How SVMs can estimate quantiles and the median

Ingo Steinwart
Information Sciences Group CCS-3
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
ingo@lanl.gov

Andreas Christmann
Department of Mathematics
Vrije Universiteit Brussel
B-1050 Brussels, Belgium
andreas.christmann@vub.ac.be

Abstract

We investigate kernel-based quantile regression based on the pinball loss and support vector regression based on the ε-insensitive loss. Conditions are given which quarantee that the set of exact minimizers contains only one function. Some results about oracle inequalities and learning rates of these methods are presented.

1 Introduction

Let P be a distribution on $X \times Y$, where X is an arbitrary set and $Y \subset \mathbb{R}$ is closed. The goal of quantile regression is then to estimate the conditional quantile, *i.e.*, the set valued function

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

where $\tau \in (0, 1)$ is a fixed constant and $\mathrm{P}(\,\cdot\,|\,x)$, $x \in X$, is the (regular) conditional probability. For conceptual simplicity (though mathematically this is not necessary) we assume throughout this paper that $F^*_{\tau,\mathrm{P}}(x)$ consists of singletons, i.e., there exists a function $f^*_{\tau,\mathrm{P}} : X \to \mathbb{R}$, called the conditional τ-quantile function, such that $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$, $x \in X$. Let us now consider the so-called τ*-pinball loss function* $L_\tau : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ defined by $L_\tau(y, t) := \psi_\tau(y - t)$, where $\psi_\tau(r) = (\tau - 1)r$, if $r < 0$, and $\psi_\tau(r) = \tau r$, if $r \geq 0$. Moreover, given a (measurable) function $f : X \to \mathbb{R}$ we define the L_τ-risk of f by $\mathcal{R}_{L_\tau,\mathrm{P}}(f) := \int_{X\times Y} L_\tau(y, f(x))\, d\mathrm{P}(x, y)$. Now recall that $f^*_{\tau,\mathrm{P}}$ minimizes the L_τ-risk, i.e. $\mathcal{R}_{L_\tau,\mathrm{P}}(f^*_{\tau,\mathrm{P}}) = \inf \mathcal{R}_{L_\tau,\mathrm{P}}(f) =: \mathcal{R}^*_{L_\tau,\mathrm{P}}$, where the infimum is taken over all measurable functions $f : X \to \mathbb{R}$. Based on this observation several estimators minimizing a (modified) empirical L_τ-risk were proposed (see [5] for a survey on both parametric and non-parametric methods) for situations where P is unknown, but i.i.d. samples $D := ((x_1, y_1), \ldots, (x_n, y_n))$ drawn from P are given. In particular, [6, 4, 10] proposed a support vector machine approach that finds a solution $f_{D,\lambda} \in H$ of

$$\arg\min_{f\in H} \lambda\|f\|^2_H + \frac{1}{n}\sum_{i=1}^{n} L_\tau(y_i, f(x_i))\,, \tag{1}$$

where $\lambda > 0$ is a regularization parameter and H is a reproducing kernel Hilbert space (RKHS) over X. Note that this optimization problem can be solved by considering the dual problem [4, 10], but since this technique is nowadays standard in machine learning we omit the details. Moreover, [10] contains an exhaustive empirical study as well some theoretical considerations.

Empirical methods estimating quantiles with the help of the pinball loss typically obtain functions f_D for which $\mathcal{R}_{L_\tau,\mathrm{P}}(f_D)$ is close to $\mathcal{R}^*_{L_\tau,\mathrm{P}}$ with high probability. However, in general this only implies that f_D is close to $f^*_{\tau,\mathrm{P}}$ in a very weak sense (see [7, Remark 3.18]), and hence there is so far only little justification for using f_D as an estimate of the quantile function. Our goal is to address this issue by showing that under certain realistic assumptions on P we have an inequality of the form

$$\|f - f^*_{\tau,\mathrm{P}}\|_{L_1(\mathrm{P}_X)} \leq c_{\mathrm{P}} \sqrt{\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}}\,. \tag{2}$$

We then use this inequality to establish an oracle inequality for SVMs defined by (1). In addition, we illustrate how this oracle inequality can be used to obtain learning rates and to justify a data-dependent method for finding the hyper-parameter λ and H. Finally, we generalize the methods for establishing (2) to investigate the role of ϵ in the ϵ-insensitive loss used in standard SVM regression.

2 Main results

In the following X is an arbitrary, non-empty set equipped with a σ-algebra, and $Y \subset \mathbb{R}$ is a closed non-empty set. Given a distribution P on $X \times Y$ we further assume throughout this paper 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 this is not a restriction at all.

Definition 2.1 *A distribution Q on $\mathbb{R}$ is said to have a τ-quantile of type $\alpha > 0$ if there exists a τ-quantile $t^* \in \mathbb{R}$ and a constant $c_{\mathrm{Q}} > 0$ such that for all $s \in [0, \alpha]$ we have*

$$\mathrm{Q}\big((t^*, t^* + s)\big) \geq c_{\mathrm{Q}}\, s \qquad \textit{and} \qquad \mathrm{Q}\big((t^* - s, t^*)\big) \geq c_{\mathrm{Q}}\, s\,. \tag{3}$$

It is not difficult to see that a distribution Q having a τ-quantile of some type α has a unique τ-quantile t^*. Moreover, distributions Q having a Lebesgue density h_{Q} have a τ-quantile of type α if h_{Q} is bounded away from zero on $[t^* - \alpha, t^* + \alpha]$. In this case we can use $c_{\mathrm{Q}} := \inf\{h_{\mathrm{Q}}(t) : t \in [t^* - \alpha, t^* + \alpha]\}$ in (3). These assumptions are general enough to cover many distributions used in parametric statistics. Examples are Gaussian, Student's t, and logistic distributions (with $Y = \mathbb{R}$), Gamma and log-normal distributions (with $Y = [0, \infty)$), and uniform and Beta distributions distributions (with $Y = [0, 1]$).

The following definition describes distributions on $X \times Y$ whose conditional distributions $\mathrm{P}(\,\cdot\,|x)$, $x \in X$, have the same τ-quantile type α.

Definition 2.2 *Let $p \in (0, \infty]$, $\tau \in (0, 1)$, and $\alpha > 0$ be real numbers. A distribution P on $X \times Y$ is said to have a p-average τ-quantile of type α, if $\mathrm{P}(\,\cdot\,|x)$ is P_X-almost surely of τ-quantile type α and the function $b : X \to (0, \infty)$ is defined by $b(x) := c_{\mathrm{P}(\,\cdot\,|x)}$, where $c_{\mathrm{P}(\,\cdot\,|x)}$ is the constant in (3) satisfies $b^{-1} \in L_p(\mathrm{P}_X)$.*

Now we give some examples for distributions having p-average τ-quantiles of type α.

Example 2.3 *Let P be a distribution on $X \times \mathbb{R}$ with marginal distribution P_X and regular conditional probability $\mathrm{Q}_x\big((-\infty, y]\big) := 1/(1+e^{-z})$, $y \in \mathbb{R}$, where $z := \big(y - m(x)\big)/\sigma(x)$, $m : X \to \mathbb{R}$ describes a location shift, and $\sigma : X \to [\beta,\, 1/\beta]$ describes a scale modification for some constant $\beta \in (0, 1]$. Let us further assume that the functions m and σ are continuous. Thus Q_x is a logistic distribution having a positive and bounded Lebesgue density $h_{\mathrm{Q}_x}(y) = e^{-z}/(1 + e^{-z})^2$, $y \in \mathbb{R}$. The τ-quantile function is $t^*(x) := f^*_{\tau,\mathrm{Q}_x} = m(x) + \sigma(x)\log(\frac{\tau}{1-\tau})$, $x \in X$, and we can choose $c_{\mathrm{Q}_x} = \inf\{h_{\mathrm{Q}_x}(t) : t \in [t^*(x) - \alpha, t^*(x) + \alpha]\}$. Note that $h_{\mathrm{Q}_x}(m(x) + y) = h_{\mathrm{Q}_x}(m(x) - y)$ for all $y \in \mathbb{R}$, and $h_{\mathrm{Q}_x}(y)$ is strictly decreasing for $y \in [m(x), \infty)$. Some algebra gives $c_{\mathrm{Q}_x} = \min\big\{h_{\mathrm{Q}_x}(t^*(x) - \alpha), h_{\mathrm{Q}_x}(t^*(x) + \alpha)\big\} = \min\big\{\frac{u_1(x)}{(1+u_1(x))^2}\,,\, \frac{u_2(x)}{(1+u_2(x))^2}\big\} \in \big(c_{\alpha,\beta}\,,\, \frac{1}{4}\big)$, where $u_1(x) := \frac{1-\tau}{\tau} e^{-\alpha/\sigma(x)}$, $u_2(x) := \frac{1-\tau}{\tau} e^{\alpha/\sigma(x)}$ and $c_{\alpha,\beta} > 0$ can be chosen independent of x, because $\sigma(x) \in [\beta, 1/\beta]$. Hence $b^{-1} \in L_\infty(\mathrm{P}_X)$ and P has ∞-average τ-quantile of type α.*

Example 2.4 *Let $\tilde{\mathrm{P}}$ be a distribution on $X \times Y$ with marginal distribution $\tilde{\mathrm{P}}_X$ and regular conditional probability $\tilde{\mathrm{Q}}_x := \tilde{\mathrm{P}}(\cdot\,|\,x)$ on Y. Further assume that $\tilde{\mathrm{Q}}_x$ is $\tilde{\mathrm{P}}_X$-surely of τ-quantile type $\alpha > 0$, $\tau \in (0, 1)$. Let us now define a family of distributions P with marginal distribution $\tilde{\mathrm{P}}_X$ and regular conditional distributions $\mathrm{Q}_x := \tilde{\mathrm{P}}\big((\cdot - m(x))/\sigma(x)\big|\,x\big)$, $x \in X$, where $m : X \to \mathbb{R}$ describes a location shift and $\sigma : X \to (\beta,\, 1/\beta)$ describes a scale modification for some constant $\beta \in (0, 1]$. Let us further assume that the functions m and σ are continuous. Then Q_x has by construction of P a τ-quantile of type $\alpha > 0$ given by $m(x) + \sigma(x) f^*_{\tau,\tilde{\mathrm{Q}}_x}$, because we obtain for $s \in [0, \alpha]$ the inequality $\mathrm{Q}_x\big(\big(\frac{f^*_{\tau,\tilde{\mathrm{Q}}_x} - m(x)}{\sigma(x)}, \frac{f^*_{\tau,\tilde{\mathrm{Q}}_x} - m(x)}{\sigma(x)} + s\big)\big) \geq \big(c_{\tilde{\mathrm{Q}}_x}\sigma(x)\big)s \geq (c_{\tilde{\mathrm{Q}}_x}\beta)s$ and an analogous inequality for the other probability in (3). Note that $c_{\tilde{\mathrm{Q}}_x}\beta$ is bounded if $\|b\|_\infty < \infty$.*

The following theorem shows that for distributions having an average quantile type, the conditional quantile can be estimated that approximately solves the pinball loss.

Theorem 2.5 *Let $p \in (0, \infty]$, $\tau \in (0, 1)$, and $\alpha > 0$ be real numbers. Moreover, let P be a distribution on $X \times Y$ that has a p-average τ-quantile of type α and L be the pinball loss with parameter τ. Then for all $f : X \to \mathbb{R}$ satisfying $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \le 2^{-\frac{p+2}{p+1}} \alpha^{\frac{2p}{p+1}}$ we have*

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

where $f^* : X \to \mathbb{R}$ is the τ-quantile function of P, b is the function from Def. 2.2, and $q := \frac{p}{p+1}$.

Our next goal is to establish an oracle inequality for SVMs defined by (1). To this end let us assume $Y = [-1, 1]$. Then we have $L_\tau(y, \bar{t}) \le L_\tau(y, t)$ for all $y \in Y$, $t \in \mathbb{R}$, where $\bar{t}$ denotes t clipped to the interval $[-1, 1]$, i.e., $\bar{t} := \max\{-1, \min\{1, t\}\}$. Since this yields $\mathcal{R}_{L_\tau,\mathrm{P}}(\bar{f}) \le \mathcal{R}_{L_\tau,\mathrm{P}}(f)$ for all functions $f : X \to \mathbb{R}$ we will focus on clipped functions $\bar{f}$ in the following.

In order to describe the approximation error of SVMs we need the *approximation error function* $a(\lambda) := \inf_{f\in H} \lambda\|f\|^2_H + \mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}^*_{L_\tau,\mathrm{P}}$, $\lambda > 0$. Recall that [8] showed that $\lim_{\lambda\to 0} a(\lambda) = 0$ if H is sufficiently rich, i.e., dense in $L_1(\mathrm{P}_X)$. We also need the *covering numbers*

$$\mathcal{N}\big(B_H, \varepsilon, L_2(\mu)\big) := \min\big\{n \ge 1 : \exists\, x_1, \dots, x_n \in L_2(\mu) \text{ with } B_H \subset \cup_{i=1}^n (x_i + \varepsilon B_{L_2(\mu)})\big\}\,,\ \varepsilon > 0,$$

where μ is an arbitrary probability measure on X, and B_H and $B_{L_2(\mu)}$ denote the closed unit balls of the RKHS H and the Hilbert space $L_2(\mu)$, respectively. Given a finite sequence $T = \big((x_1, y_1), \dots, (x_n, y_n)\big) \in (X \times Y)^n$ we write $T_X := (x_1, \dots, x_n)$, and we will write $\mathcal{N}(B_H, \varepsilon, L_2(T_X)) = \mathcal{N}(B_H, \varepsilon, L_2(\mu))$ if μ is an empirical measure defined by T_X. We often write $L \circ f$ instead of $L(x, y, t)$.

Theorem 2.6 *Let P be a distribution on $X \times [-1, 1]$ for which there exist constants $v \ge 1$, $\vartheta \in [0, 1]$ such that*

$$\mathbb{E}_\mathrm{P}\big(L \circ \bar{f} - L \circ f^*_{\tau,\mathrm{P}}\big)^2 \le v \left(\mathbb{E}_\mathrm{P}(L \circ \bar{f} - L \circ f^*_{\tau,\mathrm{P}})\right)^\vartheta \tag{4}$$

for all $f : X \to \mathbb{R}$. Moreover, let H be a RKHS over X for which there exist constants $p \in (0, 1)$ and $a \ge 1$ such that

$$\sup_{T\in(X\times Y)^n} \log \mathcal{N}\big(B_H, \varepsilon, L_2(T_X)\big) \le a\varepsilon^{-2p}\,, \qquad \varepsilon > 0\,. \tag{5}$$

Then there exists a constant $K_{p,v}$ depending only on p and v such that for all $\tau \ge 1$ and $\lambda > 0$ we have with probability not less than $1 - 3e^{-\tau}$ that

$$\mathcal{R}_{L,\mathrm{P}}(\bar{f}_{T,\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} \le a(\lambda) + \sqrt{\frac{a(\lambda)}{\lambda}\frac{\tau}{n}} + \left(\frac{K_{p,v}a}{\lambda^p n}\right)^{\frac{1}{2-\vartheta+p(\vartheta-1)}} + \frac{K_{p,v}a}{\lambda^p n} + 5\Big(\frac{32v\tau}{n}\Big)^{\frac{1}{2-\vartheta}} + \frac{145\tau}{n}.$$

[9] showed that oracle inequalities of the above type can be used to establish learning rates and investigate simple data-dependent parameter selection strategies. For example if we assume that there exist constants $c > 0$ and $\beta \in (0, 1]$ such that $a(\lambda) \le c\lambda^\beta$ for all $\lambda > 0$ then $\mathcal{R}_{L,\mathrm{P}}(\bar{f}_{T,\lambda_n})$ converges to $\mathcal{R}^*_{L,\mathrm{P}}$ with rate $n^{-\gamma}$ where $\gamma := \min\{\frac{\beta}{\beta(2-\vartheta+p(\vartheta-1))+p}, \frac{2\beta}{\beta+1}\}$ and $\lambda_n = n^{-\gamma/\beta}$. Moreover, [9] shows that this rate can also be achieved by selecting λ in a data-dependent way with the help of a validation set. Let us now consider how these learning rates in terms of risks translate into rates for $\|\bar{f}_{T,\lambda} - f^*_{\tau,\mathrm{P}}\|_{L_q(\mathrm{P}_X)}$. To this end we assume that P has a τ-quantile of p-average type α for $\tau \in (0, 1)$. Using the Lipschitz continuity of L_τ and Theorem 2.5 we then obtain

$$\mathbb{E}_\mathrm{P}\big(L \circ \bar{f} - L \circ f^*_{\tau,\mathrm{P}}\big)^2 \le \mathbb{E}_\mathrm{P}|\bar{f} - f^*_{\tau,\mathrm{P}}|^2 \le \|\bar{f} - f^*_{\tau,\mathrm{P}}\|^{2-q}_\infty \mathbb{E}_\mathrm{P}|\bar{f} - f^*_{\tau,\mathrm{P}}|^q \le c\big(\mathcal{R}_{L,\mathrm{P}}(\bar{f}) - \mathcal{R}^*_{L,\mathrm{P}}\big)^{q/2}$$

for all f satisfying $\mathcal{R}_{L,\mathrm{P}}(\bar{f}) - \mathcal{R}^*_{L,\mathrm{P}} \le 2^{-\frac{p+2}{p+1}} \alpha^{\frac{2p}{p+1}}$. In other words, we have a variance bound (4) for $\vartheta := q/2$. Arguing carefully to handle the assumption $\mathcal{R}_{L,\mathrm{P}}(\bar{f}) - \mathcal{R}^*_{L,\mathrm{P}} \le 2^{-\frac{p+2}{p+1}} \alpha^{\frac{2p}{p+1}}$ we then see that $\|\bar{f}_{T,\lambda} - f^*_{\tau,\mathrm{P}}\|_{L_q(\mathrm{P}_X)}$ can converge as fast as $n^{-\gamma}$, where $\gamma := \min\{\frac{\beta}{\beta(4-q+p(q-2))+2p}, \frac{\beta}{\beta+1}\}$.

To illustrate the latter let us assume that H is a Sobolev space $W^m(X)$ of order $m \in \mathbb{N}$ over X, where X is the unit ball in $\mathbb{R}^d$. Recall from [3] that H satisfies (5) for $p := d/(2m)$ if $m > d/2$ and in this case H also consists of continuous functions. Further assume that we are in the ideal situation $f^*_{\tau,\mathrm{P}} \in W^m(X)$ which implies $\beta = 1$. Then the learning rate for $\|\bar{f}_{T,\lambda} - f^*_{\tau,\mathrm{P}}\|_{L_q(\mathrm{P}_X)}$ becomes $n^{-1/(4-q(1-p))}$, which for ∞-average type distributions reduces to $n^{-2m/(6m+d)} \approx n^{-1/3}$.

Let us now consider the well-known ϵ-insensitive loss function is defined by $L(y,t) := \max\{0, |y-t| - \epsilon\}$ for $y, t \in \mathbb{R}$, where $\epsilon \geq 0$.

Theorem 2.7 *Let* P *be a distribution on* $X \times \mathbb{R}$ *which has a unique median, i.e., a unique* $(1/2)$*-quantile* $f^*_{1/2,\mathrm{P}}$*. Further assume that all conditional distributions* $\mathrm{P}(\cdot|x)$*,* $x \in X$*, are atom-free and symmetric, i.e.* $\mathrm{P}(h(x) + A|x) = \mathrm{P}(h(x) - A|x)$ *for all* $x \in X$*,* $A \subset \mathbb{R}$ *measurable and a suitable function* $h : X \to \mathbb{R}$*. If for an* $\epsilon > 0$ *the conditional distributions have a positive mass concentrated 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.*

Note that using [7] one can show that for distributions specified in the above theorem the SVM using the ϵ-insensitive loss approximates $f^*_{1/2,\mathrm{P}}$ whenever the SVM is $\mathcal{R}_{L,\mathrm{P}}$-consistent, i.e. $\mathcal{R}_{L,\mathrm{P}}(f_{T,\lambda}) \to \mathcal{R}^*_{L,\mathrm{P}}$ in probability, see [2]. More advanced results in the sense of Theorem 2.5 seem also possible, but are out of the scope of this paper.

3 Proofs

Let us first recall some notions from [7] who investigated surrogate loss functions in general and the quations how approximate risk minimizers approximate the exact risk minimizing functions in particular. To this end let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a measurable function which we call a loss in the following. For a distribution P and a function $f : X \to \mathbb{R}$ the L-risk is then defined by $\mathcal{R}_{L,\mathrm{P}}(f) := \int_{X\times Y} L(x, y, f(x))\, d\mathrm{P}(x, y)$, and, as usual, the minimal L-risk, or L-Bayes risk, is denoted by $\mathcal{R}^*_{L,\mathrm{P}} := \inf \mathcal{R}_{L,\mathrm{P}}(f)$, where the infimum is taken over all (measurable) functions $f : X \to \mathbb{R}$. In addition, given a distribution Q on Y the inner L-risk of Q was defined by Steinwart [7] as $\mathcal{C}_{L,\mathrm{Q},x}(t) := \int_Y L(x, y, t)\, d\mathrm{Q}(y)$, $x \in X$, $t \in \mathbb{R}$, and the minimal inner L-risks are denoted by $\mathcal{C}^*_{L,\mathrm{Q},x} := \inf \mathcal{C}_{L,\mathrm{Q},x}(t)$, $x \in X$, where the infimum is taken over all $t \in \mathbb{R}$. Moreover, following [7] we usually omit the indexes x or Q if L happens to be independent of x or y, respectively. Now note that we immediately obtain

$$\mathcal{R}_{L,\mathrm{P}}(f) = \int_X \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big)\, d\mathrm{P}_X(x)\,, \tag{6}$$

and [7, Thm. 3.2] further shows that $x \mapsto \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}$ is measurable if the σ-algebra on X is complete. Furthermore, in this case it was shown that the intuitive formula $\mathcal{R}^*_{L,\mathrm{P}} = \int_X \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}\, d\mathrm{P}_X(x)$ holds, i.e. the Bayes L-risk is obtained by minimizing the inner risks and subsequently integrating with respect to the marginal distribution P_X. Based on this observation the basic idea in [7] is to consider both steps separately. In particular, it turned out that the sets of ε*-approximate minimizers* $\mathcal{M}_{L,\mathrm{Q},x}(\varepsilon) := \big\{t \in \mathbb{R} : \mathcal{C}_{L,\mathrm{Q},x}(t) < \mathcal{C}^*_{L,\mathrm{Q},x} + \varepsilon\big\}$, $\varepsilon \in [0, \infty]$, and the set of *exact minimizers* $\mathcal{M}_{L,\mathrm{Q},x}(0^+) := \bigcap_{\varepsilon>0} \mathcal{M}_{L,\mathrm{Q},x}(\varepsilon)$ play a crucial role. Often we omit the subscripts x and Q in these definitions, if L happens to be independent of x or y, respectively.

Let us now assume that we have two loss functions $L_{\mathrm{tar}} : X \times Y \times \mathbb{R} \to [0, \infty]$ and $L_{\mathrm{sur}} : X \times Y \times \mathbb{R} \to [0, \infty]$, and the goal is to estimate the excess L_{tar}-risk by the excess L_{sur}-risk. This issue was thoroughly investigated in [7], where the main device was the so-called *calibration function* $\delta_{\max}(\,\cdot\,, \mathrm{Q}, x) : [0, \infty] \to [0, \infty]$ defined by

$$\delta_{\max}(\varepsilon, \mathrm{Q}, x) := \begin{cases} \inf_{t\in\mathbb{R}\setminus\mathcal{M}_{L_{\mathrm{tar}},\mathrm{Q},x}(\varepsilon)} \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} & \text{if } \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} < \infty\,, \\ \infty & \text{if } \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} = \infty\,, \end{cases}$$

for all $\varepsilon \in [0, \infty]$. In the following we sometimes write $\delta_{\max,L_{\mathrm{tar}},L_{\mathrm{sur}}}(\varepsilon, \mathrm{Q}, x) := \delta_{\max}(\varepsilon, \mathrm{Q}, x)$ whenever it is needed to explicitly mention the target and surrogate losses. In addition, we will follow our convention which omits x or Q if L is independent of one of them. Now recall that in [7, Lem. 2.9] the inequality

$$\delta_{\max}\big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{Q},x}, \mathrm{Q}, x\big) \;\leq\; \mathcal{C}_{L_{\mathrm{sur}},\mathrm{Q},x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x}\,, \qquad t \in \mathbb{R} \tag{7}$$

was established for situations where $\mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{Q},x} < \infty$ and $\mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{Q},x} < \infty$. Before we use this inequality to establish an inequality between the excess risks of L_{tar} and L_{sur}, let us finally recall that the Fenchel-Legendre bi-conjugate $g^{**} : I \to [0,\infty]$ of a function $g : I \to [0,\infty]$ defined on an interval I is the largest convex function $h : I \to [0,\infty]$ satisfying $h \leq g$. In addition, we write $g^{**}(\infty) := \lim_{t\to\infty} g^{**}(t)$ if $I = [0,\infty)$. With these preparations we can now establish the following result which is a generalization of [7, Thm. 2.18].

Theorem 3.1 *Let P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} < \infty$ and $\mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} < \infty$ and assume that there exist $p \in (0,\infty]$ and functions $b : X \to [0,\infty]$ and $\delta : [0,\infty) \to [0,\infty)$ such that*

$$\delta_{\max}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) \geq b(x)\,\delta(\varepsilon)\,, \qquad\qquad \varepsilon \geq 0,\ x \in X, \tag{8}$$

and $b^{-1} \in L_p(\mathrm{P}_X)$. Then for $\bar\delta := \delta^{p/(p+1)} : [0,\infty) \to [0,\infty)$, and all $f : X \to \mathbb{R}$ we have

$$\bar\delta^{**}\big(\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}\big) \;\leq\; \|b^{-1}\|^{p/(p+1)}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}}\big)^{p/(p+1)}\,.$$

Proof: Let us first consider the case $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) < \infty$. Since $\bar\delta^{**}$ is convex and satisfies $\bar\delta^{**}(\varepsilon) \leq \bar\delta(\varepsilon)$ for all $\varepsilon \in [0,\infty)$ we see by Jensen's inequality that

$$\bar\delta^{**}\big(\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}\big) \;\leq\; \int_X \bar\delta\big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big)\,d\mathrm{P}_X(x) \tag{9}$$

Moreover, using (7) and (8) we obtain

$$b(x)\,\delta\big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big) \;\leq\; \mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}$$

for P_X-almost all $x \in X$ and all $t \in \mathbb{R}$. By (9), the definition of $\bar\delta$ and Hölder's inequality in the form of $\|\cdot\|_{1/q} \leq \|\cdot\|_p \cdot \|\cdot\|_1$, where $q := (p+1)/p$, we thus find that $\bar\delta^{**}\big(\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}\big)$ is less than or equal to

$$\begin{aligned}
& \Big(\int_X \big(b(x)\big)^{-1/q}\big(\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big)^{1/q}\,d\mathrm{P}_X(x)\Big)^{q/q} \\
\leq\;& \Big(\int_X b^{-p} d\mathrm{P}_X\Big)^{1/(pq)}\Big(\int_X \big(\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big)\,d\mathrm{P}_X(x)\Big)^{1/q} \\
\leq\;& \|b^{-1}\|^{1/q}_{L_p(\mathrm{P}_X)}\big(\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}}\big)^{1/q}\,.
\end{aligned}$$

Combining this estimate with our first estimate then gives the assertion. Let us finally deal with the case $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) = \infty$. If $\bar\delta^{**}(\infty) = 0$ there is nothing to proof and hence we restrict our considerations to the case $\bar\delta^{**}(\infty) > 0$. Following the proof of [7, Thm. 2.13] we then see that there exist constants $c_1, c_2 \in (0,\infty)$ satisfying $t \leq c_1\bar\delta^{**}(t) + c_2$ for all $t \in [0,\infty]$. From this we obtain

$$\begin{aligned}
\infty \;=\;& \mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{tar}},\mathrm{P}} \leq c_1 \int_X \bar\delta^{**}\big(\mathcal{C}_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L_{\mathrm{tar}},\mathrm{P}(\,\cdot\,|x),x}\big)\,d\mathrm{P}_X(x) + c_2 \\
\leq\;& c_1 \int_X \big(b(x)\big)^{-\frac{1}{q}}\Big(\mathcal{C}_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\big(f(x)\big) - \mathcal{C}^*_{L_{\mathrm{sur}},\mathrm{P}(\,\cdot\,|x),x}\Big)^{\frac{1}{q}}\,d\mathrm{P}_X(x) + c_2\,,
\end{aligned}$$

where the last step is analogous to our considerations in the case $\mathcal{R}_{L_{\mathrm{tar}},\mathrm{P}}(f) < \infty$. Using $b^{-1} \in L_p(\mathrm{P}_X)$ and Hölder's inequality we then conclude from the above estimate that $\mathcal{R}_{L_{\mathrm{sur}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{sur}},\mathrm{P}} = \infty$, i.e. the assertion is shown. ■

Our next goal is to determine the inner risks and their minimizers for the pinball loss. To this end recall (see, e.g., [1, Theorem 23.8]) that given a distribution Q on $\mathbb{R}$ and a *non-negative* function $g : X \to [0,\infty)$ we have

$$\int_{\mathbb{R}} g\,d\mathrm{Q} = \int_0^\infty \mathrm{Q}(g \geq s)\,ds\,. \tag{10}$$

Proposition 3.2 *Let $\tau \in (0,1)$ and Q be a distribution on $\mathbb{R}$ with $\mathcal{C}^*_{L_\tau,\mathrm{Q}} < \infty$ and t^* a τ-quantile of Q. Then there exists $q_+, q_- \in [0,\infty)$ such that $q_+ + q_- = \mathrm{Q}(\{t^*\})$, and for all $t \geq 0$ we have*

$$\mathcal{C}_{L_\tau,\mathrm{Q}}(t^* + t) - \mathcal{C}_{L_\tau,\mathrm{Q}}(t^*) \;=\; tq_+ + \int_0^t \mathrm{Q}\big((t^*, t^* + s)\big)\,ds\,, \quad \textit{and} \tag{11}$$

$$\mathcal{C}_{L_\tau,\mathrm{Q}}(t^* - t) - \mathcal{C}_{L_\tau,\mathrm{Q}}(t^*) \;=\; tq_- + \int_0^t \mathrm{Q}\big((t^* - s, t^*)\big)\,ds\,. \tag{12}$$

Proof: Let us consider the distribution $\mathrm{Q}^{(t^*)}$ defined by $\mathrm{Q}^{(t^*)}(A) := \mathrm{Q}(t^* + A)$ for all measurable $A \subset \mathbb{R}$. Then it is not hard to see that 0 is a τ-quantile of $\mathrm{Q}^{(t^*)}$. Moreover, we obviously have $\mathcal{C}_{L_\tau,\mathrm{Q}}(t^* + t) = \mathcal{C}_{L_\tau,\mathrm{Q}^{(t^*)}}(t)$ and hence we may assume without loss of generality that $t^* = 0$. Then our assumptions together with $\mathrm{Q}((-\infty, 0]) + \mathrm{Q}([0,\infty)) = 1 + \mathrm{Q}(\{0\})$ yield $\tau \le \mathrm{Q}((-\infty, 0]) \le \tau + \mathrm{Q}(\{0\})$, i.e., there exists a q_+ satisfying $0 \le q_+ \le \mathrm{Q}(\{0\})$ and

$$\mathrm{Q}((-\infty, 0]) = \tau + q_+ \,. \tag{13}$$

Let us now compute the inner risks of L_τ. To this end observe we first assume $t \ge 0$. Then we have $\int_{y<t}(y-t)\,d\mathrm{Q}(y) = \int_{y<0} y\,d\mathrm{Q}(y) - t\mathrm{Q}((-\infty,t)) + \int_{0\le y<t} y\,d\mathrm{Q}(y)$ and $\int_{y\ge t}(y-t)\,d\mathrm{Q}(y) = \int_{y\ge 0} y\,d\mathrm{Q}(y) - t\mathrm{Q}([t,\infty)) - \int_{0\le y<t} y\,d\mathrm{Q}(y)$ and hence we obtain

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

Moreover, using (10) we find $t\mathrm{Q}([0,t)) - \int_{0\le y<t} y\,d\mathrm{Q}(y) = \int_0^t \mathrm{Q}([0,t))ds - \int_0^t \mathrm{Q}([s,t))\,ds = t\mathrm{Q}(\{0\}) + \int_0^t \mathrm{Q}((0,s))ds$, and since (13) implies $\mathrm{Q}((-\infty,0)) + \mathrm{Q}(\{0\}) = \mathrm{Q}((-\infty,0]) = \tau + q_+$ we thus obtain (11). Now (12) can be derived from (11) by considering the pinball loss with parameter $1-\tau$ and the distribution $\bar{\mathrm{Q}}$ defined by $\bar{\mathrm{Q}}(A) := \mathrm{Q}(-A)$, $A \subset \mathbb{R}$ measurable. This further yields a real number q_- satisfying $0 \le q_- \le \mathrm{Q}(\{0\})$ and $\mathrm{Q}([0,\infty) = 1 - \tau + q_-$. By combining this with (13) we then find $q_+ + q_- = \mathrm{Q}(\{0\})$. ■

For the proof of Theorem 2.5 we have to recall yet a few more concepts from [7]. To this end let us now assume that our loss function is independent of x, i.e., we consider a measurable function $L : Y \times \mathbb{R} \to [0,\infty]$. We write $\mathcal{Q}_{\min}(L) := \big\{\mathrm{Q} \in \mathcal{Q}_{\min}(L) : \exists\, t^*_{L,\mathrm{Q}} \in \mathbb{R}$ such that $\mathcal{M}_{L,\mathrm{Q}}(0^+) = \{t^*_{L,\mathrm{Q}}\}\big\}$, i.e. $\mathcal{Q}_{\min}(L)$ contains the distributions on Y whose inner L-risks have exactly one exact minimizer. Furthermore, note that this definition immediately yields $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$ for all $\mathrm{Q} \in \mathcal{Q}_{\min}(L)$. Following [7] we now define the *self-calibration loss function* of L by

$$\breve{L}(\mathrm{Q}, t) := |t - t^*_{L,\mathrm{Q}}|\,, \qquad \mathrm{Q} \in \mathcal{Q}_{\min}(L),\ t \in \mathbb{R}\,. \tag{14}$$

This loss is a so-called template loss in the sense of [7], i.e., for a given distribution P on $X \times Y$, where X has a complete σ-algebra and $\mathrm{P}(\,\cdot\,|x) \in \mathcal{Q}_{\min}(L)$ for P_X-almost all $x \in X$, the P-instance $\breve{L}_\mathrm{P}(x,t) := |t - t^*_{L,\mathrm{P}(\,\cdot\,|x)}|$ is measurable (and hence a loss function). [7] extended the definition of inner risks to template functions and in the case of the self-calibration loss function this extension becomes $\mathcal{C}_{\breve{L},\mathrm{Q}}(t) := \breve{L}(\mathrm{Q},t)$. Based on this [7] defined minimial inner risks and their (approximate) minimizers in the obvious way. Based on this the self-calibration function was defined by $\delta_{\max,\breve{L},L}(\varepsilon,\mathrm{Q}) = \inf_{t\in\mathbb{R};\,|t-t^*_{L,\mathrm{Q}}|\ge\varepsilon} \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}$. As shown in [7] this self-calibration function has two important properties: first it satisfies

$$\delta_{\max,\breve{L},L}\big(|t - t^*_{L,\mathrm{Q}}|, \mathrm{Q}\big) \le \mathcal{C}_{L,\mathrm{Q}}(t) - \mathcal{C}^*_{L,\mathrm{Q}}\,, \qquad t \in \mathbb{R}, \tag{15}$$

i.e. it measures how well approximate L-risk minimizers t approximate the true minimizer $t^*_{L,\mathrm{Q}}$, and second it equals the calibration function of the P-instance $\breve{L}_\mathrm{P}$, i.e.

$$\delta_{\max,\breve{L}_\mathrm{P},L}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = \delta_{\max,\breve{L},L}(\varepsilon, \mathrm{P}(\,\cdot\,|x))\,, \qquad \varepsilon \in [0,\infty],\ x \in X. \tag{16}$$

In other words, the self-calibration function can be utilized in Theorem 3.1.

Proof of Theorem 2.5: Let Q be a distribution on $\mathbb{R}$ with $\mathcal{C}^*_{L,\mathrm{Q}} < \infty$ and t^* be the *only* τ-quantile of Q. Then the formulas of Proposition 3.2 show

$$\delta_{\max,\breve{L},L}(\varepsilon,\mathrm{Q}) = \min\Big\{\varepsilon q_+ + \int_0^\varepsilon \mathrm{Q}\big((t^*, t^*+s)\big)\,ds,\ \varepsilon q_- + \int_0^\varepsilon \mathrm{Q}\big((t^*-s, t^*)\big)\,ds\Big\}\,, \quad \varepsilon \ge 0,$$

where q_+ and q_- are the real numbers defined in Proposition 3.2. Let us additionally assume that the τ-quantile t^* is of type α. For the function[1] $\delta : [0,\infty) \to [0,\infty)$ defined by $\delta(\varepsilon) := \varepsilon^2/2$, if $\varepsilon \in$

[1] Note the similarity to Huber's loss function for regression.

$[0, \alpha]$, and $\delta(\varepsilon) := \alpha\varepsilon - \alpha^2/2$, if $\varepsilon > \alpha$ a simple calculation then yields $\delta_{\max,\breve{L},L}(\varepsilon, \mathrm{Q}) \geq c_{\mathrm{Q}}\delta(\varepsilon)$, where c_{Q} is the constant satisfying (3). For $q = p/(p+1)$ let us further define $\bar{\delta} : [0, \infty) \to [0, \infty)$ by $\bar{\delta}(\varepsilon) := \delta^q(\varepsilon^{1/q})$, $\varepsilon \geq 0$. In view of Theorem 3.1 we then need to find a convex function $\hat{\delta} : [0, \infty) \to [0, \infty)$ such that $\hat{\delta} \leq \bar{\delta}$. To this end we define $\hat{\delta}(\varepsilon) := s_p^p\varepsilon^2$, if $\varepsilon \in \big[0, s_p a_p\big]$, and $\hat{\delta}(\varepsilon) := a_p\big(\varepsilon - s_p^{p+2}a_p\big)$, if $\varepsilon > s_p a_p$, where $a_p := \alpha^{p/(p+1)}$ and $s_p := 2^{-1/(p+1)}$. An easy calculation shows that $\hat{\delta} : [0, \infty) \to [0, \infty)$ is continuously differentiable, and its derivative is increasing, thus $\hat{\delta}$ is convex. Moreover, we have $\hat{\delta}' \leq \bar{\delta}'$ and hence $\hat{\delta} \leq \bar{\delta}$ by the fundamental theorem of calculus which implies $\hat{\delta} \leq \bar{\delta}^{**}$. We find the assertion by (15), (16), and Theorem 3.1. ■

Proof of Theorem 2.6: This is a direct consquence of [9, Theorem 2.1]. ■

The proof of Theorem 2.7 follows immediately from the following lemma.

Lemma 3.3 *Assume that the regular conditional distribution* Q *of* Y *given* x *is symmetric around the median* t^*. *Further assume that* Q *is atom-free and that* $\mathrm{Q}\big([t^* + \epsilon - \delta, t^* + \epsilon + \delta]\big) > 0$ *for all* $\delta > 0$. *Then the inner L-risk* $\mathcal{C}_{L,\mathrm{Q}}(t)$ *for the* ϵ*-insensitive loss with* $\epsilon \geq 0$ *fulfills*

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

The set of exact minimizers only contains the median, i.e. $\mathcal{M}_{L,\mathrm{Q},x}(0^+) = \{t^*\}$.

Proof: Because $L(y,t) = L(-y,-t)$ for all $y, t \in \mathbb{R}$ we only have to consider $t \geq 0$. Analogous to the proof of Proposition 3.2 we may assume w.l.o.g. that Q is symmetric around $t^* = 0$. For later use we note that for $0 \leq a \leq b \leq \infty$ Equation (10) yields

$$\int_a^b y\,d\mathrm{Q}(y) = a\mathrm{Q}([a,b]) + \int_a^b \mathrm{Q}([s,b])ds\,. \tag{17}$$

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

$$\mathcal{C}_{L,\mathrm{Q}}(t) = \int_0^{t-\epsilon} \mathrm{Q}(-\infty, t-\epsilon]ds - \int_0^{t+\epsilon} \mathrm{Q}[t+\epsilon,\infty)ds + \int_{\epsilon-t}^{t+\epsilon} y\,d\mathrm{Q}(y) + 2\int_{t+\epsilon}^{\infty} y\,d\mathrm{Q}(y)\,. \tag{18}$$

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 (17) implies

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

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

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

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

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

The symmetry of Q implies $\int_0^{t-\epsilon} \mathrm{Q}\big([\epsilon - t, t-\epsilon]\big)\, ds = 2\int_0^{t-\epsilon} \mathrm{Q}\big([0, t-\epsilon]\big)\, ds$, and we get

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

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

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

By $\int_0^{t-\epsilon} \mathrm{Q}\big([s,\infty)\big)\, ds + \int_0^{t+\epsilon} \mathrm{Q}\big([s,\infty)\big)\, ds = 2\int_0^{t-\epsilon} \mathrm{Q}\big([s,\infty)\big)\, ds + \int_{t-\epsilon}^{t+\epsilon} \mathrm{Q}\big([s,\infty)\big)\, ds$ we obtain $\mathcal{C}_{L,\mathrm{Q}}(t) = 2\int_0^{t-\epsilon} \mathrm{Q}\big([0,\infty)\big)\, ds + 2\int_{t+\epsilon}^{\infty} \mathrm{Q}\big([s,\infty)\big)\, ds + \int_{t-\epsilon}^{t+\epsilon} \mathrm{Q}\big([s,\infty)\big)\, ds$, if $t \geq \epsilon$.

Let us now consider the case $t \in [0, \epsilon]$. Analogously we obtain from (18) that $\mathcal{C}_{L,\mathrm{Q}}(t)$ equals

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

Combining this with $\int_0^{\epsilon-t} \mathrm{Q}\big([\epsilon - t, t+\epsilon]\big)\, ds - \int_0^{\epsilon-t} \mathrm{Q}\big([\epsilon - t,\infty)\big)\, ds = -\int_0^{\epsilon-t} \mathrm{Q}\big([\epsilon + t,\infty)\big)\, ds$ and $\int_0^{\epsilon+t} \mathrm{Q}\big([\epsilon + t,\infty)\big)\, ds - \int_0^{\epsilon-t} \mathrm{Q}\big([\epsilon + t,\infty)\big)\, ds = \int_{\epsilon-t}^{\epsilon+t} \mathrm{Q}\big([\epsilon + t,\infty)\big)\, ds$ we get

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

Hence $\mathcal{C}_{L,\mathrm{Q}}(0) = 2\int_{\epsilon}^{\infty} \mathrm{Q}\big([s,\infty)\big)\, ds$. 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 3.3 follow by using the same arguments. Hence one exact minimizer of $\mathcal{C}_{L,\mathrm{Q}}(\cdot)$ is the median $t^* = 0$. Finally, since Q does not have atoms the function $s \mapsto \mathrm{Q}[s,\infty)$ is continuous and hence the fundamental theorem of calculus shows that the derivative of $\mathcal{C}_{L,\mathrm{Q}}(\cdot) : [0,\infty) \to \mathbb{R}$ is given by $\mathcal{C}'_{L,\mathrm{Q}}(t) = \mathrm{Q}\big([\epsilon - t,\infty)\big) - \mathrm{Q}\big([\epsilon + t,\infty)\big)$. Since $\mathcal{C}_{L,\mathrm{Q}}(\cdot) : [0,\infty) \to \mathbb{R}$ is convex we see that $t \in (0,\infty)$ minimizes $\mathcal{C}_{L,\mathrm{Q}}(\cdot)$ if and only if $\mathrm{Q}[\epsilon - t,\infty) = \mathrm{Q}[\epsilon + t,\infty)$. The latter is only satisfied if Q does not have a positive mass around ϵ. I.e. if $\mathrm{Q}\big([\epsilon - \delta, \epsilon + \delta]\big) > 0$ for all $\delta > 0$, then the set of exact minimizers is $\mathcal{M}_{L,\mathrm{Q},x}(0^+) = \{0\}$. ■

References

[1] H. Bauer. *Measure and Integration Theory*. De Gruyter, Berlin, 2001.

[2] A. Christmann and I. Steinwart. Consistency and robustness of kernel based regression. *To appear in: Bernoulli*, 2007.

[3] D.E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, 1996.

[4] C. Hwang and J. Shim. A simple quantile regression via support vector machine. In *Advances in Natural Computation: First International Conference (ICNC)*, pages 512 –520. Springer, 2005.

[5] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[6] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[7] I. Steinwart. How to compare different loss functions. *Constr. Approx.*, 26:225–287, 2007.

[8] I. Steinwart, D. Hush, and C. Scovel. Function classes that approximate the Bayes risk. In *Proceedings of the 19th Annual Conference on Learning Theory, COLT 2006*, pages 79–93. Springer, 2006.

[9] I. Steinwart, D. Hush, and C. Scovel. An oracle inequality for clipped regularized risk minimizers. In *Advances in Neural Information Processing Systems 20, to appear*, 2007.

[10] I. Takeuchi, Q.V. Le, T.D. Sears, and A.J. Smola. Nonparametric quantile estimation. *J. Mach. Learn. Res.*, 7:1231–1264, 2006.