_{\min}, t^*_{\max}]$, we thus find

$$\mathcal{M}_{\breve{L},\mathrm{Q}}(\varepsilon) := \{t \in \mathbb{R} : \breve{L}(\mathrm{Q}, t) < \varepsilon\} = (t^*_{\min} - \varepsilon, t^*_{\max} + \varepsilon)$$

for all $\varepsilon > 0$. Since this gives $\delta_{\max,\breve{L},L}(\varepsilon, \mathrm{Q}) = \inf_{t \notin \mathcal{M}_{\breve{L},\mathrm{Q}}(\varepsilon)} \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}$, we obtain

$$\delta_{\max,\breve{L},L}(\varepsilon, \mathrm{Q}) = \min\{\mathcal{C}_{L,\mathrm{Q}}(t^*_{\min} - \varepsilon), \mathcal{C}_{L,\mathrm{Q}}(t^*_{\max} + \varepsilon)\} - \mathcal{C}^*_{L,\mathrm{Q}}. \tag{21}$$

Let us first consider the case $q \in (1, \infty)$. For $\varepsilon \in [0, \alpha_{\mathrm{Q}}]$, (13) and (4) then yield

$$\mathcal{C}_{L,\mathrm{Q}}(t^*_{\max} + \varepsilon) - \mathcal{C}^*_{L,\mathrm{Q}} = \varepsilon q_+ + \int_0^\varepsilon \mathrm{Q}\big((t^*_{\max}, t^*_{\max} + s)\big)\, \mathrm{d}s \geq b_{\mathrm{Q}} \int_0^\varepsilon s^{q-1}\, \mathrm{d}s = q^{-1} b_{\mathrm{Q}} \varepsilon^q,$$

and, for $\varepsilon \in [\alpha_{\mathrm{Q}}, 2]$, (13) and (4) yield

$$\mathcal{C}_{L,\mathrm{Q}}(t^*_{\max} + \varepsilon) - \mathcal{C}^*_{L,\mathrm{Q}} \geq b_{\mathrm{Q}} \int_0^{\alpha_{\mathrm{Q}}} s^{q-1}\, \mathrm{d}s + b_{\mathrm{Q}} \int_{\alpha_{\mathrm{Q}}}^\varepsilon \alpha_{\mathrm{Q}}^{q-1}\, \mathrm{d}s = q^{-1} b_{\mathrm{Q}} \big(q \alpha_{\mathrm{Q}}^{q-1} \varepsilon - \alpha_{\mathrm{Q}}^q (q-1)\big).$$

For $\varepsilon \in [0, 2]$, we have thus shown $\mathcal{C}_{L,\mathrm{Q}}(t^*_{\max} + \varepsilon) - \mathcal{C}^*_{L,\mathrm{Q}} \geq q^{-1} b_{\mathrm{Q}} \delta(\varepsilon)$, where δ is the function defined in Lemma 4.2 for $\alpha := \alpha_{\mathrm{Q}}$.

Furthermore, in the case $q = 1$ and $t^*_{\min} \neq t^*_{\max}$, Proposition 4.1 shows $q_+ = \mathrm{Q}(\{t^*_{\max}\})$, and hence (13) yields $\mathcal{C}_{L,\mathrm{Q}}(t^*_{\max} + \varepsilon) - \mathcal{C}^*_{L,\mathrm{Q}} \geq \varepsilon q_+ \geq b_{\mathrm{Q}} \varepsilon$ for all $\varepsilon \in [0, 2] = [0, \alpha_{\mathrm{Q}}]$ by the definition of b_{Q} and α_{Q}. In the case $q = 1$ and $t^*_{\min} = t^*_{\max}$, (15) yields $q_+ = \mathrm{Q}((-\infty, t^*]) - \tau \geq b_{\mathrm{Q}}$ by the definition of b_{Q}, and hence (13) again gives $\mathcal{C}_{L,\mathrm{Q}}(t^*_{\max} + \varepsilon) - \mathcal{C}^*_{L,\mathrm{Q}} \geq b_{\mathrm{Q}} \varepsilon$ for all $\varepsilon \in [0, 2]$. Finally, using (14) instead of (13), we can analogously show $\mathcal{C}_{L,\mathrm{Q}}(t^*_{\min} - \varepsilon) - \mathcal{C}^*_{L,\mathrm{Q}} \geq q^{-1} b_{\mathrm{Q}} \delta(\varepsilon)$ for all $\varepsilon \in [0, 2]$ and $q \geq 1$. By (21) we thus conclude that

$$\delta_{\max,\breve{L},L}(\varepsilon, \mathrm{Q}) \geq q^{-1} b_{\mathrm{Q}} \delta(\varepsilon)$$

for all $\varepsilon \in [0, 2]$. Now the assertion follows from Lemma 4.2. □

Proof of Theorem 2.7. For fixed $x \in X$ we write $\varepsilon := \operatorname{dist}(f(x), \mathcal{M}_{L,\mathrm{P}(\cdot|x)}(0^+))$. By Lemma 4.3 and (20) we obtain, for P_X-almost all $x \in X$,

$$\begin{aligned} \big|\operatorname{dist}\big(f(x), \mathcal{M}_{L,\mathrm{P}(\cdot|x)}(0^+)\big)\big|^q &\leq q 2^{q-1} \gamma^{-1}(x) \delta_{\max,\breve{L},L}(\varepsilon, \mathrm{P}(\cdot|x)) \\ &\leq q 2^{q-1} \gamma^{-1}(x) \big(\mathcal{C}_{L,\mathrm{P}(\cdot|x)}(f(x)) - \mathcal{C}^*_{L,\mathrm{P}(\cdot|x)}\big). \end{aligned}$$

By taking the $\frac{p}{p+1}$th power on both sides, integrating and finally applying Hölder's inequality, we then obtain the assertion. □

Proof of Theorem 2.8. Let $f : X \to [-1, 1]$ be a function. Since $F^*_{\tau,\mathrm{P}}(x)$ is closed, there then exists a P_X-almost surely uniquely determined function $f^*_{\tau,\mathrm{P}} : X \to [-1, 1]$ that satisfies both

$$\begin{aligned} f^*_{\tau,\mathrm{P}}(x) &\in F^*_{\tau,\mathrm{P}}(x), \\ |f(x) - f^*_{\tau,\mathrm{P}}(x)| &= \operatorname{dist}(f(x), F^*_{\tau,\mathrm{P}}(x)) \end{aligned}$$

for P_X-almost all $x \in X$. Let us write $r := \frac{pq}{p+1}$. We first consider the case $r \le 2$, that is, $\frac{2}{q} \le \frac{p}{p+1}$. Using the Lipschitz continuity of the pinball loss L and Theorem 2.7 we then obtain

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}(L \circ f - L \circ f^*_{\tau,\mathrm{P}})^2 &\le \mathbb{E}_{\mathrm{P}_X}|f - f^*_{\tau,\mathrm{P}}|^2 \\ &\le \|f - f^*_{\tau,\mathrm{P}}\|_\infty^{2-r} \mathbb{E}_{\mathrm{P}_X}|f - f^*_{\tau,\mathrm{P}}|^r \\ &\le 2^{2-r/q} q^{r/q} \|\gamma^{-1}\|_{L_p(\mathrm{P}_X)}^{r/q} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\bigr)^{r/q}.\end{aligned}$$

Since $\frac{r}{q} = \frac{p}{p+1} = \vartheta$, we thus obtain the assertion in this case. Let us now consider the case $r > 2$. The Lipschitz continuity of L and Theorem 2.7 yield

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}(L \circ f - L \circ f^*_{\tau,\mathrm{P}})^2 &\le \bigl(\mathbb{E}_{\mathrm{P}}(L \circ f - L \circ f^*_{\tau,\mathrm{P}})^r\bigr)^{2/r} \\ &\le (\mathbb{E}_{\mathrm{P}_X}|f - f^*_{\tau,\mathrm{P}}|^r)^{2/r} \\ &\le \bigl(2^{1-1/q} q^{1/q} \|\gamma^{-1}\|_{L_p(\mathrm{P}_X)}^{1/q} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\bigr)^{1/q}\bigr)^2 \\ &= 2^{2-2/q} q^{2/q} \|\gamma^{-1}\|_{L_p(\mathrm{P}_X)}^{2/q} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\bigr)^{2/q}.\end{aligned}$$

Since for $r > 2$ we have $\vartheta = 2/q$, we again obtain the assertion. □

Proof of Theorem 3.1. As shown in [22], Lemma 2.2, (5) is equivalent to the entropy assumption (6), which in turn implies (see [22], Theorem 2.1, and [24], Corollary 7.31)

$$\mathbb{E}_{D_X \sim \mathrm{P}_X^n} e_i\bigl(\mathrm{id}: H \to L_2(\mathrm{D}_X)\bigr) \le c\sqrt{a}\, i^{-1/(2\varrho)}, \qquad i \ge 1, \tag{22}$$

where D_X denotes the empirical measure with respect to $D_X = (x_1, \dots, x_n)$ and $c \ge 1$ is a constant only depending on ϱ. Now the assertion follows from [24], Theorem 7.23, by considering the function $f_0 \in H$ that achieves $\lambda\|f_0\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}} = A(\lambda)$. □

References

[1] Bartlett, P.L., Bousquet, O. and Mendelson, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554

[2] Bartlett, P.L., Jordan, M.I. and McAuliffe, J.D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032

[3] Bauer, H. (2001). *Measure and Integration Theory*. Berlin: De Gruyter. MR1897176

[4] Bennett, C. and Sharpley, R. (1988). *Interpolation of Operators*. Boston: Academic Press. MR0928802

[5] Birman, M.Š. and Solomyak, M.Z. (1967). Piecewise-polynomial approximations of functions of the classes W_p^α (Russian). *Mat. Sb.* **73** 331–355. MR0217487

[6] Blanchard, G., Bousquet, O. and Massart, P. (2008). Statistical performance of support vector machines. *Ann. Statist.* **36** 489–531. MR2396805

[7] Caponnetto, A. and De Vito, E. (2007). Optimal rates for regularized least squares algorithm. *Found. Comput. Math.* **7** 331–368. MR2335249

[8] Carl, B. and Stephani, I. (1990). *Entropy, Compactness and the Approximation of Operators*. Cambridge: Cambridge Univ. Press. MR1098497
[9] Christmann, A., Van Messem, A. and Steinwart, I. (2009). On consistency and robustness properties of support vector machines for heavy-tailed distributions. *Stat. Interface*. **2** 311–327. MR2540089
[10] Cucker, F. and Zhou, D.X. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge: Cambridge Univ. Press. MR2354721
[11] Edmunds, D.E. and Triebel, H. (1996). *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge: Cambridge Univ. Press. MR1410258
[12] Hwang, C. and Shim, J. (2005). A simple quantile regression via support vector machine. In *Advances in Natural Computation: First International Conference (ICNC)* 512 –520. Berlin: Springer.
[13] Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press. MR2268657
[14] Mammen, E. and Tsybakov, A. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618
[15] Massart, P. (2000). Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse, VI. Sr., Math.* **9** 245–303. MR1813803
[16] Mendelson, S. (2001). Geometric methods in the analysis of Glivenko–Cantelli classes. In *Proceedings of the 14th Annual Conference on Computational Learning Theory* (D. Helmbold and B. Williamson, eds.) 256–272. New York: Springer. MR2042040
[17] Mendelson, S. (2001). Learning relatively small classes. In *Proceedings of the 14th Annual Conference on Computational Learning Theory* (D. Helmbold and B. Williamson, eds.) 273–288. New York: Springer. MR2042041
[18] Mendelson, S. and Neeman, J. (2010). Regularization in kernel learning. *Ann. Statist.* **38** 526–565.
[19] Rosset, S. (2009). Bi-level path following for cross validated solution of kernel quantile regression. *J. Mach. Learn. Res.* **10** 2473–2505. MR2576326
[20] Schölkopf, B., Smola, A.J., Williamson, R.C. and Bartlett, P.L. (2000). New support vector algorithms. *Neural Comput.* **12** 1207–1245.
[21] Steinwart, I. (2007). How to compare different loss functions. *Constr. Approx.* **26** 225–287. MR2327600
[22] Steinwart, I. (2009). Oracle inequalities for SVMs that are based on random entropy numbers. *J. Complexity*. **25** 437–454. MR2555510
[23] Steinwart, I. and Christmann, A. (2008). How SVMs can estimate quantiles and the median. In *Advances in Neural Information Processing Systems 20* (J.C. Platt, D. Koller, Y. Singer and S. Roweis, eds.) 305–312. Cambridge, MA: MIT Press.
[24] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. New York: Springer. MR2450103
[25] Steinwart, I., Hush, D. and Scovel, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory* (S. Dasgupta and A. Klivans, eds.) 79–93. Available at http://www.cs.mcgill.ca/~colt2009/papers/038.pdf#page=1.
[26] Takeuchi, I., Le, Q.V., Sears, T.D. and Smola, A.J. (2006). Nonparametric quantile estimation. *J. Mach. Learn. Res.* **7** 1231–1264. MR2274404
[27] Temlyakov, V. (2006). Optimal estimators in learning theory. *Banach Center Publications, Inst. Math. Polish Academy of Sciences* **72** 341–366. MR2325756
[28] Tsybakov, A.B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002
[29] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500

Received November 2008 and revised January 2010