Machine Learning (2019) 108:203–227
https://doi.org/10.1007/s10994-018-5762-9

Learning rates for kernel-based expectile regression

Muhammad Farooq[1,2] · **Ingo Steinwart**[2]

Received: 2 March 2017 / Accepted: 27 August 2018 / Published online: 20 September 2018

Abstract
Conditional expectiles are becoming an increasingly important tool in finance as well as in other areas of applications. We analyse a support vector machine type approach for estimating conditional expectiles and establish learning rates that are minimax optimal modulo a logarithmic factor if Gaussian RBF kernels are used and the desired expectile is smooth in a Besov sense. As a special case, our learning rates improves the best known rates for kernel-based least squares regression in aforementioned scenario. Key ingredients of our statistical analysis are a general calibration inequality for the asymmetric least squares loss, a corresponding variance bound as well as an improved entropy number bound for Gaussian RBF kernels.

Keywords Support vector machines · Self-calibration inequality · Variance bound · Entropy number bound · Learning rates

1 Introduction

Given i.i.d samples $D := ((x_1, y_1), \ldots, (x_n, y_n))$ drawn from some unknown probability distribution P on $X \times Y$, where X is an arbitrary set and $Y \subset \mathbb{R}$, the goal to explore the conditional distribution of Y given $x \in X$ beyond the center of the distribution can be achieved, e.g., by using quantile regression, see Koenker and Bassett Jr. (1978), or expectile regression. Recall that the τ-expectile denoted by $\mu_{\tau,Q}$, where $Q := \mathrm{P}(Y|x)$, is the unique solution of

$$\tau \int_{\mu_{\tau,Q}}^{\infty} (y - \mu_{\tau,Q}) dQ(y) = (1-\tau) \int_{-\infty}^{\mu_{\tau,Q}} (\mu_{\tau,Q} - y) dQ(y),$$

Editor: Steve Hanneke.

✉ Muhammad Farooq
muhammad.farooq@uog.edu.pk; muhammad.farooq@mathematik.uni-stuttgart.de

Ingo Steinwart
ingo.steinwart@mathematik.uni-stuttgart.de

[1] Department of Statistics, University of Gujrat, Gujrat, Pakistan

[2] Institute of Stochastics and Applications, Faculty 8: Mathematics and Physics, University of Stuttgart, 70569 Stuttgart, Germany

provided that $|Q|_1 := \int_Y y\, dQ(y) < \infty$, see Newey and Powell (1987). Algorithmically, expectiles are computed by minimizing expectation of the asymmetric least squares (ALS) loss function

$$L_\tau(y,t) = \begin{cases} (1-\tau)(y-t)^2, & \text{if } \ y < t, \\ \tau(y-t)^2, & \text{if } \ y \geqslant t, \end{cases} \tag{1}$$

for all $t \in \mathbb{R}$ and a fixed $\tau \in (0,1)$, see primarily Newey and Powell (1987) and also Efron (1991) and Abdous and Remillard (1995) for further references. These expectiles have attracted considerable attention due to successful application in many areas, for instance, in demography (see, Schnabel and Eilers 2009), in education (see, Sobotka et al. 2013) and extensively in finance, see for instance Wang et al. (2011), Hamidi et al. (2014), Xu et al. (2016) and Kim and Lee (2016). In fact, it has recently been shown (see, e.g. Bellini et al. 2014; Steinwart et al. 2014) that expectiles are the only coherent and elicitable risk measures, and thus they have been suggested as potentially better alternative to Value-at-Risk (VaR) measures, see e.g. Taylor (2008), Ziegel (2016) and Bellini et al. (2014). For more applications of expectiles, we refer the interested readers to, e.g. Aragon et al. (2005), Guler et al. (2014) and references therein.

As already mentioned above, for a predictor $f : X \to \mathbb{R}$ and any $\tau \in (0,1)$, the τ-expectile can be computed with the help of asymmetric risks

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) := \int_X \int_Y L_\tau(y, f(x))\, P(dy|x)\, d\mathrm{P}_X(x), \tag{2}$$

To be more precise, there exists a P_X-almost surely unique $f^*_{L_\tau,\mathrm{P}}$ satisfying

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

provided that $|\mathrm{P}|_2 := \left(\int_{X\times Y} y^2 d\mathrm{P}(x,y)\right)^{1/2} < \infty$. Here we note that $f^*_{L_\tau,\mathrm{P}}(x)$ equals the τ-expectile of the conditional distribution $\mathrm{P}(\cdot|x)$ for P_X-almost all $x \in X$. The corresponding empirical estimator of $f^*_{L_\tau,\mathrm{P}}$ denoted by $f_\mathrm{D} : X \to \mathbb{R}$ is obtained, for example, with the help of empirical L_τ-risks

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

where D is the empirical measures associated to data D.

A typical way to access the quality of an estimator f_D is to measure its distance to the target function $f^*_{L_\tau,\mathrm{P}}$, e.g. in terms of $\|f_\mathrm{D} - f^*_{L_\tau,\mathrm{P}}\|_{L_2(\mathrm{P}_X)}$. For estimators obtained by some empirical risk minimization scheme, however, one can hardly ever estimate this L_2-norm directly. Instead, the standard tools of statistical learning theory give bounds on the excess risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f_\mathrm{D}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$. Therefore, our *first goal* in this paper is to establish a so-called calibration inequality that relates both quantities. To be more precise, we will show in Theorem 3 that

$$\|f_\mathrm{D} - f^*_{L_\tau,\mathrm{P}}\|_{L_2(\mathrm{P}_X)} \leq c_\tau^{-1/2} \sqrt{\mathcal{R}_{L_\tau,\mathrm{P}}(f_\mathrm{D}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}}, \tag{3}$$

holds for all $f_\mathrm{D} \in L_2(\mathrm{P}_X)$ and some constant c_τ only depending on τ. In particular, (3) provides rates for $\|f_\mathrm{D} - f^*_{L_\tau,\mathrm{P}}\|_{L_2(\mathrm{P}_X)}$ as soon as we have established rates for $\mathcal{R}_{L_\tau,\mathrm{P}}(f_\mathrm{D}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$. Furthermore, it is common knowledge in statistical learning theory that bounds on $\mathcal{R}_{L_\tau,\mathrm{P}}(f_\mathrm{D}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$ can be improved if so-called variance bounds are available. We will see

in Lemma 4 that (3) leads to an optimal variance bound for L_τ whenever Y is bounded. Note that both (3) and the variance bound are independent of the considered expectile estimation method. In fact, both results are key ingredients for the statistical analysis of any expectile estimation method based on some form of empirical risk minimization.

Some semiparametric and nonparametric methods for expectile regression have already been proposed in literature, however, in almost all cases the focus has been put on the *computation* of f_D, see for instance Sobotka and Kneib (2012), Yao and Tong (1996) and Yang and Zou (2015) for further details. In fact, to the best of our knowledge, the only two papers dealing with the *statistical analysis* of expectile estimation methods are Zhang (1994) and Yang et al. (2017). However, both papers only established *consistency* results for expectile regression, so that it seems fair to say that our paper is the very first one on *learning rates* for expectile regression.

The expectile estimation method we consider in this paper belongs to the family of so-called kernel-based regularized empirical risk minimization, methods, which are also known as support vector machine (SVM) methods. Recall that given a regularization parameter $\lambda > 0$, a fixed $\tau \in (0, 1)$ and a reproducing kernel Hilbert space (RKHS) H over X with bounded, measurable kernel $k : X \times X \to \mathbb{R}$, an SVM builds a predictor $f_{D,\lambda}$ by solving an optimization problem of the form

$$f_{D,\lambda} = \arg\min_{f \in H} \left(\lambda \|f\|_H^2 + \mathcal{R}_{L_\tau, D}(f) \right). \tag{4}$$

Note that learning methods of the form (4) but with different loss functions have attracted many theoretical and algorithmic considerations, see for example Wu et al. (2006), Bauer et al. (2007), and Tacchetti et al. (2013) as well as the articles mentioned below for least squares regression, Takeuchi et al. (2006), Steinwart and Christmann (2011) and Eberts and Steinwart (2013) for quantile regression, and Glasmachers and Igel (2006), Blanchard et al. (2008), and Steinwart et al. (2011) for classification with hinge loss. Recently, Farooq and Steinwart (2017) proposed an algorithm for solving (4) considering the ALS loss and Gaussian RBF kernels

$$k_\gamma(x, x') := \exp(-\gamma^{-2}\|x - x'\|^2), \quad x, x' \in \mathbb{R}^d,$$

where $\gamma > 0$, and obtained the unique solution for $f_{D,\lambda}$ of the form

$$f_{D,\lambda} := \sum_{i=1}^{n} (\alpha_i^* - \beta_i^*) K(x_i, \cdot),$$

where $\alpha_i^* \geq 0, \beta_i^* \geq 0$ for all $i = i, \ldots, n$. This paper also provides detailed statistical support to the empirical findings of Farooq and Steinwart (2017).

Since $2L_{1/2}$ equals the least squares loss, any statistical analysis of (4) in the case $\tau = 1/2$ also provides results for SVMs using the least squares loss. The latter have already been extensively investigated in the literature. For example, learning rates for generic kernels can be found in Cucker and Smale (2002), De Vito et al. (2005), Caponnetto and De Vito (2007), Steinwart et al. (2009) Mendelson and Neeman (2010) and references therein. Among these articles, only Cucker and Smale (2002), Steinwart et al. (2009) and Mendelson and Neeman (2010) obtain learning rates in minimax sense under some specific assumptions. For example, Cucker and Smale (2002) assume that the target function $f^*_{L_{1/2},P} \in H$, while Steinwart et al. (2009) and Mendelson and Neeman (2010) establish optimal learning rates for the case in which H does not contain the target function. In addition, Eberts and Steinwart (2013) have recently established (essentially) asymptotically optimal learning rates for least squares

SVMs using Gaussian RBF kernels under the assumption that the target function $f^*_{L_{1/2},P}$ is contained in some Sobolev or Besov space $B^\alpha_{2,\infty}$ with smoothness index α. A key ingredient of this work is to control the capacity of RKHS $H_\gamma(X)$ for Gaussian RBF kernel k_γ on the closed unit Euclidean ball $X \subset \mathbb{R}^d$ by an entropy number bound

$$e_i\left(\mathrm{id} : H_\gamma(X) \to l_\infty(X)\right) \le c_{p,d}(X)\gamma^{-\frac{d}{p}} i^{-\frac{1}{p}}, \tag{5}$$

see Steinwart and Christmann (2008, Theorem 6.27), which holds for all $\gamma \in (0, 1]$ and $p \in (0, 1]$. Unfortunately, the constant $c_{p,d}(X)$ derived from Steinwart and Christmann (2008, Theorem 6.27) depends on p in an unknown manner. As a consequence, Eberts and Steinwart (2013) were only able to show learning rates of the form

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

for all $\xi > 0$. To address this issue, we use Lemma 4.5 in van der Vaart and van Zanten (2009) to derive the following new entropy number bound

$$e_i\left(\mathrm{id} : H_\gamma(X) \to l_\infty(X)\right) \le (3K)^{\frac{1}{p}} \left(\frac{d+1}{ep}\right)^{\frac{d+1}{p}} \gamma^{-\frac{d}{p}} i^{-\frac{1}{p}},$$

which holds for all $p \in (0, 1]$, $\gamma \in (0, 1]$ and some constant K only depending on d. In other words, we establish an upper bound for $c_{p,d}(X)$ whose dependence on p is explicitly known. Using this new bound, we are then able to find improved learning rates of the form

$$(\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}} n^{-\frac{2\alpha}{2\alpha+d}}.$$

Clearly these new rates replace the nuisance factor n^ξ of Eberts and Steinwart (2013) learning rates by some logarithmic term, and up to this logarithmic factor our new rates are minimax optimal, see Györfi et al. (2002) for further details. In addition, our new rates also hold for $\tau \neq 1/2$, that is for general expectiles.

The rest of this paper is organized as follows: In Sect. 2, some properties of the ALS loss function are established including the self-calibration inequality and variance bound. Section 3 presents oracle inequalities and learning rates for learning problem (4) considering Gaussian RBF kernels. The proofs of our results can be found in Sect. 4.

2 Properties of the ALS loss: self-calibration and variance bounds

This section contains some properties of the ALS loss function i.e. convexity, local Lipschitz continuity, a self-calibration inequality, a supremum bound and a variance bound. Throughout this and subsequent sections, we assume that X is an arbitrary, non-empty set equipped with σ-algebra, and $Y \subset \mathbb{R}$ denotes a closed non-empty set. In addition, we assume that P is the probability distribution on $X \times Y$, $\mathrm{P}(\cdot|x)$ is a regular conditional probability distribution on Y given $x \in X$ and Q is some distribution on Y. Furthermore, $L_\tau : Y \times \mathbb{R} \to [0, \infty)$ is the ALS loss defined by (1) and $f : X \to \mathbb{R}$ is a measurable function.

It is trivial to show that L_τ is convex in t. This convexity further ensures that the optimization problem (4) is efficiently solvable. Moreover, by Steinwart and Christmann (2008, Lemma 2.13) convexity of L_τ implies convexity of corresponding risks (2). In the following, we recall Steinwart and Christmann (2008, Definition 2.22) which present the idea of clipping to restrict the prediction t to the domain $Y = [-M, M]$ where $M > 0$.

Definition 1 We say that a loss $L : Y \times \mathbb{R} \to [0, \infty)$ can be clipped at $M > 0$, if, for all $(y, t) \in Y \times \mathbb{R}$, we have

$$L(y, \widehat{t}) \leq L(y, t), \tag{6}$$

where $\widehat{t}$ denotes the clipped value of t at $\pm M$, that is

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

Moreover, we say that L can be clipped if t can be clipped at some $M > 0$.

Recall that this clipping assumption has already been utilized while establishing learning rates for SVMs, see for instance Chen et al. (2004), Steinwart et al. (2006) and Steinwart et al. (2011) for hinge loss, and Christmann and Steinwart (2007) and Steinwart and Christmann (2011) for pinball loss. It is trivial to show by convexity of L_τ together with Lemma 2.23 in Steinwart and Christmann (2008) that L_τ can be clipped at M and has at least one global minimizer in $[-M, M]$. This also implies that $\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}) \leq \mathcal{R}_{L_\tau,\mathrm{P}}(f)$ for every $f : X \to \mathbb{R}$. In other words, the clipping operation potentially reduces the risks. We therefore bound the risk $\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda})$ of the clipped decision function rather than the risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D},\lambda})$, which we will see in Sect. 3. From a practical point of view, this means that the *training* algorithm for (4) remains unchanged and the *evaluation* of the resulting decision function requires only a slight change. For further details on algorithmic advantages of clipping for SVMs using the hinge loss and the ALS loss, we refer the reader to Steinwart et al. (2011) and Farooq and Steinwart (2017) respectively.

We further recall Steinwart and Christmann (2008, Definition 2.18) that a loss function is called locally Lipschitz continuous if for all $M > 0$ there exists a constant c_M such that

$$\sup_{y \in Y} |L(y, t) - L(y, t')| \leq c_M |t - t'|, \quad t, t' \in [-M, M].$$

In the following we denote for a given $M > 0$ the smallest such constant c_M by $|L|_{1,M}$ and show that the ALS loss is locally Lipschitz continuous.

Lemma 2 *Let $Y \subseteq [-M, M]$ with $M > 0$ and $t \in Y$, then the loss function $L_\tau : Y \times \mathbb{R} \to [0, \infty)$ is locally Lipschitz continuous with Lipschitz constant*

$$|L_\tau|_{1,M} = C_\tau \, 4M,$$

where $C_\tau := \max\{\tau, 1 - \tau\}$.

For later use note that L_τ being locally Lipschitz continuous implies that L_τ is also a *Nemitski loss* in the sense of Definition 18 in Steinwart and Christmann (2008), and by Steinwart and Christmann (2008, Lemma 2.13 and 2.19), this further implies that the corresponding risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f)$ is convex and locally Lipschitz continuous.

Empirical methods of estimating expectile using L_τ loss typically lead to the function f_{D} for which $\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D}})$ is close to $\mathcal{R}^*_{L_\tau,\mathrm{P}}$ with high probability. The convexity of L_τ then ensures that f_{D} approximates $f^*_{L_\tau,\mathrm{P}}$ in a weak sense, namely in probability P_X (see Steinwart 2007, Remark 3.18). However, no guarantee on the speed of this convergence can be given, even if we know the convergence rate of $\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D}}) \to \mathcal{R}^*_{L_\tau,\mathrm{P}}$. The following theorem addresses this issue by establishing a so-called self-calibration inequality for the excess L_τ-risk.

Theorem 3 *Let L_τ be the ALS loss function defined by (1) and* P *be the distribution on $X \times Y$. Moreover, assume that $f^*_{L_\tau,\mathrm{P}}(x) < \infty$ is the conditional τ-expectile for fixed $\tau \in (0,1)$ and $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathrm{P}_X)$ for P_X-almost all $x \in X$. Then, for all measurable $f : X \to \mathbb{R}$, we have*

$$C_\tau^{-1}(\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}) \le \|f - f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)} \le c_\tau^{-1}(\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}), \tag{7}$$

where $c_\tau := \min\{\tau, 1-\tau\}$ and C_τ is defined in Lemma 2.

Note that the right-hand side of the inequality (7) in particular ensures that $f_{\mathrm{D}} \to f^*_{L_\tau,\mathrm{P}}$ in $L_2(\mathrm{P}_X)$ whenever $\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D}}) \to \mathcal{R}^*_{L_\tau,\mathrm{P}}$. In addition, the convergence rates can be directly translated. The inequality on the left of (7) shows that modulo constants the calibration inequality is sharp. We will use this left inequality in the proof of Theorem 6 in order to establish bound for the approximation error function for Gaussian RBF kernels

At the end of this section, we denote $L_\tau \circ f$ by a function $(x, y) \mapsto L_\tau(y, f(x))$ and present in the following supremum and variance bounds of L_τ-loss.

Lemma 4 *Let $X \subset \mathbb{R}^d$ be non-empty set, $Y \subseteq [-M, M]$ be a closed subset where $M > 0$, and* P *be a distribution on $X \times Y$. Additionally, we assume that $L_\tau : Y \times \mathbb{R} \to [0, \infty)$ is the ALS loss and $f^*_{L_\tau,\mathrm{P}}(x)$ is the conditional τ-expectile for fixed $\tau \in (0,1)$. Then for all $f : X \to [-M, M]$ we have*

(i) $\|L_\tau \circ f - L_\tau \circ f^*_{L_\tau,\mathrm{P}}\|_\infty \le 4\, C_\tau\, M^2$.
(ii) $\mathbb{E}_P(L_\tau \circ f - L_\tau \circ f^*_{L_\tau,\mathrm{P}})^2 \le 16\, C_\tau^2\, c_\tau^{-1}\, M^2(\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}})$.

Like the calibration inequality established in Theorem 3 these two bounds in Lemma 4 are also important for analyzing the statistical properties of any L_τ-based empirical risk minimization scheme.

3 Oracle inequalities and learning rates

In this section, we first introduce some notions related to kernels. We assume that $k : X \times X \to \mathbb{R}$ is a measurable, symmetric and positive definite kernel with associated RKHS H. Additionally, we assume that k is bounded, that is, $\|k\|_\infty := \sup_{x \in X} \sqrt{k(x,x)} \le 1$, which implies that H consists of bounded functions with $\|f\|_\infty \le \|k\|_\infty \|f\|_H$ for all $f \in H$. In practice, we often consider SVMs that are equipped with well-known Gaussian RBF kernels for input domain $X \subset \mathbb{R}^d$, (see Steinwart et al. 2011; Farooq and Steinwart 2017). Recall that the latter are defined by

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

where γ is called the width parameter that is usually determined in a data dependent way, i.e. by cross validation. By Steinwart and Christmann (2008, Corollary 4.58) the kernel k_γ is universal on every compact set $X \in \mathbb{R}^n$ and in particular strictly positive definite. In addition, the RKHS H_γ of kernel k_γ is dense in $L_p(\mu)$ for all $p \in [1, \infty)$ and all distributions μ on X, see Steinwart and Christmann (2008, Proposition 4.60).

One requirement for establishing learning rates is to control the capacity of RKHS H. One way to do this is to estimate eigenvalues of a linear operator induced by kernel k. Given a kernel k and a distribution μ on X, we define the integral operator $T_k : L_2(\mu) \to L_2(\mu)$ by

$$T_k f(\cdot) := \int_X k(x, \cdot) f(x) d\mu(x) \tag{8}$$

for μ-almost all $x \in X$. In the following, we assume that $\mu = \mathrm{P}_X$. Recall Steinwart and Christmann (2008, Theorem 4.27) that T_k is compact, positive, self-adjoint and nuclear, and thus has at most countably many non-zero (and non-negative) eigenvalues $\lambda_i(T_k)$. Ordering these eigenvalues (with geometric multiplicities) and extending the corresponding sequence by zeros, if there are only finitely many non-zero eigenvalues, we obtain the *extended sequence of eigenvalues* $(\lambda_i(T_k))_{i\geq 1}$ that satisfies $\sum_{i=1}^{\infty} \lambda_i(T_k) < \infty$ (see Steinwart and Christmann 2008, Theorem 7.29). This summability implies that for some constant $a > 1$ and $i \geq 1$, we have $\lambda_i(T_k) \leq ai^{-1}$. By Steinwart et al. (2009), this eigenvalues assumption can converge even faster to zero, that is, for $p \in (0, 1)$, we have

$$\lambda_i(T_k) \leq ai^{-\frac{1}{p}}, \quad i \geq 1. \tag{9}$$

It turns out that the speed of convergence of $\lambda_i(T_k)$ influences learning rates for SVMs. For instance, Blanchard et al. (2008) used (9) to establish learning rates for SVMs using hinge loss, and Caponnetto and De Vito (2007) and Mendelson and Neeman (2010) for SVMs using least square loss.

Another way to control the capacity of RKHS H is based on the concept of *covering numbers* or its dual called *entropy numbers*. To recall the latter, let $T : E \to F$ be a bounded, linear operator between the Banach spaces E and F, and $i \geq 1$ be an integer. Then the i-th (dyadic) entropy number of T is defined by

$$e_i(T) := \inf \left\{ \epsilon > 0 : \exists x_1, \ldots, x_{2^{i-1}} \text{ such that } T B_E \subset \cup_{j=1}^{2^{i-1}} (x_j + \epsilon B_F) \right\},$$

see Steinwart and Christmann (2008, Definition A.5.26). In the Hilbert space case, the eigenvalues and entropy number decay are closely related. For example, Steinwart (2009) showed that (9) is equivalent (modulo a constant only depending on p) to

$$e_i\left(\mathrm{id} : H \to L_2(\mathrm{P}_X)\right) \leq \sqrt{a} i^{-\frac{1}{2p}}, \quad i \geq 1, \tag{10}$$

It is further shown by Steinwart (2009) that (10) implies a bound on average entropy numbers, that is, for empirical distribution associated to the data set $D_X := (x_1, \cdots, x_n) \in X^n$, the average entropy number is

$$\mathbb{E}_{D_X \sim \mathrm{P}_X^n} e_i\left(\mathrm{id} : H \to L_2(\mathrm{P}_X)\right) \leq a i^{-\frac{1}{2p}}, \quad i \geq 1,$$

which is used in Steinwart and Christmann (2008, Theorem 7.24) to establish the general oracle inequality for SVMs. A bound of the form (10) was also established by Steinwart and Christmann (2008, Theorem 6.27) for Gaussian RBF kernels and certain distributions P_X having unbounded support. To be more precise, let $X \subset \mathbb{R}^d$ be a closed unit Euclidean ball. Then for all $\gamma \in (0, 1]$ and $p \in (0, 1)$, there exists a constant $c_{p,d}(X)$ such that

$$e_i\left(\mathrm{id} : H_\gamma(X) \to l_\infty(X)\right) \leq c_{p,d}(X) \gamma^{-\frac{d}{p}} i^{-\frac{1}{p}}, \tag{11}$$

which has been used by Eberts and Steinwart (2013) to establish leaning rates for least squares SVMs. Note that the constant $c_{p,d}(X)$ depends on p in an unknown manner. To address this issue, we use the result of van der Vaart and van Zanten (2009, Lemma 4.5) and derive an improved entropy number bound in the following theorem. As a result we obtain an upper bound for $c_{p,d}(X)$ whose dependence on p is explicitly known. We will further see in Corollary 8 that this improved bound is one factor to achieve better learning rates than the one obtained by Eberts and Steinwart (2013).

Theorem 5 *Let $X \subseteq \mathbb{R}^d$ be a closed Euclidean ball. Then there exists a constant $K > 0$, such that, for all $p \in (0,1)$, $\gamma \in (0,1]$ and $i \geq 1$, we have*

$$e_i\left(\mathrm{id} : H_\gamma(X) \to l_\infty(X)\right) \leq (3K)^{\frac{1}{p}} \left(\frac{d+1}{ep}\right)^{\frac{d+1}{p}} \gamma^{-\frac{d}{p}} i^{-\frac{1}{p}}. \tag{12}$$

Another requirement for establishing learning rates is to bound the *approximation error function* considering Gaussian RKHS H_γ. If the distribution P is such that $\mathcal{R}^*_{L_\tau,\mathrm{P}} < \infty$, then the approximation error function $\mathcal{A}_\gamma : [0,\infty) \to [0,\infty)$ is defined by

$$\mathcal{A}_\gamma(\lambda) := \inf_{f \in H_\gamma} \lambda \|f\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}. \tag{13}$$

For $\lambda > 0$, the approximation error function $\mathcal{A}_\gamma(\lambda)$ quantifies how well an infinite sample L_2-SVM with RKHS H_γ, that is, $\lambda\|f\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(f)$ approximates the optimal risk $\mathcal{R}^*_{L_\tau,\mathrm{P}}$. By Steinwart and Christmann (2008, Lemma 5.15), one can show that $\lim_{\lambda\to 0} \mathcal{A}_\gamma(\lambda) = 0$ since H_γ is dense in $L_2(P_X)$. In general, however, the speed of convergence can not be faster than $O(\lambda)$ and this rate is achieved, if and only if, there exists an $f \in H_\gamma$ such that $\mathcal{R}_{L_\tau,\mathrm{P}}(f) = \mathcal{R}^*_{L_\tau,\mathrm{P}}$ (see Steinwart and Christmann 2008, Lemma 5.18).

In order to bound $\mathcal{A}_\gamma(\lambda)$, we first need to know one important feature of the target function $f^*_{L_\tau,\mathrm{P}}$, namely, the *regularity* which, roughly speaking, measures the smoothness of the target function. Different function spaces norms e.g. Hölder norms, Besov norms or Triebel-Lizorkin norms can be used to capture this regularity. In this work, following Eberts and Steinwart (2013), see also Meister and Steinwart (2016), we assume that the target function $f^*_{L_\tau,\mathrm{P}}$ is in a Sobolev or a Besov space. Recall Tartar (2007, Definition 5.1) and Adams and Fournier (2003, Definitions 3.1 and 3.2) that for any integer $k \geq 0$, $1 \leq p \leq \infty$ and a subset $\Omega \subset \mathbb{R}^d$ with non-empty interior, the Sobolev space $W^k_p(\Omega)$ of order k is defined by

$$W^k_p(\Omega) := \left\{ f \in L_p(\Omega) : D^{(\alpha)} f \in L_p(\Omega) \text{ exists for all } \alpha \in \mathbb{N}^d_0 \text{ with } |\alpha| \leq k \right\},$$

with the norm

$$\|f\|_{W^k_p(\Omega)} := \begin{cases} \left(\sum_{|\alpha|\leq k} \|D^{(\alpha)} f\|^p_{L_p(\Omega)} \right)^{\frac{1}{p}}, & \text{if} \quad p \in [1,\infty), \\ \max \sum_{|\alpha|\leq k} \|D^{(\alpha)} f\|_{L_\infty(\Omega)}, & \text{if} \quad p = \infty, \end{cases}$$

where $D^{(\alpha)}$ is the α-th weak partial derivative for multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d_0$ of modulus $|\alpha| = |\alpha_1| + \cdots + |\alpha_d|$. In other words, the Sobolev space is the space of functions with sufficiently many derivatives and equipped with a norm that measures both the size and the regularity of the contained functions. Note that $W^k_p(\Omega)$ is a Banach space (Tartar 2007, Lemma 5.2). Moreover, by Adams and Fournier (2003, Theorem 3.6), $W^k_p(\Omega)$ is separable if $p \in [1,\infty)$, and is uniformly convex and reflexive if $p \in (1,\infty)$. Furthermore, for $p = 2$, $W^k_2(\Omega)$ is a separable Hilbert space that we denote by $H_k(\Omega)$. Despite the aforementioned advantages, Sobolev spaces can not be immediately applied when α is non-integral or when $p < 1$, however, the smoothness spaces for these extended parameters are also needed when engaging nonlinear approximation. This shortcoming of Sobolev spaces is covered by Besov spaces that bring together all functions for which the modulus of smoothness have a common behavior. Let us first recall DeVore and Sharpley (1993, Section 2) and DeVore and Popov (1988, Section 2) that for a subset $\Omega \subset \mathbb{R}^d$ with non-empty interior, a function $f : \Omega \to \mathbb{R}$

with $f \in L_p(\Omega)$ for all $p \in (0, \infty]$ and $s \in \mathbb{N}$, the modulus of smoothness of order s of a function f is defined by

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

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

$$\triangle_h^s(f,x,\Omega) := \begin{cases} \sum_{i=0}^{s} \binom{r}{i}(-1)^{r-i} f(x+ih) & \text{if } \ x, x+h, \ldots, x+sh \in \Omega, \\ 0, & \text{otherwise,} \end{cases}$$

for $h \in \mathbb{R}^d$, is used to measure the smoothness. Note that $w_{s,L_p(\Omega)}(f,t) \to 0$ as $t \to 0$, which means that the faster this convergence to 0 the smoother is the function f. For more details on properties of the modulus of smoothness, we refer the reader to Nikol'skii (2012, Chapter 4.2). Now for $0 < p, q \le \infty, \alpha > 0, s := \lfloor \alpha \rfloor + 1$, the Besov space $B^\alpha_{p,q}(\Omega)$ based on modulus of smoothness for domain $\Omega \subset \mathbb{R}^d$, see for instance DeVore (1998, Section 4.5), Nikol'skii (2012, Chapter 4.3) and DeVore and Sharpley (1993, Section 2), is defined by

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

where the semi-norm $|\cdot|_{B^\alpha_{p,q}(\Omega)}$ is given by

$$|f|_{B^\alpha_{p,q}(\Omega)} := \left(\int_0^\infty (t^{-\alpha} w_{s,L_p(\Omega)}(f,t))^q \frac{dt}{t} \right)^{\frac{1}{q}}, \qquad q \in (0,\infty),$$

and for $q = \infty$, the semi-norm $|\cdot|_{B^\alpha_{p,q}(\Omega)}$ is defined by

$$|f|_{B^\alpha_{p,q}(\Omega)} := \sup_{t>0}(t^{-\alpha} w_{s,L_p(\Omega)}(f,t)).$$

In other words, Besov spaces are collections of functions f with common smoothness. For more general definition of Besov-like spaces, we refer to Meister and Steinwart (2016, Section 4.1). Note that $\|f\|_{B^\alpha_{p,q}(\Omega)} := \|f\|_{L_p(\Omega)} + |f|_{B^\alpha_{p,q}(\Omega)}$ is the norm of $B^\alpha_{p,q}(\Omega)$, see e.g. DeVore and Sharpley (1993, Section 2) and DeVore and Popov (1988, Section 2). It is well known (see e.g. Nikol'skii 2012, Section 4.1) that $W^s_p(\Omega) \subset B^s_{p,\infty}(\Omega)$ for all $1 \le p \le \infty$, $p \ne 2$, where for $p = q = 2$ the Besov space is the same as the Sobolev space.

In the next step, we find a function $f_0 \in H_\gamma$ such that both the regularization term $\lambda \|f_0\|^2_{H_\gamma}$ and the excess risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$ are small. For this, we define the function $K_\gamma : \mathbb{R}^d \to \mathbb{R}$ (see Eberts and Steinwart 2013) by

$$K_\gamma(x) := \sum_{j=1}^{r} \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2 \pi} \right)^{\frac{d}{2}} \exp\left(-\frac{2\|x\|_2^2}{j^2 \gamma^2} \right), \tag{14}$$

for all $r \in \mathbb{N}$, $\gamma > 0$ and $x \in \mathbb{R}^d$. Additionally, we assume that there exists a function $f^*_{L_\tau,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ that satisfies $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $\mathcal{R}_{L_\tau,\mathrm{P}}(f^*_{L_\tau,\mathrm{P}}) = \mathcal{R}^*_{L_\tau,\mathrm{P}}$. Then f_0 is defined by

$$f_0(x) := K_\gamma * f^*_{L_\tau,P}(x) := \int_{\mathbb{R}} K_\gamma(x-t) f^*_{L_\tau,\mathrm{P}}(t) dt, \quad x \in \mathbb{R}.$$

With these preparation, we now establish an upper bound for the approximate error function $\mathcal{A}_\gamma(\lambda)$.

Theorem 6 *Let L_τ be the ALS loss defined by (1), P be the probability distribution on $\mathbb{R}^d \times Y$, and P_X be the marginal distribution of P on $\mathbb{R}^d$ such that $X := \operatorname{supp} \mathrm{P}_X$ satisfies $\mathrm{P}_X(\partial X) = 0$. Moreover, assume that the conditional τ-expectile $f^*_{L_\tau,\mathrm{P}}$ satisfies $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ as well as $f^*_{L_\tau,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_X)$ for some $\alpha \geq 1$. In addition, assume that k_γ is the Gaussian RBF kernel over X with associated RKHS H_γ. Then for all $\gamma \in (0, 1]$ and $\lambda > 0$, we have*

$$\|f_0\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C_1 \lambda \gamma^{-d} + C_{\tau,s} \gamma^{2\alpha},$$

where the constant $C_1 > 0$ and the constant $C_{\tau,s} > 0$ depends on s and τ.

Clearly, the upper bound of the approximation error function in Theorem 6 depends on the regularization parameter λ, the kernel width γ, and the smoothness parameter α of the target function $f^*_{L_\tau,\mathrm{P}}$. Note that in order to shrink the right-hand side we need to let $\gamma \to 0$. However, this would let the first term go to infinity unless we simultaneously let $\lambda \to 0$ with a sufficient speed. Now using Theorem 7.24 in Steinwart and Christmann (2008) together with Lemma 4, Theorem 6 and the entropy number bound (12), we establish an oracle inequality of SVMs for L_τ in the following theorem.

Theorem 7 *Consider the assumptions of Theorem 6 and additionally assume that $Y \subseteq [-M, M]$ for $M \geq 1$. Then, for all $n \geq 1, \varrho \geq 1, \gamma \in (0, 1)$ and $\lambda \in (0, e^{-2}]$, the SVM using the RKHS H_γ and the ALS loss satisfies*

$$\begin{aligned} &\lambda \|f_{\mathrm{D},\lambda,\gamma}\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\ &\quad \leq C M^2 \left(\lambda \gamma^{-d} + \gamma^{2\alpha} + (\log \lambda^{-1})^{d+1} n^{-1} \gamma^{-d} + n^{-1} \varrho \right), \end{aligned} \tag{15}$$

with probability P^n not less than $1 - 3e^{-\varrho}$. Here $C > 0$ is some constant independent of λ, γ, n and ϱ.

It is well known that there exists a relationship between Sobolev spaces and the scale of Besov spaces, that is, $B^\alpha_{p,u}(\mathbb{R}^d) \hookrightarrow W^\alpha_p(\mathbb{R}^d) \hookrightarrow B^\alpha_{p,v}(\mathbb{R}^d)$, whenever $1 \leq u \leq \min\{p, 2\}$ and $\max\{p, 2\} \leq v \leq \infty$ (see e.g. Edmunds and Triebel 2008, pp. 25 and 44). In particular, for $p = u = v = 2$, we have $W^\alpha_2(\mathbb{R}^d) = B^\alpha_{2,2}(\mathbb{R}^d)$ with equivalent norms. In addition, by Eberts and Steinwart (2013, p. 7) we have $B^\alpha_{p,q}(\mathbb{R}^d) \subset B^\alpha_{p,q}(\mathrm{P}_X)$. Thus, Theorem 7 also holds for decision functions $f^*_{L_\tau,\mathrm{P}} : \mathbb{R}^d \to \mathbb{R}$ with $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $f^*_{L_\tau,\mathrm{P}} \in W^\alpha_2(\mathbb{R}^d)$.

By assuming some suitable values for λ and γ that depends on data size n, the smoothness parameter α, and the dimension d, we obtain learning rates for learning problem (4) in the following corollary.

Corollary 8 *Under the assumptions of Theorem 7 and with*

$$\begin{aligned} \lambda_n &= (\log n)^{\delta_1} n^{-1}, \\ \gamma_n &= (\log n)^{\delta_2} n^{-\frac{1}{2\alpha+d}}, \end{aligned}$$

where $\delta_1 := d + 1$ and $\delta_2 := \frac{d+1}{2\alpha+d}$, we have, for all $n \geq 3$ and $\varrho \geq 1$,

$$\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq 4 C M^3 \varrho (\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}} n^{-\frac{2\alpha}{2\alpha+d}} \tag{16}$$

with probability P^n not less than $1 - 3e^{-\varrho}$.

Note that learning rates in Corollary 8 depend on the choice of λ_n and γ_n, where the kernel width γ_n requires knowing α which, in practice, is not available. However, Steinwart and Christmann (2008, Chapter 7.4), Steinwart et al. (2009) and Eberts and Steinwart (2013) showed that one can achieve the same learning rates adaptively, i.e. without knowing α. Let us recall Definition 6.28 in Steinwart and Christmann (2008) that describes a method to select λ and γ, which in some sense is a simplification of the cross-validation method.

Definition 9 Let H_γ be a RKHS over X and $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be the sequences of finite subsets $\Lambda_n, \Gamma_n \subset (0, 1]$. We define for a data set $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times \mathbb{R})^n$

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

where $m = \lfloor \frac{n}{2} \rfloor + 1$ and $n \geq 4$. Then use D_1 as a training set to compute the SVM decision function

$$f_{\mathrm{D}_1,\lambda,\gamma} := \arg\min_{f \in H_\gamma} \lambda \|f\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{D}_1}(f), \quad (\lambda, \gamma) \in (\Lambda_n, \Gamma_n),$$

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

$$\mathcal{R}_{L_\tau,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}) = \min_{(\lambda,\gamma)\in(\Lambda_n,\Gamma_n)} \mathcal{R}_{L_\tau,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}).$$

Every learning method that produce the resulting decision functions $\widehat{f}_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}$ is called a training validation SVM with respect to (Λ, Γ).

In the next Theorem, we use this training-validation SVM (TV-SVM) approach for suitable candidate sets $\Lambda_n := (\lambda_1, \ldots, \lambda_r)$ and $\Gamma_n := (\gamma_1, \ldots, \gamma_s)$ with $\lambda_r = \gamma_s = 1$, and establish following learning rates similar to (16).

Theorem 10 *With the assumptions of Theorem 7, let $\Lambda := (\Lambda_n)$ be a sequence of finite subset $\Lambda_n \in (0, e^{-2}]$ such that $(\log n)^{-(d+1)} n^{-1} \leq \lambda_i \leq (\log n)^{d+1} n^{-1}$ for all $n \geq 3$, and $\Gamma := (\Gamma_n)$ be the sequences of finite subsets $\Gamma_n \subset (0, 1]$ such that Λ_n is an δ_n-net of $(0, 1]$ where $\delta_n > 0$. In addition, we assume that the cardinalities $|\Lambda_n|$ and $|\Gamma_n|$ are polynomially growing in n. Then for all $\varrho \geq 1$, the TV-SVM produces $f_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}$ that satisfies*

$$\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C M^3 \varrho (\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}} n^{-\frac{2\alpha}{2\alpha+d}}$$

with probability P^n not less than $1 - 3e^{-\varrho}$, where $C > 0$ is a constant independent of n and ϱ.

So far we have only considered the case of bounded noise with known bounds, that is, $Y \subseteq [-M, M]$ where $M > 0$. In practice, M is usually unknown and in this situation, one can still achieve the same learning rates by simply increasing M slowly. However, more interesting is the case of unbounded noise. In the following we treat this case for distributions for which there exist constants $c \geq 1$ and $l > 0$ such that

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

for all $\varrho > 1$. In other words, the tails of the response variable Y decay sufficiently fast. Different examples are given by Eberts and Steinwart (2013) to show that such an assumption is realistic. For instance, if $\mathrm{P}(.|x) \sim N(\mu(x), 1)$, the assumption (17) is satisfied for $l = \frac{1}{2}$,

and for the case where $\mathrm{P}(.|x)$ has the density whose tails decay like $e^{-|t|}$, the assumption (17) holds for $l = 1$ (see Eberts and Steinwart 2013, Examples 3.7 and 3.8).

With this additional assumption, we present learning rates for the case of unbounded noise in the following theorem.

Theorem 11 *Let $Y \subset \mathbb{R}$ and P be a probability distribution on $\mathbb{R}^d \times Y$ such that $X := \operatorname{supp} \mathrm{P}_X \subset B_{l_2^d}$. Moreover, assume that the τ-expectile $f^*_{L_\tau,\mathrm{P}}$ satisfies $f^*_{L_\tau,\mathrm{P}}(x) \in [-1, 1]$ for P_X-almost all $x \in X$, and both $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ and $f^*_{L_\tau,\mathrm{P}} \in B^\alpha_{2,\infty}(\mathrm{P}_X)$ for some $\alpha \geq 1$. In addition, assume that (17) holds for all $\varrho \geq 1$. We define*

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

where $c_1 > 0$ and $c_2 > 0$ are user-specified constants. Moreover, for some fixed $\tilde{\varrho} \geq 1$ and $n \geq 3$ we define $\varrho := \tilde{\varrho} + \ln n$ and $M_n := 2c\varrho^l$. Furthermore, we consider the SVM that clips decision function $f_{\mathrm{D},\lambda_n,\gamma_n}$ at M_n after training. Then there exists a $C > 0$ independent of n, p and $\tilde{\varrho}$ such that

$$\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda_n,\gamma_n}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C\tilde{\varrho}^{3l+1} (\log n)^{3l+\frac{2\alpha(d+1)}{2\alpha+d}} n^{-\frac{2\alpha}{2\alpha+d}} \tag{18}$$

holds with probability P^n not less than $1 - 2e^{-\tilde{\varrho}}$.

Note that the assumption (17) on the tail of the distribution does not influence learning rates achieved in the Corollary 8. Furthermore, we can also achieve the same rates adaptively using TV-SVM approach considered in Theorem 10 provided that we have an upper bound of the unknown parameter l, which depends on the distribution P.

Let us now compare our results with the oracle inequalities and learning rates established by Eberts and Steinwart (2013) for least square SVMs. This comparison is justifiable because a) the least square loss is a special case of L_τ-loss for $\tau = 0.5$, b) the target function $f^*_{L_\tau,P}$ is assumed to be in the Sobolev or Besov space similar to Eberts and Steinwart (2013), and c) the supremum and the variance bounds for L_τ with $\tau = 0.5$ are the same as the ones used by Eberts and Steinwart (2013). Furthermore, recall that Eberts and Steinwart (2013) used the entropy number bounds (11) to control the capacity of the RKHS H_γ which contains a constant $c_{p,d}(X)$ depending on p in an unknown manner. As a result, they obtained a leading constant C in their oracle inequality, see Eberts and Steinwart (2013, Theorem 3.1) for which no upper bound can be determined explicitly. We cope this problem by establishing an improved entropy number bound (12) which not only provides the upper bound for $c_{p,d}(X)$ but also helps to determine the value of the constant C in the oracle inequality (15) explicitly. As a consequence we can improve their learning rates of the form $n^{-\frac{2\alpha}{2\alpha+d}+\xi}$, where $\xi > 0$, by

$$(\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}} n^{-\frac{2\alpha}{2\alpha+d}}. \tag{19}$$

In other words, the nuisance parameter n^ξ of learning rates from Eberts and Steinwart (2013) is replaced by the logarithmic term $(\log n)^{d+1}$. Moreover, our learning rates, up to this logarithmic term, are minimax optimal, see e.g. the discussion in Eberts and Steinwart (2013). Finally note that unlike Eberts and Steinwart (2013) we have not only established learning rates for the least squares case for which $\tau = 0.5$ but actually for all $\tau \in (0, 1)$.

4 Proofs

4.1 Proofs of Section 2

Proof of Lemma 2 We define $\psi : \mathbb{R} \to \mathbb{R}$ by

$$\psi(r) := \begin{cases} (1-\tau)r^2, & \text{if } r < 0, \\ \tau r^2, & \text{if } r \geqslant 0. \end{cases}$$

Clearly, ψ is convex and thus by Lemma A.6.5 in Steinwart and Christmann (2008) ψ is locally Lipschitz continuous. Moreover, for $y \in [-M, M]$ (see Steinwart and Christmann 2008, Lemma A.6.8) we obtain

$$\begin{aligned} |L_\tau|_{1,M} &= \sup_{y\in[-M,M]} |\psi(y-\cdot)|_{1,M} \\ &= \sup_{y\in[-M,M]} \sup_{t\in[-M,M]} |\psi'(y-t)|_{1,M} \\ &= \max\{\tau, 1-\tau\} \sup_{y\in[-M,M]} \sup_{t\in[M,-M]} |2(y-t)| \\ &= C_\tau\, 4M, \end{aligned}$$

where $C_\tau := \max\{\tau, 1-\tau\}$. A simple consideration shows that this estimate is also sharp. □

In order to prove Theorem 3 recall that the risk $\mathcal{R}_{L_\tau,\mathrm{P}}(f)$ in (2) uses regular conditional probability $\mathrm{P}(Y|x)$, which enable us to compute $\mathcal{R}_{L_\tau,\mathrm{P}}(f)$ by treating the *inner* and the *outer* integrals separately. Following Steinwart and Christmann (2008, Definitions 3.3, 3.4), we therefore use *inner* L_τ*-risks* as a key ingredient for establishing self-calibration inequalities.

Definition 12 Let $L_\tau : Y \times \mathbb{R} \to [0, \infty)$ be the ALS loss function defined by (1) and Q be a distribution on $Y \subseteq [-M, M]$. Then the *inner* L_τ*-risks* of Q are defined by

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

and the *minimal inner* L_τ*-risk* is

$$\mathcal{C}^*_{L_\tau,Q} := \inf_{t\in\mathbb{R}} \mathcal{C}_{L_\tau,Q}(t).$$

In the latter definition, the *inner risks* $\mathcal{C}_{L_\tau,Q}(\cdot)$ for a suitable classes of distributions Q on Y are considered as a template for $\mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}(\cdot)$. From this, we immediately can obtain the risk of function f, i.e.

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) = \int_X \mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}(f(x))\, d\mathrm{P}_X(x).$$

Moreover, by Steinwart and Christmann (2008, Lemma 3.4), the optimal risk $\mathcal{R}^*_{L_\tau,\mathrm{P}}$ can be obtained by minimal *inner* L_τ*-risks*, that is,

$$\mathcal{R}^*_{L_\tau,\mathrm{P}} = \int_X \mathcal{C}^*_{L_\tau,\mathrm{P}(\cdot|x)}\, d\mathrm{P}_X(x).$$

Consequently, the *excess* L_τ*-risk* when $\mathcal{R}^*_{L_\tau,\mathrm{P}} < \infty$ is obtained by

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}} = \int_X \mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}(f(x)) - \mathcal{C}^*_{L_\tau,\mathrm{P}(\cdot|x)}\, d\mathrm{P}_X(x). \tag{20}$$

Besides some technical advantages, this approach makes the analysis rather independent of the specific distribution P. In the following theorem, we use this approach and establish the lower and the upper bound of excess inner L_τ-risks.

Theorem 13 *Let L_τ be the ALS loss function defined by (1) and Q be a distribution on $\mathbb{R}$ with $|Q|_1 < \infty$ and $\mathcal{C}^*_{L_\tau,Q} < \infty$ holds. Then for all $t \in \mathbb{R}$ and all $\tau \in (0,1)$ we have*

$$c_\tau (t - t^*)^2 \le \mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}^*_{L_\tau,Q} \le C_\tau (t - t^*)^2, \tag{21}$$

where $c_\tau := \min\{\tau, 1-\tau\}$ and C_τ is defined in Lemma 2.

Proof Let us fix $\tau \in (0,1)$. Since the distribution Q on $\mathbb{R}$ has finite first moment, that is, $|Q|_1 < \infty$, we obtain following Newey and Powell (1987) the τ-expectile t^* as a unique solution of

$$\tau \int_{y \ge t^*} (y - t^*) dQ(y) = (1-\tau) \int_{y < t^*} (t^* - y) dQ(y). \tag{22}$$

For establishing bound for excess inner risks of L_τ with respect to Q, we fix a $t \ge t^*$. Then we have

$$\begin{aligned}
&\int_{y<t} (y-t)^2 dQ(y) \\
&\quad = \int_{y<t} (y - t^* + t^* - t)^2 dQ(y) \\
&\quad = \int_{y<t} (y-t^*)^2 dQ(y) + 2(t^* - t) \int_{y<t} (y - t^*) dQ(y) + (t^* - t)^2 Q((-\infty, t)) \\
&\quad = \int_{y<t^*} (y-t^*)^2 dQ(y) + \int_{t^* \le y < t} (y-t^*)^2 dQ(y) + (t^* - t)^2 Q((-\infty, t)) \\
&\qquad + 2(t^* - t) \int_{y<t^*} (y - t^*) dQ(y) + 2(t^* - t) \int_{t^* \le y < t} (y - t^*) dQ(y),
\end{aligned}$$

and

$$\begin{aligned}
&\int_{y \ge t} (y-t)^2 dQ(y) \\
&\quad = \int_{y \ge t^*} (y-t^*)^2 dQ(y) - \int_{t^* \le y < t} (y-t^*)^2 dQ(y) + (t^* - t)^2 Q([t, \infty)) \\
&\qquad + 2(t^* - t) \int_{y \ge t^*} (y - t^*) dQ(y) - 2(t^* - t) \int_{t^* \le y < t} (y - t^*) dQ(y).
\end{aligned}$$

By Definition 12 and using (22), we obtain

$$
\begin{aligned}
&\mathcal{C}_{L_\tau,Q}(t)\\
&\quad=(1-\tau)\int_{y<t}(y-t)^2dQ(y)+\tau\int_{y\geq t}(y-t)^2dQ(y)\\
&\quad=\tau\int_{y<t^*}(y-t^*)^2dQ(y)+(1-\tau)\int_{y\geq t^*}(y-t^*)^2dQ(y)\\
&\qquad+2(t^*-t)\left(\tau\int_{y<t^*}(y-t^*)dQ(y)+(1-\tau)\int_{y\geq t^*}(y-t^*)dQ(y)\right)\\
&\qquad+(t^*-t)^2(1-\tau)Q((-\infty,t))+(t^*-t)^2\tau Q([t,\infty))\\
&\qquad+(1-2\tau)\int_{t^*\leq y<t}(y-t^*)^2dQ(y)+2(1-2\tau)\int_{t^*\leq y<t}(y-t^*)dQ(y)\\
&\quad=\mathcal{C}_{L_\tau,Q}(t^*)+(t^*-t)^2(1-\tau)Q((-\infty,t))+(t^*-t)^2\tau Q([t,\infty))\\
&\qquad+(1-2\tau)\int_{t^*\leq y<t}(y-t^*)^2+2(t^*-t)(y-t^*)dQ(y),
\end{aligned}
$$

and this leads to the following excess inner L_τ-risk

$$
\begin{aligned}
&\mathcal{C}_{L_\tau,Q}(t)-\mathcal{C}_{L_\tau,Q}(t^*)\\
&\quad=(t^*-t)^2(1-\tau)Q((-\infty,t^*))+(t^*-t)^2(1-\tau)Q([t^*,t))+(t^*-t)^2\tau Q([t,\infty))\\
&\qquad+(1-2\tau)\int_{t^*\leq y<t}(y-t^*)^2+2(t^*-t)(y-t^*)dQ(y)\\
&\quad=(t^*-t)^2\left((1-\tau)Q((-\infty,t^*))+\tau Q([t,\infty))\right)\\
&\qquad-\tau\int_{t^*\leq y<t}(y-t^*)^2+2(t^*-t)(y-t^*)dQ(y)\\
&\qquad+(t^*-t)^2(1-\tau)Q([t^*,t))+(1-\tau)\int_{t^*\leq y<t}(y-t^*)^2+2(t^*-t)(y-t^*)dQ(y)\\
&\quad=(t^*-t)^2\left((1-\tau)Q((-\infty,t^*))+\tau Q([t,\infty))\right)\\
&\qquad-\tau\int_{t^*\leq y<t}(y-t^*)(y+t^*-2t)dQ(y)\\
&\qquad+(1-\tau)\int_{t^*\leq y<t}(y-t^*)^2+2(t^*-t)(y-t^*)+(t^*-t)^2dQ(y)\\
&\quad=(t^*-t)^2\left((1-\tau)Q((-\infty,t^*))+\tau Q([t,\infty))\right)\\
&\qquad+\tau\int_{t^*\leq y<t}(y-t^*)(2t-t^*-y)dQ(y)\\
&\qquad+(1-\tau)\int_{t^*\leq y<t}(y-t)^2dQ(y).
\end{aligned}
\tag{23}
$$

Let us define $c_\tau := \min\{\tau, 1-\tau\}$, then (23) leads to the following lower bound of excess inner L_τ-risk when $t \geq t^*$:

$$\begin{aligned}
&\mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}_{L_\tau,Q}(t^*) \\
&\quad \geq c_\tau (t^* - t)^2 \left(Q((-\infty, t^*)) + Q([t, \infty))\right) \\
&\qquad + c_\tau \int_{t^* \leq y < t} (y - t^*)(2t - t^* - y) + (y - t)^2 dQ(y) \\
&\quad = c_\tau (t^* - t)^2 \left(Q((-\infty, t^*)) + Q([t, \infty))\right) + c_\tau \int_{t^* \leq y < t} (t^*)^2 + 2tt^* + t^2 dQ(y) \\
&\quad = c_\tau (t^* - t)^2 \left(Q((-\infty, t^*)) + Q([t, \infty))\right) + c_\tau (t^* - t)^2 Q([t^*, t)) \\
&\quad = c_\tau (t^* - t)^2. \qquad (24)
\end{aligned}$$

Likewise, the excess inner L_τ-risk when $t < t^*$ is

$$\begin{aligned}
&\mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}_{L_\tau,Q}(t^*) \\
&\quad = (t^* - t)^2 \left((1-\tau) Q((-\infty, t) + \tau) Q([t^*, \infty))\right) + \tau \int_{t \leq y < t^*} (y - t)^2 dQ(y) \\
&\qquad + (1-\tau) \int_{t \leq y < t^*} (t^* - y)(y + t^* - 2t) dQ(y), \qquad (25)
\end{aligned}$$

that also leads to the lower bound (24). Now, for the proof of upper bound of the excess inner L_τ-risks, we define $C_\tau := \max\{\tau, 1-\tau\}$. Then (23) leads to the following upper bound of excess inner L_τ-risks when $t \geq t^*$:

$$\begin{aligned}
\mathcal{C}_{L_\tau,Q}(t) - \mathcal{C}_{L_\tau,Q}(t^*) &\leq C_\tau (t^* - t)^2 \left(Q((-\infty, t^*)) + Q([t, \infty))\right) \\
&\quad + C_\tau \int_{t^* \leq y < t} \left((y - t^*)(2t - t^* - y) + (y - t)^2\right) dQ(y) \\
&= C_\tau (t^* - t)^2. \qquad (26)
\end{aligned}$$

Analogously, for the case of $t < t^*$, (25) also leads to the upper bound (26) for excess inner L_τ-risks. □

Proof of Theorem 3 For a fixed $x \in X$, we write $t := f(x)$ and $t^* := f^*_{L_\tau,\mathrm{P}}(x)$. By Theorem 13, for $Q := \mathrm{P}(\cdot|x)$, we then immediately obtain

$$\begin{aligned}
&C_\tau^{-1} (\mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}(f(x)) - \mathcal{C}^*_{L_\tau,\mathrm{P}(\cdot|x)}) \\
&\quad \leq |f(x) - f^*_{L_\tau,\mathrm{P}}(x)|^2 \leq c_\tau^{-1} (\mathcal{C}_{L_\tau,\mathrm{P}(\cdot|x)}(f(x)) - \mathcal{C}^*_{L_\tau,\mathrm{P}(\cdot|x)}).
\end{aligned}$$

Integrating with respect to P_X leads to the assertion. □

Proof of Lemma 4 (i) Since L_τ can be clipped at M and the conditional τ-expectile satisfies $f^*_{L_\tau,\mathrm{P}}(x) \in [-M, M]$ almost surely. Then

$$\begin{aligned}
\|L_\tau(y, f(x)) - L_\tau(y, f^*_{L_\tau,\mathrm{P}}(x))\|_\infty &\leq \max\{\tau, 1-\tau\} \sup_{y,t \in [-M,M]} (y - t)^2 \\
&= C_\tau \, 4M^2,
\end{aligned}$$

for all $f : X \to [-M, M]$ and all $(x, y) \in X \times Y$.

(ii) Using the locally Lipschitz continuity of the loss L_τ and Theorem 3, we obtain

$$\begin{aligned}\mathbb{E}_P(L_\tau \circ f - L_\tau \circ f^*_{L_\tau,P})^2 &\le |L_\tau|^2_{1,M}\, \mathbb{E}_{P_X}|f - f^*_{L_\tau,P}|^2 \\ &\le 16 c_\tau^{-1} C_\tau^2 M^2 \left(\mathcal{R}_{L_\tau,P}(f) - \mathcal{R}^*_{L_\tau,P}\right).\end{aligned}$$

□

4.2 Proofs of Section 3

Proof of Theorem 5 By van der Vaart and van Zanten (2009, Lemma 4.5), the $\|\cdot\|_\infty$-log covering numbers of unit ball $B_\gamma(X)$ of the Gaussian RKHS $H_\gamma(X)$ for all $\gamma \in (0, 1)$ and $\varepsilon \in (0, \frac{1}{2})$ satisfy

$$\ln \mathcal{N}(B_\gamma(X), \|\cdot\|_\infty, \varepsilon) \le K \left(\log \frac{1}{\varepsilon}\right)^{d+1} \gamma^{-d}, \tag{27}$$

where $K > 0$ is a constant depending only on d. From this, we obtain

$$\sup_{\varepsilon\in(0,\frac{1}{2})} \varepsilon^p \ln \mathcal{N}(B_\gamma(X), \|\cdot\|_\infty, \varepsilon) \le K\gamma^{-d} \sup_{\varepsilon\in(0,\frac{1}{2})} \varepsilon^p \left(\log \frac{1}{\varepsilon}\right)^{d+1}.$$

Let $h(\varepsilon) := \varepsilon^p \left(\log \frac{1}{\varepsilon}\right)^{d+1}$. In order to obtain the optimal value of $h(\varepsilon)$, we differentiate it with respect to ε

$$\frac{dh(\varepsilon)}{d\varepsilon} = p\varepsilon^{p-1}\left(\log \frac{1}{\varepsilon}\right)^{d+1} - \varepsilon^p(d+1)\left(\log \frac{1}{\varepsilon}\right)^d \frac{1}{\varepsilon},$$

and set $\frac{dh(\varepsilon)}{d\varepsilon} = 0$ which gives $\log \frac{1}{\varepsilon} = \frac{d+1}{p}$, and hence

$$\varepsilon^* = \frac{1}{e^{\frac{d+1}{p}}}.$$

By plugging ε^* into $h(\varepsilon)$, we obtain

$$h(\varepsilon^*) = \left(\frac{d+1}{ep}\right)^{d+1},$$

and consequently, $\|\cdot\|_\infty$-log covering numbers (27) are

$$\ln \mathcal{N}(B_\gamma(X), \|\cdot\|_\infty, \varepsilon) \le K \left(\frac{d+1}{ep}\right)^{d+1} \gamma^{-d}\varepsilon^{-p} = \left(\frac{a^{\frac{1}{p}}}{\varepsilon}\right)^p,$$

where $a := K\left(\frac{d+1}{ep}\right)^{d+1}\gamma^{-d}$. Now, by inverse implication of Lemma 6.21 in Steinwart and Christmann (2008), see also Steinwart and Christmann (2008, Exercise 6.8), the bound on entropy number of the Gaussian RBF kernel is

$$e_i\left(\mathrm{id} : \mathcal{H}_\gamma(X) \to l_\infty(X)\right) \le (3a)^{\frac{1}{p}} i^{-\frac{1}{p}} = (3K)^{\frac{1}{p}} \left(\frac{d+1}{ep}\right)^{\frac{d+1}{p}} \gamma^{-\frac{d}{p}} i^{-\frac{1}{p}},$$

for all $i \ge 1$, $\gamma \in (0, 1)$. □

Proof of Theorem 6 The assumption $f^*_{L_\tau,\mathrm{P}} \in L_2(\mathbb{R}^d)$ and Theorem 2.3 in Eberts and Steinwart (2013) immediately yield that $f_0 := K_\gamma * f^*_{L_\tau,\mathrm{P}} \in H_\gamma$, i.e. f_0 is contained in RKHS H_γ. Furthermore, the latter theorem also yields the following upper bound for the regularization term

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

In the next step, we bound the excess risk. By Eberts and Steinwart (2013, Theorem 2.2), the upper bound for $L_2(\mathrm{P}_X)$-distance between f_0 and $f^*_{L_\tau,\mathrm{P}}$ is

$$\|f_0 - f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)} = \|K_\gamma * f^*_{L_\tau,\mathrm{P}} - f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)} \leq C_{s,2}\|g\|_{L_2(\mathbb{R}^d)}c^2\gamma^{2\alpha}, \tag{28}$$

where $C_{s,2} :=:= \sum_{i=0}^{\lceil 2s\rceil}\binom{\lceil 2s\rceil}{i}(2d)^{\frac{i}{2}}\prod_{j=1}^{i}(j-\frac{1}{2})^{\frac{1}{2}}$ (see Eberts and Steinwart 2013, p. 27) is constant only depending on s and $g \in L_2(\mathbb{R}^d)$ is the Lebesgue density of P_X. Now using Theorem 13 together with (28), we obtain

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \leq C_\tau\,\|f_0 - f^*_{L_\tau,P}\|^2_{L_2(\mathrm{P}_X)} = C_{\tau,s}\gamma^{2\alpha},$$

where $C_{\tau,s} := c^2\,C_\tau\,C_{s,2}\,\|g\|_{L_2(\mathbb{R}^d)}$. With these results, we finally obtain

$$\begin{aligned}\inf_{f\in H_\gamma}\lambda\|f\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}} &\leq \lambda\|f_0\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\tau,\mathrm{P}},\\ &\leq C_1\lambda\gamma^{-d} + C_{\tau,s}\gamma^{2\alpha},\end{aligned}$$

where $C_1 := (\sqrt{\pi})^{-d}(2^r-1)^2\|f^*_{L_\tau,\mathrm{P}}\|^2_{L_2(\mathbb{R}^d)}$. □

In order to prove the main oracle inequality given in Theorem 7, we need the following lemma.

Lemma 14 *The function* $h : (0, \frac{1}{2}] \to \mathbb{R}$ *defined by*

$$h(p) := \left(\frac{\sqrt{2}-1}{\sqrt{2}-2^{\frac{2p-1}{2p}}}\right)^p,$$

is convex. Moreover, we have $\sup_{p\in(0,\frac{1}{2}]} h(p) = 1$.

Proof By considering the linear transformation $t := 2p$, it is suffices to show that the function $g : (0, 1] \to \mathbb{R}$ defined by

$$g(t) := \left(\frac{\sqrt{2}-1}{\sqrt{2}-2^{1-\frac{1}{t}}}\right)^{\frac{t}{2}},$$

is convex. To solve the latter, we first compute the first and second derivative of $g(t)$ with respect to t, that is:

$$g'(t) = \frac{1}{2}\left(\frac{\sqrt{2}-1}{\sqrt{2}-2^{1-\frac{1}{t}}}\right)^{\frac{t}{2}}\left(\log\left(\frac{\sqrt{2}-1}{\sqrt{2}-2^{1-\frac{1}{t}}}\right) + \frac{2^{1-\frac{1}{t}}\log 2}{t\left(\sqrt{2}-2^{1-\frac{1}{t}}\right)}\right),$$

and

$$g''(t) = \left(\frac{\sqrt{2}-1}{\sqrt{2}-2^{1-\frac{1}{t}}}\right)^{\frac{t}{2}} \left(\frac{1}{2}\log\left(\frac{\sqrt{2}-1}{\sqrt{2}-2^{1-\frac{1}{t}}}\right) + \frac{2^{1-\frac{1}{t}}\log 2}{2t\left(\sqrt{2}-2^{1-\frac{1}{t}}\right)}\right)^2 + \left(\frac{\sqrt{2}-1}{\sqrt{2}-2^{1-\frac{1}{t}}}\right)^{\frac{t}{2}} \left(\frac{\left(2^{1-\frac{1}{t}}\right)^2(\log 2)^2}{2t^3\left(\sqrt{2}-2^{1-\frac{1}{t}}\right)^2} + \frac{2^{1-\frac{1}{t}}(\log 2)^2}{2t^3\left(\sqrt{2}-2^{1-\frac{1}{t}}\right)}\right) \tag{29}$$

Since $t \in (0, 1]$, it is not hard to see that all terms in $g''(t)$ are strictly positive. Thus $g''(t) > 0$ and hence $g(t)$ is convex. Furthermore, by convexity of $g(t)$, it is easy to find that

$$\sup_{t\in(0,1]} g(t) = \max\left\{\lim_{t\to 0} g(t), g(1)\right\} = 1.$$

□

Proof of Theorem 7 The assumption $f^*_{L_\tau,\mathrm{P}} \in L_\infty(\mathbb{R}^d)$ and Theorem 2.3 in Eberts and Steinwart (2013) yield that

$$|K_\gamma * f^*_{L_\tau,\mathrm{P}}(x)| \le (2^s - 1)\|f^*_{L_\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)},$$

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

$$\begin{aligned} L_\tau\left(y, K_\gamma * f^*_{L_\tau,\mathrm{P}}(x)\right) &\le C_\tau(M + \|K_\gamma * f^*_{L_\tau,\mathrm{P}}\|_\infty)^2 \\ &\le 4C_\tau\left(M + 2^s\|f^*_{L_\tau,\mathrm{P}}\|_{L_\infty(\mathbb{R}^d)}\right)^2 := B_0, \end{aligned}$$

and hence we conclude that $B_0 \ge 4C_\tau M^2 = B$. Now, by plugging the result of Theorem 6 together with $a = (3K)^{\frac{1}{2p}}\left(\frac{d+1}{ep}\right)^{\frac{d+1}{2p}}$ from Theorem 5 and $V = 16\,c_\tau^{-1}\,C_\tau^2\,M^2$ from Lemma 4, into Theorem 7.23 in Steinwart and Christmann (2008) we obtain

$$\begin{aligned} &\lambda\|f_{\mathrm{D},\lambda,\gamma}\|^2_{H_\gamma} + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\ &\le 9\,C_1\lambda\gamma^{-d} + 9\,C_{\tau,s}\gamma^{2\alpha} + 3K(p)\,K\left(\frac{d+1}{e}\right)^{d+1}\frac{\gamma^{-d}}{p^{d+1}\lambda^p n} \\ &\quad + \left(3456M^2\,C_\tau^2\,c_\tau^{-1} + 60\left(M + 2^s\|f^*_{L_\tau,P}\|_{L_\infty(\mathbb{R}^d)}\right)^2\right)\frac{\varrho}{n}, \\ &\le 9\,C_1\lambda\gamma^{-d} + 9\,C_{\tau,s}\gamma^{2\alpha} + C_d\,K(p)\,\frac{\gamma^{-d}}{p^{d+1}\lambda^p n} + C_2\frac{\varrho}{n}, \end{aligned} \tag{30}$$

where C_1 and $C_{\tau,s}$ are from Theorem 6, $K(p)$ is a constant given in Theorem 7.23, Steinwart and Christmann (2008) that depends on p, $C_2 := 3456\,M^2\,C_\tau^2\,c_\tau^{-1} + 60(M + 2^s\|f^*_{L_\tau,P}\|_{L_\infty(\mathbb{R}^d)})^2$, and $C_d := 3K\left(\frac{d+1}{e}\right)^{d+1}$ is a constant only depending on d. Let us assume that $p := \frac{1}{\log\lambda^{-1}}$. Since $\lambda \le e^{-2}$ and $\lambda^p = e^{-1}$, the result (30) becomes

$$\begin{aligned} &\lambda\|f_{\mathrm{D},\lambda,\gamma}\|^2_{H} + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\ &\le 9\,C_1\lambda\gamma^{-d} + 9\,C_{\tau,s}\gamma^{2\alpha} + C_d\,e\,K(p)\,(\log\lambda^{-1})^{d+1}\,\frac{\gamma^{-d}}{n} + C_2\frac{\varrho}{n} \end{aligned} \tag{31}$$

We now consider the constant $K(p)$ in more details. From the proof of Theorem 7.23 in Steinwart and Christmann (2008) the constant $K(p)$ for $\vartheta = 1$ is

$$K(p) := \max\left\{2700 \cdot 2^{2p}\, C_1^2(p)|L_\tau|_{1,M}^{2p}\, V^{1-p}, 90 \cdot (120)^p\, C_2^{1+p}(p)|L_\tau|_{1,M}^{2p}\, B^{1-p}, 2B\right\} \tag{32}$$

where the constants $C_1(p)$ and $C_2(p)$ derived from the proof of Theorem 7.16 in Steinwart and Christmann (2008) are

$$C_1(p) := \frac{2\sqrt{\ln 256}C_p^p}{(\sqrt{2}-1)(1-p)2^{p/2}} \quad \text{and} \quad C_2(p) := \left(\frac{8\sqrt{\ln 16}C_p^p}{(\sqrt{2}-1)(1-p)4^p}\right)^{\frac{2}{1+p}},$$

and by Steinwart and Christmann (2008, Lemma 7.15), we have

$$C_p := \frac{\sqrt{2}-1}{\sqrt{2}-2^{\frac{2p-1}{2p}}} \cdot \frac{1-p}{p}.$$

In order to bound $K(p)$ for $p \in (0, \frac{1}{2}]$, we first need to bound the constants $C_1(p)$ and $C_2(p)$. Let us start with C_p and obtain the following bound of it for $p \in (0, \frac{1}{2}]$.

$$C_p^p = \left(\frac{\sqrt{2}-1}{\sqrt{2}-2^{\frac{2p-1}{2p}}}\right)^p \left(\frac{1-p}{p}\right)^p \le e \max_{p\in(0,\frac{1}{2}]} \left(\frac{\sqrt{2}-1}{\sqrt{2}-2^{\frac{2p-1}{2p}}}\right)^p = e,$$

where we used $\left(\frac{1-p}{p}\right)^p = \left(\frac{1}{p}-1\right)^p \le e$ for all $p \in (0, \frac{1}{2}]$, and Lemma 14. Now the bound for $C_1(p)$ is the following:

$$C_1(p) \le \max_{p\in(0,\frac{1}{2}]} \frac{2\sqrt{\ln 256}\, C_p^p}{(\sqrt{2}-1)(1-p)2^{p/2}} \le \frac{4\,e\,\sqrt{\ln 256}}{\sqrt{2}-1} \max_{p\in(0,\frac{1}{2}]} \frac{1}{2^{p/2}} \le 46\,e.$$

Analogously, the bound for the constant $C_2(p)$ is:

$$C_2^{1+p}(p) \le \max_{p\in(0,\frac{1}{2}]} \left(\frac{8\sqrt{\ln 16}\, C_p^p}{(\sqrt{2}-1)(1-p)4^p}\right)^2 \le \frac{256\,e^2 \ln(16)}{(\sqrt{2}-1)^2} \max_{p\in(0,\frac{1}{2}]} \frac{1}{4^{2p}} \le 1035\,e^2.$$

By plugging $C_1(p)$ and $C_2(p)$ into (32), together with the Lipschitz constant $|L_\tau|_{1,M} = 4C_\tau\, M$ from Lemma 2 and the supremum bound B and variance bound V from Lemma 4 we thus obtain

$$\begin{aligned} K &\le 3 \max\left\{4 \cdot 10^7\, C_\tau^3\, c_\tau^{-1}\, M^3, 2 \cdot 10^9\, C_\tau^2 M^3, 8\, C_\tau\, M^2\right\} \\ &\le 2 \cdot 10^9\, C_\tau^3\, c_\tau^{-1}\, M^3, \end{aligned}$$

and by plugging this result into (31), we obtain

$$\begin{aligned} &\lambda \|f_{D,\lambda,\gamma}\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\ &\quad \le C M^2 \left(\lambda\gamma^{-d} + \gamma^{2\alpha} + (\log \lambda^{-1})^{d+1}\gamma^{-d} n^{-1} + \varrho\, n^{-1}\right), \end{aligned}$$

where C is a constant independent of λ, γ, n and ϱ. □

Proof of Corollary 8 For all $n \geq 1$, Theorem 7 yields

$$\begin{aligned}&\lambda \| f_{\mathrm{D},\lambda,\gamma} \|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\ &\quad \leq CM^3 \left(\lambda_n \gamma_n^{-d} + \gamma_n^{2\alpha} + (\log \lambda_n^{-1})^{d+1} n^{-1} \gamma_n^{-d} + n^{-1} \varrho \right),\end{aligned}$$

with probability P^n not less than $1 - 3e^{-\varrho}$ and a constant $c > 0$. Using the sequences $\lambda_n = (\log n)^{\delta_1} n^{-1}$ and $\gamma_n = (\log n)^{\delta_2} n^{-\frac{1}{2\alpha+d}}$, we obtain for $n \geq 3$:

$$\begin{aligned}&\lambda \| f_{\mathrm{D},\lambda,\gamma} \|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} \\ &\quad \leq CM^3 \left(\left((\log n)^{\delta_1 - d\delta_2} + (\log n)^{2\alpha\delta_2} \right) n^{-\frac{2\alpha}{2\alpha+d}} \right. \\ &\qquad \left. + (\log n - \delta_1 \log\log n)^{d+1} (\log n)^{-d\delta_2} n^{-\frac{2\alpha}{2\alpha+d}} + n^{-1}\varrho \right) \\ &\quad \leq CM^3 \left(\left((\log n)^{\delta_1 - d\delta_2} + (\log n)^{2\alpha\delta_2} \right) n^{-\frac{2\alpha}{2\alpha+d}} + (\log n)^{d+1-d\delta_2} n^{-\frac{2\alpha}{2\alpha+d}} + n^{-1}\varrho \right) \\ &\quad \leq 3CM^3 \varrho \, (\log n)^{\max\{\delta_1 - d\delta_2, 2\alpha\delta_2, d+1-d\delta_2\}} n^{-\frac{2\alpha}{2\alpha+d}} + n^{-1}\varrho. \end{aligned} \tag{33}$$

Now, some simple calculations show that $\delta_1 - d\delta_2 = d + 1 - d\delta_2 = (d+1) \cdot \frac{2\alpha}{2\alpha+d}$ and $2\alpha\delta_2 = (d+1) \cdot \frac{2\alpha}{2\alpha+d}$. This proves the assertion. □

Before we proof the Theorem 10, we need the following technical lemma.

Lemma 15 *Let $n \geq 3$ and $\Lambda_n \subset (0,1]$ be a finite set such that there exists a $\lambda_i \in \Lambda_n$ with $(\log n)^{-(d+1)} n^{-1} \leq \lambda_i \leq (\log n)^{d+1} n^{-1}$. Moreover assume that $\delta_n \geq 0$ and $\Gamma_n \subset (0,1]$ is a finite δ_n-net of $(0,1]$. Then for $d \geq 1$ and $\alpha \geq 1$ we have*

$$\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \left(\lambda\gamma^{-d} + \gamma^{2\alpha} + (\log \lambda^{-1})^{d+1} \gamma^{-d} n^{-1} \right) \leq c (\log n)^{d+1} \left(n^{-\frac{2\alpha}{2\alpha+d}} + \delta_n^{2\alpha} \right),$$

where c is a constant independent of n, δ_n, Λ_n and Γ_n.

Proof Let us assume that $\Lambda_n = \{\lambda_1, \ldots, \lambda_r\}$ and $\Gamma_n = \{\gamma_1, \ldots, \gamma_s\}$, and $\lambda_{i-1} < \lambda_i$ for all $i = 2, \ldots, r$ and $\gamma_{j-1} < \gamma_j$ for all $j = 2, \ldots, s$. We thus obtain

$$\begin{aligned}&\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \left(\lambda\gamma^{-d} + \gamma^{2\alpha} + (\log \lambda^{-1})^{d+1} \gamma^{-d} n^{-1} \right) \\ &\quad \leq \inf_{\gamma\in\Gamma_n} \left(\lambda_i \gamma^{-d} + \gamma^{2\alpha} + (\log \lambda_i^{-1})^{d+1} \gamma^{-d} n^{-1} \right) \\ &\quad \leq \inf_{\gamma\in\Gamma_n} \left((\log n)^{d+1} \gamma^{-d} n^{-1} + \gamma^{2\alpha} + (\log n - (d+1)\log\log n)^{d+1} \gamma^{-d} n^{-1} \right) \\ &\quad \leq \inf_{\gamma\in\Gamma_n} \left(2(\log n)^{d+1} \gamma^{-d} n^{-1} + \gamma^{2\alpha} \right) \end{aligned} \tag{34}$$

It is not hard to see that the function $\gamma \mapsto 2(\log n)^{d+1} \gamma^{-d} n^{-1} + \gamma^{2\alpha}$ is optimal at $\gamma_n^* := c_1 (\log n)^{\frac{d+1}{2\alpha+d}} n^{-\frac{1}{2\alpha+d}}$, where $c_1 > 0$ is a constant only depending on α and d. Furthermore, with $\gamma_0 = 0$, we see that $\gamma_j - \gamma_{j-1} \leq 2\delta_n$ for all $j = 1, \ldots, s$. In addition, there exits an index $j \in \{1, \ldots, s\}$ such that $\gamma_{j-1} \leq \gamma_n^* \leq \gamma_j$. Consequently, we have $\gamma_n^* \leq \gamma_j \leq \gamma_n^* + 2\delta_n$.

Using this result in (34), we obtain

$$\begin{aligned}
&\inf_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n}\left(\lambda\gamma^{-d}+\gamma^{2\alpha}+(\log\lambda^{-1})^{d+1}\gamma^{-d}n^{-1}\right)\\
&\quad\le 2(\log n)^{d+1}\gamma_j^{-d}n^{-1}+\gamma_j^{2\alpha}\\
&\quad\le 2(\log n)^{d+1}(\gamma_n^*)^{-d}n^{-1}+(\gamma_n^*+2\delta_n)^{2\alpha}\\
&\quad\le 2(\log n)^{d+1}(\gamma_n^*)^{-d}n^{-1}+c_\alpha(\gamma_n^*)^{2\alpha}+c_\alpha\delta_n^{2\alpha}\\
&\quad\le c\left((\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}}n^{-\frac{2\alpha}{2\alpha+d}}+\delta_n^{2\alpha}\right),
\end{aligned}$$

where $c := 2c_1^{-d}+c_\alpha c_1^{2\alpha}+c_\alpha$ is a constant depending only on α and d. □

Proof of Theorem 10 This proof is the repetition of the proof given by Eberts and Steinwart (2013, Theorem 3.6) for the least squares loss. However, for the sake of completeness, we present here in the case of the L_τ-loss. Let us define $m := \lfloor\frac{n}{2}\rfloor+1 \ge \frac{n}{2}$, then for all $(\lambda,\gamma)\in\Lambda_n\times\Gamma_n$, Theorem 7 yields

$$\begin{aligned}
\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma})-\mathcal{R}^*_{L_\tau,\mathrm{P}} &\le \frac{c_1}{2}\left(\lambda\gamma^{-d}+\gamma^{2\alpha}+\frac{(\log\lambda^{-1})^{d+1}}{\gamma^d m}+\frac{\varrho}{m}\right)\\
&\le c_1\left(\lambda\gamma^{-d}+\gamma^{2\alpha}+\frac{(\log\lambda^{-1})^{d+1}}{\gamma^d n}+\frac{\varrho}{n}\right),
\end{aligned}$$

with probability P^m not less than $1-3|\Lambda_n\times\Gamma_n|\,e^{-\varrho}$. Now define $n-m\ge\frac{n}{2}-1\ge\frac{n}{4}$ and $\varrho_n := \varrho+\ln(1+|\Lambda_n\times\Gamma_n|)$, then by using Theorem 7.2 in Steinwart and Christmann (2008) and Lemma 15, we obtain

$$\begin{aligned}
&\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{\mathrm{D}_2},\gamma_{\mathrm{D}_2}})-\mathcal{R}^*_{L_\tau,\mathrm{P}}\\
&\quad\le 6\inf_{(\lambda,\gamma)\in\Lambda_n,\Gamma_n}\left(\mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma})-\mathcal{R}^*_{L_\tau,\mathrm{P}}\right)+512M^2c_\tau^{-1}\frac{\varrho_n}{n-m}\\
&\quad\le 6c_1\inf_{(\lambda,\gamma)\in\Lambda_n,\Gamma_n}\left(\lambda\gamma^{-d}+\gamma^{2\alpha}+\frac{(\log\lambda^{-1})^{d+1}}{\gamma^d n}+\frac{\varrho}{n}\right)+2048M^2c_\tau^{-1}\frac{\varrho_n}{n}\\
&\quad\le 6c_1c\left((\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}}n^{-\frac{2\alpha}{2\alpha+d}}+\delta_n^{2\alpha}\right)+2048M^2c_\tau^{-1}\frac{\varrho_n}{n}\\
&\quad\le \varrho M^2\left(6c_1c+6cc_1\delta_n^{2\alpha}+6c_1+2048c_\tau^{-1}\varrho_n\right)(\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}}n^{-\frac{2\alpha}{2\alpha+d}}\\
&\quad\le c_2M^3\varrho(\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}}n^{-\frac{2\alpha}{2\alpha+d}},
\end{aligned}$$

with probability P^n not less than $1-3(1+|\Lambda_n\times\Gamma_n|)e^{-\varrho}$. □

Proof of Theorem 11 By (17), we obtain

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

This implies that

$$P^n\left(\left\{D\in(X\times Y)^n:\max_{i\in\{1,\ldots,n\}}\{|y_i|\}\le c(\tilde{\varrho}+\ln n)^l\right\}\right)\ge 1-e^{-\tilde{\varrho}}.$$

This leads us to conclude with probability P^n not less than $1-e^{-\tilde{\varrho}}$ that the SVM for ALS loss with belatedly clipped decision function at M_n is actually a clipped regularized empirical risk minimization (CR-ERM) in the sense of Definition 7.18 in Steinwart and Christmann (2008). Consequently, Theorem 7.20 in Steinwart and Christmann (2008) holds for $\tilde{Y} := \{-M_n, M_n\}$ modulo a set of probability P^n not less than $1-e^{-\tilde{\varrho}}$. From Theorem 7, we then obtain

$$\begin{aligned}&\lambda\|f_{\mathrm{D},\lambda,\gamma}\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}}\\ &\quad\le C M_n^3\left(\lambda\gamma^{-d} + \gamma^{2\alpha} + (\log\lambda^{-1})^{d+1}n^{-1}\gamma^{-d} + n^{-1}\bar{\varrho}\right).\end{aligned}$$

with probability P^n not less than $1-e^{-\bar{\varrho}}-e^{-\tilde{\varrho}}$. As in the proof of Corollary (8) and by using the inequality $(a+b)^c \le (2ab)^c$, for $a, b \ge 1$ and $c > 0$, we finally obtain

$$\begin{aligned}\lambda\|f_{D,\lambda,\gamma}\|_{H_\gamma}^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(\widehat{f}_{D,\lambda,\gamma}) - \mathcal{R}^*_{L_\tau,\mathrm{P}} &\le C\bar{\varrho}M_n^3(\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}}n^{-\frac{2\alpha}{2\alpha+d}}\\ &= C\bar{\varrho}\left(2c(\tilde{\varrho}+\log n)^l\right)^3(\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}}n^{-\frac{2\alpha}{2\alpha+d}}\\ &\le C\bar{\varrho}\,8c^3\,(2\tilde{\varrho}\,\log n)^{3l}\,(\log n)^{\frac{2\alpha(d+1)}{2\alpha+d}}n^{-\frac{2\alpha}{2\alpha+d}}\\ &\le \hat{C}\bar{\varrho}\tilde{\varrho}^{3l}(\log n)^{3l+\frac{2\alpha(d+1)}{2\alpha+d}}n^{-\frac{2\alpha}{2\alpha+d}},\end{aligned}$$

for all $n \ge 3$ with probability P^n not less than $1-e^{-\bar{\varrho}}-e^{-\tilde{\varrho}}$. Choosing $\bar{\varrho} = \tilde{\varrho}$ leads to the assertion. □

Acknowledgements We gratefully thank to the anonymous referees for their valuable comments, particularly, the referee who pointed us to improve the logarithmic term in learning rates. This research is supported by Higher Education Commission (HEC) Pakistan (PS/OS-I/Batch-2012/Germany/2012/3449) and German Academic Exchange Service (DAAD) scholarship program/-ID50015451.

References

Abdous, B., & Remillard, B. (1995). Relating quantiles and expectiles under weighted-symmetry. *Annals of the Institute of Statistical Mathematics*, *47*, 371–384. https://doi.org/10.1007/bf00773468.

Adams, R. A., & Fournier, J. J. F. (2003). *Sobolev spaces* (2nd ed.). New York: Academic Press. https://doi.org/10.1016/s0079-8169(03)x8001-0.

Aragon, Y., Casanova, S., Chambers, R., & Leconte, E. (2005). Conditional ordering using nonparametric expectiles. *Journal of Official Statistics*, *21*, 617–633.

Bauer, F., Pereverzev, S., & Rosasco, L. (2007). On regularization algorithms in learning theory. *Journal of Complexity*, *23*, 52–72. https://doi.org/10.1016/j.jco.2006.07.001.

Bellini, F., Klar, B., Müller, A., & Gianin, R. E. (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, *54*, 41–48. https://doi.org/10.1016/j.insmatheco.2013.10.015.

Blanchard, G., Bousquet, O., & Massart, P. (2008). Statistical performance of support vector machines. *The Annals of Statistics*,. https://doi.org/10.1214/009053607000000839.

Caponnetto, A., & De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, *7*, 331–368. https://doi.org/10.1007/s10208-006-0196-8.

Chen, D., Wu, Q., Ying, Y., & Zhou, D. (2004). Support vector machine soft margin classifiers: Error analysis. *Journal of Machine Learning Research*, *5*, 1143–1175.

Christmann, A., & Steinwart, I. (2007). How SVMs can estimate quantiles and the median. In: *Advances in neural information processing systems* (pp. 305–312).

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, *39*, 1–49. https://doi.org/10.1090/s0273-0979-01-00923-5.

De Vito, E., Caponnetto, A., & Rosasco, L. (2005). Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, *5*, 59–85. https://doi.org/10.1007/s10208-004-0134-1.

DeVore, R. A. (1998). Nonlinear approximation. *Acta Numerica*, *7*, 51–150.

DeVore, R. A., & Popov, V. A. (1988). Interpolation of Besov spaces. *Transactions of the American Mathematical Society*, *305*, 397–414.

DeVore, R. A., & Sharpley, R. C. (1993). Besov spaces on domains in $\mathbb{R}^d$. *Transactions of the American Mathematical Society*, *335*, 843–864.

Eberts, M., & Steinwart, I. (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electronic Journal of Statistics*, *7*, 1–42. https://doi.org/10.1214/12-ejs760.

Edmunds, D. E., & Triebel, H. (2008). *Function spaces, entropy numbers, differential operators*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511662201.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistical Science*, *1*, 93–125.

Farooq, M., & Steinwart, I. (2017). An SVM-like approach for expectile regression. *Computational Statistics & Data Analysis*, *109*, 159–181. https://doi.org/10.1016/j.csda.2016.11.010.

Glasmachers, T., & Igel, C. (2006). Maximum-gain working set selection for SVMs. *Journal of Machine Learning Research*, *7*, 1437–1466.

Guler, K., Ng, P.T., & Xiao, Z. (2014). *Mincer-Zarnovitz quantile and expectile regressions for forecast evaluations under asymmetric loss functions*. Northern Arizona University, The WA Franke College of Business Working Paper Series 14-01.

Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. New York: Springer.

Hamidi, B., Maillet, B., & Prigent, J. L. (2014). A dynamic autoregressive expectile for time-invariant portfolio protection strategies. *Journal of Economic Dynamics and Control*, *46*, 1–29. https://doi.org/10.1016/j.jedc.2014.05.005.

Kim, M., & Lee, S. (2016). Nonlinear expectile regression with application to value-at-risk and expected shortfall estimation. *Computational Statistics & Data Analysis*, *94*, 1–19. https://doi.org/10.1016/j.csda.2015.07.011.

Koenker, R., & Bassett, G, Jr. (1978). Regression quantiles. *Econometrica*, *46*, 33–50. https://doi.org/10.2307/1913643.

Meister, M., & Steinwart, I. (2016). Optimal learning rates for localized SVMs. *Journal of Machine Learning Research*, *17*, 1–44.

Mendelson, S., & Neeman, J. (2010). Regularization in kernel learning. *The Annals of Statistics*, *38*, 526–565. https://doi.org/10.1214/09-aos728.

Newey, W. K., & Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, *55*, 819–847. https://doi.org/10.2307/1911031.

Nikol'skii, S. M. (2012). *Approximation of functions of several variables and imbedding theorems* (Vol. 205). Berlin: Springer. https://doi.org/10.1007/978-3-642-65711-5.

Schnabel, S., & Eilers, P. (2009). An analysis of life expectancy and economic production using expectile frontier zones. *Demographic Research*, *21*, 109–134. https://doi.org/10.4054/demres.2009.21.5.

Sobotka, F., & Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, *56*, 755–767. https://doi.org/10.1016/j.csda.2010.11.015.

Sobotka, F., Radice, R., Marra, G., & Kneib, T. (2013). Estimating the relationship between women's education and fertility in Botswana by using an instrumental variable approach to semiparametric expectile regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *62*, 25–45. https://doi.org/10.1111/j.1467-9876.2012.01050.x.

Steinwart, I. (2007). How to compare different loss functions and their risks. *Constructive Approximation*, *26*, 225–287. https://doi.org/10.1007/s00365-006-0662-3.

Steinwart, I. (2009). Oracle inequalities for support vector machines that are based on random entropy numbers. *Journal of Complexity*, *25*, 437–454. https://doi.org/10.1016/j.jco.2009.06.002.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York: Springer. https://doi.org/10.1007/978-0-387-77242-4.

Steinwart, I., & Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, *17*, 211–225. https://doi.org/10.3150/10-bej267.

Steinwart, I., Hush, D., & Scovel, C. (2006). An oracle inequality for clipped regularized risk minimizers. In: *Advances in neural information processing systems* (pp. 1321–1328).

Steinwart, I., Hush, D. R., & Scovel, C. (2009). Optimal rates for regularized least squares regression. In: *22nd Annual conference on learning theory* (pp. 79–93).

Steinwart, I., Hush, D., & Scovel, C. (2011). Training SVMs without offset. *Journal of Machine Learning Research*, *12*, 141–202.

Steinwart, I., Pasin, C., Williamson, R., & Zhang, S. (2014). Elicitation and identification of properties. In: M. F. Balcan & C. Szepesvari (Eds.), *JMLR workshop and conference proceedings*. Volume 35: Proceedings of the 27th conference on learning theory 2014 (pp. 482–526).

Tacchetti, A., Mallapragada, P. K., Santoro, M., & Rosasco, L. (2013). GURLS: A least squares library for supervised learning. *Journal of Machine Learning Research*, *14*, 3201–3205.
Takeuchi, I., Le, Q. V., Sears, T. D., & Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, *7*, 1231–1264.
Tartar, L. (2007). *An introduction to sobolev spaces and interpolation spaces*. Berlin: Springer. https://doi.org/10.1007/978-3-540-71483-5.
Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Economics*, *6*, 231–252. https://doi.org/10.1093/jjfinec/nbn001.
van der Vaart, A. W., & van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, *37*, 2655–2675. https://doi.org/10.1214/08-aos678.
Wang, Y., Wang, S., & Lai, K. K. (2011). Measuring financial risk with generalized asymmetric least squares regression. *Applied Soft Computing*, *11*(8), 5793–5800. https://doi.org/10.1016/j.asoc.2011.02.018.
Wu, Q., Ying, Y., & Zhou, D. X. (2006). Learning rates of least square regularized regression. *Foundations of Computational Mathematics*, *6*, 171–192. https://doi.org/10.1007/s10208-004-0155-9.
Xu, Q., Liu, X., Jiang, C., & Yu, K. (2016). Nonparametric conditional autoregressive expectile model via neural network with applications to estimating financial risk. *Applied Stochastic Models in Business and Industry*, *32*, 882–908. https://doi.org/10.1002/asmb.2212.
Yang, Y., & Zou, H. (2015). Nonparametric multiple expectile regression via ER-Boost. *Journal of Statistical Computation and Simulation*, *85*, 1442–1458. https://doi.org/10.1080/00949655.2013.876024.
Yang, Y., Zhang, T., & Zou, H. (2017). Flexible expectile regression in reproducing kernel Hilbert spaces. *Technometrics*, *1*, 1–10. https://doi.org/10.1080/00401706.2017.1291450.
Yao, Q., & Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics*, *6*, 273–292. https://doi.org/10.1080/10485259608832675.
Zhang, B. (1994). Nonparametric regression expectiles. *Journal of Nonparametric Statistics*, *3*, 255–275. https://doi.org/10.1080/10485259408832586.
Ziegel, J. F. (2016). Coherence and elicitability. *Mathematical Finance*, *26*, 901–918. https://doi.org/10.1111/mafi.12080.