9

Support Vector Machines for Regression

Overview. *Regression is, besides classification, one of the main areas where support vector machines are applied. This chapter presents results on the learning properties of SVMs when applied to regression problems such as estimating conditional means, medians, or quantiles.*

Prerequisites. *Knowledge of loss functions, kernels, and stability of infinite-sample SVMs is needed from Chapters 2 and 4 and Section 5.3, respectively. Some results from measure theory, integration, and functional analysis from the appendix are used.*

In this chapter, we investigate under which conditions support vector machines are able to learn in the sense of L-risk consistency in regression problems with emphasis on the case of an unbounded output space. An introduction into SVMs for regression problems is given in Section 9.1. Section 9.2 considers the case of general loss functions. Section 9.3 covers the special case of SVMs designed to estimate conditional quantile functions, and Section 9.4 contains some numerical results for such SVMs.

9.1 Introduction

The goal in non-parametric regression is to estimate a functional relationship between an input random variable X and an output random variable Y under the assumption that the joint distribution P of (X, Y) is (almost) completely *unknown*. In order to solve this problem, we assume the existence of a set of observations (x_i, y_i) from independent and identically distributed random variables (X_i, Y_i), $i = 1, \dots, n$, all of which have the distribution P on $X \times Y$ with corresponding Borel σ-algebra, where $Y \subset \mathbb{R}$ is Polish and e.g. $X = \mathbb{R}^d$. Denote by D the corresponding empirical distribution. The aim is to build a predictor $f : X \to \mathbb{R}$ on the basis of these observations such that $f(X)$ is a "good" approximation of Y. In this chapter, we investigate SVMs for regression problems, defined as minimizers of the regularized L-risk

$$f_{\mathrm{P},\lambda} = \arg \inf_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) + \lambda \|f\|_H^2 \,. \tag{9.1}$$

Replacing P by D in (9.1) gives the corresponding empirical problem. Following the interpretation that a small risk is desired, one tries to find a predictor whose risk is close to the optimal risk given by the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$ (see Definition 2.3). This is a much stronger requirement than convergence in probability of $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n})$ to $\mathcal{R}_{L,\mathrm{P},H} := \inf_{f\in H}\mathcal{R}_{L,\mathrm{P}}(f)$, $n \to \infty$, because it is not obvious whether $\mathcal{R}_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$ even for large Hilbert spaces H.

In this chapter, we will assume that the loss function L is chosen in advance. If a suitable choice of L is not known for a specific application, Chapter 3, on surrogate loss functions, may be helpful in choosing one. We restrict attention in this chapter to convex loss functions because in this case we are not faced with computationally NP-hard problems and we know that $f_{\mathrm{P},\lambda}$ exists and is unique. We have to fix a reproducing kernel Hilbert space H with corresponding kernel k and a regularization constant $\lambda > 0$. However, we are additionally faced with the question of how to specify the output space Y. The obvious choice in Chapter 8, on binary classification problems, was $Y = \{-1, +1\}$. In regression problems, the choice of Y is less obvious and depends on the application. Two cases are important: the *unbounded case*, where $Y = \mathbb{R}$ or Y is equal to an unbounded interval, and the *bounded case*, where Y is a bounded interval, say $[a, b]$ with $-\infty < a < b < \infty$. Although in some regression problems the output variable can only take values in some bounded interval, suitable numbers for a and b are often unknown. In this situation, many practitioners prefer to choose the unbounded case.[1]

Of course, a natural question is whether the risk $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n})$ actually tends to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$ if n increases. If $\mathcal{R}_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$ and the output space Y is bounded, this can be assured by concentration inequalities, and the techniques from Chapters 6 and 7 are applicable. However, these techniques do not work for the unbounded case. Therefore, in this chapter attention is restricted to the unbounded case and we will present some results on the L-risk consistency of SVMs in regression problems for this more challenging situation.

Traditionally, most research on non-parametric regression considered the least squares loss $L(y, t) := (y - t)^2$ mainly because of historical and computational reasons. However, both from a theoretical and from a practical point of view, there are situations in which a different loss function is more appropriate. We mention three cases:

i) Regression problems not described by least squares loss. It is well-known that the least squares risk is minimized by the conditional mean of Y given x; see Example 2.6. However, in many situations one is actually not interested in this mean but for example in the conditional median instead. Now recall that the conditional median is the minimizer of $\mathcal{R}^*_{L,\mathrm{P}}$, where L is the absolute value loss (i.e., $L(y, t) := |y - t|$), and the same statement holds for conditional quantiles if one replaces the absolute value loss by

[1] In many parametric regression models, the output variable is assumed to have a Gaussian, Gamma, or log-Gaussian distribution; hence $Y = \mathbb{R}$ in the first situation and $Y = (0, \infty)$ in the last two cases.

an *asymmetric* variant known as the pinball loss treated in Example 2.43; see also Proposition 3.9.

ii) Surrogate losses. If the conditional distributions of Y given x are known to be symmetric, basically all distance-based loss functions of the form $L(y,t) = \psi(y-t)$, where $\psi : \mathbb{R} \to [0,\infty)$ is convex, symmetric, and has its only minimum at 0, can be used to estimate the conditional mean of Y given x; see Section 3.7. In this case, a less steep surrogate such as the absolute value loss, Huber's loss, or the logistic loss may be more suitable if one expects outliers in the y-direction, as we will discuss in Chapter 10.

iii) Algorithmic aspects. If the goal is to estimate the conditional median of Y given x, then the ϵ-insensitive loss given by $L_\epsilon(y,t) = \max\{|y-t| - \epsilon, 0\}$, $y,t \in \mathbb{R}$, $\epsilon \in (0,\infty)$, promises algorithmic advantages in terms of sparseness compared with the absolute loss; see Chapter 11 for details.

In Section 9.2, a general result on L-risk consistency of SVMs based on a distance-based loss function $L(y,t) = \psi(y-t)$ is given. Section 9.3 covers SVMs based on the pinball loss for quantile regression and Section 9.4 gives some numerical results for this particular case. In Section 9.5 median regression based on the ϵ-insensitive loss is considered.

9.2 Consistency

In this section, we assume that $L : Y \times \mathbb{R} \to [0,\infty)$ is a convex distance-based loss in the sense of Definition 2.32. In other words, we have a representing function $\psi : \mathbb{R} \to [0,\infty)$ with $\psi(0) = 0$ and $L(t,y) = \psi(y-t)$ for all $y,t \in \mathbb{R}$. Recall that L-risk consistency is defined as the convergence in probability of

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}\,, \quad n \to \infty,$$

where (λ_n) is a suitably chosen data-independent null sequence of regularization parameters with $\lambda_n > 0$. For technical reasons, we will additionally assume that the reproducing kernel Hilbert space H is separable and that its kernel is measurable, so the canonical feature map Φ becomes measurable by Lemma 4.25. We can now formulate our main result on the learnability of SVMs under a natural tail assumption on P. Note that no symmetry assumption on L is made.

Theorem 9.1. *Let X be a complete measurable space, $Y \subset \mathbb{R}$ be closed, L be a continuous, convex, distance-based loss function of growth type $p \in [1,\infty)$, and $H \subset L_p(\mathrm{P}_X)$ be a dense, separable RKHS with a bounded and measurable kernel k and canonical feature map $\Phi : X \to H$. We write $p^* := \max\{2p, p^2\}$ and fix a sequence (λ_n) of positive numbers with $\lambda_n \to 0$ and $\lambda_n^{p^*} n \to \infty$. Then*

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}\,, \quad n \to \infty, \tag{9.2}$$

in probability for all $|D| = n$ and for all distributions $\mathrm{P} \in \mathcal{M}_1(X \times Y)$ with $|\mathrm{P}|_p < \infty$.

Recall that the RKHS of a Gaussian RBF kernel is dense in $L_p(\mathrm{P}_X)$ by Theorem 4.63. Hence this RKHS fulfills the assumption of Theorem 9.1 if $X = \mathbb{R}^d$.

In order to prove Theorem 9.1, we need the following lemma to bound the probability of $|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})| \le \varepsilon$ for $|D| \to \infty$.

Lemma 9.2. *Let Z be a measurable space, P be a distribution on Z, H be a separable Hilbert space, and $g : Z \to H$ be a measurable function with $\|g\|_q := (\mathbb{E}_{\mathrm{P}}\|g\|_H^q)^{1/q} < \infty$ for some $q \in (1,\infty)$. We write $q^* := \min\{1/2, 1/q'\}$, where q' fulfills $\frac{1}{q} + \frac{1}{q'} = 1$. Then there exists a universal constant $c_q > 0$ such that, for all $\varepsilon > 0$ and all $n \ge 1$, we have*

$$\mathrm{P}^n\left((z_1,\dots,z_n) \in Z^n : \left\| \frac{1}{n}\sum_{i=1}^n g(z_i) - \mathbb{E}_{\mathrm{P}} g \right\|_H \ge \varepsilon \right) \le c_q \left(\frac{\|g\|_q}{\varepsilon\, n^{q^*}} \right)^q .$$

For the proof of Lemma 9.2, we have to recall some basics from local Banach space theory and Rademacher sequences. We refer to Section 7.3 and Section A.8 for details. A sequence of independent, symmetric $\{-1,+1\}$-valued random variables (ε_i) is called a *Rademacher sequence.* Now let E be a separable Banach space, (X_i) be an i.i.d. sequence of E-valued random variables with expectation 0, and (ε_i) be a Rademacher sequence that is independent from the sequence (X_i). The distribution of ε_i is denoted by ν. Using the symmetrization argument given in Theorem A.8.1, we have for all $1 \le p < \infty$ and all $n \ge 1$ that

$$\mathbb{E}_{\mathrm{P}^n} \left\| \sum_{i=1}^n X_i \right\|^p \le 2^p\, \mathbb{E}_{\mathrm{P}^n} \mathbb{E}_{\nu^n} \left\| \sum_{i=1}^n \varepsilon_i X_i \right\|^p , \tag{9.3}$$

where the left expectation is with respect to the product distribution P^n of $(X_1,\dots,X_n)$, whereas the right expectation is also with respect to the product distribution ν^n of $(\varepsilon_1,\dots,\varepsilon_n)$. Furthermore, for $n = 2$, we obtain

$$\begin{aligned}
\mathbb{E}_{\nu^2} \|\varepsilon_1 x_1 + \varepsilon_2 x_2\|^2 &= \frac{1}{2}\left(\|x_1 + x_2\|^2 + \|x_1 - x_2\|^2 \right) \\
&= \frac{1}{2}\left(\|x_1\|^2 + 2\langle x_1, x_2\rangle + \|x_2\|^2 + \|x_1\|^2 - 2\langle x_1, x_2\rangle + \|x_2\|^2 \right) \\
&= \|x_1\|^2 + \|x_2\|^2 .
\end{aligned}$$

An induction over n therefore shows that

$$\mathbb{E}_{\nu^n} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|^2 = \sum_{i=1}^n \|x_i\|^2 \tag{9.4}$$

for all $n \ge 1$ and all finite sequences $x_1,\dots,x_n$. Furthermore, Kahane's inequality (see Theorem A.8.3) ensures that

$$\left(\mathbb{E}_{\nu^n}\Big\|\sum_{i=1}^{n}\varepsilon_i x_i\Big\|^p\right)^{1/p} \le c_{p,q}\,\left(\mathbb{E}_{\nu^n}\Big\|\sum_{i=1}^{n}\varepsilon_i x_i\Big\|^q\right)^{1/q}$$

for all $p, q \in (0,\infty)$, $n \ge 1$, all Banach spaces E, all $x_1, \dots, x_n \in E$, and constants $c_{p,q}$ only depending on p and q. Now we can proceed with the following proof.

Proof of Lemma 9.2. Define $h : Z^n \to H$,

$$h(z_1, \dots, z_n) := \frac{1}{n}\sum_{i=1}^{n} g(z_i) - \mathbb{E}_{\mathrm{P}} g\,, \qquad (z_1, \dots, z_n) \in Z^n.$$

Markov's inequality yields

$$\mathrm{P}^n\big(\|h\|_H \ge \varepsilon\big) \le \varepsilon^{-q}\,\mathbb{E}_{\mathrm{P}^n}\|h\|_H^q\,.$$

Hence it remains to estimate $\mathbb{E}_{\mathrm{P}^n}\|h\|_H^q$. By (9.3), we have

$$\mathbb{E}_{\mathrm{P}^n}\left\|\sum_{i=1}^{n} g(Z_i) - \mathbb{E}_{\mathrm{P}} g\right\|_H^q \le 2^q\,\mathbb{E}_{\mathrm{P}^n}\mathbb{E}_{\nu^n}\left\|\sum_{i=1}^{n}\varepsilon_i\big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big)\right\|_H^q\,. \qquad (9.5)$$

If $q \in (1,2]$, we obtain with Kahane's inequality, see Theorem A.8.3, that

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}^n}\|h\|_H^q &\le 2^q n^{-q}\,\mathbb{E}_{\mathrm{P}^n}\mathbb{E}_{\nu^n}\left\|\sum_{i=1}^{n}\varepsilon_i\big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big)\right\|_H^q \\
&\le 2^q n^{-q}\,\mathbb{E}_{\mathrm{P}^n}\left(\mathbb{E}_{\nu^n}\left\|\sum_{i=1}^{n}\varepsilon_i\big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big)\right\|_H^2\right)^{q/2} \\
&= 2^q n^{-q}\,\mathbb{E}_{\mathrm{P}^n}\left(\sum_{i=1}^{n}\big\|(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big\|_H^2\right)^{q/2} \\
&\le 2^q n^{-q}\,\mathbb{E}_{\mathrm{P}^n}\sum_{i=1}^{n}\big\|(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big\|_H^q \\
&\le 4^q n^{1-q}\,\mathbb{E}_{\mathrm{P}}\,\|g\|_H^q\,.
\end{aligned}$$

From this we obtain the assertion for $q \in (1,2]$. Now assume that $q \in (2,\infty)$. By (9.5) and Kahane's inequality, there is a universal constant $c_q > 0$ with

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}^n} \|h\|_H^q &\leq 2^q n^{-q} \mathbb{E}_{\mathrm{P}^n} \mathbb{E}_{\nu^n} \left\| \sum_{i=1}^n \varepsilon_i \big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big) \right\|_H^q \\
&\leq c_q n^{-q} \mathbb{E}_{\mathrm{P}^n} \left(\mathbb{E}_{\nu^n} \left\| \sum_{i=1}^n \varepsilon_i \big(g(Z_i) - \mathbb{E}_{\mathrm{P}} g\big) \right\|_H^2 \right)^{q/2} \\
&\leq c_q n^{-q} \mathbb{E}_{\mathrm{P}^n} \left(\sum_{i=1}^n \left\| g(Z_i) - \mathbb{E}_{\mathrm{P}} g \right\|_H^2 \right)^{q/2} \\
&\leq c_q n^{-q} \left(\sum_{i=1}^n \big(\mathbb{E}_{\mathrm{P}} \|g(Z_i) - \mathbb{E}_{\mathrm{P}} g\|_H^q \big)^{2/q} \right)^{q/2} \\
&\leq 2^q c_q n^{-q/2} \, \mathbb{E}_{\mathrm{P}} \|g\|_H^q ,
\end{aligned}$$

where (9.4) is used in the third step. The assertion follows for $q \in (2, \infty)$. □

Proof of Theorem 9.1. Because $H \subset L_p(\mathrm{P}_X)$ is dense, we obtain from Lemma 2.38 (i) and from the assumption $|\mathrm{P}|_p < \infty$ that L is a P-integrable Nemitski loss of order $p \in [1, \infty)$. Hence Theorem 5.31 gives

$$\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}} \,. \tag{9.6}$$

To avoid handling too many constants, let us assume $\|k\|_\infty = 1$, $|\mathrm{P}|_p = 1$, and $c := c_{L,p} := 2^{-(p+2)}$ for the upper order constant of L. This yields $\mathcal{R}_{L,\mathrm{P}}(0) \leq 1$. Furthermore, we assume without loss of generality that $\lambda_n \leq 1$ for all $n \geq 1$. Using (5.4), we obtain $\|f_{\mathrm{P},\lambda_n}\|_\infty \leq \|f_{\mathrm{P},\lambda_n}\|_H \leq \lambda_n^{-1/2}$. It follows from Lemma 5.15 that

$$\lim_{\lambda \to 0} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{\mathrm{P},\lambda}) = \mathcal{R}^*_{L,\mathrm{P},H}$$

because $A_2(\lambda)$ is continuous and $A_2(0) = 0$. Combining this with (9.6) yields

$$\lim_{\lambda_n \to 0} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) = \mathcal{R}^*_{L,\mathrm{P}} \,.$$

Therefore, we obtain L-risk consistency if we can show that

$$|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n})| \to 0$$

holds in probability for $n \to \infty$.

For $n \in \mathbb{N}$ and $\lambda_n > 0$, let $h_n : X \times Y \to \mathbb{R}$ be the function obtained by Corollary 5.11. Our assumptions give for $p' := p/(p-1)$ that there is a constant $c(p, L)$ such that

$$\|h_n\|_{L_{p'}(\mathrm{P})} \leq c(p, L) \lambda_n^{-(p-1)/2} \,. \tag{9.7}$$

Moreover, for $g \in H$ with $\|f_{\mathrm{P},\lambda_n} - g\|_H \leq 1$, we have

$$\|g\|_\infty \;\leq\; \|f_{\mathrm{P},\lambda_n}\|_\infty + \|f_{\mathrm{P},\lambda_n} - g\|_\infty \;\leq\; 2\lambda_n^{-1/2} \,.$$

First, we consider the case $p > 1$. By Lemma 2.38 (ii) with $q := p-1$, there exists a constant $c_{p,L} > 0$ only depending on L and p such that

$$\begin{aligned}
&\bigl|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| \\
&\leq c_{p,L}\left(|\mathrm{P}|_{p-1}^{p-1} + \|f_{\mathrm{P},\lambda_n}\|_\infty^{p-1} + \|g\|_\infty^{p-1} + 1\right)\|f_{\mathrm{P},\lambda_n} - g\|_\infty \\
&\leq \tilde{c}_{p,L}\,\lambda_n^{-(p-1)/2}\,\|f_{\mathrm{P},\lambda_n} - g\|_H
\end{aligned} \tag{9.8}$$

for all measurable $g \in H$ with $\|f_{\mathrm{P},\lambda_n} - g\|_H \leq 1$. Let $\varepsilon \in (0,1]$ and $D \in (X \times Y)^n$ be a training set of length n with empirical distribution D such that

$$\|\mathbb{E}_{\mathrm{P}} h_n \Phi - \mathbb{E}_{\mathrm{D}} h_n \Phi\|_H \leq \lambda_n^{(p+1)/2}\,\varepsilon / \,\tilde{c}_{p,L}\,. \tag{9.9}$$

Then Corollary 5.11 gives $\|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D},\lambda_n}\|_H \leq \lambda_n^{(p-1)/2}\,\varepsilon/\tilde{c}_{p,L} \leq 1$ for n large enough, and hence (9.8) yields

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n})\bigr| \leq \varepsilon\,. \tag{9.10}$$

Second, we consider the special case $p = 1$. The loss function L is by assumption convex and of upper growth type 1 and therefore Lipschitz continuous by Lemma 2.36 (iv). Hence we obtain

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| \;\leq\; |\psi|_1\,\|f_{\mathrm{P},\lambda_n} - g\|_\infty \;\leq\; |\psi|_1\,\|f_{\mathrm{P},\lambda_n} - g\|_H \tag{9.11}$$

for all $g \in H$ with $\|f_{\mathrm{P},\lambda_n} - g\|_H \leq 1$. As L is of growth type $p \in [1,\infty)$ we have $|\psi|_1 > 0$. Now let $\varepsilon \in (0, |\psi|_1^{-1}]$ and $D \in (X \times Y)^n$ be a training set of length n with corresponding empirical distribution D such that

$$\|\mathbb{E}_{\mathrm{P}} h_n \Phi - \mathbb{E}_{\mathrm{D}} h_n \Phi\|_H \leq \lambda_n\,\varepsilon \,/\, |\psi|_1\,. \tag{9.12}$$

Corollary 5.11 and (9.12) give $\|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D},\lambda_n}\|_H \leq \varepsilon/|\psi|_1 \leq 1$ such that (9.11) yields the validity of (9.10) also for the case $p = 1$.

Let us now estimate the probability of D satisfying (9.9). Define $q := p/(p-1)$ if $p > 1$ and $q := 2$ if $p = 1$. Then we have $q^* := \min\{1/2, 1-1/q\} = \min\{1/2, 1/p\} = p/p^*$. Further define $c := \max\{c(p,L), \tilde{c}_{p,L}, |\psi|_1\}$. Combining Lemma 9.2 and (9.7) yields

$$\begin{aligned}
&\mathrm{P}^n\bigl(D \in (X \times Y)^n \,:\, \|\mathbb{E}_{\mathrm{P}} h_n \Phi - \mathbb{E}_{\mathrm{D}} h_n \Phi\|_H \leq c^{-1}\lambda_n^{(p+1)/2}\varepsilon\bigr) \\
&\geq 1 - \hat{c}_{p,L}\left(\frac{\|h_n\|_q}{\varepsilon\,\lambda_n^{(p+1)/2}\,n^{q^*}}\right)^q \geq 1 - \hat{c}_{p,L}\left(\frac{c(p,L)}{\varepsilon\,\lambda_n^p\,n^{p/p^*}}\right)^q,
\end{aligned} \tag{9.13}$$

where $\hat{c}_{p,L}$ is a constant only depending on p and L. Now using $\lambda_n^p n^{p/p^*} = (\lambda_n^{p^*} n)^{p/p*} \to \infty$, if $n \to \infty$, we find that the probability of sample sets D satisfying (9.9) converges to 1 if $|D| = n \to \infty$. As we have seen above, this implies that (9.10) holds true with probability tending to 1. Now, since $\lambda_n \to 0$, we additionally have $|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}}| \leq \varepsilon$ for all sufficiently large values of n, and hence we obtain the assertion of L-risk consistency. □

Remark 9.3. Note that Theorem 9.1 in particular shows that the SVM for regression using the least squares loss function is *weakly universally consistent* in the sense of Györfi *et al.* (2002, p.13). Furthermore, it is worthwhile to note that under the assumptions above on L, H, and (λ_n) we can even characterize the distributions P for which SVMs based on (9.1) are L-risk consistent. Indeed, if $|\mathrm{P}|_p = \infty$, then SVMs are trivially L-risk consistent for P whenever $\mathcal{R}_{L,\mathrm{P}} = \infty$. Conversely, if $|\mathrm{P}|_p = \infty$ and $\mathcal{R}_{L,\mathrm{P}} < \infty$, then SVMs cannot be L-risk consistent for P since $\mathcal{R}_{L,\mathrm{P}}(f) = \infty$ for all $f \in H$. ◁

Remark 9.4. In some sense, it seems natural to consider only consistency for distributions satisfying the tail condition $|\mathrm{P}|_p < \infty$ as was done for example in Györfi *et al.* (2002) for least squares methods. In this sense, Theorem 9.1 gives consistency for all reasonable distributions. Note that the characterization above shows that SVMs are in general *not* robust against small violations of this tail assumption in regression problems. Indeed, let P be a distribution with $|\mathrm{P}|_p < \infty$ and $\tilde{\mathrm{P}}$ be a distribution with $|\tilde{\mathrm{P}}|_p = \infty$ and $\mathcal{R}_{L,\tilde{\mathrm{P}}}(f^*) < \infty$ for some $f^* \in L_p(\mathrm{P})$ (see Problem 2.6). Then every mixture distribution $\mathrm{Q}_\varepsilon := (1-\varepsilon)\mathrm{P} + \varepsilon\tilde{\mathrm{P}}$, $\varepsilon \in (0,1)$, satisfies both $|\mathrm{Q}_\varepsilon|_p = \infty$ and $\mathcal{R}_{L,\mathrm{Q}_\varepsilon} < \infty$. Thus an SVM for regression defined by (9.1) is not consistent for any of the small perturbations Q_ε of P, while it is consistent for the distribution P. Hence some integrability conditions for the robustness results in Section 10.3 and Section 10.4 seem to be necessary if Y is unbounded, see also Remark 10.19(iii). ◁

9.3 SVMs for Quantile Regression

This section gives a mathematical justification for using support vector machines based on the pinball loss function $L := L_{\tau\text{-pin}}$ for quantile regression. The results can informally be described in the following manner.

i) SVMs based on the pinball loss allow the non-parametric estimation of conditional quantiles.(i)

ii) Such SVMs are L-risk consistent under weak assumptions on P and k.

iii) If these SVMs are L-risk consistent, then the empirical decision function $f_{\mathrm{D},\lambda}$ approximates the conditional quantile function.

Consider a random sample (x_i, y_i), $1 \leq i \leq n$, from independent and identically distributed random variables (X_i, Y_i) each with unknown probability distribution P on some measurable space $X \times Y$ with corresponding Borel σ-algebra. For technical reasons, we assume throughout this section that $Y \subset \mathbb{R}$ is closed and is thus Polish. Recall that in this case P can be split up into the marginal distribution P_X and the regular conditional probability $\mathrm{P}(\,\cdot\,|\,x)$, $x \in X$, on Y, which exists by Ulam's theorem (Theorem A.3.15); see also Lemma A.3.16.

The goal of quantile regression is to estimate the (set-valued) *τ-quantile function*

$$F^*_{\tau,\mathrm{P}}(x) := \big\{q \in \mathbb{R} : \mathrm{P}(Y \leq q \,|\, x) \geq \tau \text{ and } \mathrm{P}(Y \geq q \,|\, x) \geq 1-\tau\big\},\ x \in X, \tag{9.14}$$

where $\tau \in (0,1)$ is a fixed constant. For conceptual simplicity, we assume throughout this section that $F^*_{\tau,\mathrm{P}}(x)$ consists of singletons,[2] so that there exists a unique conditional quantile function $f^*_{\tau,\mathrm{P}} : X \to \mathbb{R}$ defined by $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$, $x \in X$; see Proposition 3.9. Now recall that the distance-based pinball loss function $L : Y \times \mathbb{R} \to [0,\infty)$ is given by

$$L(y,t) = \psi_\tau(r) = \begin{cases} (\tau-1)r & \text{if } r < 0, \\ \tau r & \text{if } r \geq 0, \end{cases}$$

where $r := y - t$ (see Example 2.43) has the property that $f^*_{\tau,\mathrm{P}}(x)$ is a τ-quantile, $\tau \in (0,1)$, of $\mathrm{P}(Y\,|\,x)$ for all $x \in X$ if and only if $f^*_{\tau,\mathrm{P}}(x)$ minimizes the inner L-risk of $\mathrm{P}(y|x)$,

$$\int_Y L(y, f^*_{\tau,\mathrm{P}}(x))\, d\mathrm{P}(y|x) = \inf_{q(x)\in\mathbb{R}} \int_Y L(y, q(x))\, d\mathrm{P}(y|x)\,. \tag{9.15}$$

It is easy to check that the pinball loss function has for each $\tau \in (0,1)$ the following properties (see Problem 9.1): ψ_τ is strictly convex, $\psi_\tau(0) = 0$, $\lim_{|r|\to\infty} \psi_\tau(r) = \infty$, and ψ_τ is Lipschitz continuous with Lipschitz constant $|\psi_\tau|_1 = \max\{\tau, 1-\tau\} \leq 1$. Furthermore, we have

$$\min\{\tau, 1-\tau\}\,|r| \leq \psi_\tau(r) \leq |\psi_\tau|_1\,|r|\,, \quad r \in \mathbb{R}. \tag{9.16}$$

Koenker and Bassett (1978) proposed the estimator

$$\hat{f}_\tau = \arg\inf_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \psi_\tau(y_i - x_i^\mathsf{T}\theta)$$

if $f^*_{\tau,\mathrm{P}}$ is assumed to be a linear function and $X = \mathbb{R}^d$. Schölkopf *et al.* (2000, p. 1216) and Takeuchi *et al.* (2006) proposed to relax the assumption of a linear quantile function by using the kernel trick.

Definition 9.5. *Let X be a measurable space, $Y \subset \mathbb{R}$, and $\mathrm{P} \in \mathcal{M}_1(X \times Y)$. Furthermore, let L be the pinball loss function with $\tau \in (0,1)$, H be a separable RKHS of a measurable kernel k with canonical feature map $\Phi : X \to H$, and $\lambda > 0$. A support vector machine for quantile regression is defined by*

$$f_{\mathrm{P},\lambda} := \arg\inf_{f\in H} \mathbb{E}_\mathrm{P}\, L(Y, f(X)) + \lambda\|f\|_H^2.$$

For any fixed data set $D = \{(x_i,y_i), 1 \leq i \leq n\} \subset X \times Y$, we thus obtain the estimator $f_{\mathrm{D},\lambda}$. We obtain $f_{\mathrm{D},\lambda} = \hat{f}_\tau$ if we choose the linear kernel $k(x,x') := \langle x, x'\rangle$ and $\lambda := 0$. We can now formulate our result on L-risk consistency for SVMs for quantile regression.

[2] This assumption can be relaxed; see Theorem 3.61 and Corollary 3.65.

Theorem 9.6. *Let X be a complete measurable space, $Y \subset \mathbb{R}$ be closed, L be the pinball loss with $\tau \in (0,1)$, and H be a separable RKHS of a bounded measurable kernel k on X such that H is dense in $L_1(\mu)$ for all distributions μ on X. Let (λ_n) be a sequence of strictly positive numbers with $\lambda_n \to 0$.*

i) If $\lambda_n^2 n \to \infty$, then

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}\,, \quad n \to \infty, \tag{9.17}$$

in probability for all $|D| = n$ and for all $\mathrm{P} \in \mathcal{M}_1(X \times Y)$ with $|\mathrm{P}|_1 < \infty$.

ii) If $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$, then (9.17) holds even almost surely.

We mention that Theorem 9.6(i) is a direct consequence of Theorem 9.1 for $p = 1$, but the following proof uses a better concentration inequality because the pinball loss function is Lipschitz continuous.

Proof. (i). To avoid handling too many constants, let us assume $\|k\|_\infty = 1$. This implies $\|f\|_\infty \le \|k\|_\infty \|f\|_H \le \|f\|_H$ for all $f \in H$. Now we use the Lipschitz continuity of $L_{\tau\text{-pin}}$, $|\psi_\tau|_1 \le 1$, and Lemma 2.19 to obtain

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| \le |\psi_\tau|_1 \, \|f_{\mathrm{P},\lambda_n} - g\|_H\,, \ g \in H. \tag{9.18}$$

For $n \in \mathbb{N}$ and $\lambda_n > 0$, we now write $h_{\tau,n} : X \times Y \to \mathbb{R}$ for the function h obtained by Corollary 5.11. Let $\Phi : X \to H$ be the canonical feature map. We have $f_{\mathrm{P},\lambda_n} = -(2\lambda_n)^{-1}\mathbb{E}_{\mathrm{P}} h_{\tau,n}\Phi$, and for all distributions Q on $X \times Y$ with $|\mathrm{Q}|_1 < \infty$, we have

$$\|f_{\mathrm{P},\lambda_n} - f_{\mathrm{Q},\lambda_n}\|_H \le \lambda_n^{-1} \, \|\mathbb{E}_{\mathrm{P}} h_{\tau,n}\Phi - \mathbb{E}_{\mathrm{Q}} h_{\tau,n}\Phi\|_H .$$

Note that $\|h_{\tau,n}\|_\infty \le |\psi_\tau|_1$. Moreover, let $\varepsilon \in (0,1)$ and D be a training set of n data points and empirical distribution D such that

$$\|\mathbb{E}_{\mathrm{P}} h_{\tau,n}\Phi - \mathbb{E}_{\mathrm{D}} h_{\tau,n}\Phi\|_H \le \lambda_n \varepsilon. \tag{9.19}$$

Then Corollary 5.11 gives $\|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D}_n,\lambda_n}\|_H \le \varepsilon$ and hence (9.18) yields

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n})\bigr| \le \|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D},\lambda_n}\|_H \le \varepsilon. \tag{9.20}$$

Let us now estimate the probability of D satisfying (9.19). To this end, we first observe that $\lambda_n n^{1/2} \to \infty$ implies that for all sufficiently large n we have $\lambda_n \varepsilon \ge n^{-1/2}$. Moreover, Corollary 5.11 shows $\|h_{\tau,n}\|_\infty \le 1$, and our assumption $\|k\|_\infty = 1$ thus yields $\|h_{\tau,n}\Phi\|_\infty \le 1$. Consequently, Hoeffding's inequality in Hilbert spaces (see Theorem 6.15) yields for $B = 1$ and $\xi = \frac{3}{8}\varepsilon^2\lambda_n^2 n/(\varepsilon\lambda_n + 3)$ the bound

$$\begin{aligned}
&\mathrm{P}^n\bigl(D \in (X \times Y)^n : \|\mathbb{E}_{\mathrm{P}} h_{\tau,n}\Phi - \mathbb{E}_{\mathrm{D}} h_{\tau,n}\Phi\|_H \le \lambda_n \varepsilon\bigr)\\
&\ge \mathrm{P}^n\Bigl(D \in (X \times Y)^n : \|\mathbb{E}_{\mathrm{P}} h_{\tau,n}\Phi - \mathbb{E}_{\mathrm{D}} h_{\tau,n}\Phi\|_H \le \bigl(\sqrt{2\xi} + 1\bigr)n^{-1/2} + \frac{4\xi}{3n}\Bigr)\\
&\ge 1 - \exp\Bigl(-\frac{3}{8} \cdot \frac{\varepsilon^2\lambda_n^2 n}{\varepsilon\lambda_n + 3}\Bigr)
\end{aligned} \tag{9.21}$$

for all sufficiently large values of n. Note that (9.21) is stronger than (9.13). Using $\lambda_n n^{1/2} \to \infty$, $\lambda_n \to 0$, and the same argumentation as in the proof of Theorem 9.1, we obtain that (9.20) holds true with probability tending to 1 and that $|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}}| \le \varepsilon$ for all sufficiently large n, which gives the assertion.

(ii). In order to show the second assertion, we define $\varepsilon_n := (\ln(n+1))^{-1/2}$, and $\delta_n := \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon_n$, $n \ge 1$. Moreover, for an infinite sample $D_\infty := ((x_1,y_1),(x_2,y_2),\dots) \in (X \times Y)^\infty$, we write $D_n := ((x_1,y_1),\dots,(x_n,y_n))$. With these notations, we define

$$A_n := \bigl\{ D_\infty \in (X \times Y)^\infty : \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}} > \delta_n \bigr\}, \qquad n \in \mathbb{N}.$$

Now, our estimates above together with $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$ yield

$$\sum_{n\in\mathbb{N}} \mathrm{P}^\infty(A_n) \;\le\; \sum_{n\in\mathbb{N}} \exp\left(-\frac{3}{8}\cdot\frac{\varepsilon_n^2\lambda_n^2 n}{\varepsilon_n\lambda_n + 3}\right) \;<\; \infty,$$

and hence we obtain by the Borel-Cantelli lemma (Lemma A.4.7) that

$$\mathrm{P}^\infty\bigl(\bigl\{D_\infty \in (X\times Y)^\infty \,\big|\, \exists\, n_0\, \forall\, n \ge n_0 : \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}} \le \delta_n\bigr\}\bigr) = 1.$$

The assertion follows because $\lambda_n \to 0$ implies $\delta_n \to 0$. □

In order to formulate our result on consistency of the estimated quantile function itself, we need some additional notations. Let $f, g : X \to \mathbb{R}$ be measurable functions. We write

$$\|f\|_{L_0(\mathrm{P}_X)} := \mathbb{E}_{\mathrm{P}_X} \min\{1, |f(X)|\} \tag{9.22}$$

and define $d(f,g) := \|f-g\|_{L_0(\mathrm{P}_X)}$. Note that d is a *translation-invariant* metric on the space of all measurable functions defined on X and that d describes the convergence in probability P_X (see Problem 9.2).

The next result shows that f_{D,λ_n} approximates the conditional quantile function in terms of $\|\cdot\|_{L_0(\mathrm{P}_X)}$.

Theorem 9.7. *Let X be a complete measurable space, $Y \subset \mathbb{R}$ be closed, L be the pinball loss with $\tau \in (0,1)$, H be a separable RKHS of a bounded measurable kernel k on X such that H is dense in $L_1(\mu)$ for all distributions μ on X, and (λ_n) be a sequence of strictly positive numbers with $\lambda_n \to 0$.*

i) If $\lambda_n^2 n \to \infty$, then

$$\|f_{\mathrm{D},\lambda_n} - f^*_{\tau,\mathrm{P}}\|_{L_0(\mathrm{P}_X)} \to 0 \tag{9.23}$$

in probability for $n \to \infty$ for all $\mathrm{P} \in \mathcal{M}_1(X \times Y)$ with $|\mathrm{P}|_1 < \infty$.

ii) If $\lambda_n^{2+\delta} n \to \infty$ for some $\delta > 0$, then (9.23) holds even almost surely.

Proof. (i). We have already seen in Theorem 9.6(i) that f_{D,λ_n} satisfies $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}$ in probability for $n \to \infty$. The existence of a unique minimizer $f^*_{\tau,\mathrm{P}}$ is guaranteed by the general assumption of this section, and Corollary 3.62 yields the assertion.

(ii). Combine Theorem 9.6(ii) with Corollary 3.62. □

It is interesting to note that the assumption $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$ is only needed to formulate Theorem 9.7 in terms of $\|\cdot\|_{L_0(\mathrm{P}_X)}$. However, Theorem 3.63 provides a framework to replace $\|\cdot\|_{L_0(\mathrm{P}_X)}$ by a more general notion of closeness if the assumption $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$ is violated. Note that Theorem 9.6 established for $\tau = 1/2$ the convergence in probability of

$$\mathbb{E}_{\mathrm{P}}|Y - f_{\mathrm{D}_n,\lambda_n}(X)| - \mathbb{E}_{\mathrm{P}}|Y - f^*_{\tau,\mathrm{P}}(X)| \to 0, \quad n \to \infty, \tag{9.24}$$

which naturally raises the question of whether we have the convergence of

$$\mathbb{E}_{\mathrm{P}}|f_{\mathrm{D}_n,\lambda_n}(X) - f^*_{\tau,\mathrm{P}}(X)| \to 0, \quad n \to \infty \tag{9.25}$$

in probability. Of course, the inverse triangle inequality $\big||a| - |b|\big| \le |a-b|$ immediately shows that (9.25) implies (9.24), but since for general $a, b, c \in \mathbb{R}$ the inequality $|a-c| - |b-c| \ge |a-c|$ is false, we conjecture that without additional assumptions on P the convergence in (9.25) does not follow from (9.24). However, Example 3.67 shows that we can actually replace $\|\cdot\|_{L_0(\mathrm{P}_X)}$ by some (quasi)-norm $\|\cdot\|_{L_p(\mathrm{P}_X)}$ for certain distributions P. By (3.74), we have the following inequality for distributions P of $\mathcal{Q}^{\alpha}_{\tau}$-type. Assume the function $b : X \to [0,\infty)$ defined by $b(x) := c_{\mathrm{P}(\,\cdot\,|x)}$, $x \in X$, where $c_{\mathrm{P}(\,\cdot\,|x)}$ is determined by (3.73). If b satisfies $b^{-1} \in L_p(\mathrm{P}_X)$ for some $p \in (0,\infty]$, then

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_q(\mathrm{P}_X)} \le \sqrt{2}\, \|b^{-1}\|^{1/2}_{L_p(\mathrm{P}_X)} \left(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\right)^{1/2}$$

for all functions $f : X \to \mathbb{R}$ such that $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \le 2^{-\frac{p+2}{p+1}} \alpha^{\frac{2p}{p+1}}$, where $f^*_{L,\mathrm{P}}(x) = f^*_{\tau,\mathrm{P}}$ and $q := \frac{p}{p+1}$.

Another interesting question is whether we can establish convergence rates in Theorem 9.6 or Theorem 9.7. It is well-known in statistical learning theory that such convergence rates require additional assumptions on the distribution P due to the no-free-lunch theorem (see Corollary 6.8), for example in terms of the approximation properties of H with respect to $f^*_{\tau,\mathrm{P}}$. Moreover, the techniques used in the proofs of Theorem 9.6 and Theorem 9.7 are tuned to provide consistency under rather minimal assumptions on X, Y, P, and H, but in general these techniques are too weak to obtain good convergence results. Some results on learning rates of SVMs for quantile regression are given by Steinwart and Christmann (2008).

9.4 Numerical Results for Quantile Regression

In this section, we will illustrate that SVMs for quantile regression with different values of the quantile level τ can offer valuable information that is not obtainable by just considering one regression function for the center (conditional mean or conditional median).[3] It will also be shown by a small simulation that the asymptotic results obtained in the previous section can offer reasonable approximations for small to moderate sample sizes.

[3] Portions of this section are based on material originally published in "A. Christmann and I. Steinwart (2008), 'Consistency of kernel based quantile

Example: LIDAR Data Set

Let us start with a simple example for the application of SVMs for quantile regression. We analyze data concerning the so-called LIDAR technique. LIDAR is the abbreviation of LIght Detection And Ranging. This technique uses the reflection of laser-emitted light to detect chemical compounds in the atmosphere. We consider the logarithm of the ratio of light received from two laser sources as the response variable $Y = \texttt{logratio}$, whereas the single explanatory variable $X = \texttt{range}$ is the distance traveled before the light is reflected back to its source. We refer to Ruppert *et al.* (2003) for more details on this data set.

A scatterplot of the data set consisting of $n = 221$ observations is shown in Figure 9.1 together with the fitted curves based on SVMs using the pinball loss function using the Gaussian RBF kernel for the median and the lower and upper 5 percent quantiles. The KBQR clearly shows that the relationship between both variables is non-linear, almost constant for values of `range` below 550 and decreasing for higher values of `range`. However, KBQR also shows that the variability of `logratio` is non-constant and much greater for values of `range`, say, above 600 than for values below this.

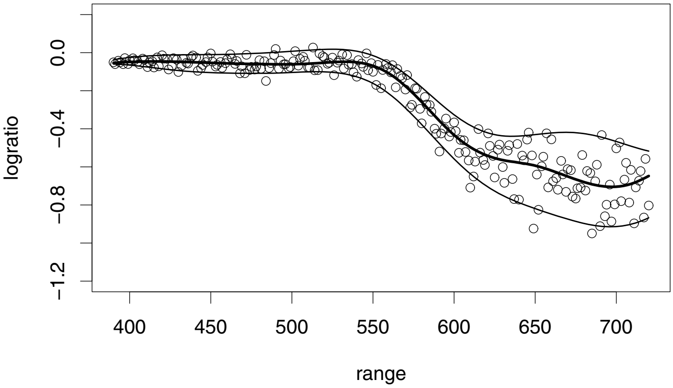

Fig. 9.1. LIDAR data set ($n = 221$). SVMs for quantile regression based on the Gaussian RBF kernel with $\gamma^2 = 0.5$ and $\lambda = \frac{1}{700}n^{-1/3}$ (resulting from a grid search). Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

Simulation Results

Now we describe a small simulation and its results to investigate how well the asymptotic results derived in Section 9.3 on the consistency of SVMs for

regression.' *Appl. Stoch. Models Bus. Ind.*, **24**, 171–183.

quantile regression work for small to moderate sample sizes. We consider $n \in \{221, 1000, 4000\}$ and use the same parameter settings for the Gaussian RBF kernel as in the previous subsection; i.e., $\gamma^2 = 0.5$ and $\lambda_n = \frac{1}{700} n^{-1/3}$. The number of replications in the simulation was set to 1000. For each replication $\ell \in \{1, \ldots, 1000\}$, we independently generated n data points $x_i^{(\ell)}$ for `range` from a continuous uniform distribution with support $(390, 720)$. Furthermore, for each replication, we generated n data points $y_i^{(\ell)}$ for `logratio` according to independent normal distributions with conditional expectations

$$\mu(x_i^{(\ell)}) := -0.05 - 0.7\big(1 + \exp(-(x_i^{(\ell)} - 600)/10)\big)^{-1}$$

and conditional variances

$$\sigma^2(x_i^{(\ell)}) := \big(0.01 + 0.1\big(1 + \exp(-(x_i^{(\ell)} - 600)/50)\big)\big)^2 ,$$

respectively. The true conditional τ-quantile curves are thus given by $f^*_{\tau,\mathrm{P}}(x) = \mu(x) + u_\tau \sigma(x)$, $\tau \in (0,1)$, where u_τ defines the τ-quantile of a normal distribution with mean 0 and variance 1. The curves for the conditional medians and the conditional lower and upper 5 percent quantiles are shown in Figure 9.2 to illustrate that this model generates data sets similar to the LIDAR data set; see Figure 9.1.

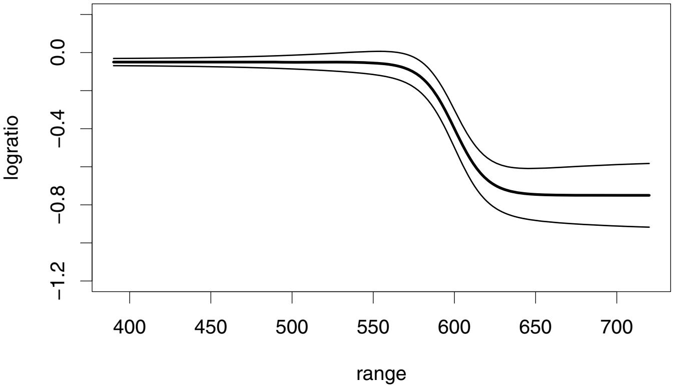

Fig. 9.2. True quantile regression curves for the simulation. Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

We use two criteria to measure how well the KBQR estimates approximate the true conditional quantiles. Our first criterion is

$$\mathrm{IMSE}_\tau := \frac{1}{1000} \sum_{\ell=1}^{1000} \frac{1}{n} \sum_{i=1}^{n} \Big(f_{\mathrm{D}_n,\lambda_n}(x_i^{(\ell)}) - f^*_{\tau,\mathrm{P}}(x_i^{(\ell)}) \Big)^2 ,$$

which is an empirical version of the integrated mean squared error. To measure the worst-case behavior of the KBQR estimates, we use the criterion

$$\mathrm{mBias}_\tau := \frac{1}{1000} \sum_{\ell=1}^{1000} \max_{1 \le i \le n} \left| f_{\mathrm{D}_n, \lambda_n}(x_i^{(\ell)}) - f^*_{\tau,\mathrm{P}}(x_i^{(\ell)}) \right| ,$$

which is an empirical version of the maximum bias. The random number generation and the plots were made with the statistical software `R` (R Development Core Team, 2006). The program `mySVM` (Rüping, 2000) was used for the computation of the KBQR estimates.

The boxplots[4] given in Figures 9.3 and 9.4 show that SVMs based on the pinball loss function perform quite well with respect to both criteria under the circumstances considered, because both criteria have relatively small values and their values decrease with increasing sample sizes. A considerable improvement is obtained by increasing the sample size by a factor of around 4. The boxplots also show that the variability of the estimated conditional median is considerably smaller than the variability of the estimated conditional quantiles for $\tau \in \{0.05, 0.95\}$. The simulations indicate that the consistency results derived in Section 9.3 can be useful even for moderate sample sizes.

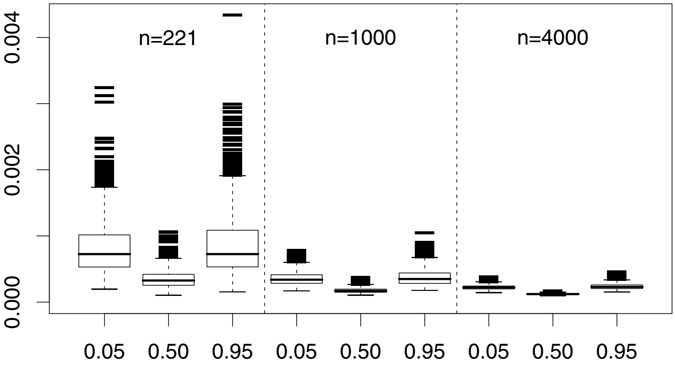

Fig. 9.3. Simulation results for the criterion IMSE_τ for SVMs for quantile regression based on the Gaussian RBF kernel with $\gamma^2 = 0.5$ and $\lambda_n = \frac{1}{700} n^{-1/3}$. Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

[4] The box is defined by the 0.25 and 0.75 quantiles, and the median is the line inside the box. The whiskers give additional information about the variability. Values outside the whiskers are shown as small lines.

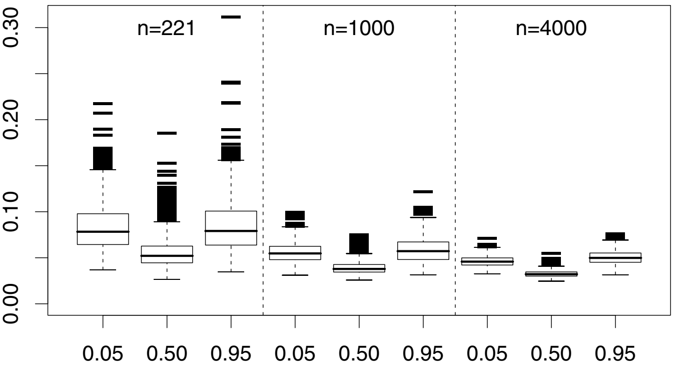

Fig. 9.4. Simulation results for the criterion mBias_τ for SVMs for quantile regression based on the Gaussian RBF kernel with $\gamma^2 = 0.5$ and $\lambda_n = \frac{1}{700}n^{-1/3}$. Considered quantile levels: $\tau = 0.05, 0.50$, and 0.95.

9.5 Median Regression with the eps-Insensitive Loss (*)

In this section, we give simple conditions for the distribution P that guarantee that the set of exact minimizers of support vector machines based on the ϵ-insensitive loss function contains only one function. This fact is at first glance surprising because this loss function equals zero in the entire interval $[-\epsilon, +\epsilon]$.

In this section, we will assume that P is a distribution on $X \times Y$, where X is an arbitrary set and $Y \subset \mathbb{R}$ is closed. Further, we assume that the σ-algebra on X is complete with respect to the marginal distribution P_X of P; i.e., every subset of a P_X-zero set is contained in the σ-algebra. Since the latter can always be ensured by increasing the original σ-algebra in a suitable manner, we note that the assumption of a complete σ-algebra on X is no restriction at all. Recall from Section 3.7 that a distribution $\mathrm{Q} \in \mathcal{M}_1(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is symmetric, if there exists some $c \in \mathbb{R}$ such that $\mathrm{Q}(c + A) = \mathrm{Q}(c - A)$ for all $A \in \mathcal{B}(\mathbb{R})$ with $A \subset [0, \infty)$.

Theorem 9.8. *Let* P *be a distribution on* $X \times \mathbb{R}$ *that has a unique median* $f^*_{1/2,\mathrm{P}}$. *Further, assume that all conditional distributions* $\mathrm{P}(\,\cdot\,|\,x)$, $x \in X$, *are atom-free and symmetric. If for an* $\epsilon > 0$ *the conditional distributions have a positive mass for intervals around* $f^*_{1/2,\mathrm{P}} \pm \epsilon$, *then* $f^*_{1/2,\mathrm{P}}$ *is the only minimizer of* $\mathcal{R}_{L,\mathrm{P}}(\cdot)$ *where* L *is the* ϵ*-insensitive loss.*

The proof of Theorem 9.8 follows immediately from the following lemma.

Lemma 9.9. *Let* Q *be a symmetric, atom-free distribution on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *with median* $q^* = 0$. *Then, for* $\epsilon > 0$ *and* L *the* ϵ*-insensitive loss, we have*

$$\mathcal{C}_{L,\mathrm{Q}}(0) = \mathcal{C}^*_{L,\mathrm{Q}} = 2\int_\epsilon^\infty \mathrm{Q}([s, \infty)),$$

and if $\mathcal{C}_{L,\mathrm{Q}}(0) < \infty$, *we further have*

$$\begin{aligned} \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(0) &= \int_{\epsilon-t}^{\epsilon} \mathrm{Q}([s,\epsilon])\,ds + \int_{\epsilon}^{\epsilon+t} \mathrm{Q}([\epsilon,s])\,ds, && \text{if } t \in [0,\epsilon], \\ \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(\epsilon) &= \int_{0}^{t-\epsilon} \mathrm{Q}([s,\infty))\,ds - \int_{2\epsilon}^{\epsilon+t} \mathrm{Q}([s,\infty))\,ds \\ &\quad +2\int_0^{t-\epsilon} \mathrm{Q}([0,s])\,ds \geq 0, && \text{if } t > \epsilon. \end{aligned}$$

In particular, if $Q([\epsilon-\delta,\epsilon+\delta]) = 0$ *for some* $\delta > 0$, *then* $\mathcal{C}_{L,\mathrm{Q}}(\delta) = \mathcal{C}^*_{L,\mathrm{Q}}$.

Proof. Because $L(y,t) = L(-y,-t)$ for all $y,t \in \mathbb{R}$, we only have to consider $t \geq 0$. Recall that, given a distribution Q on $\mathbb{R}$ and a *non-negative* measurable function $g : X \to [0,\infty)$, we have

$$\int_{\mathbb{R}} g\,d\mathrm{Q} = \int_0^\infty \mathrm{Q}(g \geq s)\,ds\,; \tag{9.26}$$

see Lemma A.3.11. For later use, we note that for $0 \leq a \leq b \leq \infty$ equation (9.26) yields

$$\int_a^b y\,d\mathrm{Q}(y) = a\mathrm{Q}([a,b]) + \int_a^b \mathrm{Q}([s,b])\,ds\,. \tag{9.27}$$

Moreover, the definition of L implies

$$\mathcal{C}_{L,\mathrm{Q}}(t) = \int_{-\infty}^{t-\epsilon} t - y - \epsilon\,d\mathrm{Q}(y) + \int_{t+\epsilon}^{\infty} y - \epsilon - t\,d\mathrm{Q}(y)\,.$$

Using the symmetry of Q yields

$$-\int_{-\infty}^{t-\epsilon} y\,d\mathrm{Q}(y) = \int_{\epsilon-t}^{\infty} y\,d\mathrm{Q}(y),$$

and hence we obtain

$$\begin{aligned} \mathcal{C}_{L,\mathrm{Q}}(t) &= \int_0^{t-\epsilon} \mathrm{Q}((-\infty,t-\epsilon])\,ds - \int_0^{t+\epsilon} \mathrm{Q}([t+\epsilon,\infty))\,ds \\ &\quad + \int_{\epsilon-t}^{t+\epsilon} y\,d\mathrm{Q}(y) + 2\int_{t+\epsilon}^{\infty} y\,d\mathrm{Q}(y)\,. \end{aligned} \tag{9.28}$$

Let us first consider the case $t \geq \epsilon$. Then the symmetry of Q yields

$$\int_{\epsilon-t}^{t+\epsilon} y\,d\mathrm{Q}(y) = \int_{t-\epsilon}^{t+\epsilon} y\,d\mathrm{Q}(y),$$

and hence (9.27) implies

$$
\begin{aligned}
\mathcal{C}_{L,\mathrm{Q}}(t) = & \int_0^{t-\epsilon} \mathrm{Q}([\epsilon - t, \infty))ds + \int_0^{t-\epsilon} \mathrm{Q}([t-\epsilon, t+\epsilon])\, ds + \int_{t-\epsilon}^{t+\epsilon} \mathrm{Q}([s, t+\epsilon])\, ds \\
& +2\int_{t+\epsilon}^{\infty} \mathrm{Q}([s, \infty))\, ds + \int_0^{t+\epsilon} \mathrm{Q}([t+\epsilon, \infty))\, ds.
\end{aligned}
$$

Using

$$
\int_{t-\epsilon}^{t+\epsilon} \mathrm{Q}([s, t+\epsilon))\, ds = \int_0^{t+\epsilon} \mathrm{Q}([s, t+\epsilon))\, ds - \int_0^{t-\epsilon} \mathrm{Q}([s, t+\epsilon))\, ds,
$$

we further obtain

$$
\begin{aligned}
& \int_{t-\epsilon}^{t+\epsilon} \mathrm{Q}([s, t+\epsilon))\, ds + \int_0^{t+\epsilon} \mathrm{Q}([t+\epsilon, \infty))\, ds + \int_{t+\epsilon}^{\infty} \mathrm{Q}([s, \infty))\, ds \\
= & \int_0^{\infty} \mathrm{Q}([s, \infty))\, ds - \int_0^{t-\epsilon} \mathrm{Q}([s, t+\epsilon))\, ds.
\end{aligned}
$$

From this and

$$
\int_0^{t-\epsilon} \mathrm{Q}([t-\epsilon, t+\epsilon])\, ds - \int_0^{t-\epsilon} \mathrm{Q}([s, t+\epsilon])\, ds = -\int_0^{t-\epsilon} \mathrm{Q}([s, t-\epsilon])\, ds,
$$

it follows that $\mathcal{C}_{L,\mathrm{Q}}(t)$ equals

$$
-\int_0^{t-\epsilon} \mathrm{Q}([s, t-\epsilon])\, ds + \int_0^{t-\epsilon} \mathrm{Q}([\epsilon - t, \infty))\, ds + \int_{t+\epsilon}^{\infty} \mathrm{Q}([s, \infty))\, ds + \int_0^{\infty} \mathrm{Q}([s, \infty))\, ds.
$$

The symmetry of Q implies

$$
\int_0^{t-\epsilon} \mathrm{Q}([\epsilon - t, t-\epsilon])\, ds = 2\int_0^{t-\epsilon} \mathrm{Q}([0, t-\epsilon])\, ds,
$$

such that

$$
\begin{aligned}
& -\int_0^{t-\epsilon} \mathrm{Q}([s, t-\epsilon])\, ds + \int_0^{t-\epsilon} \mathrm{Q}([\epsilon - t, \infty))\, ds \\
= & \; 2\int_0^{t-\epsilon} \mathrm{Q}([0, s))\, ds + \int_0^{t-\epsilon} \mathrm{Q}([s, \infty))\, ds.
\end{aligned}
$$

This and

$$
\int_{t+\epsilon}^{\infty} \mathrm{Q}([s, \infty))\, ds + \int_0^{\infty} \mathrm{Q}([s, \infty))\, ds = 2\int_{t+\epsilon}^{\infty} \mathrm{Q}([s, \infty))\, ds + \int_0^{t+\epsilon} \mathrm{Q}([s, \infty))\, ds
$$

shows that $\mathcal{C}_{L,\mathrm{Q}}(t)$ equals

$$
2\int_0^{t-\epsilon} \mathrm{Q}([0, s))\, ds + \int_0^{t-\epsilon} \mathrm{Q}([s, \infty))\, ds + 2\int_{t+\epsilon}^{\infty} \mathrm{Q}([s, \infty))\, ds + \int_0^{t+\epsilon} \mathrm{Q}([s, \infty))\, ds.
$$

By

$$\int_0^{t-\epsilon} Q([s,\infty))\,ds + \int_0^{t+\epsilon} Q([s,\infty))\,ds = 2\int_0^{t-\epsilon} Q([s,\infty))\,ds + \int_{t-\epsilon}^{t+\epsilon} Q([s,\infty))\,ds,$$

we obtain

$$\mathcal{C}_{L,\mathrm{Q}}(t) = 2\int_0^{t-\epsilon} Q([0,\infty))\,ds + 2\int_{t+\epsilon}^{\infty} Q([s,\infty))\,ds + \int_{t-\epsilon}^{t+\epsilon} Q([s,\infty))\,ds$$

if $t \geq \epsilon$. Let us now consider the case $t \in [0,\epsilon]$. Analogously, we get from (9.28) that $\mathcal{C}_{L,\mathrm{Q}}(t)$ equals

$$\begin{aligned}&\int_0^{\epsilon-t} Q([\epsilon-t,t+\epsilon])\,ds + \int_{\epsilon-t}^{\epsilon+t} Q([s,t+\epsilon])\,ds + 2\int_{\epsilon+t}^{\infty} Q([s,\infty))\,ds\\ &+2\int_0^{\epsilon+t} Q([\epsilon+t,\infty))\,ds - \int_0^{\epsilon-t} Q([\epsilon-t,\infty))\,ds - \int_0^{\epsilon+t} Q([\epsilon+t,\infty))\,ds.\end{aligned}$$

Combining this with

$$\int_0^{\epsilon-t} Q([\epsilon-t,t+\epsilon])\,ds - \int_0^{\epsilon-t} Q([\epsilon-t,\infty))\,ds = -\int_0^{\epsilon-t} Q([\epsilon+t,\infty))\,ds$$

and

$$\int_0^{\epsilon+t} Q([\epsilon+t,\infty))\,ds - \int_0^{\epsilon-t} Q([\epsilon+t,\infty))\,ds = \int_{\epsilon-t}^{\epsilon+t} Q([\epsilon+t,\infty))\,ds,$$

we get

$$\begin{aligned}\mathcal{C}_{L,\mathrm{Q}}(t) &= \int_{\epsilon-t}^{\epsilon+t} Q([\epsilon+t,\infty))\,ds + \int_{\epsilon-t}^{\epsilon+t} Q([s,t+\epsilon])\,ds + 2\int_{\epsilon+t}^{\infty} Q([s,\infty))\,ds\\ &= \int_{\epsilon-t}^{\epsilon+t} Q([s,\infty))\,ds + 2\int_{\epsilon+t}^{\infty} Q([s,\infty))\,ds\\ &= \int_{\epsilon-t}^{\infty} Q([s,\infty))\,ds + \int_{\epsilon+t}^{\infty} Q([s,\infty))\,ds.\end{aligned}$$

Hence

$$\mathcal{C}_{L,\mathrm{Q}}(0) = 2\int_{\epsilon}^{\infty} Q([s,\infty))\,ds.$$

The expressions for $\mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(0)$, $t \in (0,\epsilon]$, and $\mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(\epsilon)$, $t > \epsilon$, given in Lemma 9.9 follow by using the same arguments. Hence one exact minimizer of $\mathcal{C}_{L,\mathrm{Q}}(\cdot)$ is the median $t^* = 0$. The last assertion is a direct consequence of the formula for $\mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}_{L,\mathrm{Q}}(0)$ in the case $t \in (0,\epsilon]$. □

9.6 Further Reading and Advanced Topics

For additional information, we refer to Poggio and Girosi (1990), Wahba (1990), Vapnik (1995, 1998), Schölkopf and Smola (2002), and the references cited by these authors. For ν-support vector regression, which is strongly related to support vector regression based on the ϵ-insensitive loss function but allows the tube width to adapt automatically to the data, see Schölkopf *et al.* (2000), Smola and Schölkopf (2004), and Chen *et al.* (2005). For a brief description of kernel ridge regression and Gaussian processes, see Cristianini and Shawe-Taylor (2000, Section 6.2). We refer to Wahba (1999) for the relationship between SVMs and Gaussian processes.

Section 9.2 on the L-risk consistency of SVMs for regression was based on Christmann and Steinwart (2007). An unbounded output space Y instead of a bounded one makes proofs of L-risk consistency of SVMs harder. This is also true for investigating robustness properties; see Chapter 10. This was one reason why we treated Nemitski loss functions in Chapters 2 and 5. Equation (9.4) shows that the reproducing kernel Hilbert space H is a Banach space with type and cotype two. For details, we refer to Diestel *et al.* (1995).

Quantile regression for linear models was proposed by Koenker and Bassett (1978). A recent textbook on this topic is Koenker (2005). We refer to Koenker (1986) for strong consistency of regression quantiles and related empirical processes, He and Liang (2000) for quantile regression in errors-in-variables models, Portnoy (2003) for censored regression quantiles, and Koenker and Xiao (2006) for quantile autoregression. SVMs for quantile regression were proposed by Schölkopf *et al.* (2000, p. 1216) and Takeuchi *et al.* (2006). The latter article also describes algorithmic aspects for the efficient computation of SVMs for quantile regression. It is possible that the estimated conditional quantile functions intersect for different values of $\tau \in (0,1)$. For a discussion of this crossing problem and how to overcome this unfavorable property, we refer to Example 11.9, He (1997), and Takeuchi *et al.* (2006). For non-parametric generalizations of quantile regression based on splines, we refer to Koenker *et al.* (1994) and He and Ng (1999).

Sections 9.3 and 9.4 on the L-risk consistency of SVMs for quantile regression are based on Christmann and Steinwart (2008). We conjecture[5] that it is possible to get rid of the assumption $|\mathrm{P}|_1 < \infty$ if one changes the regularized minimization problem to

$$f_{\mathrm{P},\lambda} := \arg \inf_{f \in H} \mathbb{E}_{\mathrm{P}} L^*(Y, f(X)) + \lambda \|f\|_H^2 ,$$

where $L^*(y,t) := L(y,t) - L(y,0)$. However, loss functions that can take on negative values are not treated in this textbook because many practitioners do not accept negative losses.

Section 9.5 on the uniqueness of the SVM solution based on the ϵ-insensitive loss function is based on Steinwart and Christmann (2008).

[5] As a result of discussions with Ursula Gather and Xuming He

For non-parametric regression with constraints such as monotonicity or convexity, we refer to Smola and Schölkopf (1998), Takeuchi *et al.* (2006), Hall and Huang (2001), and Dette *et al.* (2006).

9.7 Summary

This chapter gave a mathematical justification for the informal notion that support vector machines are "able to learn" in non-parametric regression models. It was shown that SVMs are L-risk consistent for a broad class of convex loss functions under weak assumptions even for the case of an unbounded output space. Support vector machines based on the pinball loss function lead to kernel-based quantile regression. SVMs based on this loss function are L-risk consistent under weak assumptions on P and k. Furthermore, if this SVM is L-risk consistent, we also obtained a consistent estimator for the conditional quantile function. Finally, conditions were derived under which SVMs based on the ϵ-insensitive loss function, which is equal to zero in the interval $[-\epsilon, +\epsilon]$, allow a *unique* estimation of the conditional median function.

9.8 Exercises

9.1. Pinball loss ($\star$)
Prove that the pinball loss function $L_{\tau\text{-pin}}(y-t) = \psi_\tau(y-t)$, $y, t \in \mathbb{R}$, has for each $\tau \in (0,1)$ the following properties.

i) ψ_τ is strictly convex and satisfies both $\psi_\tau(0) = 0$ and $\lim_{|r|\to\infty} \psi_\tau(r) = \infty$.
ii) ψ_τ is Lipschitz continuous with Lipschitz constant $|\psi_\tau|_1 = \max\{\tau, 1-\tau\}$.
iii) For all $r \in \mathbb{R}$, we have $\min\{\tau, 1-\tau\}\,|r| \le \psi_\tau(r) \le |\psi_\tau|_1\,|r|$.

9.2. Translation-invariant metric ($\star\star$)
Let P be a distribution on $X \times Y$ with $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ both closed sets. Let $f, g : X \to \mathbb{R}$ be measurable functions. Define

$$\|f\|_{L_0(\mathrm{P}_X)} := \|f\|_0 := \mathbb{E}_{\mathrm{P}_X} \min\{1, |f(X)|\}$$

and $d(f,g) := \|f-g\|_0$; see (9.22). Show that d is a *translation-invariant* metric on the space of all measurable functions defined on X and d describes the convergence in probability P_X.

Hint: Use Chebyshev's inequality.

9.3. Least squares loss ($\star\star$)
Specialize Theorem 9.1 for the least squares loss function. Investigate in which sense the conditional mean function is approximated.

9.4. Comparison least squares loss and logistic loss $(\star\star)$
Compare the results of the previous exercise concerning L_{LS} with the results for the logistic loss function for the special case of a symmetric conditional distribution of Y given x. Consider the bounded and the unbounded cases.

9.5. Consistency for Lipschitz-continuous loss functions $(\star\star)$
Generalize Theorem 9.6 based on the pinball loss function $L_{\tau\text{-pin}}$ to general Lipschitz-continuous loss functions.

9.6. Quantile regression $(\star\star\star\star)$
Generalize the results of Section 9.3 to the case of non-unique quantiles.

9.7. A variance bound for the pinball loss $(\star\star\star)$
For fixed $\tau \in (0,1)$, let L be the τ-pinball loss. Moreover, let $Y := [-1,1]$, X be a measurable space, and P be a distribution on $X \times Y$ such that $\mathrm{P}(\,\cdot\,|x)$ is of type $\mathcal{Q}_\tau^\alpha$ for some $\alpha > 0$ and all $x \in X$. In other words, (3.73) is satisfied for $\mathrm{Q} := \mathrm{P}(\,\cdot\,|x)$, $x \in X$. We write $b(x) := c_{\mathrm{P}(\,\cdot\,|x)}$, $x \in X$, for the corresponding constants in (3.73) and assume that $b \in L_q(\mathrm{P}_X)$ for some $q \in (0,\infty]$. Show that there exists a constant $V \geq 1$ such that

$$\mathbb{E}_{\mathrm{P}}(L\circ \widehat{f} - L\circ f^*_{\tau,\mathrm{P}})^2 \leq V\cdot\big(\mathbb{E}_{\mathrm{P}}(L\circ \widehat{f} - L\circ f^*_{\tau,\mathrm{P}})\big)^{\vartheta} \tag{9.29}$$

for all measurable $f : X \to \mathbb{R}$, where $\vartheta := q/(2q+2)$, $\widehat{t} := \max\{-1,\min\{1,t\}\}$ denotes the clipping operation (2.14) at $M := 1$, and $L\circ f(x,y) := L(y,f(x))$.

Hint: For clipped functions with “small” excess risk, use the Lipschitz-continuity and (3.74), whereas for clipped functions with “large” excess risk, use the fact that the left-hand side of (9.29) is never larger than 4.

9.8. Oracle inequality for quantile regression $(\star\star\star)$
For fixed $\tau \in (0,1)$, let L be the τ-pinball loss. Moreover, let $Y := [-1,1]$, X be a measurable space, P be a distribution on $X \times Y$, and H be the separable RKHS of a measurable kernel. Assume that (9.29) holds and that there are constants $p \in (0,1)$ and $a \geq 1$ such that the entropy numbers satisfy

$$e_i(\mathrm{id} : H \to L_2(\mathrm{P}_X)) \leq a\, i^{-1/(2p)}\,, \qquad i \geq 1.$$

Show that there is a constant K only depending on p, ϑ, and V such that for all $\tau, \lambda > 0$, and $n \geq 1$ we have with probability not less than $1 - 3e^{-\tau}$ that

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} \leq 9A_2(\lambda) + \frac{15\tau}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}} + K\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{c(p,\vartheta)} + K\tau n^{-\frac{1}{2-\vartheta}}\,,$$

where $c(p,\vartheta) := 1/(2 - p - \vartheta + \vartheta p)$ and $A_2(\lambda)$ denotes the approximation error function from Definition 5.14. Use this oracle inequality to derive L-risk learning rates for SVMs based on the pinball loss in the sense of the discussion following Theorem 7.23. Moreover, interpret these rates with the help of (3.74) if P satisfies the assumptions of Exercise 9.7, and discuss the adaptivity of a TV-SVM. Finally, apply the discussion in Section 5.6 if H is a Sobolev space and $f^*_{\tau,\mathrm{P}}$ is contained in another Sobolev space.

Hint: Use Theorem 7.23 together with Corollary 7.31.