CONSISTENCY OF KERNEL BASED QUANTILE REGRESSION

By Andreas Christmann[1] and Ingo Steinwart[2]

[1] *Vrije Universiteit Brussel and* [2] *Los Alamos National Laboratory*

Quantile regression is used in many areas of applied research and business. Examples are actuarial, financial or biometrical applications. We show that a non-parametric generalization of quantile regression based on kernels shares with support vector machines the property of consistency to the Bayes risk. We further use this consistency to prove that the non-parametric generalization approximates the conditional quantile function which gives the mathematical justification for kernel based quantile regression.

1. Introduction

Consider a random sample (x_i, y_i) from independent and identically distributed random variables (X_i, Y_i) each with unknown probability distribution P on $X \times Y$, $1 \leq i \leq n$. For technical reasons we assume throughout this work that X and Y are closed subsets of $\mathbb{R}^m$ and $\mathbb{R}$, respectively. Recall that in this case P can be split up into the marginal distribution P_X and the regular conditional probability $\mathrm{P}(\,\cdot\,|X = x)$, $x \in X$, on Y.

The goal of quantile regression is to estimate the conditional quantile, *i.e.* the set valued function

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

where $\tau \in (0, 1)$ is a fixed constant. For conceptual simplicity (though mathematically this is not necessary) we assume throughout this paper that $F^*_{\tau,\mathrm{P}}(x)$ consists of singletons, so that there exists a unique conditional quantile function $f^*_{\tau,\mathrm{P}} : X \to \mathbb{R}$ defined by $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$, $x \in X$. Now recall that the so-called pinball loss function

$$L_\tau : \mathbb{R} \to [0, \infty), \quad L_\tau(r) := r(\tau - \mathbf{1}_{\{r<0\}}) = \begin{cases} (\tau - 1)r & \text{if } \tau < 0, \\ \tau r & \text{if } \tau \geq 0, \end{cases}$$

has the property that $q^* = f^*_{\tau,\mathrm{P}}(x)$ if and only if q^* minimizes the conditional L_τ risk, *i.e.*

$$\mathbb{E}_{\mathrm{P}_{Y|X=x}} L_\tau(Y - q^*) = \inf_{q \in \mathbb{R}} \mathbb{E}_{\mathrm{P}_{Y|X=x}} L_\tau(Y - q)\,. \tag{1}$$

Based on this fact Koenker and Bassett (1978) proposed the estimator

$$\hat{f}_\tau = \arg \inf_{\theta \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} L_\tau\big(y_i - \langle x_i, \theta\rangle\big),$$

for cases in which $f^*_{\tau,\mathrm{P}}$ is a linear function. In this paper we consider a kernel based generalization of $\hat{f}_\tau$ which does not require this linearity assumption on $f^*_{\tau,\mathrm{P}}$. In order to

1. *AMS 2000 subject classification. Primary 62G08, 62G35; secondary 68Q32, 62G20.*
2. *Keywords and Phrases.* Consistency, convex risk minimization, empirical risk minimization, kernel, non-parametric, quantile regression.

introduce this generalization let $\lambda > 0$ be a regularization parameter and H a reproducing kernel Hilbert space (RKHS) of a kernel $k : X \times X \to \mathbb{R}$. Recall, that the reproducing property gives $f(x) = \langle f, \Phi(x) \rangle$ for all $f \in H$ and $x \in X$, where $\Phi : X \to H$ is the canonical feature map defined by $\Phi(x) := k(\cdot, x)$, $x \in X$. Throughout this paper we additionally assume that k is measurable and H is separable, so that Φ becomes Borel measurable by Petti's measurability theorem (see Diestel and Uhl, 1977). Takeuchi *et al.* (2006) proposed for $\tau \in (0, 1)$ the kernel based quantile regression (KBQR) estimator which is defined by

$$f_{\mathrm{P},\lambda} := \arg\min_{f \in H} \mathbb{E}_{\mathrm{P}} L_\tau\big(Y - f(X)\big) + \lambda \|f\|_H^2 \,. \tag{2}$$

For any fixed data set $D_n = \{(x_i, y_i), 1 \le i \le n\} \subset X \times Y$ we obtain the estimator

$$f_{\mathrm{D}_n,\lambda} := \arg\min_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L_\tau\big(y_i - f(x_i)\big) + \lambda \|f\|_H^2, \tag{3}$$

where $\mathrm{D}_n = \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}$ denotes the empirical distribution. Note that we obtain $f_{\mathrm{D}_n,\lambda} = \hat{f}_\tau$, if we choose the linear kernel $k(x, x') := \langle x, x' \rangle$ and $\lambda := 0$.

Our first main result is Theorem 5 which shows that kernel based quantile regression is risk consistent to the Bayes risk under rather weak assumptions, *i.e.*

$$\mathbb{E}_{\mathrm{P}} L_\tau\big(Y - f_{\mathrm{D}_n,\lambda_n}(X)\big) \to \inf\big\{\mathbb{E}_{\mathrm{P}} L_\tau\big(Y - f(X)\big) \,\big|\, f : X \to \mathbb{R} \text{ measurable}\big\} \tag{4}$$

holds in probability for $n \to \infty$ for suitable sequences of positive regularization parameters (λ_n). Note that the infimum on the right hand side of (4) is with respect to *all* measurable functions and not only with respect to all functions in the RKHS H. Our second main result which is Theorem 6 shows that whenever KBQR is Bayes risk consistent it also satisfies

$$\|f_{\mathrm{D}_n,\lambda_n} - f^*_{\tau,\mathrm{P}}\|_0 \to 0$$

where $\|\cdot\|_0$ denotes a translation invariant metric describing the convergence in probability. Together both results give a mathematical justification for using KBQR in non-parametric quantile regression problems.

The rest of the paper is organized as follows. Section 2 presents conditions which ensures the existence of $f_{\mathrm{P},\lambda}$. These results will be used to prove our main theorems which are presented in Section 3. All proofs are given in the appendix.

2. Existence and uniqueness of infinite-sample KBQR

For any distribution P on $X \times Y$ and any measurable map $f : X \to \mathbb{R}$ we define the L_τ-risk of f with respect to P by

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) := \mathbb{E}_{\mathrm{P}} L_\tau\big(Y - f(X)\big) = \int_X \int_Y L_\tau\big(y - f(x)\big)\, d\mathrm{P}_{Y|X=x}(y)\, d\mathrm{P}_X(x) \,,$$

where we recall that the regular conditional probability $\mathrm{P}(\cdot|X = x)$ exists because Y is closed (and thus a Polish space). Moreover, note that the above integral is always defined

since L_τ is non-negative and continuous, but in general it is not finite. In order to find a condition which ensures $\mathcal{R}_{L_\tau,\mathrm{P}}(f) < \infty$ we define

$$|\mathrm{P}|_1 := \int_{X\times Y} |y| \, d\mathrm{P}(x,y) \,.$$

Now we can formulate a sufficient condition ensuring $\mathcal{R}_{L_\tau,\mathrm{P}}(f) < \infty$.

Proposition 1 *Let* P *be a distribution on* $X \times Y$ *with* $|\mathrm{P}|_1 < \infty$ *and* $f : X \to \mathbb{R}$ *be a function with* $f \in L_1(\mathrm{P})$*. Then we have* $\mathcal{R}_{L_\tau,\mathrm{P}}(f) < \infty$*.*

The following lemma presents in some sense an inverse statement of the above proposition.

Lemma 2 *Let* $f : X \to \mathbb{R}$ *be a measurable function and* P *be a distribution on* $X \times Y$ *with* $\mathcal{R}_{L_\tau,\mathrm{P}}(f) < \infty$*. Then we have* $|\mathrm{P}|_1 < \infty$ *if and only if* $f \in L_1(\mathrm{P})$*.*

The next result ensures the existence of a solution $f_{\mathrm{P},\lambda}$. In order to formulate it recall that a kernel $k : X \times X \to \mathbb{R}$ of a RKHS H is called *bounded* if

$$\|k\|_\infty := \sup_{x\in X} \sqrt{k(x,x)} < \infty.$$

For such kernels it is well known that the reproducing property yields $\|f\|_\infty \le \|k\|_\infty \cdot \|f\|_H$ for all $f \in H$. In particular, if P is a distribution on $X \times Y$ with $|\mathrm{P}|_1 < \infty$ then the objective function in (2) is always finite by Proposition 1, *i.e.* we have

$$R^{reg}_{L_\tau,P,\lambda}(f) := \mathcal{R}_{L_\tau,\mathrm{P}}(f) + \lambda\|f\|^2_H < \infty$$

for all $f \in H$. With these preparations we can now establish the existence and uniqueness of $f_{\mathrm{P},\lambda}$.

Proposition 3 *Let* P *be a distribution on* $X \times Y$ *with* $|\mathrm{P}|_1 < \infty$*,* H *be an RKHS of a bounded kernel* k*, and* $\lambda > 0$*. Then there exists a unique minimizer* $f_{\mathrm{P},\lambda} \in H$ *of*

$$f \mapsto \mathcal{R}^{reg}_{L_\tau,\mathrm{P},\lambda}(f)$$

and we have $\|f_{\mathrm{P},\lambda}\|_H \le \sqrt{|\mathrm{P}|_1/\lambda}$*.*

3. Main results

Our first goal in this section is to present a result that establishes risk consistency of kernel based quantile regression, *i.e.* we will show that (4) holds in probability for $n \to \infty$ and suitable sequences of positive regularization parameters (λ_n). Of course, for such convergence to hold it is necessary that the used RKHS H is rich enough in the sense of

$$\mathcal{R}^*_{L_\tau,\mathrm{P},H} := \inf_{f\in H} \mathcal{R}_{L_\tau,\mathrm{P}}(f) = \inf\bigl\{\mathcal{R}_{L_\tau,\mathrm{P}}(f) \,|\, f : X \to \mathbb{R} \text{ measurable}\bigr\} =: \mathcal{R}^*_{L_\tau,\mathrm{P}} \,. \qquad (5)$$

The following proposition which is essentially taken from Steinwart *et al.* (2006) translates this richness in an easier to handle denseness assumption.

Proposition 4 *Let H be the RKHS of a bounded kernel $k : X \times X \to \mathbb{R}$ and μ be a distribution on X. Then the following statements are equivalent:*

i) H is dense in $L_1(\mu)$.

ii) Equation (5) holds for all distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$ and $|\mathrm{P}|_1 < \infty$.

Note that it was shown by Steinwart *et al.* (2006) that e.g. the popular Gaussian radial basis function (RBF) kernel defined by $k(x, x') = \exp(-\gamma\|x - x'\|^2)$ for $\gamma > 0$ fixed and $x, x' \in \mathbb{R}^m$ satisfies condition *i)* of Proposition 4 for *all* distributions μ on $\mathbb{R}^m$. Obviously, this kernel is also bounded, because $|k(x, x')| \leq 1$ for all $x, x' \in \mathbb{R}^m$. Moreover, for *compact* X and continuous kernels k on X condition *i)* of Proposition 4 is satisfied for all distributions μ on X if k is *universal* in the sense of Steinwart (2001), *i.e.* if its RKHS is dense in the space $C(X)$ of continuous functions mapping X to $\mathbb{R}$. Examples of such kernels including the Gaussian RBF kernel are described by Steinwart (2001). Finally, note that polynomial kernels $k(x, x') = (c + \langle x, x' \rangle)^m$, $m \geq 1$, $c \geq 0$, $x, x' \in \mathbb{R}^m$, are also popular in practice, but they are neither bounded nor dense for general measures μ.

We can now formulate our first main result.

Theorem 5 *Let $X \subset \mathbb{R}^m$ be a closed subset and H be the RKHS of a bounded measurable kernel k on X such that H is dense in $L_1(\mu)$ for all distributions μ on X. Furthermore, let (λ_n) be a sequence of strictly positive numbers with $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$. Then the KBQR estimator defined by (3) using λ_n for sample sets of length n is risk consistent in the sense of (4) for all distributions P with $|\mathrm{P}|_1 < \infty$.*

In order to formulate our second main result let us introduce some more notations. To this end let P be a distribution on $X \times Y$ and $f, g : X \to \mathbb{R}$ be measurable functions. We write

$$\|f\|_{L_0(\mathrm{P}_X)} := \|f\|_0 := \int_X \min\bigl\{1, |f|\bigr\}\, d\mathrm{P}_X$$

and $d(f, g) := \|f - g\|_0$. It is elementary to check that d is a *translation invariant* metric on the space of all measurable functions defined on X, and furthermore a simple application of Chebyshev's inequality shows that d describes the convergence in probability P_X.

The following result shows that under the assumptions of Theorem 5 the KBQR estimator approximates the conditional quantile function in terms of $\|\cdot\|_{L_0(\mathrm{P}_X)}$.

Theorem 6 *Let $X \subset \mathbb{R}^m$ be a closed subset and H be the RKHS of a bounded measurable kernel k on X such that H is dense in $L_1(\mu)$ for all distributions μ on X. Furthermore, let (λ_n) be a sequence of strictly positive numbers with $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$. Then the KBQR estimator defined by (3) satisfies*

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

in probability for $n \to \infty$ and all distributions P on $X \times Y$ with $|\mathrm{P}|_1 < \infty$.

It is interesting to note that the assumption $F^*_{\tau, \mathrm{P}}(x) = \{f^*_{\tau, \mathrm{P}}(x)\}$ is only needed to formulate Theorem 6 in terms of $\|\cdot\|_0$. However, Theorem 3.16 of Steinwart (2005) which is

used in the proof of Theorem 6 actually provides a framework to replace $\|\cdot\|_0$ by a more general notion of closedness if the assumption $F^*_{\tau,\mathrm{P}}(x) = \{f^*_{\tau,\mathrm{P}}(x)\}$ is violated.

In some sense the convergence with respect to $\|\cdot\|_0$ is rather weak and one may wonder whether it can be replaced by some stronger notion of convergence. For example, note that for $\tau = 1/2$ Theorem 5 established the convergence

$$\mathbb{E}_P\big|Y - f_{\mathrm{D}_n,\lambda_n}(X)\big| - \mathbb{E}_P\big|Y - f^*_{\tau,\mathrm{P}}(X)\big| \to 0\,, \quad n \to \infty, \tag{6}$$

which naturally raises the question whether we actually have

$$\mathbb{E}_P\big|f_{\mathrm{D}_n,\lambda_n}(X) - f^*_{\tau,\mathrm{P}}(X)\big| \to 0. \tag{7}$$

Of course, the inverse triangle inequality $\big||a| - |b|\big| \le |a-b|$ immediately shows that (7) implies (6), 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 (7) does not follow from the one in (6). In this direction it is also interesting to note that the framework developed by Steinwart (2005) suggests that for certain classes of distributions P we can actually replace $\|\cdot\|_{L_0(\mathrm{P}_X)}$ by some (quasi)-norm $\|\cdot\|_{L_p(\mathrm{P}_X)}$. However, such considerations are out of the scope of the paper.

Another interesting question is whether we can establish convergence rates in Theorem 5 or Theorem 6. Of course, it is well-known in learning theory that such convergence rates require additional assumptions on the distribution P, *e.g.* in terms of the approximation properties of H with respect to $f^*_{\tau,\mathrm{P}}$. Moreover, the techniques used in the proofs of Theorem 5 or Theorem 6 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 rates in the statisitical anaysis. Because of these reasons, convergence rates are out of the scope of this paper, too.

4. Conclusion

In this paper we proved that kernel based quantile regression proposed by Takeuchi *et al.* (2006) is risk consistent, *i.e.* the L_τ-risk of the KBQR estimator converges in probability to the Bayes risk which is defined as the smallest L_τ-risk for all measurable functions. A similar result was recently obtained by Christmann and Steinwart (2005) for support vector regression (see Schölkopf and Smola, 2002, for an introduction). Moreover, we have shown that the KBQR estimator converges in probability to the conditional quantile function which provides a mathematical justification of this method.

It might be possible to get rid of the assumption $|\mathrm{P}|_1 < \infty$ when considering KBQR if one changes the empirical regularized minimization problem (3) to

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

where $L^*_\tau(y,t) := L_\tau(y-t) - L_\tau(y)$ for $y, t \in \mathbb{R}$. However, loss functions which can take on negative values are beyond the scope of this paper.

For another non-parametric generalization of $\hat{f}_\tau$ based on splines we refer to Koenker *et al.* (1994) and He and Ng (1999).

Appendix

The appendix contains the proofs of our results. We begin by summarizing some properties of the pinball loss function in the following lemma whose trivial proof is omitted for brevity's sake.

Lemma 7 *For each $\tau \in (0,1)$ the pinball loss function L_τ satisfies the following statements:*

i) *L_τ is convex and satisfies both $L_\tau(0) = 0$ and $\lim_{|r|\to\infty} L_\tau(r) = \infty$.*

ii) *L_τ is Lipschitz continuous with Lipschitz constant $|L_\tau|_1 = \max\{\tau, 1-\tau\}$, i.e. $|L_\tau(r) - L_\tau(r')| \leq |L_\tau|_1 \cdot |r - r'|$ for all $r, r' \in \mathbb{R}$.*

iii) *For all $r \in \mathbb{R}$ we have $\min\{\tau, 1-\tau\}\,|r| \leq L_\tau(r) \leq |L_\tau|_1\,|r|$.*

Proof of Proposition 1. By part *iii)* of Lemma 7 we have

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f) = \mathbb{E}_\mathrm{P} L_\tau\big(Y - f(X)\big) \leq |L_\tau|_1 \mathbb{E}_\mathrm{P}\big(|Y| + |f(X)|\big) \leq |\mathrm{P}|_1 + \|f\|_{L_1(\mathrm{P})} \;<\; \infty\,. \qquad \square$$

Proof of Lemma 2. For all $a, b \in \mathbb{R}$ we have $|a-b| \geq |a| - |b|$. Now let us assume that we know $f \in L_1(\mathrm{P})$. Part *iii)* of Lemma 7 then implies

$$\infty \;>\; \mathcal{R}_{L_\tau,\mathrm{P}}(f) \;\geq\; \min\{\tau, 1-\tau\}\mathbb{E}_\mathrm{P}\big(|Y - f(X)|\big) \;\geq\; \min\{\tau, 1-\tau\}\mathbb{E}_\mathrm{P}\big(|Y| - |f(X)|\big)\,.$$

From this we immediately get $|\mathrm{P}|_1 < \infty$. The converse implication can be shown analogously. $\square$

Proof of Proposition 3. Our proof follows DeVito *et al.* (2004) in a streamlined fashion. Combining part *iii)* of Lemma 7 with Lemma 2 of Steinwart *et al.* (2006) we see that $\mathcal{R}_{L_\tau,\mathrm{P}} : L_1(\mathrm{P}_X) \to \mathbb{R}$ is continuous. Furthermore, $\mathrm{id} : H \to L_1(\mathrm{P}_X)$ is continuous since k is bounded and hence $\mathcal{R}^{reg}_{L_\tau,\mathrm{P},\lambda} : H \to \mathbb{R}$ is continuous. This map is also convex, and the set $\{f \in H : \mathcal{R}^{reg}_{L_\tau,\mathrm{P},\lambda}(f) \leq \delta_{\mathrm{P},\lambda}\}$ is bounded and non-empty, because it contains $0 \in H$. Therefore, Ekeland and Turnbull (1983, Prop. II.4.6) ensures the existence of $f_{\mathrm{P},\lambda}$. The uniqueness follows from the strict convexity of $\mathcal{R}^{reg}_{L_\tau,\mathrm{P},\lambda}$. The last assertion is trivial. $\square$

Our next goal is to obtain a representation of $f_{\mathrm{P},\lambda}$. To this end we need the notion of subdifferentials which is recalled in the following definition.

Definition 8 (Subdifferential) *Let H be a Hilbert space, $F : H \to \mathbb{R} \cup \{\infty\}$ be a convex function and $w \in H$ with $F(w) \neq \infty$. Then the subdifferential of F at w is defined by*

$$\partial F(w) := \big\{w^* \in H : \langle w^*, v - w\rangle \leq F(v) - F(w) \;\; \text{for all } v \in H\big\}\,.$$

With the help of the subdifferential ∂L_τ we can now recall a result shown by DeVito *et al.* (2004) (in a slightly generalized form) which in turn is a generalization of a representation derived by Steinwart (2003).

Proposition 9 *Let* P *be a distribution on* $X \times Y$ *with* $|\mathrm{P}|_1 < \infty$, *k be a bounded, measurable kernel* k *over* X *with separable RKHS* H, *and* $\Phi : X \to H$ *be the canonical feature map of* k. *Then for all* $\lambda > 0$ *there exists a bounded and measurable function* $h_\lambda : X \times Y \to \mathbb{R}$ *such that* $h_\lambda(x, y) \in \partial L_\tau(y - f_{\mathrm{P},\lambda}(x))$ *for all* $(x, y) \in X \times Y$ *and*

$$f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\, \mathbb{E}_{\mathrm{P}} h_\lambda \Phi \,. \tag{8}$$

With the help of Proposition 9 we can now state the following stability result.

Theorem 10 *Let* P, H, Φ, *and* h_λ *be as in Proposition 9. Then we have* $\|h_\lambda\|_\infty \leq |L_\tau|_1$ *and for all distributions* Q *on* $X \times Y$ *with* $|\mathrm{Q}|_1 < \infty$ *we have*

$$\|f_{\mathrm{P},\lambda} - f_{\mathrm{Q},\lambda}\|_H \leq \frac{1}{\lambda}\, \|\mathbb{E}_{\mathrm{P}} h_\lambda \Phi - \mathbb{E}_{\mathrm{Q}} h_\lambda \Phi\|_H \,. \tag{9}$$

Proof. Let us first show the upper bound for $\|h_\lambda\|_\infty$. To this end we observe

$$|h_\lambda(x, y)| \;\leq\; |\partial L_\tau(y - f_{\mathrm{P},\lambda}(x))| \;\leq\; \big|L_\tau(y - \cdot)_{|\,[-f_{\mathrm{P},\lambda}(x),\, f_{\mathrm{P},\lambda}(x)]}\big|_1 \;\leq\; |L_\tau|_1 \,,$$

and hence we deduce $\|h_\lambda\|_\infty \leq |L_\tau|_1$. In order to prove (9) we first observe that the definition of the subdifferential yields

$$h(x, y)\big(f_{\mathrm{Q},\lambda}(x) - f_{\mathrm{P},\lambda}(x)\big) \leq L_\tau\big(y - f_{\mathrm{Q},\lambda}(x)\big) - L_\tau\big(y - f_{\mathrm{P},\lambda}(x)\big) \,,$$

and hence

$$\mathbb{E}_{\mathrm{Q}} L_\tau\big(Y - f_{\mathrm{P},\lambda}(X)\big) + \langle f_{\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda} \,,\, \mathbb{E}_{\mathrm{Q}} h\Phi\rangle \leq \mathbb{E}_{\mathrm{Q}} L_\tau\big(Y - f_{\mathrm{Q},\lambda}(X)\big) \,. \tag{10}$$

Moreover an easy calculation shows

$$\lambda\, \|f_{\mathrm{P},\lambda}\|_H^2 + 2\lambda\, \langle f_{\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda} \,,\, f_{\mathrm{P},\lambda}\rangle + \lambda \|f_{\mathrm{P},\lambda} - f_{\mathrm{Q},\lambda}\|_H^2 = \lambda \|f_{\mathrm{Q},\lambda}\|_H^2 \,. \tag{11}$$

Combining (10) and (11) it follows

$$\begin{aligned} \mathcal{R}^{reg}_{L_\tau,\mathrm{Q},\lambda}(f_{\mathrm{P},\lambda}) + \langle f_{\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda} \,,\, \mathbb{E}_{\mathrm{Q}} h\Phi + 2\lambda f_{\mathrm{P},\lambda}\rangle + \lambda \|f_{\mathrm{P},\lambda} - f_{\mathrm{Q},\lambda}\|_H^2 \;&\leq\; \mathcal{R}^{reg}_{L_\tau,\mathrm{Q},\lambda}(f_{\mathrm{Q},\lambda}) \\ &\leq\; \mathcal{R}^{reg}_{L_\tau,\mathrm{Q},\lambda}(f_{\mathrm{P},\lambda}) \,. \end{aligned}$$

Therefore by using the representation $f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\, \mathbb{E}_{\mathrm{P}} h\Phi$ we obtain

$$\begin{aligned} \lambda \|f_{\mathrm{P},\lambda} - f_{\mathrm{Q},\lambda}\|_H^2 \;&\leq\; \langle f_{\mathrm{P},\lambda} - f_{\mathrm{Q},\lambda} \,,\, \mathbb{E}_{\mathrm{Q}} h\Phi - \mathbb{E}_{\mathrm{P}} h\Phi\rangle \\ &\leq\; \|f_{\mathrm{P},\lambda} - f_{\mathrm{Q},\lambda}\|_H \,\cdot\, \|\mathbb{E}_{\mathrm{Q}} h\Phi - \mathbb{E}_{\mathrm{P}} h\Phi\|_H \,. \end{aligned}$$

From this we easily obtain the assertion. □

Proof of Proposition 4. The implication *i*) ⇒ *ii*) immediately follows from Theorem 3 of Steinwart *et al.* (2006) and the converse implication can be easily established by combining Theorem 8 with a straighforward modification of Example 5 of Steinwart *et al.* (2006). □

In order to prove Theorem 5 we need some preliminary results. Our first lemma shows that the influence of the regularization term $\lambda\|f_{\mathrm{P},\lambda}\|_H^2$ used in the definition of KBQR vanishes for $\lambda \to 0$.

Lemma 11 *Let H be an RKHS over X with bounded kernel k and P be a distribution on $X \times Y$ such that $|\mathrm{P}|_1 < \infty$. Then we have*

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

Proof. For $\varepsilon > 0$ we fix an $f_\varepsilon \in H$ such that $\mathcal{R}_{L_\tau,\mathrm{P}}(f_\varepsilon) \leq \mathcal{R}^*_{L_\tau,\mathrm{P},H} + \varepsilon$. Then for all $\lambda < \varepsilon\|f_\varepsilon\|_H^{-2}$ we have

$$\mathcal{R}_{L_\tau,\mathrm{P},H} \leq \lambda\|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{P},\lambda}) \leq \lambda\|f_\varepsilon\|_H^2 + \mathcal{R}_{L_\tau,\mathrm{P}}(f_\varepsilon) \leq 2\varepsilon + \mathcal{R}^*_{L_\tau,\mathrm{P},H}\,. \qquad \square$$

The next lemma gives a simple but useful approximation of $\bigl|\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}_{L_\tau,\mathrm{P}}(g)\bigr|$.

Lemma 12 *Let P be a distribution on $X \times Y$ with $|\mathrm{P}|_1 < \infty$. For all bounded measurable functions $f, g : X \to Y$ we have*

$$\bigl|\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}_{L_\tau,\mathrm{P}}(g)\bigr| \leq |L_\tau|_1 \,\|f - g\|_\infty\,.$$

Proof. The Lipschitz continuity of L_τ immediately gives

$$\bigl|\mathcal{R}_{L_\tau,\mathrm{P}}(f) - \mathcal{R}_{L_\tau,\mathrm{P}}(g)\bigr| \leq \int \bigl|L_\tau\bigl(y - f(x)\bigr) - L_\tau\bigl(y - g(x)\bigr)\bigr|\,d\mathrm{P}(x,y) \leq |L_\tau|_1\,\|f - g\|_\infty\,. \quad \square$$

Under the assumptions of Lemma 11 and Proposition 4 we immediately see that $\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) \to \mathcal{R}^*_{L_\tau,\mathrm{P}}$ holds for $\lambda_n \to 0$. Therefore, we obtain risk consistency whenever we can show that $|\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D}_n,\lambda_n})| \to 0$ holds in probability for $n \to \infty$ and suitable null sequences (λ_n). Our main tool for ensuring this convergence will be Theorem 10 which in particular describes the behavior of $\|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D}_n,\lambda_n}\|_\infty$ if we let Q be an empirical measure based on a sample set of length n. Lemma 12 showed how the norm of this difference can be used to estimate $|\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D}_n,\lambda_n})|$.

Let us now deal with the stochastic analysis of $|\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D}_n,\lambda})| \to 0$. To this end we need the following theorem which can be found in Chapter 3 of Yurinsky (1995).

Theorem 13 (Hoeffding's inequality in Hilbert spaces) *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, H be a separable Hilbert space, and $B > 0$. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to H$ be independent H-valued, bounded random variables with $\|\xi_i\|_\infty \leq B$ for all $i = 1, \ldots, n$. Then for all $\varepsilon \geq n^{-1/2}$ we have*

$$\mathrm{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \bigl(\xi_i - \mathbb{E}_\mathrm{P}\xi_i\bigr)\right\|_H \geq \varepsilon\right) \leq \exp\left(-\frac{3}{8}\cdot\frac{\varepsilon^2 n}{\varepsilon B + 3B^2}\right).$$

Proof of Theorem 5. To avoid handling with too many constants let us assume $\|k\|_\infty = 1$. Obviously, this implies $\|f\|_\infty \leq \|k\|_\infty \|f\|_H \leq \|f\|_H$ for all $f \in H$ and hence Lemma 12 implies

$$\bigl|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| \leq |L_\tau|_1\,\|f_{\mathrm{P},\lambda_n} - g\|_\infty \leq \|f_{\mathrm{P},\lambda_n} - g\|_H \qquad (12)$$

for all $g \in H$. For $n \in \mathbb{N}$ and $\lambda_n > 0$ we now write $h_n : X \times Y \to \mathbb{R}$ for the function we obtain by Proposition 9 and Theorem 10. Moreover, let $\varepsilon > 0$ and D_n be a training set of length n with empirical distribution D_n such that

$$\|\mathbb{E}_{\mathrm{P}} h_n \Phi - \mathbb{E}_{\mathrm{D}_n} h_n \Phi\|_H \leq \lambda_n \varepsilon \,. \tag{13}$$

Then Theorem 10 gives $\|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D}_n,\lambda_n}\|_H \leq \varepsilon$ and hence (12) yields

$$\left|\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_n}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_n,\lambda_n})\right| \leq \|f_{\mathrm{P},\lambda_n} - f_{\mathrm{D}_n,\lambda_n}\|_H \leq \varepsilon \,. \tag{14}$$

Let us now estimate the probability of D_n satisfying (13). To this end we first observe that $\lambda_n n^{1/2} \to \infty$ implies that for all sufficiently large n we have $\lambda_n \varepsilon \geq n^{-1/2}$. Moreover, Theorem 10 shows $\|h_n\|_\infty \leq 1$ and our assumption $\|k\|_\infty = 1$ thus yields $\|h_n \Phi\|_\infty \leq 1$. Consequently, Theorem 13 yields

$$P^n\big(D_n \in (X \times Y)^n \,:\, \|\mathbb{E}_{\mathrm{P}} h_n \Phi - \mathbb{E}_{\mathrm{D}_n} h_n \Phi\|_H \leq \lambda_n \varepsilon\big) \geq 1 - \exp\left(-\frac{3}{8} \cdot \frac{\varepsilon^2 \lambda_n^2 n}{\varepsilon \lambda_n + 3}\right)$$

for all sufficiently large n. Using $\lambda_n n^{1/2} \to \infty$ and $\lambda_n \to 0$ we thus find that the probability of sample sets D_n satisfying (13) converges to 1 if $|D_n| = n \to \infty$. As we have seen above this implies that (14) 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 n and hence we finally obtain the assertion. □

Proof of Theorem 6. We have already seen in Theorem 5 that the KBQR estimator satisfies

$$\mathcal{R}_{L_\tau,\mathrm{P}}(f_{\mathrm{D}_n,\lambda_n}) \to \mathcal{R}^*_{L_\tau,\mathrm{P}}$$

in probability for $n \to \infty$. Moreover, $(y,t) \mapsto L_\tau(y,t)$ is a supervised convex loss function in the sense of Steinwart (2005) whose conditional risks have a unique minimizer, namely $f^*_{\tau,\mathrm{P}}$. Consequently, Theorem 3.16 of Steinwart (2005) in the form of Remark 3.18 yields the assertion. □

References

Christmann, A. and Steinwart, I. (2005). Consistency and robustness of kernel based regression. University of Dortmund, SFB-475, TR-01/05. Submitted.

DeVito, E., Rosasco, L., Caponnetto, A., Piana, M., and Verri, A. (2004). Some properties of regularized kernel methods. *Journal of Machine Learning Research*, **5**, 1363–1390.

Diestel, J. and Uhl, J. (1977). *Vector Measures*. American Mathematical Society, Providence.

Ekeland, I. and Turnbull, T. (1983). *Infinite-dimensional Optimization and Convexity*. Chicago Lectures in Mathematics. The University of Chicago Press.

He, X. and Ng, P. (1999). COBS: Qualitatively constrained smoothing via linear programming. *Computational Statistics*, **14**, 315–337.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33–50.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, **81**, 673–680.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, **2**, 67–93.

Steinwart, I. (2003). Sparseness of support vector machines. *Journal of Machine Learning Research*, **4**, 1071–1105.

Steinwart, I. (2005). How to compare different loss functions. *Constr. Approx.*, **accepted**.

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

Takeuchi, I., Le, Q., Sears, T., and Smola, A. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, **7**, 1231–1264.

Yurinsky, V. (1995). *Sums and Gaussian Vectors*, volume 1617 of *Lecture Notes in Math.* Springer.

Andreas Christmann
Vrije Universiteit Brussel
Department of Mathematics
Pleinlaan 2
B–1050 Brussel
BELGIUM
E-mail:
andreas.christmann@vub.ac.be

Ingo Steinwart
CCS-3
Mail Stop B256
Los Alamos National Laboratory
Los Alamos, NM 87545
USA
E-mail:
ingo@lanl.gov