Supplementary: Joint quantile regression in vector-valued RKHSs

Maxime Sangnier **Olivier Fercoq** **Florence d'Alché-Buc**
LTCI, CNRS, Télécom ParisTech
Université Paris-Saclay
75013, Paris, France
{`maxime.sangnier, olivier.fercoq, florence.dalche`}
`@telecom-paristech.fr`

A Detailed proofs

A.1 Generalization

This section describes a proof of the generalization bound given in the corpus of the paper. The result is based on a concentration inequality à la Bartlett and Mendelson [1] for vector-valued functions [4], along with a bound on the Rademacher complexity of operator-valued kernel based hypothesis sets [7]. Before stating these two critical properties, let us remind the definition of the Rademacher complexity, used to quantify the complexity of a class of functions.

Definition A.1 (Rademacher complexity [4]). *Let $(X_i)_{1 \le i \le n} \in \mathcal{X}^n$ be an independent and identically distributed (*iid*) sample of random variables and $(\epsilon_{i,j})_{\substack{1 \le i \le n \\ 1 \le j \le p}} \in \{-1, 1\}^{n \times p}$ be $n \times p$ independent Rademacher variables (*i.e. *uniformly distributed on $\{-1, 1\}$). Let now $\mathcal{F} \subset (\mathbb{R}^p)^{\mathcal{X}}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}^p$. The Rademacher complexity (or average) of the class $\mathcal{F}$ is defined as:*

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{\substack{1 \le i \le n \\ 1 \le j \le p}} \epsilon_{i,j} f_j(X_i) \right],$$

where the expectation is computed jointly on $(X_i)_{1 \le i \le n}$ and $(\epsilon_{i,j})_{\substack{1 \le i \le n \\ 1 \le j \le p}}$.

Proposition A.1 (Concentration for Lipschitz hypotheses). *Let $X \in \mathcal{X}$ and $(X_i)_{1 \le i \le n} \in \mathcal{X}^n$ be* iid *random variables, $\mathcal{F} \subset (\mathbb{R}^p)^{\mathcal{X}}$ a class of functions. Let $\phi \colon \mathbb{R}^p \to [a, b]$ ($a, b \in \mathbb{R}$) be a Lipschitz continuous mapping with Lipschitz constant L_ϕ:*

$$\forall (\boldsymbol{z}, \boldsymbol{z}') \in \mathbb{R}^p \colon |\phi(\boldsymbol{z}) - \phi(\boldsymbol{z}')| \le L_\phi \left\| \boldsymbol{z} - \boldsymbol{z}' \right\|_{\ell_2}.$$

Let $\delta \in (0, 1]$, then with probability at least $1 - \delta$:

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}\left[\phi(f(X))\right] - \frac{1}{n} \sum_{i=1}^{n} \phi(f(X_i)) \right) \le 2\sqrt{2} L_\phi \mathcal{R}_n(\mathcal{F}) + (b - a) \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. Thanks to the assumptions above, with probability at least $1 - \delta$, we have [1, 3]:

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}\left[\phi(f(X))\right] - \frac{1}{n} \sum_{i=1}^{n} \phi(f(X_i)) \right) \le 2\mathcal{R}_n(\Phi) + (b - a) \sqrt{\frac{\log(1/\delta)}{2n}},$$

where $\Phi = \{\phi \circ f : f \in \mathcal{F}\}$. Then, using [5, Corollary 1], we obtain:

$$\mathcal{R}_n(\Phi) \le \sqrt{2} L_\phi \mathcal{R}_n(\mathcal{F}).$$

Gathering both equations concludes the proof. □

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Proposition A.2 (Bound on the Rademacher average [7, Theorem 3.1])**.** *Assume that there exists* $\kappa \in \mathbb{R}_+$ *such that:* $\sup_{\mathbf{x}\in\mathcal{X}} \operatorname{tr}(K(\mathbf{x},\mathbf{x})) \leq \kappa$ *and let* $\mathcal{F} = \{f \in \mathcal{K}_K, \|f\|_{\mathcal{K}} \leq c\}$ *for a given* $c \in \mathbb{R}_+$*. Then:*

$$\mathcal{R}_n(\mathcal{F}) \leq c\sqrt{\frac{\kappa}{n}}.$$

Theorem A.3 (Generalization)**.** *Let* $\boldsymbol{\tau} \in (0,1)^p$*,* $((X_i,Y_i))_{1\leq i\leq n} \in (\mathcal{X}\times\mathcal{Y})^n$ *be* iid *random variables (independent from* (X,Y)*),* $a \in \mathbb{R}_+$ *such that* $\sup_{y\in\mathcal{Y}} |y| \leq a$ *and* $\boldsymbol{b} \in \mathcal{Y}^p$*. Let* $\mathcal{H} = \{f + \boldsymbol{b} \colon f \in \mathcal{K}_K / \|f\|_{\mathcal{K}} \leq c\}$*, for a given* $c \in \mathbb{R}_+$*, be the class of hypotheses. Moreover, assume that there exists* $\kappa \in \mathbb{R}_+$ *such that:* $\sup_{\mathbf{x}\in\mathcal{X}} \operatorname{tr}(K(\mathbf{x},\mathbf{x})) \leq \kappa$ *and, for a hypothesis* h*, let us denote*

$$\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\boldsymbol{\tau}}(Y_i\mathbb{1} - h(X_i)),$$

the random variable associated to the empirical risk. Let $\delta \in (0,1]$*, then with probability at least* $1-\delta$*:*

$$\forall h \in \mathcal{H}\colon R(h) \leq \hat{R}_n(h) + 2\sqrt{2}c\sqrt{\frac{p\kappa}{n}} + (2pa + c\sqrt{p\kappa})\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. Let $\mathcal{F} = \{f \in \mathcal{K}_K, \|f\|_{\mathcal{K}} \leq c\}$. The proof begins with the following lemma.

Lemma A.4. *Under the assumptions of Theorem A.3:* $\forall (f,\mathbf{x}) \in \mathcal{F}\times\mathcal{X}, \|f(\mathbf{x})\|_{\ell_2} \leq c\sqrt{\kappa}$*.*

Proof of Lemma A.4.

$$\begin{aligned} &\forall (f,\mathbf{x},\boldsymbol{y}) \in \mathcal{F}\times\mathcal{X}\times\mathbb{R}^p, \\ &\qquad \langle f(\mathbf{x}) \mid \boldsymbol{y}\rangle_{\ell_2} = \langle K_{\mathbf{x}}\boldsymbol{y} \mid f\rangle_{\mathcal{K}} \\ &\qquad\qquad \leq \|f\|_{\mathcal{K}} \|K_{\mathbf{x}}\boldsymbol{y}\|_{\mathcal{K}} \\ &\qquad\qquad \leq c\sqrt{\langle K_{\mathbf{x}}\boldsymbol{y} \mid K_{\mathbf{x}}\boldsymbol{y}\rangle_{\mathcal{K}}} \\ &\qquad\qquad = c\sqrt{\langle \boldsymbol{y} \mid K(\mathbf{x},\mathbf{x})\boldsymbol{y}\rangle_{\ell_2}}. \end{aligned}$$

From the properties of operator-valued kernels, we know that $K(\mathbf{x},\mathbf{x})$ is symmetric positive semi-definite. Thus, denoting $(\lambda_j)_{1\leq j\leq p}$ its (non-negative) eigenvalues and $(e_j)_{1\leq j\leq p}$ the corresponding orthonormal basis, we obtain when $\|\boldsymbol{y}\|_{\ell_2} \leq 1$:

$$\begin{aligned} \langle \boldsymbol{y} \mid K(\mathbf{x},\mathbf{x})\boldsymbol{y}\rangle_{\ell_2} &= \sum_{i,j} \langle \boldsymbol{y} \mid e_i\rangle_{\ell_2} \langle \boldsymbol{y} \mid e_j\rangle_{\ell_2} \langle e_i \mid K(\mathbf{x},\mathbf{x})e_j\rangle_{\ell_2} \\ &= \sum_j \langle \boldsymbol{y} \mid e_j\rangle_{\ell_2}^2 \lambda_j \\ &\leq \sum_j \lambda_j && \text{(since } \|\boldsymbol{y}\|_{\ell_2} \leq 1\text{)} \\ &= \operatorname{tr}(K(\mathbf{x},\mathbf{x})). \end{aligned}$$

Thus: $\forall (f,\mathbf{x}) \in \mathcal{F}\times\mathcal{X}, \|f(\mathbf{x})\|_{\ell_2} = \sup_{\|\boldsymbol{y}\|_{\ell_2}\leq 1} \langle f(\mathbf{x}) \mid \boldsymbol{y}\rangle_{\ell_2} \leq c\sqrt{\kappa}$. □

In order to apply Proposition A.1, let us observe that the loss function $\ell_{\boldsymbol{\tau}}$ is $\sqrt{p}$-Lipschitz:

$$\begin{aligned}
&\forall(\boldsymbol{r}, \boldsymbol{r}') \in \mathbb{R}^p, \\
&\qquad \ell_{\boldsymbol{\tau}}(\boldsymbol{r}) = \ell_{\boldsymbol{\tau}}(\boldsymbol{r} - \boldsymbol{r}' + \boldsymbol{r}') \\
&\qquad = \sum_{j=1}^{p} \max\left(\tau_j(r_j - r'_j + r'_j), (\tau_j - 1)(r_j - r'_j + r'_j)\right) \\
&\qquad = \sum_{j=1}^{p} \begin{cases} \tau_j(r_j - r'_j + r'_j) & \text{if } r_j - r'_j + r'_j \geq 0 \\ (\tau_j - 1)(r_j - r'_j + r'_j) & \text{if } r_j - r'_j + r'_j \leq 0 \end{cases} \\
&\qquad = \sum_{j=1}^{p} \begin{cases} \tau_j(r_j - r'_j) + \tau_j r'_j & \text{if } r_j - r'_j + r'_j \geq 0 \\ (\tau_j - 1)(r_j - r'_j) + (\tau_j - 1)r'_j & \text{if } r_j - r'_j + r'_j \leq 0 \end{cases} \\
&\qquad \leq \sum_{j=1}^{p} \begin{cases} |r_j - r'_j| + \tau_j r'_j & \text{if } r_j - r'_j + r'_j \geq 0 \\ |r_j - r'_j| + (\tau_j - 1)r'_j & \text{if } r_j - r'_j + r'_j \leq 0 \end{cases} \\
&\qquad \leq \sum_{j=1}^{p} \left(|r_j - r'_j| + \max(\tau_j r'_j, (\tau_j - 1)r'_j)\right) \\
&\qquad = \|\boldsymbol{r} - \boldsymbol{r}'\|_{\ell_1} + \ell_{\boldsymbol{\tau}}(\boldsymbol{r}').
\end{aligned}$$

Switching $\boldsymbol{r}$ and $\boldsymbol{r}'$ we get $|\ell_{\boldsymbol{\tau}}(\boldsymbol{r}) - \ell_{\boldsymbol{\tau}}(\boldsymbol{r}')| \leq \|\boldsymbol{r} - \boldsymbol{r}'\|_{\ell_1}$. Since by Cauchy-Schwarz inequality $\|\boldsymbol{r} - \boldsymbol{r}'\|_{\ell_1} \leq \sqrt{p}\, \|\boldsymbol{r} - \boldsymbol{r}'\|_{\ell_2}$, we obtain that $\ell_{\boldsymbol{\tau}}$ is $\sqrt{p}$-Lipschitz.

In addition, $\ell_{\boldsymbol{\tau}}$ is bounded for the residuals of interest:

$$\begin{aligned}
&\forall(f, \mathbf{x}, \boldsymbol{y}) \in \mathcal{F} \times \mathcal{X} \times \mathcal{Y}^p, \\
&\qquad 0 \leq \ell_{\boldsymbol{\tau}}(\boldsymbol{y} - f(\mathbf{x}) - \boldsymbol{b}) \leq \|\boldsymbol{y} - f(\mathbf{x}) - \boldsymbol{b}\|_{\ell_1} \\
&\qquad\qquad \leq \|\boldsymbol{y} - \boldsymbol{b}\|_{\ell_1} + \|f(\mathbf{x})\|_{\ell_1} \\
&\qquad\qquad \leq 2pa + \sqrt{p}\, \|f(\mathbf{x})\|_{\ell_2} \\
&\qquad\qquad \leq 2pa + c\sqrt{p\kappa}.
\end{aligned}$$

Let $\mathcal{U} = \{u \colon (\mathbf{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathbb{R}^p \mapsto \boldsymbol{y} - f(\mathbf{x}) - \boldsymbol{b}, f \in \mathcal{F}\}$. By Proposition A.1 we have with probability at least $1 - \delta$:

$$\sup_{u \in \mathcal{U}} \left(\mathbb{E}\left[\ell_{\boldsymbol{\tau}}(u(X, Y))\right] - \frac{1}{n} \sum_{i=1}^{n} \ell_{\boldsymbol{\tau}}(u(X_i, Y_i)) \right) \leq 2\sqrt{2p}\mathcal{R}_n(\mathcal{U}) + (2pa + c\sqrt{p\kappa}) \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Let $(\epsilon_{i,j})_{\substack{1\leq i\leq n\\1\leq j\leq p}}$ be an *iid* sample of Rademacher random variables. Then:

$$\begin{aligned}
\mathcal{R}_n(\mathcal{U}) &= \mathbb{E}\left[\sup_{u\in\mathcal{U}}\left(\frac{1}{n}\sum_{\substack{1\leq i\leq n\\1\leq j\leq p}}\epsilon_{i,j}u_j(X_i,Y_i)\right)\right]\\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{\substack{1\leq i\leq n\\1\leq j\leq p}}\epsilon_{i,j}(Y_i-f_j(X_i)-b_j)\right)\right]\\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{\substack{1\leq i\leq n\\1\leq j\leq p}}\epsilon_{i,j}f_j(X_i)\right)\right]\\
&\quad+\mathbb{E}\left[\frac{1}{n}\sum_{\substack{1\leq i\leq n\\1\leq j\leq p}}\epsilon_{i,j}Y_i\right]+b_j\mathbb{E}\left[\frac{1}{n}\sum_{\substack{1\leq i\leq n\\1\leq j\leq p}}\epsilon_{i,j}\right]\\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_{i,j}f_j(X_i)\right)\right]\\
&\quad+\frac{1}{n}\sum_{\substack{1\leq i\leq n\\1\leq j\leq p}}\mathbb{E}\left[\epsilon_{i,j}\right]\mathbb{E}\left[Y_i\right]+b_j\frac{1}{n}\sum_{\substack{1\leq i\leq n\\1\leq j\leq p}}\mathbb{E}\left[\epsilon_{i,j}\right]\\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{\substack{1\leq i\leq n\\1\leq j\leq p}}\epsilon_{i,j}f_j(X_i)\right)\right]\\
&= \mathcal{R}_n(\mathcal{F})\\
&\leq c\sqrt{\frac{\kappa}{n}} \qquad\qquad \text{(Proposition A.2).}
\end{aligned}$$

This concludes the proof. □

A.2 Quantile deviation

Given a vector of probabilities $\boldsymbol{\tau}\in(0,1)^p$ and a quantile estimator $\hat{h}\colon\mathcal{X}\to\mathbb{R}^p$, we are interested in controlling the deviation of $\mathbb{E}\left[\mathbb{P}\left(Y\leq\hat{h}_j(X)\mid X\right)\right]$ from τ_j (for a particular $j\in\mathbb{N}_p$). For this purpose, we would like to derive a uniform bound using the scalar counterpart of Proposition A.1 (which is identical but substituting $\sqrt{2}$ by 1 [3]). Since such a bound is true for all hypothesis h, we do not require $\mathbb{E}\left[\mathbb{P}\left(Y\leq h_j(X)\mid X\right)\right]$ to be close to τ_j, but to its empirical twin $\frac{1}{n}\sum_{i=1}^n \mathrm{I}_{\mathbb{R}_-}\left(Y_i-h_j(X_i)\right)$, where $\mathrm{I}_{\mathbb{R}_-}$ is the indicator function of the set $\mathbb{R}_-$. For a quantile estimator $\hat{h}$, the quantile property states that τ_j is sufficiently close to $\frac{1}{n}\sum_{i=1}^n \mathrm{I}_{\mathbb{R}_-}\left(y_i-\hat{h}_j(\mathbf{x}_i)\right)$.

Now, remark that:

$$\mathbb{E}\left[\mathbb{P}\left(Y\leq h_j(X)\mid X\right)\right]=\mathbb{E}\left[\mathbb{E}\left[\mathrm{I}_{\mathbb{R}_-}\left(Y-h_j(X)\right)\mid X\right]\right]=\mathbb{E}\left[\mathrm{I}_{\mathbb{R}_-}\left(Y-h_j(X)\right)\right].$$

Thus, the two quantities to compare clearly appear as an expected and an empirical costs based on the loss function $\mathrm{I}_{\mathbb{R}_-}$. Unfortunately, that loss function is not Lipschitz continuous. In order to circumvent that pitfall, we introduce an artificial margin $\epsilon>0$ and two ramp functions Γ_ϵ^- and Γ_ϵ^+ (see definition in Theorem A.5). These surrogate mappings are $\frac{1}{\epsilon}$-Lipschitz and respectively lower and upper bound $\mathrm{I}_{\mathbb{R}_-}$.

Finally, Theorem A.5 states that $\mathbb{E}\left[\mathbb{P}\left(Y\leq h_j(X)\mid X\right)\right]$ is uniformly bounded by the *empirical quantile levels* $\frac{1}{n}\sum_{i=1}^n\Gamma_\epsilon^-(Y_i-h_j(X_i))$ and $\frac{1}{n}\sum_{i=1}^n\Gamma_\epsilon^+(Y_i-h_j(X_i))$ to an additive bias in $O(1/\sqrt{n})$.

Theorem A.5 (Quantile deviation)**.** *Let* $\boldsymbol{\tau} \in (0,1)^p$*,* $((X_i, Y_i))_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ *be* iid *random variables (independent from* (X, Y)*) and* $\boldsymbol{b} \in \mathbb{R}^p$*. Let* $\mathcal{H} = \{f + \boldsymbol{b} \colon f \in \mathcal{K}_K / \|f\|_{\mathcal{K}} \leq c\}$*, for a given* $c \in \mathbb{R}_+$*, be the class of hypotheses. Moreover, assume that there exists* $\kappa \in \mathbb{R}_+$ *such that:* $\sup_{\mathbf{x} \in \mathcal{X}} \operatorname{tr}(K(\mathbf{x}, \mathbf{x})) \leq \kappa$*. Let* $\epsilon > 0$ *be an artificial margin,*

$$\Gamma_\epsilon^+ \colon r \in \mathbb{R} \mapsto \operatorname{proj}_{[0,1]}\left(1 - \frac{r}{\epsilon}\right) \quad \textit{and} \quad \Gamma_\epsilon^- \colon r \in \mathbb{R} \mapsto \operatorname{proj}_{[0,1]}\left(-\frac{r}{\epsilon}\right),$$

two ramp functions, $j \in \mathbb{N}_p$ *and* $\delta \in (0, 1]$*. Then with probability at least* $1 - \delta$*:*

$$\forall h \in \mathcal{H} \colon \frac{1}{n} \sum_{i=1}^n \Gamma_\epsilon^-(Y_i - h_j(X_i)) - \Delta \leq \mathbb{E}\left[\mathbb{P}\left(Y \leq h_j(X) \mid X\right)\right] \leq \frac{1}{n} \sum_{i=1}^n \Gamma_\epsilon^+(Y_i - h_j(X_i)) + \Delta,$$

where $\Delta = \frac{2c}{\epsilon}\sqrt{\frac{\kappa}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$.

Proof. First, let us remind that: $\mathbb{E}\left[\mathbb{P}\left(Y \leq h_j(X) \mid X\right)\right] = \mathbb{E}\left[\mathrm{I}_{\mathbb{R}_-}\left(Y - h_j(X)\right)\right]$, and $\forall r \in \mathbb{R}, \Gamma_\epsilon^-(r) \leq \mathrm{I}_{\mathbb{R}_-}(r) \leq \Gamma_\epsilon^+(r)$. Thus,

$$\mathbb{E}\left[\Gamma_\epsilon^-\left(Y - h_j(X)\right)\right] \leq \mathbb{E}\left[\mathbb{P}\left(Y \leq h_j(X) \mid X\right)\right] \leq \mathbb{E}\left[\Gamma_\epsilon^+\left(Y - h_j(X)\right)\right].$$

Then, remarking that Γ_ϵ^+ is $\frac{1}{\epsilon}$-Lipschitz, we obtain (by the same reasoning as for the proof of Theorem A.3 and using [3, Theorem 1]):

$$\forall h \in \mathcal{H} \colon \mathbb{E}\left[\Gamma_\epsilon^+\left(Y - h_j(X)\right)\right] \leq \frac{1}{n} \sum_{i=1}^n \Gamma_\epsilon^+(Y_i - h_j(X_i)) + \frac{2c}{\epsilon}\sqrt{\frac{\kappa}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with probability at least $1 - \delta$. Respectively, for $-\Gamma_\epsilon^-$, with probability at least $1 - \delta$:

$$\forall h \in \mathcal{H} \colon -\mathbb{E}\left[\Gamma_\epsilon^-\left(Y - h_j(X)\right)\right] \leq -\frac{1}{n} \sum_{i=1}^n \Gamma_\epsilon^-(Y_i - h_j(X_i)) + \frac{2c}{\epsilon}\sqrt{\frac{\kappa}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Gathering everything with the union bound concludes the proof. □

B Dual formulation

In this section, we derive a dual problem for learning a joint quantile regressor *with non-crossing constraints*. These last constraints are set thanks to a matrix $\boldsymbol{A}$ defined below. Considering $\boldsymbol{A}$ as the null-matrix gives a dual formulation for the learning problem *without non-crossing constraints*.

Let $C \in \mathbb{R}_+$, $\boldsymbol{\tau} \in (0,1)^p$, such that $\tau_j > \tau_{j+1}$ ($\forall j \in \mathbb{N}_{p-1}$) and the finite difference operator, embodied by the matrix:

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -1 \end{pmatrix} \in \mathbb{R}^{(p-1) \times p}.$$

The primal problem we are interested in (with the associated dual variables) is:

$$\begin{aligned} \underset{\substack{f \in \mathcal{H}, \boldsymbol{b} \in \mathbb{R}^p \\ \boldsymbol{\xi}, \boldsymbol{\xi}^* \in (\mathbb{R}^p)^n,}}{\text{minimize}} \quad & \frac{1}{2}\|f\|_{\mathcal{K}}^2 + C \sum_{i=1}^n \langle \boldsymbol{\tau} \mid \boldsymbol{\xi}_i \rangle_{\ell_2} + C \sum_{i=1}^n \langle \mathbb{1} - \boldsymbol{\tau} \mid \boldsymbol{\xi}_i^* \rangle_{\ell_2} \\ \text{s.t.} \quad & \left\{ \begin{array}{lll} \forall i \in \mathbb{N}_n \colon & \boldsymbol{y}_i - f(\mathbf{x}_i) - \boldsymbol{b} = \boldsymbol{\xi}_i - \boldsymbol{\xi}_i^* & : \boldsymbol{\alpha}_i \in \mathbb{R}^p \\ & \boldsymbol{\xi}_i \succcurlyeq 0 & : \boldsymbol{\beta}_i \in \mathbb{R}_+^p \\ & \boldsymbol{\xi}_i^* \succcurlyeq 0 & : \boldsymbol{\gamma}_i \in \mathbb{R}_+^p \\ & \boldsymbol{A}(f(\mathbf{x}_i) + b) \succcurlyeq 0 & : \boldsymbol{\delta}_i \in \mathbb{R}_+^{p-1}. \end{array} \right. \end{aligned}$$

The last constraint enforces the regressors not to cross on the training points (hard non-crossing constraints). Let us write the Lagrangian function:

$$\begin{aligned}
\mathfrak{L}(f,\boldsymbol{b},\boldsymbol{\xi}_i,\boldsymbol{\xi}_i^*,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i,\boldsymbol{\gamma}_i,\boldsymbol{\delta}_i) = {} & \frac{1}{2}\|f\|_{\mathcal{K}}^2 + C\sum_{i=1}^n \langle \boldsymbol{\tau} \mid \boldsymbol{\xi}_i\rangle_{\ell_2} + C\sum_{i=1}^n \langle \mathbb{1}-\boldsymbol{\tau} \mid \boldsymbol{\xi}_i^*\rangle_{\ell_2} \\
& + \sum_{i=1}^n \langle \boldsymbol{\alpha}_i \mid \boldsymbol{y}_i\rangle_{\ell_2} - \sum_{i=1}^n \langle \boldsymbol{\alpha}_i \mid f(\mathbf{x}_i)\rangle_{\ell_2} - \sum_{i=1}^n \langle \boldsymbol{\alpha}_i \mid \boldsymbol{b}\rangle_{\ell_2} \\
& - \sum_{i=1}^n \langle \boldsymbol{\alpha}_i \mid \boldsymbol{\xi}_i\rangle_{\ell_2} + \sum_{i=1}^n \langle \boldsymbol{\alpha}_i \mid \boldsymbol{\xi}_i^*\rangle_{\ell_2} \\
& - \sum_{i=1}^n \langle \boldsymbol{\beta}_i \mid \boldsymbol{\xi}_i\rangle_{\ell_2} - \sum_{i=1}^n \langle \boldsymbol{\gamma}_i \mid \boldsymbol{\xi}_i^*\rangle_{\ell_2} \\
& - \sum_{i=1}^n \langle \boldsymbol{\delta}_i \mid \boldsymbol{A}(f(\mathbf{x}_i)+b)\rangle_{\ell_2} \\
= {} & \frac{1}{2}\|f\|_{\mathcal{K}}^2 - \left\langle \sum_{i=1}^n K_{\mathbf{x}_i}(\boldsymbol{\alpha}_i + \boldsymbol{A}^\top\boldsymbol{\delta}_i) \mid f \right\rangle_{\mathcal{H}} + \sum_{i=1}^n \langle C\boldsymbol{\tau} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_i \mid \boldsymbol{\xi}_i\rangle_{\ell_2} \\
& + \sum_{i=1}^n \langle C(\mathbb{1}-\boldsymbol{\tau}) + \boldsymbol{\alpha}_i - \boldsymbol{\gamma}_i \mid \boldsymbol{\xi}_i^*\rangle_{\ell_2} - \left\langle \sum_{i=1}^n (\boldsymbol{\alpha}_i + \boldsymbol{A}^\top\boldsymbol{\delta}_i) \mid \boldsymbol{b} \right\rangle_{\ell_2} \\
& + \sum_{i=1}^n \langle \boldsymbol{\alpha}_i \mid \boldsymbol{y}_i\rangle_{\ell_2}.
\end{aligned}$$

First order optimality conditions for the primal variables give:

$$\begin{cases}
\nabla_f \mathfrak{L}(f,\boldsymbol{b},\boldsymbol{\xi}_i,\boldsymbol{\xi}_i^*,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i,\boldsymbol{\gamma}_i,\boldsymbol{\delta}_i) = f - \displaystyle\sum_{i=1}^n K_{\mathbf{x}_i}(\boldsymbol{\alpha}_i + \boldsymbol{A}^\top\boldsymbol{\delta}_i) = 0 \\
\nabla_b \mathfrak{L}(f,\boldsymbol{b},\boldsymbol{\xi}_i,\boldsymbol{\xi}_i^*,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i,\boldsymbol{\gamma}_i,\boldsymbol{\delta}_i) = -\displaystyle\sum_{i=1}^n (\boldsymbol{\alpha}_i + \boldsymbol{A}^\top\boldsymbol{\delta}_i) = 0 \\
\nabla_{\boldsymbol{\xi}_i} \mathfrak{L}(f,\boldsymbol{b},\boldsymbol{\xi}_i,\boldsymbol{\xi}_i^*,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i,\boldsymbol{\gamma}_i,\boldsymbol{\delta}_i) = C\boldsymbol{\tau} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_i = 0 \\
\nabla_{\boldsymbol{\xi}_i^*} \mathfrak{L}(f,\boldsymbol{b},\boldsymbol{\xi}_i,\boldsymbol{\xi}_i^*,\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i,\boldsymbol{\gamma}_i,\boldsymbol{\delta}_i) = C(\mathbb{1}-\boldsymbol{\tau}) + \boldsymbol{\alpha}_i - \boldsymbol{\gamma}_i = 0.
\end{cases}$$

That is:

$$\begin{cases}
f = \displaystyle\sum_{i=1}^n K_{\mathbf{x}_i}(\boldsymbol{\alpha}_i + \boldsymbol{A}^\top\boldsymbol{\delta}_i) \\
0 = \displaystyle\sum_{i=1}^n (\boldsymbol{\alpha}_i + \boldsymbol{A}^\top\boldsymbol{\delta}_i) \\
\boldsymbol{\beta}_i = C\boldsymbol{\tau} - \boldsymbol{\alpha}_i \\
\boldsymbol{\gamma}_i = C(\mathbb{1}-\boldsymbol{\tau}) + \boldsymbol{\alpha}_i.
\end{cases}$$

Recalling that $\boldsymbol{\beta}_i \succcurlyeq 0$ and $\boldsymbol{\gamma}_i \succcurlyeq 0$, we obtain: $C(\boldsymbol{\tau} - \mathbb{1}) \preccurlyeq \boldsymbol{\alpha}_i \preccurlyeq C\boldsymbol{\tau}$. Then, by substitution of the first order equations in the Lagrangian function, the linear expressions in the primal variables vanish and the quadratic part becomes:

$$\begin{aligned}
\frac{1}{2}\|f\|_{\mathcal{K}}^2 - \left\langle \sum_{i=1}^n K_{\mathbf{x}_i}(\boldsymbol{\alpha}_i + \boldsymbol{A}^\top\boldsymbol{\delta}_i) \mid f \right\rangle_{\mathcal{H}} &= \frac{1}{2}\sum_{i,j=1}^n \left\langle K_{\mathbf{x}_i}(\boldsymbol{\alpha}_i + \boldsymbol{A}^\top\boldsymbol{\delta}_i) \mid K_{\mathbf{x}_j}(\boldsymbol{\alpha}_j + \boldsymbol{A}^\top\boldsymbol{\delta}_j) \right\rangle_{\mathcal{H}} \\
&\quad - \sum_{i,j=1}^n \left\langle K_{\mathbf{x}_i}(\boldsymbol{\alpha}_i + \boldsymbol{A}^\top\boldsymbol{\delta}_i) \mid K_{\mathbf{x}_j}(\boldsymbol{\alpha}_j + \boldsymbol{A}^\top\boldsymbol{\delta}_j) \right\rangle_{\mathcal{H}} \\
&= -\frac{1}{2}\sum_{i,j=1}^n \left\langle (\boldsymbol{\alpha}_i + \boldsymbol{A}^\top\boldsymbol{\delta}_i) \mid K(\mathbf{x}_i,\mathbf{x}_j)(\boldsymbol{\alpha}_j + \boldsymbol{A}^\top\boldsymbol{\delta}_j) \right\rangle_{\ell_2}.
\end{aligned}$$

Gathering every thing, the dual problem writes:

$$\begin{aligned}
&\underset{\substack{\boldsymbol{\alpha}_i \in \mathbb{R}^p, \boldsymbol{\delta}_i \in \mathbb{R}^{p-1} \\ \forall i \in \mathbb{N}_n}}{\text{maximize}} \quad -\frac{1}{2} \sum_{i,j=1}^{n} \left\langle (\boldsymbol{\alpha}_i + \boldsymbol{A}^\top \boldsymbol{\delta}_i) \mid K(\mathbf{x}_i, \mathbf{x}_j)(\boldsymbol{\alpha}_j + \boldsymbol{A}^\top \boldsymbol{\delta}_j) \right\rangle_{\ell_2} + \sum_{i=1}^{n} \langle \boldsymbol{\alpha}_i \mid \boldsymbol{y}_i \rangle_{\ell_2} \\
&\text{s. t.} \quad \begin{cases} \forall i \in \mathbb{N}_n \colon C(\boldsymbol{\tau} - \mathbb{1}) \preccurlyeq \boldsymbol{\alpha}_i \preccurlyeq C\boldsymbol{\tau} \\ \qquad\qquad \boldsymbol{\delta}_i \succcurlyeq 0 \\ \displaystyle\sum_{i=1}^{n} (\boldsymbol{\alpha}_i + \boldsymbol{A}^\top \boldsymbol{\delta}_i) = 0_{\mathbb{R}^p}. \end{cases}
\end{aligned}$$

In order to simplify the previous problem, let $\boldsymbol{u}_i = \boldsymbol{\alpha}_i + \boldsymbol{A}^\top \boldsymbol{\delta}_i$ and remark that $\boldsymbol{A}\boldsymbol{y}_i = y_i(\boldsymbol{A}\mathbb{1}) = 0$. The new dual problem then becomes:

$$\begin{aligned}
&\underset{\substack{\boldsymbol{u}_i \in \mathbb{R}^p, \boldsymbol{\delta}_i \in \mathbb{R}^{p-1} \\ \forall i \in \mathbb{N}_n}}{\text{maximize}} \quad -\frac{1}{2} \sum_{i,j=1}^{n} \langle \boldsymbol{u}_i \mid K(\mathbf{x}_i, \mathbf{x}_j) \boldsymbol{u}_j \rangle_{\ell_2} + \sum_{i=1}^{n} \langle \boldsymbol{u}_i \mid \boldsymbol{y}_i \rangle_{\ell_2} \\
&\text{s. t.} \quad \begin{cases} \forall i \in \mathbb{N}_n \colon C(\boldsymbol{\tau} - \mathbb{1}) \preccurlyeq \boldsymbol{u}_i - \boldsymbol{A}^\top \boldsymbol{\delta}_i \preccurlyeq C\boldsymbol{\tau} \\ \qquad\qquad \boldsymbol{\delta}_i \succcurlyeq 0 \\ \displaystyle\sum_{i=1}^{n} \boldsymbol{u}_i = 0_{\mathbb{R}^p}. \end{cases}
\end{aligned}$$

Primal variables are recovered thanks to first order conditions. First, $f = \sum_{i=1}^n K_{\mathbf{x}_i} \boldsymbol{u}_i$. Second, the intercept $\boldsymbol{b}$ can be obtained either by detecting couples $(i, \ell) \in \mathbb{N}_n \times \mathbb{N}_p$ such that $C(\tau_\ell - 1) < (u_i)_\ell - (\boldsymbol{A}^\top \boldsymbol{\delta}_i)_\ell < C\tau_\ell$ (in this case $b_\ell = y_i - f_\ell(\mathbf{x}_i)$), or by remarking that $\boldsymbol{b}$ is a dual vector for the linear constraint $\sum_{i=1}^n \boldsymbol{u}_i = 0_{\mathbb{R}^p}$ (if one uses a primal-dual algorithm to solve the previous optimization problem).

When non-crossing constraints are dismissed ($\boldsymbol{A} = 0_{\mathbb{R}^{(p-1) \times p}}$), the regressor $h = f + \boldsymbol{b}$ satisfies the quantile property. Thus, knowing f, the intercepts b_ℓ can be recovered as τ_ℓ-quantiles of $(y_i - f_\ell(\mathbf{x}_i)_{1 \le i \le n}$.

C Algorithmic details

This section details an augmented Lagrangian scheme for estimating quantile regressors. We start with the dual formulation of the learning problem (without non-crossing constraints):

$$\begin{aligned}
&\underset{\boldsymbol{\alpha} \in (\mathbb{R}^p)^n}{\text{minimize}} \quad \frac{1}{2} \sum_{i,j=1}^{n} \langle \boldsymbol{\alpha}_i \mid K(\mathbf{x}_i, \mathbf{x}_j) \boldsymbol{\alpha}_j \rangle_{\ell_2} - \sum_{i=1}^{n} y_i \langle \boldsymbol{\alpha}_i \mid \mathbb{1} \rangle_{\ell_2} \\
&\text{s. t.} \quad \begin{cases} \forall i \in \mathbb{N}_n \colon C(\boldsymbol{\tau} - \mathbb{1}) \preccurlyeq \boldsymbol{\alpha}_i \preccurlyeq C\boldsymbol{\tau} \\ \displaystyle\sum_{i=1}^{n} \boldsymbol{\alpha}_i = 0_{\mathbb{R}^p}. \end{cases}
\end{aligned}$$

The method consists in solving the saddle point problem with an additional squared penalty [2]:

$$\begin{aligned}
\underset{\boldsymbol{b} \in \mathbb{R}^p}{\text{maximize}} \; \underset{\boldsymbol{\alpha} \in (\mathbb{R}^p)^n}{\text{minimize}} \quad & \frac{1}{2} \sum_{i,j=1}^{n} \langle \boldsymbol{\alpha}_i \mid K(\mathbf{x}_i, \mathbf{x}_j) \boldsymbol{\alpha}_j \rangle_{\ell_2} - \sum_{i=1}^{n} y_i \langle \boldsymbol{\alpha}_i \mid \mathbb{1} \rangle_{\ell_2} \\
& + \left\langle \boldsymbol{b} \mid \sum_{i=1}^{n} \boldsymbol{\alpha}_i \right\rangle_{\ell_2} + \frac{\mu}{2} \left\| \sum_{i=1}^{n} \boldsymbol{\alpha}_i \right\|_{\ell_2}^2 \\
\text{s. t.} \quad & \forall i \in \mathbb{N}_n \colon C(\boldsymbol{\tau} - \mathbb{1}) \preccurlyeq \boldsymbol{\alpha}_i \preccurlyeq C\boldsymbol{\tau},
\end{aligned}$$

where μ is a positive scalar. The next step is to split the optimization program into an outer problem (depending only on the variable $\boldsymbol{b}$) and an inner one (depending on $\boldsymbol{\alpha}$). This latter problem is:

$$\begin{aligned}
&\underset{\boldsymbol{\alpha} \in (\mathbb{R}^p)^n}{\text{minimize}} \quad \frac{1}{2} \sum_{i,j=1}^{n} \langle \boldsymbol{\alpha}_i \mid (K(\mathbf{x}_i, \mathbf{x}_j) + \mu \boldsymbol{I}) \boldsymbol{\alpha}_j \rangle_{\ell_2} + \sum_{i=1}^{n} \langle \boldsymbol{\alpha}_i \mid \boldsymbol{b} - y_i \mathbb{1} \rangle_{\ell_2} \\
&\text{s. t.} \quad \forall i \in \mathbb{N}_n \colon C(\boldsymbol{\tau} - \mathbb{1}) \preccurlyeq \boldsymbol{\alpha}_i \preccurlyeq C\boldsymbol{\tau},
\end{aligned} \tag{1}$$

Algorithm 1 Augmented Lagrangian algorithm

```
Initialize $\mu \leftarrow 10$, $\boldsymbol{b} \leftarrow 0_{\mathbb{R}^p}$.
repeat
    Solve Optimization Problem (1).
    Make a gradient step on $\boldsymbol{b}$ with step size $\mu$.
until $\|\sum_{i=1}^n \boldsymbol{\alpha}_i\|_{\ell_2}^2$ is small enough
```

Table 1: Empirical quantile loss ×100 (the closer to 0, the better).

Data set	IND.	IND. (NC)	MTFL	JQR
caution	4.50 ± 39.08	**3.33** ± 37.84	7.00 ± 32.57	○○○ 7.17 ± 36.40
ftcollinssnow	1.43 ± 38.49	1.79 ± 38.12	1.25 ± 38.50	○○○ **0.54** ± 35.77
highway	10.83 ± 70.12	10.83 ± 71.20	**7.50** ± 62.96	○○○ 15.00 ± 67.11
heights	-1.15 ± 9.88	-1.14 ± 9.88	**-0.76** ± 9.39	○○○ -1.26 ± 9.31
sniffer	-6.58 ± 26.45	-6.58 ± 27.59	**-3.95** ± 27.57	○○○ -4.08 ± 29.23
snowgeese	**-0.00** ± 44.03	1.07 ± 43.41	1.43 ± 50.03	○○○ -6.43 ± 44.94
ufc	0.58 ± 12.13	0.89 ± 11.97	**-0.31** ± 13.59	●●○ -1.79 ± 13.42
birthwt	2.02 ± 29.80	2.02 ± 29.80	2.11 ± 34.55	○○○ **0.88** ± 33.86
crabs	-2.42 ± 20.52	-1.25 ± 22.08	-2.17 ± 22.21	○○○ **-0.50** ± 21.75
GAGurine	1.95 ± 17.43	1.74 ± 17.39	**0.89** ± 16.71	○○○ 1.84 ± 16.89
geyser	1.22 ± 18.84	1.17 ± 18.66	**0.22** ± 19.20	○○○ 1.61 ± 19.04
gilgais	0.95 ± 18.35	0.95 ± 18.19	-0.64 ± 16.12	○○○ **0.18** ± 20.22
topo	**-19.38** ± 70.18	-19.38 ± 71.34	-20.00 ± 64.18	○○○ -20.31 ± 65.73
BostonHousing	7.40 ± 18.26	7.30 ± 17.94	5.72 ± 16.72	○○○ **5.30** ± 15.74
CobarOre	26.67 ± 72.49	25.42 ± 72.58	-22.08 ± 77.16	○○○ **20.00** ± 72.62
engel	-3.66 ± 18.71	-3.73 ± 18.72	-3.73 ± 19.73	○○○ **-2.39** ± 17.18
mcycle	**1.37** ± 29.20	1.75 ± 30.73	4.75 ± 28.42	○○○ 6.25 ± 30.84
BigMac2003	-4.76 ± 51.10	**-0.00** ± 45.24	1.67 ± 52.19	○○○ 0.24 ± 45.76
UN3	4.44 ± 19.11	4.21 ± 19.23	**2.62** ± 17.06	○○○ 4.76 ± 20.68
cpus	2.38 ± 20.30	3.57 ± 19.63	**1.51** ± 15.02	○○○ 1.67 ± 31.78

where $\boldsymbol{I}$ is the identity matrix. This inner optimization problem is a quadratic program with a box constraint. This is quite easily solvable. Thus, following [2], we can learn quantile estimators thanks to the simple alternate scheme described in Algorithm 1. In practice, the inner solver used in Algorithm 1 in order to get an approximate solution for Problem (1) is the primal dual-dual coordinate descent proposed in [6] with 10^4 as the maximum number of iterations.

D Numerical results

D.1 Quantile regression

Another criterion for assessing quantile regression methods is the quantile loss $\sum_{j=1}^p \left[\left[\frac{1}{n}\sum_{i=1}^n \mathrm{I}_{\mathbb{R}_-}(y_i - h_j(\mathbf{x}_i))\right] - \tau_j\right]$, where $\mathrm{I}_{\mathbb{R}_-}$ is the indicator function of the set $\mathbb{R}_-$. This loss measures the deviation of the estimators h_j to the prescribed quantile levels τ_j.

However, the quantile loss is quite an equivocal criterion, since it measures *how much* the *unconditional* quantile property is satisfied. This unconditional indicator is indeed the only way to get a piece of information concerning the *conditional* quantile property. For instance, Takeuchi et al. [8] empirically showed (with the same datasets) that the constant function based on the unconditional quantile estimator performs best under this criterion, even though it is expected to be a poor conditional quantile regressor. The numerical results in Table 1 follow this remark and the results previously obtained [8]. No significant ranking comes out.

D.2 Training algorithms

In order to compare the implementations of the three algorithms for solving the dual optimization problem of joint quantile regession, the following procedure has been set up: we first run QP, with a relative tolerance set to 10^{-2}, and store the optimal objective value. Then, the two other methods (AUG. LAG and PDCD) are launched and stopped when they pass the objective value reached by QP.

During the descent of PDCD, we used an efficient accumulated objective value, which is not exact since the iterate $\boldsymbol{\alpha}$ is not feasible to the linear constraint. Table 2 describes the average objective values (divided by the sample size) reached by each algorithm after projection of the best candidate

Table 2: Average objective value (divided by the sample size) reached in the second numerical experiment presented in the corpus of the paper.

Size	QP	AUG. LAG.	PDCD
250	-109.69 ± 5.41	-109.73 ± 5.41	**-109.72** ± 5.40
500	-109.84 ± 2.09	-109.85 ± 2.08	**-109.88** ± 2.11
1000	-104.13 ± 1.49	–	**-104.17** ± 1.49
2000	-106.35 ± 2.36	–	**-106.39** ± 2.38

onto the set of constraints. We can check that our approach (PDCD) always reaches a smaller objective value than the target QP. This validates our procedure.

References

[1] P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[2] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999. ISBN 9781886529007.

[3] S.M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*. 2009.

[4] A. Maurer. Bounds for Linear Multi-Task Learning. *Journal of Machine Learning Research*, 7:117–139, 2006.

[5] A. Maurer. A vector-contraction inequality for Rademacher complexities. In *Proceedings of The 27th International Conference on Algorithmic Learning Theory*, 2016.

[6] S. Shalev-Shwartz and T. Zhang. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *Journal of Machine Learning Research*, 14:567–599, 2013.

[7] V. Sindhwani, M.H. Quang, and A.C. Lozano. Scalable Matrix-valued Kernel Learning for High-dimensional Nonlinear Multivariate Regression and Granger Causality. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013.

[8] I. Takeuchi, Q.V. Le, T.D. Sears, and A.J. Smola. Nonparametric Quantile Estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.