SUPPLEMENTARY MATERIAL

Acronyms

CSC	Cost-Sensitive Classification
DLSE	Density Level-Set Estimation
e. g.	exempli gratia
∞-CSC	Infinite Cost-Sensitive Classification
i. e.	id est
i. i. d.	independent identically distributed
ITL	Infinite Task Learning
L-BFGS-B	Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm for Bound constraind optimization
MC	Monte-Carlo
MTL	Multi-Task Learning
OCSVM	One-Class Support Vector Machine
OVK	Operator-Valued Kernel
PTL	Parametric Task Learning
QMC	Quasi Monte Carlo
QR	Quantile Regression
RKHS	Reproducing Kernel Hilbert Space
r. v.	random variable
vv-RKHS	Vector-Valued Reproducing Kernel Hilbert Space
w. r. t.	with respect to

Below we provide the proofs of the results stated in the main part of the paper.

S.7 Quantile Regression

Let us recall the expression of the pinball loss (see Fig. S.3):

$$v_\theta : (y, y') \in \mathbb{R}^2 \mapsto \max\left(\theta(y - y'), (\theta - 1)(y - y')\right) \in \mathbb{R}. \tag{18}$$

Proposition S.7.1. *Let X, Y be two random variables (r. v.s) respectively taking values in $\mathcal{X}$ and $\mathbb{R}$, and $q\colon \mathcal{X} \to \mathcal{F}([0,1],\mathbb{R})$ the associated conditional quantile function. Let μ be a positive measure on $[0,1]$ such that $\int_0^1 \mathbf{E}\left[v_\theta\left(Y, q(X)(\theta)\right)\right] \mathrm{d}\mu(\theta) < \infty$. Then for $\forall h \in \mathcal{F}(\mathcal{X}; \mathcal{F}([0,\, 1]\,;\, \mathbb{R}))$*

$$R(h) - R(q) \geqslant 0,$$

where R is the risk defined in Eq. (6).

Proof. The proof is based on the one given in (Li et al., 2007) for a single quantile. Let $f \in \mathcal{F}(\mathcal{X}; \mathcal{F}([0,\, 1]\,;\, \mathbb{R}))$, $\theta \in (0,1)$ and $(x,y) \in \mathcal{X} \times \mathbb{R}$. Let also

$$s = \begin{cases} 1 \text{ if } y \leqslant f(x)(\theta) \\ 0 \text{ otherwise} \end{cases}, \qquad t = \begin{cases} 1 \text{ if } y \leqslant q(x)(\theta) \\ 0 \text{ otherwise} \end{cases}.$$

It holds that

$$\begin{aligned}\nu_\theta(y, h(x)(\theta)) - \nu_\theta(y, q(x)(\theta)) &= \theta(1-s)(y - h(x)(\theta)) + (\theta - 1)s(y - h(x)(\theta)) \\ &\quad - \theta(1-t)(y - q(x)(\theta)) - (\theta - 1)t(y - q(x)(\theta)) \\ &= \theta(1-t)(q(x)(\theta) - h(x)(\theta)) + \theta((1-t) - (1-s))h(x)(\theta) \\ &\quad + (\theta - 1)t(q(x)(\theta - h(x)(\theta))) + (\theta - 1)(t-s)h(x)(\theta) + (t-s)y \\ &= (\theta - t)(q(x)(\theta) - h(x)(\theta)) + (t-s)(y - h(x)(\theta)).\end{aligned}$$

Then, notice that

$$\mathbf{E}[(\theta - t)(q(X)(\theta) - h(X)(\theta))] = \mathbf{E}[\mathbf{E}[(\theta - t)(q(X)(\theta) - h(X)(\theta))]|X] = \mathbf{E}[\mathbf{E}[(\theta - t)|X](q(X)(\theta) - h(X)(\theta))]$$

and since q is the true quantile function,

$$\mathbf{E}[t|X] = \mathbf{E}[\mathbf{1}_{\{Y \leqslant q(X)(\theta)\}}|X] = \mathbf{P}[Y \leqslant q(X)(\theta)|X] = \theta,$$

so

$$\mathbf{E}[(\theta - t)(q(X)(\theta) - h(X)(\theta))] = 0.$$

Moreover, $(t - s)$ is negative when $q(x)(\theta) \leqslant y \leqslant h(x)(\theta)$, positive when $h(x)(\theta) \leqslant y \leqslant q(x)(\theta)$ and 0 otherwise, thus the quantity $(t - s)(y - h(x)(\theta))$ is always positive. As a consequence,

$$R(h) - R(q) = \int_{[0,1]} \mathbf{E}[\nu_\theta(Y, h(X)(\theta)) - \nu_\theta(Y, q(X)(\theta))] d\mu(\theta) \geqslant 0$$

which concludes the proof. □

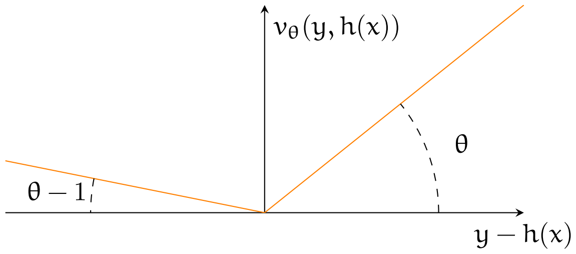

Figure S.3: Pinball loss for $\theta = 0.8$.

The Proposition S.7.1 allows us to derive conditions under which the minimization of the risk above yields the true quantile function. Under the assumption that (i) q is continuous (as seen as a function of two variables), (ii) $\mathrm{Supp}(\mu) = [0, 1]$, then the minimization of the integrated pinball loss performed in the space of continuous functions yields the true quantile function on the support of $\mathbf{P}_{X,Y}$.

S.8 Representer Propositions

Proof of Proposition 3.1. First notice that

$$J : h \in \mathcal{H}_K \mapsto \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} w_j \nu(\theta_j, y_i, h(x_i)(\theta_j)) + \frac{\lambda}{2} \|h\|^2_{\mathcal{H}_K} \in \mathbb{R} \tag{19}$$

is a proper lower semicontinuous strictly convex function (Bauschke et al., 2011, Corollary 9.4), hence J admits a unique minimizer $h^* \in \mathcal{H}_K$ (Bauschke et al., 2011, Corollary 11.17). Let

$$\mathcal{U} = \mathrm{span}\left\{ (K(\cdot, x_i) k_\Theta(\cdot, \theta_j))_{i,j=1}^{n,m} \,\middle|\, \forall x_i \in \mathcal{X}, \forall \theta_j \in \Theta \right\} \subset \mathcal{H}_K. \tag{20}$$

Then $\mathcal{U}$ is a finite-dimensional subspace of $\mathcal{H}_K$, thus closed in $\mathcal{H}_K$, and it holds that $\mathcal{U} \oplus \mathcal{U}^\perp = \mathcal{H}_K$, so h^* can be decomposed as $h^* = h^*_{\mathcal{U}} + h^*_{\mathcal{U}^\perp}$ with $h^*_{\mathcal{U}} \in \mathcal{U}$ and $h^*_{\mathcal{U}^\perp} \in \mathcal{U}^\perp$. Moreover, for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant m$,

$$h^*_{\mathcal{U}^\perp}(x_i)(\theta_j) = \langle h^*_{\mathcal{U}^\perp}(x_i), k_\Theta(\cdot, \theta_j)\rangle_{\mathcal{H}_{k_\Theta}} = \langle h^*_{\mathcal{U}^\perp}, K(\cdot, x_i) k_\Theta(\cdot, \theta_j)\rangle_{\mathcal{H}_K} = 0,$$

so $J(h^*) = J(h^*_{\mathcal{U}}) + \lambda \left\| h^*_{\mathcal{U}^\perp} \right\|^2_{\mathcal{H}_K}$. However h^* is the minimizer of J, therefore $h^*_{\mathcal{U}^\perp} = 0$ and there exist $(\alpha_{ij})_{i,j=1}^{n,m}$ such that $\forall x, \theta \in \mathcal{X} \times \Theta$, $h^*(x)(\theta) = \sum_{i,j=1}^{n,m} \alpha_{ij} k_{\mathcal{X}}(x, x_i) k_\Theta(\theta, \theta_j)$.

Derivative shapes constraints: Reminder: for a function h of one variable, we denote ∂h the derivative of h. For a function $k(\theta, \theta')$ of two variables we denote $\partial_1 k$ the derivative of k with respect to θ and $\partial_2 k$ the derivative of k with respect to θ'. From Zhou (2008), notice that if $f \in \mathcal{H}_k$, where $\mathcal{H}_k$ is a scalar-valued RKHS on a compact subset Θ of $\mathbb{R}^d$, and $k \in \mathcal{C}^2(\Theta \times \Theta)$ (in the sense of Ziemer (2012)) then $\partial f \in \mathcal{H}_k$. Hence if one add a new term of the form:

$$\lambda_{nc} \sum_{i\,=1}^{n} \sum_{j\,=1}^{m} \Omega_{nc}\left((\partial\left[h(x_i)\right])(\theta_j) \right) = \lambda_{nc} \sum_{i=1}^{n} \sum_{j=1}^{m} \Omega_{nc}\left((\partial h(x_i))(\theta_j) \right)$$

where g is a strictly monotonically increasing function and $\lambda_{nc} > 0$, a new representer theorem can be obtained by constructing the new set

$$\mathcal{U} = \operatorname{span}\ \left\{ \left(K(\cdot, x_i) k_\Theta(\cdot, \theta_j)\right)_{i,j=1}^{n,m} \,\middle|\, \forall x_i \in \mathcal{X}, \forall \theta_j \in \Theta \right\} \cup \left\{ \left(K(\cdot, x_i)(\partial_2 k_\Theta)(\cdot, \theta_j)\right)_{i,j=1}^{n,m} \,\middle|\, \forall x_i \in \mathcal{X}, \forall \theta_j \in \Theta \right\} \subset \mathcal{H}_K.$$

The proof is the same as Proposition 3.1 with the new set $\mathcal{U}$ to obtain the expansion $h(x)(\theta) = \sum_{i=1}^n \sum_{j=1}^m \alpha_{ij} k_{\mathcal{X}}(x, x_i) k_\Theta(\theta, \theta_j) + \beta_{ij} k(x, x_i)(\partial_2 k_\Theta)(\theta, \theta_j)$. For the regularization notice that for a symmetric function $(\partial_1 k)(\theta, \theta') = (\partial_2 k)(\theta', \theta)$. Hence $\langle (\partial_1 k)(\cdot, \theta'), k(\cdot, \theta)\rangle_{\mathcal{H}_k} = \langle k(\cdot, \theta'), (\partial_2 k)(\cdot, \theta)\rangle_{\mathcal{H}_k}$ and $(\partial k_{\theta'})(\theta) = (\partial^* k_\theta)(\theta')$ and

$$\begin{aligned}
\|h\|^2_{\mathcal{H}_K} &= \langle h, h\rangle_{\mathcal{H}_K} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{i'=1}^{n} \sum_{j'=1}^{m} \alpha_{ij}\alpha_{i'j'} k_{\mathcal{X}}(x_i, x_{i'}) k_\Theta(\theta_j, \theta_{j'}) + \alpha_{ij}\beta_{i'j'} k_{\mathcal{X}}(x_i, x_{i'})(\partial_2 k_\Theta)(\theta_j, \theta_{j'}) \\
&\quad + \alpha_{i'j'}\beta_{ij} k_{\mathcal{X}}(x_i, x_{i'})(\partial_1 k_\Theta)(\theta_j, \theta_{j'}) + \beta_{ij}\beta_{i'j'} k_{\mathcal{X}}(x_i, x_{i'})(\partial_1 \partial_2 k_\Theta)(\theta_j, \theta_{j'})
\end{aligned}$$

Eventually $(\partial h(x))(\theta) = \sum_{i=1}^n \sum_{j=1}^m \alpha_{ij} k_{\mathcal{X}}(x, x_i)(\partial_1 k_\Theta)(\theta, \theta_j) + \beta_{ij} k(x, x_i)(\partial_1 \partial_2 k_\Theta)(\theta, \theta_j)$. □

To prove Proposition 3.2, the following lemmas are useful.

Lemma S.8.1. *(Carmeli et al., 2010) Let* $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $k_\Theta : \Theta \times \Theta \to \mathbb{R}$ *be two scalar-valued kernels and* $K(\theta', \theta) = k_\Theta(\theta, \theta') I_{\mathcal{H}_{k_{\mathcal{X}}}}$. *Then* H_K *is isometric to* $\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_\Theta}$ *by means of the isometry* $W : f \otimes g \in \mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_\Theta} \mapsto (\theta \mapsto g(\theta) f) \in \mathcal{H}_K$.

Remark 1. *Given* $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $k_\Theta : \Theta \times \Theta \to \mathbb{R}$ *two scalar-valued kernels, we define* $K : (x, z) \in \mathcal{X} \times \mathcal{X} \mapsto k_{\mathcal{X}}(x, z) I_{\mathcal{H}_{k_\Theta}} \in \mathcal{L}(\mathcal{H}_{k_\Theta})$, $K' : (\theta, \theta') \in \Theta \times \Theta \mapsto k_\Theta(\theta, \theta') I_{\mathcal{H}_{k_{\mathcal{X}}}} \in \mathcal{L}(\mathcal{H}_{k_{\mathcal{X}}})$. *Lemma S.8.1 allows us to say that* $\mathcal{H}_K$ *and* $\mathcal{H}_{K'}$ *are isometric by means of the isometry*

$$W : h \in \mathcal{H}_{K'} \mapsto (x \mapsto (\theta \mapsto h(\theta)(x))) \in \mathcal{H}_K. \tag{21}$$

Lemma S.8.2. *Let* $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $k_\Theta : \Theta \times \Theta \to \mathbb{R}$ *be two scalar-valued kernels and* $K : (\theta, \theta') \mapsto k_\Theta(\theta, \theta') I_{\mathcal{H}_{k_{\mathcal{X}}}}$. *For* $\theta \in \Theta$, *define* $K_\theta : f \in \mathcal{H}_{k_{\mathcal{X}}} \mapsto (\theta' \mapsto K(\theta', \theta) f) \in \mathcal{H}_K$. *It is easy to see that* K^*_θ *is the evaluation operator* $K^*_\theta : h \in \mathcal{H}_K \mapsto h(\theta) \in \mathcal{H}_{k_{\mathcal{X}}}$. *Then* $\forall m \in \mathbb{N}^*, \forall (\theta_j)_{j=1}^m \in \Theta^m$,

$$\left(+_{j=1}^m \operatorname{Im}\left(K_{\theta_j}\right) \right) \oplus \left(\cap_{j=1}^m \operatorname{Ker}\left(K^*_{\theta_j}\right) \right) = \mathcal{H}_K \tag{22}$$

Proof. The statement boils down to proving that $\mathcal{V} := \left(+_{j=1}^{m} \operatorname{Im}\left(K_{\theta_j}\right)\right)$ is closed in $\mathcal{H}_K$, since it is straightforward that $\mathcal{V}^{\perp} = \left(\cap_{j=1}^{m} \operatorname{Ker}\left(K_{\theta_j}^{*}\right)\right)$. Let $(e_j)_{j=1}^{k}$ be an orthonormal basis of span $\left\{ (k_\Theta(\cdot, \theta_j))_{j=1}^{m} \right\} \subset \mathcal{H}_{k_\Theta}$. Such basis can be obtained by applying the Gram-Schmidt orthonormalization method to $(k_\Theta(\cdot, \theta_j))_{j=1}^{m}$. Then, $V = \operatorname{span}\left\{ e_j \cdot f, 1 \leqslant j \leqslant k, f \in \mathcal{H}_{k_\mathcal{X}} \right\}$. Notice also that $1 \leqslant j, l \leqslant k, \forall f, g \in \mathcal{H}_{k_\mathcal{X}}$,

$$\langle e_j \cdot f, e_l \cdot g \rangle_{\mathcal{H}_K} = \langle e_j, e_l \rangle_{\mathcal{H}_{k_\Theta}} \cdot \langle f, g \rangle_{\mathcal{H}_{k_\mathcal{X}}} \tag{23}$$

Let $(h_n)_{n \in \mathbb{N}^*}$ be a sequence in $\mathcal{V}$ converging to some $h \in \mathcal{H}_K$. By definition, one can find sequences $(f_{1,n})_{n \in \mathbb{N}^*}, \dots, (f_{k,n})_{n \in \mathbb{N}^*} \in \mathcal{H}_{k_\mathcal{X}}$ such that $\forall n \in \mathbb{N}^*$, $h_n = \sum_{j=1}^{k} e_j \cdot f_{n,j}$. Let $p, q \in \mathbb{N}^*$. It holds that, using the orthonormal property of $(e_j)_{j=1}^{k}$ and Eq. (23), $\|h_p - h_q\|_{\mathcal{H}_K}^2 = \left\|\sum_{j=1}^{k} e_j (f_{j,p} - f_{j,q})\right\|_{\mathcal{H}_K}^2 = \sum_{j=1}^{k} \|f_{j,p} - f_{j,q}\|_{\mathcal{H}_{k_\mathcal{X}}}^2$. $(h_n)_{n \in \mathbb{N}^*}$ being convergent, it is a Cauchy sequence, thus so are the sequences $(f_{j,n})_{n \in \mathbb{N}^*}$. But $\mathcal{H}_{k_\mathcal{X}}$ is a complete space, so these sequences are convergent in $\mathcal{H}_{k_\mathcal{X}}$, and by denoting $f_j = \lim_{n \to \infty} f_{j,n}$, one gets $h = \sum_{j=1}^{k} e_k \cdot f_j$. Therefore $h \in \mathcal{V}$, $\mathcal{V}$ is closed and the orthogonal decomposition Eq. (22) holds. □

Lemma S.8.3. *Let* $k_\mathcal{X}, k_\Theta$ *be two scalar kernels and* $K : (\theta, \theta') \mapsto k_\Theta(\theta, \theta') I_{\mathcal{H}_{k_\mathcal{X}}}$. *Let also* $m \in \mathbb{N}^*$ *and* $(\theta_j)_{j=1}^{m} \in \Theta^m$, *and* $\mathcal{V} = \left(+_{j=1}^{m} \operatorname{Im}\left(K_{\theta_j}\right)\right)$. *Then* $I : \mathcal{V} \to \mathbb{R}$ *defined as* $I(h) = \sum_{j=1}^{m} \|h(\theta_j)\|_{\mathcal{H}_{k_\mathcal{X}}}^2$ *is coercive.*

Proof. Notice first that if there exists θ_j such that $k_\Theta(\theta_j, \theta_j) = 0$, then $\operatorname{Im}\left(K_{\theta_j}\right) = 0$, so without loss of generality, we assume that $k_\Theta(\theta_j, \theta_j) > 0$ $(1 \leqslant j \leqslant m)$. Notice that I is the quadratic form associated to the $L : \mathcal{H}_K \to \mathcal{H}_K$ linear mapping $L(h) = \sum_{j=1}^{m} K_{\theta_j} K_{\theta_j}^{*}$. Indeed, $\forall h \in \mathcal{V}$, $I(h) = \sum_{j=1}^{m} \langle K_{\theta_j}^{*} h, K_{\theta_j}^{*} h \rangle_{\mathcal{H}_{k_\mathcal{X}}} = \sum_{j=1}^{m} \langle h, K_{\theta_j} K_{\theta_j}^{*} h \rangle_{\mathcal{H}_K} = \langle h, Lh \rangle_{\mathcal{H}_K}$. Moreover, $\forall 1 \leqslant j \leqslant m$, $K_{\theta_j} K_{\theta_j}^{*}$ has the same eigenvalues as $K_{\theta_j}^{*} K_{\theta_j}$, and $\forall f \in \mathcal{H}_{k_\mathcal{X}}$, $K_{\theta_j}^{*} K_{\theta_j} f = k_\Theta(\theta_j, \theta_j) f$, so that the only possible eigenvalue is $k_\Theta(\theta_j, \theta_j)$. Let $h \in \mathcal{V}$, $h \neq 0$. Because of the Eq. (22), h cannot be simultaneously in all $\operatorname{Ker}\left(K_{\theta_j}^{*}\right)$, and there exists i_0 such that $I(h) \geqslant k_\Theta(\theta_{i_0}, \theta_{i_0}) \|h\|_{\mathcal{H}_K}^2$. Let $\gamma = \min_{1 \leqslant j \leqslant m} k_\Theta(\theta_j, \theta_j)$. By assumption $\gamma > 0$, and it holds that $\forall h \in \mathcal{V}$, $I(h) \geqslant \gamma \|h\|_{\mathcal{H}_K}^2$, which proves the coercivity of I. □

Proof of Proposition 3.2. Let $K : (x, z) \in \mathcal{X} \times \mathcal{X} \mapsto k_\mathcal{X}(x, z) I_{\mathcal{H}_{k_\Theta}} \in \mathcal{L}(\mathcal{H}_{k_\Theta})$, $K' : (\theta, \theta') \in \Theta \times \Theta \mapsto k_\Theta(\theta, \theta') I_{\mathcal{H}_{k_\mathcal{X}}} \in \mathcal{L}(\mathcal{H}_{k_\mathcal{X}})$, and define

$$J \colon \begin{cases} \mathcal{H}_K \times \mathcal{H}_{k_b} & \to \mathbb{R} \\ (h, t) & \mapsto \frac{1}{n} \displaystyle\sum_{i,j=1}^{n,m} \frac{w_j}{\theta_j} |t(\theta_j) - h(x_i)(\theta_j)|_+ + \sum_{j=1}^{m} w_j \left(\|h(\cdot)(\theta_j)\|_{\mathcal{H}_{k_\mathcal{X}}}^2 - t(\theta_j) \right) + \frac{\lambda}{2} \|t\|_{\mathcal{H}_{k_b}}^2. \end{cases}$$

Let $\mathcal{V} = W\left(+_{j=1}^{m} \operatorname{Im}\left(K'_{\theta_j}\right)\right)$ where $W \colon \mathcal{H}_{K'} \to \mathcal{H}_K$ is defined in Eq. (21). Since W is an isometry, thanks to Eq. (22), it holds that $\mathcal{V} \oplus \mathcal{V}^{\perp} = \mathcal{H}_K$. Let $(h, t) \in \mathcal{H}_K \times \mathcal{H}_{k_b}$, there exists unique $h_{\mathcal{V}^\perp} \in \mathcal{V}^\perp$, $h_\mathcal{V} \in \mathcal{V}$ such that $h = h_\mathcal{V} + h_{\mathcal{V}^\perp}$. Notice that $J(h, t) = J(h_\mathcal{V} + h_{\mathcal{V}^\perp}, t) = J(h_\mathcal{V}, t)$ since $\forall 1 \leqslant j \leqslant m, \forall x \in \mathcal{X}$, $h_{\mathcal{V}^\perp}(x)(\theta_j) = W^{-1} h_{\mathcal{V}^\perp}(\theta_j)(x) = 0$. Moreover, J is bounded by below so that its infinimum is well-defined, and $\inf_{(h,t) \in H_K \times H_{k_b}} J(h, t) = \inf_{(h,t) \in V \times H_{k_b}} J(h, t)$. Finally, notice that J is coercive on $\mathcal{V} \times \mathcal{H}_{k_b}$ endowed with the sum of the norm (which makes it a Hilbert space): if $(h_n, t_n)_{n \in \mathbb{N}^*} \in \mathcal{V} \times \mathcal{H}_{k_b}$ is such that $\|h_n\|_{\mathcal{H}_K} + \|t_n\|_{\mathcal{H}_{k_b}} \underset{n \to \infty}{\to} +\infty$, then either $(\|h_n\|_{\mathcal{H}_K})_{n \in \mathbb{N}}$ or $(\|t_n\|_{\mathcal{H}_{k_b}})_{n \in \mathbb{N}}$ has to diverge :

- If $\|t_n\|_{\mathcal{H}_{k_b}} \underset{n \to \infty}{\to} +\infty$, since $t_n(\theta_j) = \langle t_n, k_b(\cdot, \theta_j) \rangle_{\mathcal{H}_{k_b}} \leqslant k_b(\theta_j, \theta_j) \|t_n\|_{\mathcal{H}_{k_b}} \leqslant \kappa_b \|t_n\|_{\mathcal{H}_{k_b}}$ $(\forall 1 \leqslant j \leqslant m)$, then $J(h_n, t_n) \geqslant \frac{\lambda}{2} \|t_n\|_{\mathcal{H}_{k_b}}^2 - \sum_{j=1}^{m} w_j t(\theta_j) \underset{n \to \infty}{\to} +\infty$.

- If $\|h_n\|_{\mathcal{H}_K} \underset{n \to \infty}{\to} +\infty$, according to Lemma S.8.3, $J(h_n, t_n) \underset{n \to \infty}{\to} +\infty$ as long as all w_j are strictly positive.

Thus J is coercive, so that (Bauschke et al., 2011, Proposition 11.15) allows to conclude that J has a minimizer (h^*, t^*) on $\mathcal{V} \times \mathcal{H}_{k_b}$. Then, in the same fashion as Eq. (20), define $\mathcal{U}_1 = \operatorname{span}\left\{ (K(\cdot, x_i)k_\Theta(\cdot, \theta_j))_{i,j=1}^{n,m} \right\} \subset \mathcal{V}$ and $\mathcal{U}_2 = \operatorname{span}\left\{ (k_b(\cdot, \theta_j))_{j=1}^{m} \right\} \subset \mathcal{H}_{k_b}$, and use the reproducing property to show that $(h^*, t^*) \in \mathcal{U}_1 \times \mathcal{U}_2$, so that there there exist $(\alpha_{ij})_{i,j=1}^{n,m}$ and $(\beta_j)_{j=1}^m$ such that $\forall x, \theta \in \mathcal{X} \times \Theta$, $h^*(x)(\theta) = \sum_{i,j=1}^{n,m} \alpha_{ij} k_\mathcal{X}(x, x_i) k_\theta(\theta, \theta_j)$, $t^*(\theta) = \sum_{j=1}^m \beta_j k_b(\theta, \theta_j)$.

□

S.9 Generalization error in the context of stability

The analysis of the generalization error will be performed using the notion of uniform stability introduced in (Bousquet et al., 2002). For a derivation of generalization bounds in vv-RKHS, we refer to (**kadri2016operator**). In their framework, the goal is to minimize a risk which can be expressed as

$$R_{\mathcal{S},\lambda}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h, x_i) + \lambda \|h\|_{\mathcal{H}_K}^2, \tag{24}$$

where $\mathcal{S} = ((x_1, y_1), \dots, (x_n, y_n))$ are i. i. d. inputs and $\lambda > 0$. We almost recover their setting by using losses defined as

$$\ell \colon \begin{cases} \mathbb{R} \times \mathcal{H}_K \times \mathcal{X} & \to \mathbb{R} \\ (y, h, x) & \mapsto \widetilde{V}(y, f(x)), \end{cases}$$

where $\widetilde{V}$ is a loss associated to some local cost defined in Eq. (8). Then, they study the stability of the algorithm which, given a dataset $\mathcal{S}$, returns

$$h_\mathcal{S}^* = \underset{h \in \mathcal{H}_K}{\arg\min}\ R_{\mathcal{S},\lambda}(h). \tag{25}$$

There is a slight difference between their setting and ours, since they use losses defined for some y in the output space of the vv-RKHS, but this difference has no impact on the validity of the proofs in our case. The use of their theorem requires some assumption that are listed below. We recall the shape of the OVK we use : $K : (x, z) \in \mathcal{X} \times \mathcal{X} \mapsto k_\mathcal{X}(x, z) I_{\mathcal{H}_{k_\Theta}} \in \mathcal{L}(\mathcal{H}_{k_\Theta})$, where $k_\mathcal{X}$ and k_Θ are both bounded scalar-valued kernels, in other words there exist $(\kappa_\mathcal{X}, \kappa_\Theta) \in \mathbb{R}^2$ such that $\sup\limits_{x \in X} k_\mathcal{X}(x, x) < \kappa_\mathcal{X}^2$ and $\sup\limits_{\theta \in \Theta} k_\Theta(\theta, \theta) < \kappa_\Theta^2$.

Assumption 1. *$\exists \kappa > 0$ such that $\forall x \in \mathcal{X}$, $\|K(x, x)\|_{\mathcal{L}(\mathcal{H}_{k_\Theta})} \leqslant \kappa^2$.*

Assumption 2. *$\forall h_1, h_2 \in \mathcal{H}_{k_\Theta}$, the function $(x_1, x_2) \in \mathcal{X} \times \mathcal{X} \mapsto \langle K(x_1, x_2) h_1, h_2 \rangle_{\mathcal{H}_{k_\Theta}} \in \mathbb{R}$, is measurable.*

Remark 2. *Assumptions 1, 2 are satisfied for our choice of kernel.*

Assumption 3. *The application $(y, h, x) \mapsto \ell(y, h, x)$ is σ-admissible, i. e. convex with respect to f and Lipschitz continuous with respect to $f(x)$, with σ as its Lipschitz constant.*

Assumption 4. *$\exists \xi \geqslant 0$ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\forall \mathcal{S}$ training set, $\ell(y, h_\mathcal{S}^*, x) \leqslant \xi$.*

Definition S.9.1. *Let $\mathcal{S} = ((x_i, y_i))_{i=1}^n$ be the training data. We call $\mathcal{S}^i$ the training data $\mathcal{S}^i = ((x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_n, y_n))$, $1 \leqslant i \leqslant n$.*

Definition S.9.2. *A learning algorithm mapping a dataset $\mathcal{S}$ to a function $h_\mathcal{S}^*$ is said to be β-uniformly stable with respect to the loss function ℓ if $\forall n \geqslant 1$, $\forall 1 \leqslant i \leqslant n$, $\forall \mathcal{S}$ training set, $\|\ell(\cdot, h_\mathcal{S}^*, \cdot) - \ell(\cdot, h_{\mathcal{S}^i}^*, \cdot)\|_\infty \leqslant \beta$.*

Proposition S.9.1. *(Bousquet et al., 2002) Let $\mathcal{S} \mapsto h_\mathcal{S}^*$ be a learning algorithm with uniform stability β with respect to a loss ℓ satisfying Assumption 4. Then $\forall n \geqslant 1$, $\forall \delta \in (0, 1)$, with probability at least $1 - \delta$ on the drawing of the samples, it holds that*

$$R(h_\mathcal{S}^*) \leqslant R_\mathcal{S}(h_\mathcal{S}^*) + 2\beta + (4\beta + \xi)\sqrt{\frac{\log(1/\delta)}{n}}.$$

Proposition S.9.2. *(**kadri2016operator**) Under assumptions 1, 2, 3, a learning algorithm that maps a training set $\mathcal{S}$ to the function $h_\mathcal{S}^*$ defined in Eq. (25) is β-stable with $\beta = \frac{\sigma^2 \kappa^2}{2\lambda n}$.*

S.9.1 Quantile Regression

We recall that in this setting, $\nu(\theta, y, h(x)(\theta)) = \max(\theta(y - h(x)(\theta)), (1-\theta)(y - h(x)(\theta)))$ and the loss is

$$\ell \colon \begin{cases} \mathbb{R} \times \mathcal{H}_K \times \mathcal{X} & \to \mathbb{R} \\ (y, h, x) & \mapsto \frac{1}{m} \sum_{j=1}^{m} \max(\theta_j (y - h(x)(\theta_j)), (\theta_j - 1)(y - h(x)(\theta_j))). \end{cases} \tag{26}$$

Moreover, we will assume that $|Y|$ is bounded by $B \in \mathbb{R}$ as a r. v.. We will therefore verify the hypothesis for $y \in [-B, B]$ and not $y \in \mathbb{R}$.

Lemma S.9.3. *In the case of the* QR*, the loss* ℓ *is* σ*-admissible with* $\sigma = 2\kappa_\Theta$*.*

Proof. Let $h_1, h_2 \in \mathcal{H}_K$ and $\theta \in [0,1]$. $\forall x, y \in \mathcal{X} \times \mathbb{R}$, it holds that

$$\nu(\theta, y, h_1(x)(\theta)) - \nu(\theta, y, h_2(x)(\theta)) = (\theta - t)(h_2(x)(\theta) - h_1(x)(\theta)) + (t - s)(y - h_1(x)(\theta)),$$

where $s = \mathbf{1}_{y \leqslant h_1(x)(\theta)}$ and $t = \mathbf{1}_{y \leqslant h_2(x)(\theta)}$. We consider all possible cases for t and s :

- $t = s = 0$: $|(t-s)(y - h_1(x)(\theta))| \leqslant |h_2(x)(\theta) - h_1(x)(\theta)|$
- $t = s = 1$: $|(t-s)(y - h_1(x)(\theta))| \leqslant |h_2(x)(\theta) - h_1(x)(\theta)|$
- $s = 1$,$t = 0$: $|(t-s)(y - h_1(x)(\theta))| = |h_1(x)(\theta) - y| \leqslant |h_1(x)(\theta) - h_2(x)(\theta)|$
- $s = 0$,$t = 1$: $|(t-s)(y - h_1(x)(\theta))| = |y - h_1(x)(\theta)| \leqslant |h_1(x)(\theta) - h_2(x)(\theta)|$ because of the conditions on t, s.

Thus $|\nu(\theta, y, h_1(x)(\theta)) - \nu(\theta, y, h_2(x)(\theta))| \leqslant (\theta + 1)|h_1(x)(\theta) - h_2(x)(\theta)| \leqslant (\theta + 1)\kappa_\Theta \|h_1(x) - h_2(x)\|_{\mathcal{H}_{k_\Theta}}$. By summing this expression over the $(\theta_j)_{j=1}^m$, we get that

$$|\ell(x, h_1, y) - \ell(x, h_2, y)| \leqslant \frac{1}{m} \sum_{j=1}^{m} (\theta_j + 1)\kappa_\Theta \|h_1(x) - h_2(x)\|_{\mathcal{H}_{k_\Theta}} \leqslant 2\kappa_\Theta \|h_1(x) - h_2(x)\|_{\mathcal{H}_{k_\Theta}}$$

and ℓ is σ-admissible with $\sigma = 2\kappa_\Theta$. □

Lemma S.9.4. *Let* $\mathcal{S} = ((x_1, y_1), \dots, (x_n, y_n))$ *be a training set and* $\lambda > 0$*. Then* $\forall x, \theta \in \mathcal{X} \times (0,1)$*, it holds that* $|h_{\mathcal{S}}^*(x)(\theta)| \leqslant \kappa_{\mathcal{X}} \kappa_\Theta \sqrt{\frac{B}{\lambda}}$.

Proof. Since $h_{\mathcal{S}}^*$ is the output of our algorithm and $0 \in \mathcal{H}_K$, it holds that

$$\lambda \|h_{\mathcal{S}}^*\|^2 \leqslant \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \nu(\theta_j, y_i, 0) \leqslant \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \max(\theta_j, 1 - \theta_j)|y_i| \leqslant B.$$

Thus $\|h_{\mathcal{S}}^*\| \leqslant \sqrt{\frac{B}{\lambda}}$. Moreover, $\forall x, \theta \in \mathcal{X} \times (0,1)$, $|h_{\mathcal{S}}^*(x)(\theta)| = |\langle h_{\mathcal{S}}^*(x), k_\Theta(\theta, \cdot)\rangle_{\mathcal{H}_{k_\Theta}}| \leqslant \|h_{\mathcal{S}}^*(x)\|_{\mathcal{H}_{k_\Theta}} \kappa_\Theta \leqslant \|h_{\mathcal{S}}^*\|_{\mathcal{H}_{k_\Theta}} \kappa_{\mathcal{X}} \kappa_\Theta$ which concludes the proof. □

Lemma S.9.5. *Assumption 4 is satisfied for* $\xi = 2\left(B + \kappa_{\mathcal{X}} \kappa_\Theta \sqrt{\frac{B}{\lambda}}\right)$.

Proof. Let $\mathcal{S} = ((x_1, y_1), \dots, (x_n, y_n))$ be a training set and $h_{\mathcal{S}}^*$ be the output of our algorithm. $\forall (x, y) \in \mathcal{X} \times [-B, B]$, it holds that

$$\begin{aligned} \ell(y, h_{\mathcal{S}}^*, x) = \frac{1}{m} \sum_{j=1}^{m} \max(\theta_j (y - h_{\mathcal{S}}^*(x)(\theta_j)), (\theta_j - 1)(y - h_{\mathcal{S}}^*(x)(\theta_j))) \leqslant \frac{2}{m} \sum_{j=1}^{m} |y - h_{\mathcal{S}}^*(x)(\theta_j)| \\ \leqslant \frac{2}{m} \sum_{j=1}^{m} |y| + |h_{\mathcal{S}}^*(x)(\theta_j)| \leqslant 2\left(B + \kappa_{\mathcal{X}} \kappa_\Theta \sqrt{\frac{B}{\lambda}}\right). \end{aligned}$$

□

Corollary S.9.6. *The* QR *learning algorithm defined in Eq. (10) is such that* $\forall n \geqslant 1$, $\forall \delta \in (0,1)$, *with probability at least* $1-\delta$ *on the drawing of the samples, it holds that*

$$\widetilde{R}(h_{\mathcal{S}}^*) \leqslant \widetilde{R}_{\mathcal{S}}(h_{\mathcal{S}}^*) + \frac{4\kappa_{\mathcal{X}}^2\kappa_{\Theta}^2}{\lambda n} + \left[\frac{8\kappa_{\mathcal{X}}^2\kappa_{\Theta}^2}{\lambda n} + 2\left(B + \kappa_{\mathcal{X}}\kappa_{\Theta}\sqrt{\frac{B}{\lambda}}\right)\right]\sqrt{\frac{\log(1/\delta)}{n}}. \tag{27}$$

Proof. This is a direct consequence of Proposition S.9.2, Proposition S.9.1, Lemma S.9.3 and Lemma S.9.5. □

Definition S.9.3 (Hardy-Krause variation)**.** *Let* Π *be the set of subdivisions of the interval* $\Theta = [0,1]$*. A subdivision will be denoted* $\sigma = (\theta_1, \theta_2, \dots, \theta_p)$ *and* $f\colon \Theta \to \mathbb{R}$ *be a function. We call Hardy-Krause variation of the function* f *the quantity* $\sup\limits_{\sigma\in\Pi} \sum_{i=1}^{p-1} |f(\theta_{i+1}) - f(\theta_i)|$.

Remark 3. *If* f *is continuous,* $V(f)$ *is also the limit as the mesh of* σ *goes to zero of the above quantity.*

In the following, let $f\colon \theta \mapsto \mathbf{E}_{X,Y}[\nu(\theta, Y, h_{\mathcal{S}}^*(X)(\theta))]$. This function is of primary importance for our analysis, since in the Quasi Monte-Carlo setting, the bound of Proposition 4.1 makes sense only if the function f has finite Hardy-Krause variation, which is the focus of the following lemma.

Lemma S.9.7. *Assume the boundeness of both scalar kernels* $k_{\mathcal{X}}$*and* k_{Θ}*. Assume moreover that* k_{Θ} *is* $\mathcal{C}^1$ *and that its partial derivatives are uniformly bounded by some constant* C*. Then*

$$V(f) \leqslant B + \kappa_{\mathcal{X}}\kappa_{\Theta}\sqrt{\frac{B}{\lambda}} + 2\kappa_{\mathcal{X}}\sqrt{\frac{2BC}{\lambda}}. \tag{28}$$

Proof. It holds that

$$\begin{aligned}
\sup_{\sigma\in\Pi}\sum_{i=1}^{p-1}|f(\theta_{i+1}) - f(\theta_i)| &= \sup_{\sigma\in\Pi}\sum_{i=1}^{p-1}\left|\int \nu(\theta_{i+1}, y, h_{\mathcal{S}}^*(x)(\theta_{i+1}))d\mathbf{P}_{X,Y} - \int \nu(\theta_i, y, h_{\mathcal{S}}^*(x)(\theta_i))d\mathbf{P}_{X,Y}\right| \\
&= \sup_{\sigma\in\Pi}\sum_{i=1}^{p-1}\left|\int \nu(\theta_{i+1}, y, h_{\mathcal{S}}^*(x)(\theta_{i+1})) - \nu(\theta_i, y, h_{\mathcal{S}}^*(x)(\theta_i))d\mathbf{P}_{X,Y}\right| \\
&\leqslant \sup_{\sigma\in\Pi}\sum_{i=1}^{p-1}\int|\nu(\theta_{i+1}, y, h_{\mathcal{S}}^*(x)(\theta_{i+1})) - \nu(\theta_i, y, h_{\mathcal{S}}^*(x)(\theta_i))|d\mathbf{P}_{X,Y} \\
&\leqslant \sup_{\sigma\in\Pi}\int\sum_{i=1}^{p-1}|\nu(\theta_{i+1}, y, h_{\mathcal{S}}^*(x)(\theta_{i+1})) - \nu(\theta_i, y, h_{\mathcal{S}}^*(x)(\theta_i))|d\mathbf{P}_{X,Y}.
\end{aligned}$$

The supremum of the integral is smaller than the integral of the supremum, as such

$$V(f) \leqslant \int V(f_{x,y})d\mathbf{P}_{X,Y}, \tag{29}$$

where $f_{x,y}\colon \theta \mapsto \nu(\theta, y, h_{\mathcal{S}}^*(x)(\theta))$ is the counterpart of the function f at point (x,y). To bound this quantity, let us first bound locally $V(f_{x,y})$. To that extent, we fix some (x,y) in the following. Since $f_{x,y}$ is continuous (because k_{Θ} is $\mathcal{C}^1$), then using Choquet (1969, Theorem 24.6), it holds that

$$V(f_{x,y}) = \lim_{|\sigma|\to 0}\sum_{i=1}^{p-1}|f_{x,y}(\theta_{i+1}) - f_{x,y}(\theta_i)|.$$

Moreover since $k \in \mathcal{C}^1$ and $\partial k_\theta = (\partial_1 k)(\cdot, \theta)$ has a finite number of zeros for all $\theta \in \times$, one can assume that in the subdivision considered afterhand all the zeros (in θ) of the residuals $y - h_{\mathcal{S}}^*(x)(\theta)$ are present, so that $y - h_{\mathcal{S}}^*(x)(\theta_{i+1})$ and $y - h_{\mathcal{S}}^*(x)(\theta_i)$ are always of the same sign. Indeed, if not, create a new, finer subdivision with this property and work with this one. Let us begin the proper calculation: let $\sigma = (\theta_1, \theta_2, \dots, \theta_p)$ be a subdivision of Θ, it holds that $\forall i \in \{1, \dots, p-1\}$:

$$\begin{aligned}|f_{x,y}(\theta_{i+1}) - f_{x,y}(\theta_i)| = |&\max\left(\theta_{i+1}(y - h^*_{\mathcal{S}}(x)(\theta_{i+1})), (1-\theta_{i+1})(y - h^*_{\mathcal{S}}(x)(\theta_{i+1}))\right)\\ &- \max\left(\theta_i(y - h^*_{\mathcal{S}}(x)(\theta_i)), (1-\theta_{i+1})(y - h^*_{\mathcal{S}}(x)(\theta_i))\right)|.\end{aligned}$$

We now study the two possible outcomes for the residuals:

- If $y - h(x)(\theta_{i+1}) \geqslant 0$ and $y - h(x)(\theta_i) \geqslant 0$ then

$$\begin{aligned}|f_{x,y}(\theta_{i+1}) - f_{x,y}(\theta_i)| &= |\theta_{i+1}(y - h^*_{\mathcal{S}}(x)(\theta_{i+1})) - \theta_i(y - h^*_{\mathcal{S}}(x)(\theta_i))|\\ &= |(\theta_{i+1} - \theta_i)y + (\theta_i - \theta_{i+1})h^*_{\mathcal{S}}(x)(\theta_{i+1}) + \theta_i(h^*_{\mathcal{S}}(x)(\theta_i) - h^*_{\mathcal{S}}(x)(\theta_{i+1}))|\\ &\leqslant |(\theta_{i+1} - \theta_i)y| + |(\theta_i - \theta_{i+1})h^*_{\mathcal{S}}(x)(\theta_{i+1})| + |\theta_i(h^*_{\mathcal{S}}(x)(\theta_i) - h^*_{\mathcal{S}}(x)(\theta_{i+1}))|.\end{aligned}$$

  From Lemma S.9.4, it holds that $h^*_{\mathcal{S}}(x)(\theta_{i+1}) \leqslant \kappa_{\mathcal{X}}\kappa_{\Theta}\sqrt{\frac{B}{\lambda}}$. Moreover,

$$\begin{aligned}|h^*_{\mathcal{S}}(x)(\theta_i) - h^*_{\mathcal{S}}(x)(\theta_{i+1})| &= \left|\langle h(x), k_{\Theta}(\theta_i, \cdot) - k_{\Theta}(\theta_{i+1}, \cdot)\rangle_{\mathcal{H}_{k_\Theta}}\right|\\ &\leqslant \|h(x)\|_{\mathcal{H}_{k_\Theta}} \|k_{\Theta}(\theta_i, \cdot) - k_{\Theta}(\theta_{i+1}, \cdot)\|_{\mathcal{H}_{k_\Theta}}\\ &\leqslant \kappa_{\mathcal{X}}\sqrt{\frac{B}{\lambda}}\sqrt{|k_{\Theta}(\theta_i, \theta_i) + k_{\Theta}(\theta_{i+1}, \theta_{i+1}) - 2k_{\Theta}(\theta_{i+1}, \theta_i)|}\\ &\leqslant \kappa_{\mathcal{X}}\sqrt{\frac{B}{\lambda}}\left(\sqrt{|k_{\Theta}(\theta_{i+1}, \theta_{i+1}) - k_{\Theta}(\theta_{i+1}, \theta_i)|} + \sqrt{|k_{\Theta}(\theta_i, \theta_i) - k_{\Theta}(\theta_{i+1}, \theta_i)|}\right).\end{aligned}$$

  Since k_{Θ} is $\mathcal{C}^1$, with partial derivatives uniformly bounded by C, $|k_{\Theta}(\theta_{i+1}, \theta_{i+1}) - k_{\Theta}(\theta_{i+1}, \theta_i)| \leqslant C(\theta_{i+1} - \theta_i)$ and $|k_{\Theta}(\theta_i, \theta_i) - k_{\Theta}(\theta_{i+1}, \theta_i)| \leqslant C(\theta_{i+1} - \theta_i)$ so that $|h^*_{\mathcal{S}}(x)(\theta_i) - h^*_{\mathcal{S}}(x)(\theta_{i+1})| \leqslant \kappa_{\mathcal{X}}\sqrt{\frac{2BC}{\lambda}}\sqrt{\theta_{i+1} - \theta_i}$ and overall

$$|f_{x,y}(\theta_{i+1}) - f_{x,y}(\theta_i)| \leqslant \left(B + \kappa_{\mathcal{X}}\kappa_{\Theta}\sqrt{\frac{B}{\lambda}}\right)(\theta_{i+1} - \theta_i) + \kappa_{\mathcal{X}}\sqrt{\frac{2BC}{\lambda}}\sqrt{\theta_{i+1} - \theta_i}.$$

- If $y - h(x)(\theta_{i+1}) \leqslant 0$ and $y - h(x)(\theta_i) \leqslant 0$ then $|f_{x,y}(\theta_{i+1}) - f_{x,y}(\theta_i)| = |(1 - \theta_{i+1})(y - h^*_{\mathcal{S}}(x)(\theta_{i+1})) - (1 - \theta_i)(y - h^*_{\mathcal{S}}(x)(\theta_i))| \leqslant |h^*_{\mathcal{S}}(x)(\theta_i) - h^*_{\mathcal{S}}(x)(\theta_{i+1})| + |(\theta_{i+1} - \theta_i)y| + |(\theta_i - \theta_{i+1})h^*_{\mathcal{S}}(x)(\theta_{i+1})| + |\theta_i(h^*_{\mathcal{S}}(x)(\theta_i) - h^*_{\mathcal{S}}(x)(\theta_{i+1}))|$ so that with similar arguments one gets

$$|f_{x,y}(\theta_{i+1}) - f_{x,y}(\theta_i)| \leqslant \left(B + \kappa_{\mathcal{X}}\kappa_{\Theta}\sqrt{\frac{B}{\lambda}}\right)(\theta_{i+1} - \theta_i) + 2\kappa_{\mathcal{X}}\sqrt{\frac{2BC}{\lambda}}\sqrt{\theta_{i+1} - \theta_i}. \qquad (30)$$

Therefore, regardless of the sign of the residuals $y - h(x)(\theta_{i+1})$ and $y - h(x)(\theta_i)$, one gets Eq. (30). Since the square root function has Hardy-Kraus variation of 1 on the interval $\Theta = [0, 1]$, it holds that

$$\sup_{\sigma \in \Pi} \sum_{i=1}^{p-1} |f_{x,y}(\theta_{i+1}) - f_{x,y}(\theta_i)| \leqslant B + \kappa_{\mathcal{X}}\kappa_{\Theta}\sqrt{\frac{B}{\lambda}} + 2\kappa_{\mathcal{X}}\sqrt{\frac{2BC}{\lambda}}.$$

Combining this with Eq. (29) finally gives

$$V(f) \leqslant B + \kappa_{\mathcal{X}}\kappa_{\Theta}\sqrt{\frac{B}{\lambda}} + 2\kappa_{\mathcal{X}}\sqrt{\frac{2BC}{\lambda}}.$$

□

Lemma S.9.8. *Let* R *be the risk defined in Eq. (6) for the quantile regression problem. Assume that* $(\theta)_{j=1}^m$ *have been generated via the Sobol sequence and that* k_Θ *is* $\mathcal{C}^1$ *and that its partial derivatives are uniformly bounded by some constant* C. *Then*

$$|R(h_S^*) - \widetilde{R}(h_S^*)| \leqslant \left(B + \kappa_{\mathcal{X}}\kappa_\Theta\sqrt{\frac{B}{\lambda}} + 2\kappa_{\mathcal{X}}\sqrt{\frac{2BC}{\lambda}}\right)\frac{\log(m)}{m}. \tag{31}$$

Proof. Let $f\colon \theta \mapsto \mathbf{E}_{X,Y}[v(\theta, Y, h_S^*(X)(\theta))]$. It holds that $|R(h_S^*) - \widetilde{R}(h_S^*)| \leqslant V(f)\frac{\log(m)}{m}$ according to classical Quasi-Monte Carlo approximation results, where $V(f)$ is the Hardy-Krause variation of f. Lemma S.9.7 allows then to conclude.

□

Proof of Proposition 4.1. Combine Lemma S.9.8 and Corollary S.9.6 to get an asymptotic behaviour as $n, m \to \infty$. □

S.9.2 Cost-Sensitive Classification

In this setting, the cost is $v(\theta, y, h(x)(\theta)) = \left|\frac{\theta+1}{2} - \mathbb{1}_{\{-1\}(y)}\right||1 - yh_\theta(x)|_+$ and the loss is

$$\ell\colon \begin{cases} \mathbb{R} \times \mathcal{H}_K \times \mathcal{X} & \to \mathbb{R} \\ (y, h, x) & \mapsto \frac{1}{m}\sum_{j=1}^m \left|\frac{\theta_j+1}{2} - \mathbb{1}_{\{-1\}}(y)\right|\left|1 - yh_{\theta_j}(x)\right|_+. \end{cases}$$

It is easy to verify in the same fashion as for QR that the properties above still hold, but with constants $\sigma = \kappa_\Theta$, $\beta = \frac{\kappa_{\mathcal{X}}^2\kappa_\Theta^2}{2\lambda n}$, $\xi = 1 + \frac{\kappa_{\mathcal{X}}\kappa_\Theta}{\sqrt{\lambda}}$. so that we get analogous properties to QR.

Corollary S.9.9. *The* CSC *learning algorithm defined in Eq. (10) is such that* $\forall n \geqslant 1$, $\forall \delta \in (0,1)$, *with probability at least* $1 - \delta$ *on the drawing of the samples, it holds that*

$$\widetilde{R}(h_S^*) \leqslant \widetilde{R}_S(h_S^*) + \frac{\kappa_{\mathcal{X}}^2\kappa_\Theta^2}{\lambda n} + \left(\frac{2\kappa_{\mathcal{X}}^2\kappa_\Theta^2}{\lambda n} + 1 + \frac{\kappa_{\mathcal{X}}\kappa_\Theta}{\sqrt{\lambda}}\right)\sqrt{\frac{\log(1/\delta)}{n}}.$$

S.10 Experimental remarks

We present here more details on the experimental protocol used in the main paper as well as new experiments.

S.10.1 Alternative hyperparameters sampling

Many quadrature rules such as Monte-Carlo (MC) and QMC methods are well suited for Infinite Task Learning. For instance when Θ is high dimensional, MC is typically prefered over QMC, and vice versa. If Θ is one dimensional and the function to integrate is smooth enough then a Gauss-Legendre quadrature would be preferable. In Section 3.1 of the main paper we provide a unified notation to handle MC, QMC and other quadrature rules. In the case of

- MC: $w_j = \frac{1}{m}$ and $(\theta_j)_{j=1}^m \sim \mu^{\otimes m}$.
- QMC: $w_j = m^{-1}F^{-1}(\theta_j)$ and $(\theta_j)_{j=1}^m$ is a sequence with values in $[0,1]^d$ such as the Sobol or Halton sequence, μ is assumed to be absolutely continuous w. r. t. the Lebesgue measure, F is the associated cdf.
- Quadrature rules: $((\theta_j, w_j'))_{j=1}^m$ is the indexed set of locations and weights produced by the quadrature rule, $w_j = w_j' f_\mu(\theta_j)$, μ is assumed to be absolutely continuous w. r. t. the Lebesgue measure, and f_μ denotes its corresponding probability density function.

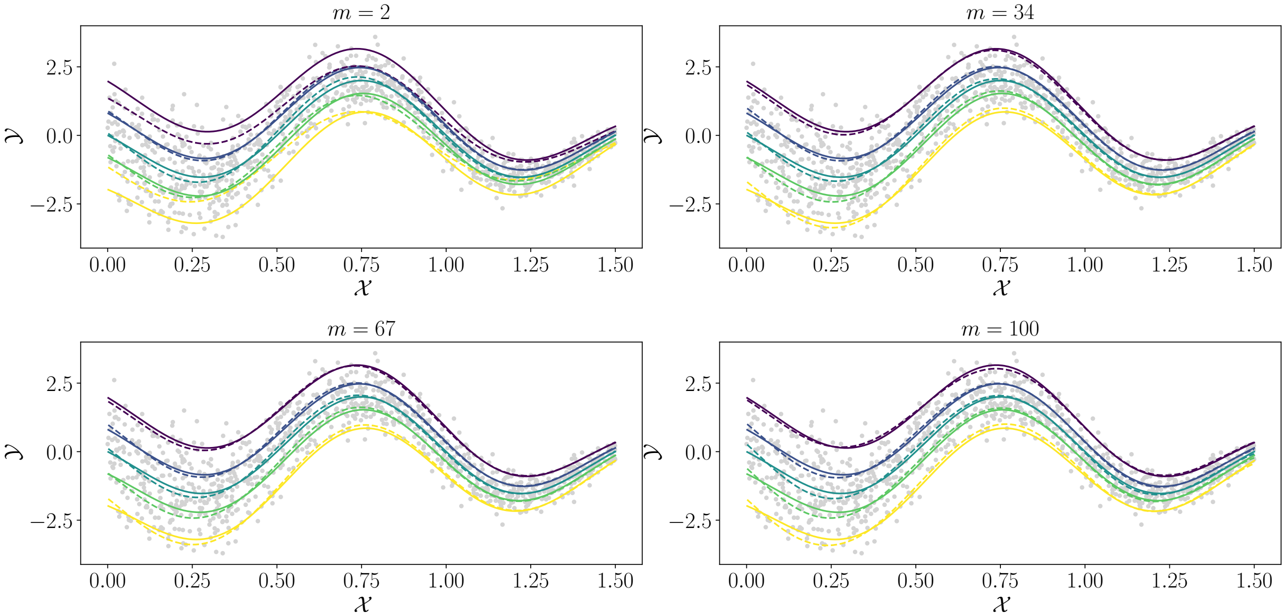

Figure S.4: Impact of the number of hyperparameters sampled.

S.10.2 Impact of the number of hyperparameters sampled

In the experiment presented on Fig. S.4, on the sine synthetic benchmark, we draw $n = 1000$ training points and study the impact of increasing m on the quality of the quantiles at $\theta \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$. We notice that when $m \geqslant 34 \approx \sqrt{1000}$ there is little benefit to draw more m samples are the quantile curves do not change on the $n_{\text{test}} = 2000$ test points.

S.10.3 Smoothifying the cost function

The resulting κ-smoothed ($\kappa \in \mathbb{R}_+$) absolute value (ψ_1^κ) and positive part (ψ_+^κ) are as follows:

$$\psi_1^\kappa(p) := \left(\kappa|\cdot|\square\frac{1}{2}|\cdot|^2\right)(p) = \begin{cases} \frac{1}{2\kappa}p^2 & \text{if } |p| \leqslant \kappa \\ |p| - \frac{\kappa}{2} & \text{otherwise,} \end{cases}$$

$$\psi_+^\kappa(p) := \left(\kappa|\cdot|_+\square\frac{1}{2}|\cdot|^2\right)(p) = \begin{cases} \frac{1}{2\kappa}|p|_+^2 & \text{if } p \leqslant \kappa \\ p - \frac{\kappa}{2} & \text{otherwise.} \end{cases}$$

where $\square$ is the classical infimal convolution (Bauschke et al., 2011). All the smoothified loss functions used in this paper have been gathered in Table S.2.

Remarks

- Minimizing the κ-smoothed pinball loss

$$v_{\theta,\kappa}(y, h(x)) = |\theta - \mathbb{1}_{\mathbb{R}_-}(y - h(x))|\psi_1^\kappa(y - h(x)),$$

  yields the quantiles when $\kappa \to 0$, the expectiles as $\kappa \to +\infty$. The intermediate values are known as M-quantiles (Breckling et al., 1988).
- In practice, the absolute value and positive part can be approximated by a smooth function by setting the smoothing parameter κ to be a small positive value; the optimization showed a robust behaviour w.r.t. this choice with a random coefficient initialization.

Impact of the Huber loss support The influence of the κ parameter is illustrated in Fig. S.5. For this experiment, 10000 samples have been generated from the sine wave dataset described in Section 5, and the model

Table S.2: Examples for objective (8). ψ_1^κ, ψ_+^κ: κ-smoothed absolute value and positive part. $h_x(\theta) := h(x)(\theta)$.

	LOSS	PENALTY
QUANTILE	$\int_{[0,1]} \left\lvert \theta - \mathbb{1}_{\mathbb{R}_-}(y - h_x(\theta)) \right\rvert \lvert y - h_x(\theta) \rvert d\mu(\theta)$	$\lambda_{nc} \int_{[0,1]} \left\lvert -\frac{dh_x}{d\theta}(\theta) \right\rvert_+ d\mu(\theta) + \frac{\lambda}{2} \lVert h \rVert_{\mathcal{H}_K}^2$
M-QUANTILE (SMOOTH)	$\int_{[0,1]} \left\lvert \theta - \mathbb{1}_{\mathbb{R}_-}(y - h_x(\theta)) \right\rvert \psi_1^\kappa (y - h_x(\theta)) \, d\mu(\theta)$	$\lambda_{nc} \int_{(0,1)} \psi_+^\kappa \left(-\frac{dh_x}{d\theta}(\theta) \right) d\mu(\theta) + \frac{\lambda}{2} \lVert h \rVert_{\mathcal{H}_K}^2$
EXPECTILES (SMOOTH)	$\int_{[0,1]} \left\lvert \theta - \mathbb{1}_{\mathbb{R}_-}(y - h_x(\theta)) \right\rvert (y - h_x(\theta))^2 \, d\mu(\theta)$	$\lambda_{nc} \int_{(0,1)} \left\lvert -\frac{dh_x}{d\theta}(\theta) \right\rvert_+^2 d\mu(\theta) + \frac{\lambda}{2} \lVert h \rVert_{\mathcal{H}_K}^2$
COST-SENSITIVE	$\int_{[-1,1]} \left\lvert \frac{\theta + 1}{2} - \mathbb{1}_{\{-1\}}(y) \right\rvert \lvert 1 - y h_x(\theta) \rvert_+ d\mu(\theta)$	$\frac{\lambda}{2} \lVert h \rVert_{\mathcal{H}_K}^2$
COST-SENSITIVE (SMOOTH)	$\int_{[-1,1]} \left\lvert \frac{\theta + 1}{2} - \mathbb{1}_{\{-1\}}(y) \right\rvert \psi_+^\kappa (1 - y h_x(\theta)) \, d\mu(\theta)$	$\frac{\lambda}{2} \lVert h \rVert_{\mathcal{H}_K}^2$
LEVEL-SET	$\int_{[\epsilon,1]} -t(\theta) + \frac{1}{\theta} \lvert t(\theta) - h_x(\theta) \rvert_+ d\mu(\theta)$	$\frac{1}{2} \int_{[\epsilon,1]} \lVert h(\cdot)(\theta) \rVert_{\mathcal{H}_{k_{\mathcal{X}}}}^2 d\mu(\theta) + \frac{\lambda}{2} \lVert t \rVert_{\mathcal{H}_{k_b}}^2$

have been trained on 100 quantiles generated from a Gauss-Legendre Quadrature. When κ is large the expectiles are learnt (dashed lines) while when κ is small the quantiles are recovered (the dashed lines on the right plot match the theoretical quantiles in plain lines). It took 225s (258 iteration, and 289 function evaluations) to train for $\kappa = 1 \cdot 10^1$, 1313s for $\kappa = 1 \cdot 10^{-1}$ (1438 iterations and 1571 function evaluations), 931s for $\kappa = 1e^{-3}$ (1169 iterations and 1271 function evaluations) and 879s for $\kappa = 0$ (1125 iterations and 1207 function evaluations). We used a GPU Tensorflow implementation and run the experiments in float64 on a computer equipped with a GTX 1070, and intel i7 7700 and 16GB of DRAM.

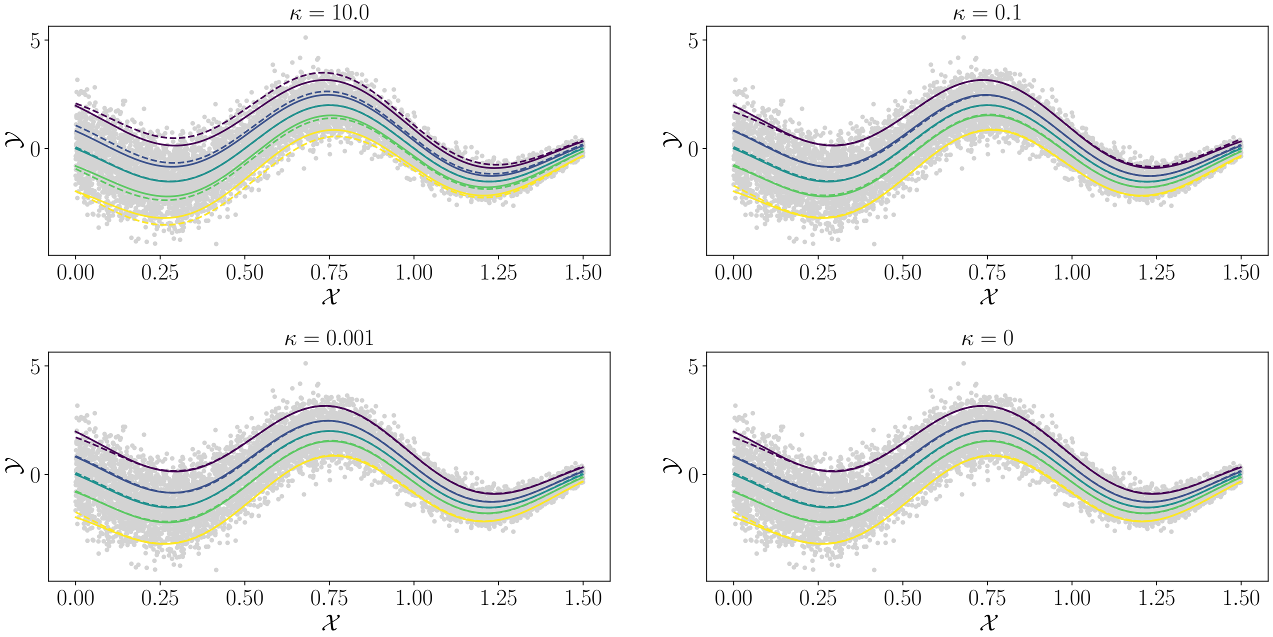

Figure S.5: Impact of the Huber loss smoothing of the pinball loss for differents values of κ.

S.10.4 Experimental protocol for QR

In this section, we give additional details regarding the choices being made while implementing the ITL method for ∞-QR.

QR real datasets For ∞-QR, $k_{\mathcal{X}}$, k_Θ were Gaussian kernels. We set a bias term $k_b = k_\Theta$. The hyperparameters optimized were λ, the weight of the ridge penalty, $\sigma_{\mathcal{X}}$, the input kernel parameter, and $\sigma_\Theta = \sigma_b$, the output kernel parameter. They were optimized in the (log)space of $\left[10^{-6}, 10^6\right]^3$. The non-crossing constraint

λ_{nc} was set to 1. The model was trained on the continuum $\Theta = (0, 1)$ using QMC and Sobol sequences. For all datasets we draw $m = 100$ quantiles form a Sobol sequence

For JQR we similarly chose two Gaussian kernels. The optimized hyperparameters were the same as for ∞-QR. The quantiles learned were $\theta \in \{ 0.1, 0.3, 0.5, 0.7, 0.9 \}$.

For the IND-QR baseline, we trained independently a non-paramatric quantile estimator as described in Takeuchi et al. (2006). A Gaussian kernel was used and its bandwidth was optimized in the (log)space of $\left[10^{-6}, 10^{6}\right]$. No non-crossing was enforced.

S.10.5 Experiments with CSC

In this section we provide numerical illustration concerning the CSC problem. We used the Iris UCI dataset with 4 attributes and 150 samples, the two synthetic SCIKIT-LEARN (Pedregosa et al., 2011) datasets Two-Moons (noise=0.4) and Circles (noise=0.1) with both 2 attributes and 1000 samples and a third synthetic SCIKIT-LEARN dataset Toy (class sep=0.5) with 20 features (4 redundant and 10 informative) and $n = 1000$ samples.

As detailed in Section 2, Cost-Sensitive Classification on a continuum $\Theta = [-1, 1]$ that we call Infinite Cost-Sensitive Classification (∞-CSC) can be tackled by our proposed technique. In this case, the hyperparameter θ controls the tradeoff between the importance of the correct classification with labels -1 and $+1$. When $\theta = -1$, class -1 is emphasized; the probability of correctly classified instances with this label (called specificity) is desired to be 1. Similarly, for $\theta = +1$, the probability of correct classification of samples with label $+1$ (called sensitivity) is ideally 1.

To illustrate the advantage of (infinite) joint learning we used two synthetic datasets Circles and Two-Moons and the UCI Iris dataset. We chose $k_{\mathcal{X}}$ to be a Gaussian kernel with bandwidth $\sigma_{\mathcal{X}} = (2\gamma_{\mathcal{X}})^{(-1/2)}$ the median of the Euclidean pairwise distances of the input points (Jaakkola et al., 1999). k_{Θ} is also a Gaussian kernel with bandwidth $\gamma_{\Theta} = 5$. We used $m = 20$ for all datasets. As a baseline we trained independently 3 Cost-Sensitive Classification classifiers with $\theta \in \{ -0.9, 0, 0.9 \}$. We repeated 50 times a random $50 - 50\%$ train-test split of the dataset and report the average test error and standard deviation (in terms of sensitivity and specificity)

Our results are illustrated in Table S.3. For $\theta = -0.9$, both independent and joint learners give the desired 100% specificity; the joint Cost-Sensitive Classification scheme however has significantly higher sensitivity value (15% vs 0%) on the dataset Circles. Similar conclusion holds for the $\theta = +0.9$ extreme: the ideal sensitivity is reached by both techniques, but the joint learning scheme performs better in terms of specificity (0% vs 12%) on the dataset Circles.

Table S.3: ∞-CSC vs Independent (IND)-CSC. Higher is better.

Dataset	Method	$\theta = -0.9$		$\theta = 0$		$\theta = +0.9$	
		sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
Two-Moons	IND	0.3 ± 0.05	0.99 ± 0.01	0.83 ± 0.03	0.86 ± 0.03	0.99 ± 0	0.32 ± 0.06
	∞-CSC	0.32 ± 0.05	0.99 ± 0.01	0.84 ± 0.03	0.87 ± 0.03	1 ± 0	0.36 ± 0.04
Circles	IND	0 ± 0	1 ± 0	0.82 ± 0.02	0.84 ± 0.03	1 ± 0	0 ± 0
	∞-CSC	0.15 ± 0.05	1 ± 0	0.82 ± 0.02	0.84 ± 0.03	1 ± 0	0.12 ± 0.05
Iris	IND	0.88 ± 0.08	0.94 ± 0.06	0.94 ± 0.05	0.92 ± 0.06	0.97 ± 0.05	0.87 ± 0.06
	∞-CSC	0.89 ± 0.08	0.94 ± 0.05	0.94 ± 0.06	0.92 ± 0.05	0.97 ± 0.04	0.90 ± 0.05
Toy	IND	0.51 ± 0.06	0.98 ± 0.01	0.83 ± 0.03	0.86 ± 0.03	0.97 ± 0.01	0.49 ± 0.07
	∞-CSC	0.63 ± 0.04	0.96 ± 0.01	0.83 ± 0.03	0.85 ± 0.03	0.95 ± 0.02	0.61 ± 0.04

References

Bauschke, H. H. and P. L. Combettes (2011). *Convex analysis and monotone operator theory in Hilbert spaces.* Springer (cit. on pp. 11, 14, 19).

Bousquet, O. and A. Elisseeff (2002). "Stability and generalization." In: *Journal of Machine Learning Research* 2, pp. 499–526 (cit. on p. 14).

Breckling, J. and R. Chambers (1988). "M-quantiles." In: *Biometrika* 75.4, pp. 761–771 (cit. on p. 19).

Carmeli, C. et al. (2010). "Vector valued reproducing kernel Hilbert spaces and universality." In: *Analysis and Applications* 8 (1), pp. 19–61 (cit. on p. 12).

Choquet, G. (1969). *Cours d'analyse: Tome II. Topologie*. Masson et Cie. (cit. on p. 16).

Jaakkola, T., M. Diekhans, and D. Haussler (1999). "Using the Fisher kernel method to detect remote protein homologies." In: *ISMB*. Vol. 99, pp. 149–158 (cit. on p. 21).

Li, Y., Y. Liu, and J. Zhu (2007). "Quantile regression in reproducing kernel Hilbert spaces." In: *Journal of the American Statistical Association* 102.477, pp. 255–268 (cit. on p. 10).

Pedregosa, F. et al. (2011). "Scikit-learn: Machine learning in Python." In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830 (cit. on p. 21).

Takeuchi, I. et al. (2006). "Nonparametric quantile estimation." In: *Journal of Machine Learning Research* 7, pp. 1231–1264 (cit. on p. 21).

Zhou, D.-X. (2008). "Derivative reproducing properties for kernel methods in learning theory." In: *Journal of computational and Applied Mathematics* 220.1-2, pp. 456–463 (cit. on p. 12).

Ziemer, W. P. (2012). *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*. Vol. 120. Springer Science & Business Media (cit. on p. 12).