JMLR: Workshop and Conference Proceedings 80:1–9, 2017 ACML 2017

Data sparse nonparametric regression with ϵ-insensitive losses: supplementary material

Maxime Sangnier MAXIME.SANGNIER@UPMC.FR
Sorbonne Universités, UPMC Univ Paris 06, CNRS, Paris, France

Olivier Fercoq OLIVIER.FERCOQ@TELECOM-PARISTECH.FR
Université Paris-Saclay, Télécom ParisTech, LTCI, Paris, France

Florence d'Alché-Buc FLORENCE.DALCHE@TELECOM-PARISTECH.FR
Université Paris-Saclay, Télécom ParisTech, LTCI, Paris, France

Editors: Yung-Kyun Noh and Min-Ling Zhang

Appendix A. Remarks

Remark 1 *It is easy to see that, for a unidimensional loss ℓ, ℓ_ϵ is obtained in the following manner:*

$$\forall \xi \in \mathbb{R}\colon \quad \ell_\epsilon(\xi) = \ell\left(\xi - \min(|\xi|, \epsilon)\operatorname{sign}(\xi)\right) = \begin{cases} 0 & \text{if } |\xi| \leq \epsilon \\ \ell\left(\xi\left(1 - \frac{\epsilon}{|\xi|}\right)\right) & \text{otherwise.} \end{cases}$$

Consequently, when a multivariate loss ℓ is separable, that is $\ell(\boldsymbol{\xi}) = \sum_{j=1}^p \ell^{(j)}(\xi_j)$ (for some unidimensional losses $\ell^{(j)}$), it is tempting to consider each component separately and to define $\ell_\epsilon = \sum_{j=1}^p \ell_\epsilon^{(j)}$. Basically, this boils down to replacing $\|\cdot\|_{\ell_2}$ by $\|\cdot\|_\infty$ in the general ϵ-loss introduced in this paper.

However, this is not a good idea since this definition would result in adding an ℓ_1-norm $\sum_{i=1}^n \|\boldsymbol{\alpha}_i\|_{\ell_1}$ instead of an ℓ_1/ℓ_2-norm in the dual. As a consequence, we would obtain sparse vectors $\boldsymbol{\alpha}_i$, which is not the data sparsity we pursue since $\boldsymbol{\alpha}_i$ could have null components but could be different from 0, forcing us to keep the points $\mathbf{x}_i$ for prediction.

Remark 2 *In the body of the text, omitting the intercept $\boldsymbol{b}$ in Problem (P2) comes down to removing the linear constraint in Problem (P3). This practice is common for support vector regression (SVR) with a Gaussian kernel, but is excluded for quantile regression (QR) (Takeuchi et al., 2006; Sangnier et al., 2016).*

Example 1 *Examples of scalar and matrix kernels are:*

$$k(\mathbf{x}, \mathbf{x}') = \left(1 + \langle \mathbf{x}, \mathbf{x}' \rangle_{\ell_2}\right)^d \qquad \textit{(polynomial)},$$

where $d > 0$ is the degree (Mohri et al., 2012), and

$$K(\mathbf{x}, \mathbf{x}') = \left[\left(1 + \langle T_i(\mathbf{x}), T_j(\mathbf{x}') \rangle_{\ell_2}\right)^d\right]_{1 \leq i,j \leq p} \qquad \textit{(transformable)},$$

where $T_i \colon \mathbb{R}^p \to \mathbb{R}^p$ are any transformations (Alvarez et al., 2012).

Remark 3 *As it is standard for coordinate descent methods, our implementation uses efficient updates for the computation of both $\sum_{j=1}^n K(\mathbf{x}_i, \mathbf{x}_j)\boldsymbol{\alpha}_j$ and $\overline{\boldsymbol{\theta}}^l$. In addition, convergence of Algorithm 1 can be assessed by duality gap (objective of (P2) minus objective of (P3) in the body of the text). Yet, even though we do not have closed-form expressions for the primal loss ℓ_ϵ, the duality gap can be over-estimated by upper-bounding ℓ_ϵ in the following manner:*

$$\forall \boldsymbol{\xi} \in \mathbb{R}^p \colon \quad \ell_\epsilon(\boldsymbol{\xi}) \leq \ell\left(\boldsymbol{\xi}\left(1 - \frac{\min(\epsilon, \|\boldsymbol{\xi}\|_{\ell_2})}{\|\boldsymbol{\xi}\|_{\ell_2}}\right)\right).$$

This is true since $\left\|\frac{\min(\epsilon, \|\boldsymbol{\xi}\|_{\ell_2})}{\|\boldsymbol{\xi}\|_{\ell_2}}\boldsymbol{\xi}\right\|_{\ell_2} \leq \epsilon$.

Remark 4 *Contrarily to QR, expectile regression involves a differentiable mapping $\ell^\star$. Consequently, it can be easily incorporated to the quadratic contribution of (P7) (see body of the text). Nevertheless, it can also be considered jointly with $\|\cdot\|_{\ell_2}$, in the same manner as for QR. In this case, the differentiable part remains the same for expectile and quantile regression, only the non-differentiable part changes. The proximal operator needed is given in the following proposition.*

Proposition 5 *Let $\psi \colon \boldsymbol{y} \in \mathbb{R}^p \mapsto \frac{1}{2}\sum_{i=1}^p \left|\tau_j - \mathrm{I}_{y_j<0}\right|^{-1} y_j^2$. Then*

$$\forall \boldsymbol{y} \in \mathbb{R}^p, \forall j \in [p]$$
$$\left[\operatorname{prox}_{\lambda\left(\|\cdot\|_{\ell_2} + \psi\right)}(\boldsymbol{y})\right]_j = \left(1 + \frac{\lambda}{\mu} + \lambda\left|\tau_j - \mathrm{I}_{y_j<0}\right|^{-1}\right)^{-1} y_j,$$

if $\|\boldsymbol{y}\|_{\ell_2} > \lambda$, where $\mu > 0$ is solution to:

$$\sum_{j=1}^p \frac{y_j^2}{\left(\mu\left(1 + \lambda\left|\tau_j - \mathrm{I}_{y_j<0}\right|^{-1}\right) + \lambda\right)^2} = 1, \tag{1}$$

(such a solution exists) and $\operatorname{prox}_{\lambda\left(\|\cdot\|_{\ell_2} + \psi\right)}(\boldsymbol{y}) = 0$ if $\|\boldsymbol{y}\|_{\ell_2} \leq \lambda$,.

Similarly to Equation 1 in the body of the text, the scaling factor μ in Proposition 5 can be easily obtained by a bisection of a Newton-Raphson method.

Appendix B. Technical details

B.1. Convexity and redefinition of ℓ_ϵ

For the sake of simplicity, let us first define:

$$\forall \boldsymbol{\xi} \in \mathbb{R}^p \colon \quad \tilde{\ell}_\epsilon(\boldsymbol{\xi}) = \inf_{\boldsymbol{u} \in \mathbb{R}^p \colon \|\boldsymbol{u}\|_{\ell_2} \leq \epsilon} \ell(\boldsymbol{\xi} - \boldsymbol{u}). \tag{2}$$

Since ℓ is convex, $(\boldsymbol{\xi}, \boldsymbol{u}) \mapsto \ell(\boldsymbol{\xi} - \boldsymbol{u}) + \chi_{\|\boldsymbol{u}\|_{\ell_2} \leq \epsilon}$ is jointly convex with respect to $\boldsymbol{\xi}$ and $\boldsymbol{u}$. Therefore, $\tilde{\ell}_\epsilon$ is convex as the coordinate infimum of a jointly convex function (Boyd and Vandenberghe, 2004).

Let us now show that $\tilde{\ell}_\epsilon = \ell_\epsilon$. First, since for any $\boldsymbol{\xi}$, Slater's constraint qualification are satisfied for (2), strong duality holds, that is,

$$\forall \boldsymbol{\xi} \in \mathbb{R}^p, \exists \lambda \geq 0 : \quad \tilde{\ell}_\epsilon(\boldsymbol{\xi}) = \inf_{\boldsymbol{u} \in \mathbb{R}^p} \ell(\boldsymbol{\xi} - \boldsymbol{u}) + \lambda \|\boldsymbol{u}\|_{\ell_2} - \lambda \epsilon,$$

and thanks to the lower semi-continuity of the objective, the infimum is attained at, let us say, $\hat{\boldsymbol{u}}$. Then, when $\|\boldsymbol{\xi}\|_{\ell_2} \leq \epsilon$, we can chose $\hat{\boldsymbol{u}} = \boldsymbol{\xi}$ and we get $\tilde{\ell}_\epsilon(\boldsymbol{\xi}) = 0$, which is the infimum of ℓ. On the other hand, when $\|\boldsymbol{\xi}\|_{\ell_2} > \epsilon$, let us consider the Karush-Kuhn-Tucker (KKT) conditions. By complementary slackness, either $\lambda = 0$ and $\|\hat{\boldsymbol{u}}\|_{\ell_2} \leq \epsilon$, or $\|\hat{\boldsymbol{u}}\|_{\ell_2} = \epsilon$. In the first situation ($\lambda = 0$ and $\|\hat{\boldsymbol{u}}\|_{\ell_2} \leq \epsilon$), $\tilde{\ell}_\epsilon(\boldsymbol{\xi}) = \inf_{\boldsymbol{u} \in \mathbb{R}^p} \ell(\boldsymbol{\xi} - \boldsymbol{u}) = \ell(\boldsymbol{\xi} - \hat{\boldsymbol{u}}) = \ell(0) = 0$ and $\hat{\boldsymbol{u}} = \boldsymbol{\xi}$ (by uniqueness of the minimizer of ℓ). Thus, $\|\boldsymbol{\xi}\|_{\ell_2} \leq \epsilon$, which is contradictory. Consequently, we have necessarily, $\|\hat{\boldsymbol{u}}\|_{\ell_2} = \epsilon$. To summarize:

$$\forall \boldsymbol{\xi} \in \mathbb{R}^p : \quad \tilde{\ell}_\epsilon(\boldsymbol{\xi}) = \begin{cases} 0 & \text{if } \|\boldsymbol{\xi}\|_{\ell_2} \leq \epsilon \\ \inf\limits_{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_{\ell_2} = \epsilon} \ell(\boldsymbol{\xi} - \boldsymbol{u}) & \text{otherwise,} \end{cases}$$

which is exactly the definition of ℓ_ϵ.

B.2. Dual and representer theorem

Since ℓ_ϵ is convex and can be replaced by (2), Problem (P2) from the body of the text can be reformulated in (Lagrange multipliers are indicated on the right):

$$\begin{aligned} &\underset{\substack{h \in \mathcal{H}, \boldsymbol{b} \in \mathbb{R}^p, \\ \forall i \in [n], \boldsymbol{\xi}_i \in \mathbb{R}^p, \boldsymbol{r}_i \in \mathbb{R}^p}}{\text{minimize}} && \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\xi}_i) \\ &\text{s.t.} && \begin{cases} \forall i \in [n], \\ \dfrac{\boldsymbol{y}_i - (h(\mathbf{x}_i) + \boldsymbol{b})}{n} = \dfrac{\boldsymbol{r}_i + \boldsymbol{\xi}_i}{n} & : \boldsymbol{\alpha}_i \in \mathbb{R}^p \\ \dfrac{\|\boldsymbol{r}_i\|_{\ell_2}^2}{2\epsilon n} \leq \dfrac{\epsilon}{2n} & : \mu_i \in \mathbb{R}_+. \end{cases} \end{aligned} \tag{P1}$$

Let us compute a dual to Problem (P1). The Lagrangian reads:

$$\begin{aligned}
&\mathfrak{L}\left(h, \boldsymbol{b}, (\boldsymbol{\xi}_i)_{1\leq i\leq n}, (\boldsymbol{r}_i)_{1\leq i\leq n}, (\boldsymbol{\alpha}_i)_{1\leq i\leq n}, (\mu_i)_{1\leq i\leq n}\right)\\
&= \frac{\lambda}{2}\left\|h\right\|_{\mathcal{H}}^2 + \frac{1}{n}\sum_{i=1}^n \ell\left(\boldsymbol{\xi}_i\right) + \frac{1}{n}\sum_{i=1}^n \left\langle \boldsymbol{\alpha}_i, \boldsymbol{y}_i - (h(\mathbf{x}_i)+\boldsymbol{b}) - \boldsymbol{r}_i - \boldsymbol{\xi}_i\right\rangle_{\ell_2}\\
&+ \frac{1}{2\epsilon n}\sum_{i=1}^n \mu_i \left\|\boldsymbol{r}_i\right\|_{\ell_2}^2 - \frac{\epsilon}{2n}\sum_{i=1}^n \mu_i\\
&= \frac{1}{n}\sum_{i=1}^n \left(\ell\left(\boldsymbol{\xi}_i\right) - \left\langle \boldsymbol{\alpha}_i, \boldsymbol{\xi}_i\right\rangle_{\ell_2}\right) + \frac{\lambda}{2}\left\|h\right\|_{\mathcal{H}}^2 - \left\langle \frac{1}{n}\sum_{i=1}^n E_{\mathbf{x}_i}^* \boldsymbol{\alpha}_i, h\right\rangle_{\mathcal{H}}\\
&- \left\langle \frac{1}{n}\sum_{i=1}^n \boldsymbol{\alpha}_i, \boldsymbol{b}\right\rangle_{\ell_2} + \frac{1}{n}\sum_{i=1}^n \left(\frac{\mu_i}{2\epsilon}\left\|\boldsymbol{r}_i\right\|_{\ell_2}^2 - \left\langle \boldsymbol{\alpha}_i, \boldsymbol{r}_i\right\rangle_{\ell_2}\right) + \frac{1}{n}\sum_{i=1}^n \left\langle \boldsymbol{\alpha}_i, \boldsymbol{y}_i\right\rangle_{\ell_2}\\
&- \frac{\epsilon}{2n}\sum_{i=1}^n \mu_i.
\end{aligned}$$

The objective function of the dual problem to (P1) is obtained by minimizing the Lagrangian with respect to the primal variables h, $\boldsymbol{b}$, $(\boldsymbol{\xi}_i)_{1\leq i\leq n}$ and $(\boldsymbol{r}_i)_{1\leq i\leq n}$. For this purpose, let us remark that minimizing on $\boldsymbol{\xi}_i$ boils down to introducing the Fenchel-Legendre transform of ℓ: $\ell^\star\colon \boldsymbol{\alpha}\in\mathbb{R}^p \mapsto \sup_{\boldsymbol{\xi}\in\mathbb{R}^p} \left\langle \boldsymbol{\alpha}, \boldsymbol{\xi}\right\rangle_{\ell_2} - \ell(\boldsymbol{\xi})$. Thus, it remains to compute:

$$\begin{aligned}
&\mathfrak{L}_D\left((\boldsymbol{\alpha}_i)_{1\leq i\leq n}, (\mu_i)_{1\leq i\leq n}\right)\\
&= \inf_{\substack{h\in\mathcal{H},\, \boldsymbol{b}\in\mathbb{R}^p,\\ \forall i\in[n],\, \boldsymbol{r}_i\in\mathbb{R}^p}} \left\{ -\frac{1}{n}\sum_{i=1}^n \ell^\star(\boldsymbol{\alpha}_i) + \frac{\lambda}{2}\left\|h\right\|_{\mathcal{H}}^2 - \left\langle \frac{1}{n}\sum_{i=1}^n E_{\mathbf{x}_i}^* \boldsymbol{\alpha}_i, h\right\rangle_{\mathcal{H}}\right.\\
&- \left\langle \frac{1}{n}\sum_{i=1}^n \boldsymbol{\alpha}_i, \boldsymbol{b}\right\rangle_{\ell_2} + \frac{1}{n}\sum_{i=1}^n \left(\frac{\mu_i}{2\epsilon}\left\|\boldsymbol{r}_i\right\|_{\ell_2}^2 - \left\langle \boldsymbol{\alpha}_i, \boldsymbol{r}_i\right\rangle_{\ell_2}\right) + \frac{1}{n}\sum_{i=1}^n \left\langle \boldsymbol{\alpha}_i, \boldsymbol{y}_i\right\rangle_{\ell_2}\\
&\left. -\frac{\epsilon}{2n}\sum_{i=1}^n \mu_i \right\}.
\end{aligned}$$

Since $\mathcal{H}$ is unbounded in all directions, the minimum of $\mathfrak{L}$ with respect to h, $\boldsymbol{b}$ and $(\boldsymbol{r}_i)_{i=1}^n$ is obtained by setting the gradients to 0, which leads to $h = \frac{1}{\lambda n}\sum_{i=1}^n E_{\mathbf{x}_i}^* \boldsymbol{\alpha}_i$, $\sum_{i=1}^n \boldsymbol{\alpha}_i = 0$ and $\boldsymbol{r}_i = \frac{\epsilon}{\mu_i}\boldsymbol{\alpha}_i, \forall i\in[n]$. Thus, the dual objective reads:

$$\begin{aligned}
&\mathfrak{L}_D\left((\boldsymbol{\alpha}_i)_{1\leq i\leq n}, (\mu_i)_{1\leq i\leq n}\right)\\
&= -\frac{1}{n}\sum_{i=1}^n \ell^\star(\boldsymbol{\alpha}_i) - \frac{1}{2\lambda n^2}\sum_{i,j=1}^n \left\langle \boldsymbol{\alpha}_i, E_{\mathbf{x}_i}E_{\mathbf{x}_j}^* \boldsymbol{\alpha}_j\right\rangle_{\ell_2} + \frac{1}{n}\sum_{i=1}^n \left\langle \boldsymbol{\alpha}_i, \boldsymbol{y}_i\right\rangle_{\ell_2}\\
&- \frac{\epsilon}{2n}\sum_{i=1}^n \left(\frac{1}{\mu_i}\left\|\boldsymbol{\alpha}_i\right\|_{\ell_2}^2 + \mu_i\right).
\end{aligned}$$

Then, the dual optimization problem consists in maximizing $\mathfrak{L}_D$ subject to the constraints $\sum_{i=1}^n \boldsymbol{\alpha}_i = 0$ and $\mu_i \geq 0, \forall i\in[n]$. Remarking that $\inf_{\forall i\in[n],\, \mu_i\in\mathbb{R}_+} \frac{1}{2}\sum_{i=1}^n \left(\frac{1}{\mu_i}\left\|\boldsymbol{\alpha}_i\right\|_{\ell_2}^2 + \mu_i\right) =$

$\sum_{i=1}^n \|\boldsymbol{\alpha}_i\|_{\ell_2}$ (Bach et al., 2012), a dual to Problem (P1) is:

$$\begin{aligned}
\underset{\forall i \in [n], \boldsymbol{\alpha}_i \in \mathbb{R}^p}{\text{minimize}} \quad & \frac{1}{n}\sum_{i=1}^n \ell^\star(\boldsymbol{\alpha}_i) + \frac{1}{2\lambda n^2}\sum_{i,j=1}^n \left\langle \boldsymbol{\alpha}_i, E_{\mathbf{x}_i}E^*_{\mathbf{x}_j}\boldsymbol{\alpha}_j \right\rangle_{\ell_2} \\
& -\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{\alpha}_i, \boldsymbol{y}_i\rangle_{\ell_2} + \frac{\epsilon}{n}\sum_{i=1}^n \|\boldsymbol{\alpha}_i\|_{\ell_2} \\
\text{s. t.} \quad & \sum_{i=1}^n \boldsymbol{\alpha}_i = 0.
\end{aligned} \tag{P2}$$

B.3. Generalization

Let $P\colon f \in \mathcal{F} \mapsto \mathbb{E}\left[\ell\left(Y - f(X)\right)\right]$ and $P_n\colon f \in \mathcal{F} \mapsto \frac{1}{n}\sum_{i=1}^n \ell_\epsilon\left(Y - f(X)\right)$, as well as respective twins P_ϵ and $P_{n,\epsilon}$ obtained by substituting ℓ_ϵ to ℓ. Let us decompose $P\hat{f}_\epsilon - Pf^\dagger$:

$$P\hat{f}_\epsilon - Pf^\dagger = \left(P\hat{f}_\epsilon - P_n\hat{f}_\epsilon\right) + \left(P_n\hat{f}_\epsilon - P_n f^\dagger\right) + \left(P_n f^\dagger - Pf^\dagger\right).$$

First, by concentration inequalities (Bartlett and Mendelson, 2002; Maurer, 2016; Sangnier et al., 2016), we have, with probability greater that $1-\delta$:

$$P\hat{f}_\epsilon - P_n\hat{f}_\epsilon \leq \sup_{f\in\mathcal{F}}\left(Pf - P_n f\right) \leq 2\sqrt{2}L\mathcal{R}_n\left(\mathcal{F}\right) + LM\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Second, let us decompose $P_n\hat{f}_\epsilon - P_n f^\dagger$:

$$P_n\hat{f}_\epsilon - P_n f^\dagger = \left(P_n\hat{f}_\epsilon - P_{n,\epsilon}\hat{f}_\epsilon\right) + \left(P_{n,\epsilon}\hat{f}_\epsilon - P_{n,\epsilon}f^\dagger\right) + \left(P_{n,\epsilon}f^\dagger - P_n f^\dagger\right).$$

By Lipschitz continuity, we have:

$$\forall \boldsymbol{\xi}, \boldsymbol{u} \in \mathbb{R}^p, \|\boldsymbol{u}\|_{\ell_2} \leq \epsilon\colon \ell(\boldsymbol{\xi}) - \ell(\boldsymbol{\xi} - \boldsymbol{u}) \leq L\left\|\boldsymbol{\xi} - (\boldsymbol{\xi} - \boldsymbol{u})\right\|_{\ell_2} \leq L\epsilon.$$

Consequently, $\ell(\boldsymbol{\xi}) - \ell_\epsilon(\boldsymbol{\xi}) \leq L\epsilon$ and $P_n\hat{f}_\epsilon - P_{n,\epsilon}\hat{f}_\epsilon \leq L\epsilon$. In addition $P_{n,\epsilon}\hat{f}_\epsilon - P_{n,\epsilon}f^\dagger \leq 0$ since $\hat{f}_\epsilon$ is a minimizer of $P_{n,\epsilon}$ over $\mathcal{F}$, and $f^\dagger \in \mathcal{F}$. Finally, $P_{n,\epsilon}f^\dagger - P_n f^\dagger \leq 0$ since ℓ upper bounds ℓ_ϵ. To summarize the second point, $P_n\hat{f}_\epsilon - P_n f^\dagger \leq L\epsilon$.

Third and last, by Hoeffding's inequality (Boucheron et al., 2013), with probability at least $1-\delta$:

$$P_n f^\dagger - Pf^\dagger \leq LM\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Gathering these three points with a union bound concludes the proof.

B.4. Algorithms

Proof (Lemma 3, body of the text) Let $\phi\colon \mu \in [0,1] \mapsto \left(1 + \frac{\lambda}{\left\|[\mu \boldsymbol{y}]_{-\boldsymbol{a}}^{\boldsymbol{b}}\right\|_{\ell_2}}\right)\mu$. First, $\phi(1) = 1 + \frac{\lambda}{\left\|[\boldsymbol{y}]_{-\boldsymbol{a}}^{\boldsymbol{b}}\right\|_{\ell_2}} \geq 1$. Second, for $\mu \geq 0$ sufficiently close to 0, $[\mu \boldsymbol{y}]_{-\boldsymbol{a}}^{\boldsymbol{b}} = \mu \boldsymbol{y}$ (since entries of $\boldsymbol{a}$ and $\boldsymbol{b}$ are positive). Therefore $\phi(0) = \lim_{\mu \downarrow 0} \left(\mu + \frac{\lambda \mu}{\mu \|\boldsymbol{y}\|_{\ell_2}}\right) = \frac{\lambda}{\|\boldsymbol{y}\|_{\ell_2}} \leq 1$. Finally, since ϕ is a continuous mapping on $[0,1]$ and $1 \in [\phi(0), \phi(1)]$, then the equation $\phi(\mu) = 1$ has a solution in $[0,1]$. ■

Proof (Proposition 4, body of the text) The proof is in two part. First, we write optimality conditions for the proximal operator of interest, then we show that $[\mu \boldsymbol{y}]_{-\boldsymbol{a}}^{\boldsymbol{b}}$ satisfies these optimality conditions when μ is appropriately defined. From now on, let $\boldsymbol{y} \in \mathbb{R}^p$.

Optimality conditions Let $\mathbf{x}^\star = \operatorname{prox}_{\lambda \|\cdot\|_{\ell_2} + \chi_{-\boldsymbol{a} \preccurlyeq \cdot \preccurlyeq \boldsymbol{b}}}(\boldsymbol{y}) = \arg\min_{-\boldsymbol{a} \preccurlyeq \mathbf{x} \preccurlyeq \boldsymbol{b}} \lambda \|\mathbf{x}\|_{\ell_2} + \frac{1}{2} \|\boldsymbol{y} - \mathbf{x}\|_{\ell_2}^2$.

1. Assume that $\mathbf{x}^\star \neq 0$. Then, $\lambda \|\cdot\|_{\ell_2} + \frac{1}{2} \|\boldsymbol{y} - \cdot\|_{\ell_2}^2$ is differentiable at $\mathbf{x}^\star$ and for each coordinate $j \in [p]$, either:

   (a) $-a_j < x_j^\star < b_j$ and $\left(1 + \frac{\lambda}{\|\mathbf{x}^\star\|_{\ell_2}}\right) x_j^\star = y_j$;

   (b) or $x_j^\star = b_j$ and $\left(1 + \frac{\lambda}{\|\mathbf{x}^\star\|_{\ell_2}}\right) x_j^\star \leq y_j$;

   (c) or $x_j^\star = -a_j$ and $\left(1 + \frac{\lambda}{\|\mathbf{x}^\star\|_{\ell_2}}\right) x_j^\star \geq y_j$.

   Gathering Conditions 1a-1c gives $\|\mathbf{x}^\star\|_{\ell_2} + \lambda \leq \|\boldsymbol{y}\|_{\ell_2}$. Since $\mathbf{x}^\star \neq 0$, we get $\lambda < \|\boldsymbol{y}\|_{\ell_2}$. Conversely, if $\|\boldsymbol{y}\|_{\ell_2} \leq \lambda$, then $\mathbf{x}^\star = 0$.

2. If $\mathbf{x}^\star = 0$, then $\forall \delta > 0$ such that $-\boldsymbol{a} \preccurlyeq \delta \boldsymbol{y} \preccurlyeq \boldsymbol{b}$, $\lambda \|\delta \boldsymbol{y}\|_{\ell_2} + \frac{1}{2} \|\boldsymbol{y} - \delta \boldsymbol{y}\|_{\ell_2}^2 \geq \frac{1}{2} \|\boldsymbol{y}\|_{\ell_2}^2$, that is $\lambda \|\boldsymbol{y}\|_{\ell_2} \geq \left(1 - \frac{\delta}{2}\right) \|\boldsymbol{y}\|_{\ell_2}^2$. Thus, by continuity when $\delta \downarrow 0$, we have $\lambda \geq \|\boldsymbol{y}\|_{\ell_2}$. To sum up, $\mathbf{x}^\star = 0$ if and only if $\|\boldsymbol{y}\|_{\ell_2} \leq \lambda$.

Proximal solution Let $\mathbf{x} = [\mu \boldsymbol{y}]_{-\boldsymbol{a}}^{\boldsymbol{b}}$, where μ is defined in Proposition 4 from the body of the text. Assume that $\|\boldsymbol{y}\|_{\ell_2} \leq \lambda$, then $\mu = 0$ and $\mathbf{x} = 0$ satisfies the optimality conditions. On the other hand, if $\|\boldsymbol{y}\|_{\ell_2} > \lambda$, then $\left(1 + \frac{\lambda}{\|\mathbf{x}\|_{\ell_2}}\right) \mu = 1$. As a result, either:

1. $-a_j < x_j < b_j$, so necessarily $x_j = \mu y_j$ (otherwise it would be clipped to b_j or $-a_j$). Therefore $\left(1 + \frac{\lambda}{\|\mathbf{x}\|_{\ell_2}}\right) x_j = \left(1 + \frac{\lambda}{\|\mathbf{x}\|_{\ell_2}}\right) (\mu y_j) = y_j$;

2. or $x_j = b_j$, meaning that $\mu y_j \geq b_j$. So $\left(1 + \frac{\lambda}{\|\mathbf{x}\|_{\ell_2}}\right) (\mu y_j) \geq \left(1 + \frac{\lambda}{\|\mathbf{x}\|_{\ell_2}}\right) b_j$, that is $\left(1 + \frac{\lambda}{\|\mathbf{x}\|_{\ell_2}}\right) x_j \leq y_j$;

3. or $x_j = -a_j$, meaning that $\mu y_j \leq -a_j$. So $\left(1 + \frac{\lambda}{\|\mathbf{x}\|_{\ell_2}}\right)(\mu y_j) \leq \left(1 + \frac{\lambda}{\|\mathbf{x}\|_{\ell_2}}\right)(-a_j)$, that is $\left(1 + \frac{\lambda}{\|\mathbf{x}\|_{\ell_2}}\right) x_j \geq y_j$.

Thus, when $\|\boldsymbol{y}\|_{\ell_2} > \lambda$, $\mathbf{x}$ satisfies the optimality conditions. This concludes the proof. ■

Corollary 6 *Let two n-tuples $\boldsymbol{A} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_n)$ and $\boldsymbol{B} = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_n)$ of vectors from $\mathbb{R}^p$ with positive entries. For any n-tuple $\boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$ of vectors from $\mathbb{R}^p$, let:*

$$(\mathbf{x}_1, \dots, \mathbf{x}_n) = \operatorname{prox}_{\lambda \|\cdot\|_{\ell_1/\ell_2} + \chi_{-\boldsymbol{A} \preccurlyeq \cdot \preccurlyeq \boldsymbol{B}}}(\boldsymbol{Y}),$$

where $\|\boldsymbol{Y}\|_{\ell_1/\ell_2} = \sum_{i=1}^n \|\boldsymbol{y}_i\|_{\ell_2}$. Then, $\forall i \in [n]$:

$$\mathbf{x}_i = \operatorname{prox}_{\lambda \|\cdot\|_{\ell_2} + \chi_{-\boldsymbol{a}_i \preccurlyeq \cdot \preccurlyeq \boldsymbol{b}_i}}(\boldsymbol{y}_i).$$

Proof This is a direct consequence of the separability of $\lambda \|\cdot\|_{\ell_1/\ell_2} + \chi_{-\boldsymbol{A} \preccurlyeq \cdot \preccurlyeq \boldsymbol{B}}$. ■

Proof (Proposition 5) The proof is similar to the one for Proposition 4 (see body of the text). Let $\boldsymbol{y} \in \mathbb{R}^p$.

Optimality conditions Let $\mathbf{x}^\star = \operatorname{prox}_{\lambda\left(\|\cdot\|_{\ell_2} + \psi\right)}(\boldsymbol{y}) = \arg\min_{\mathbf{x} \in \mathbb{R}^p} \lambda \|\mathbf{x}\|_{\ell_2} + \lambda \psi(\mathbf{x}) + \frac{1}{2} \|\boldsymbol{y} - \mathbf{x}\|_{\ell_2}^2$.

1. Assume that $\mathbf{x}^\star \neq 0$. Then, $\lambda \|\cdot\|_{\ell_2} + \lambda \psi + \frac{1}{2} \|\boldsymbol{y} - \cdot\|_{\ell_2}^2$ is differentiable at $\mathbf{x}^\star$ and for each coordinate $j \in [p]$:

$$y_j = \left(\frac{\lambda}{\|\mathbf{x}^\star\|_{\ell_2}} + \lambda \left|\tau_j - \mathrm{I}_{x_j^\star < 0}\right|^{-1} + 1\right) x_j^\star.$$

   It appears that $x_j^\star$ and y_j have same sign. Therefore, $\mathrm{I}_{x_j^\star < 0} = \mathrm{I}_{y_j < 0}$ and

$$x_j^\star = \left(\frac{\lambda}{\|\mathbf{x}^\star\|_{\ell_2}} + \lambda \left|\tau_j - \mathrm{I}_{y_j < 0}\right|^{-1} + 1\right)^{-1} y_j.$$

   Now, the previous relation implies:

$$\|\mathbf{x}^\star\|_{\ell_2}^2 = \sum_{j=1}^p \frac{y_j^2}{\left(\frac{\lambda}{\|\mathbf{x}^\star\|_{\ell_2}} + \lambda \left|\tau_j - \mathrm{I}_{y_j < 0}\right|^{-1} + 1\right)^2}.$$

   Since $\mathbf{x}^\star \neq 0$, we get:

$$1 = \sum_{j=1}^p \frac{y_j^2}{\left(\lambda + \|\mathbf{x}^\star\|_{\ell_2} \left(\lambda \left|\tau_j - \mathrm{I}_{y_j < 0}\right|^{-1} + 1\right)\right)^2}.$$

But $\|\mathbf{x}^\star\|_{\ell_2} > 0$, so:

$$1 = \sum_{j=1}^{p} \frac{y_j^2}{\left(\lambda + \|\mathbf{x}^\star\|_{\ell_2}\left(\lambda\left|\tau_j - \mathrm{I}_{y_j<0}\right|^{-1} + 1\right)\right)^2} < \sum_{j=1}^{p} \frac{y_j^2}{\lambda^2},$$

that is $\lambda < \|\boldsymbol{y}\|_{\ell_2}$. Conversely, if $\|\boldsymbol{y}\|_{\ell_2} \leq \lambda$, then $\mathbf{x}^\star = 0$.

2. If $\mathbf{x}^\star = 0$, then $\forall \delta > 0$, $\lambda \|\delta \boldsymbol{y}\|_{\ell_2} + \lambda\psi(\delta\boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{y} - \delta\boldsymbol{y}\|_{\ell_2}^2 \geq \frac{1}{2}\|\boldsymbol{y}\|_{\ell_2}^2$, that is $\lambda(\|\boldsymbol{y}\|_{\ell_2} + \delta\psi(\boldsymbol{y})) \geq \left(1 - \frac{\delta}{2}\right)\|\boldsymbol{y}\|_{\ell_2}^2$. Thus, by continuity when $\delta \downarrow 0$, we have $\lambda \geq \|\boldsymbol{y}\|_{\ell_2}$. To sum up, $\mathbf{x}^\star = 0$ if and only if $\|\boldsymbol{y}\|_{\ell_2} \leq \lambda$.

Proximal solution If $\|\boldsymbol{y}\|_{\ell_2} \leq \lambda$, then $\mathbf{x} = 0$ is satisfies trivially the optimality conditions. On the other hand, if $\|\boldsymbol{y}\|_{\ell_2} > \lambda$, then $\sum_{j=1}^p \frac{y_j^2}{\lambda^2} > 1$ and $\lim_{\mu\to+\infty} \sum_{j=1}^p \frac{y_j^2}{\left(\mu\left(1+\lambda\left|\tau_j - \mathrm{I}_{y_j<0}\right|^{-1}\right)+\lambda\right)^2} = 0$. Thus, by continuity, Equation 1 has a solution $\mu > 0$. Let μ be such a solution and let $\mathbf{x} \in \mathbb{R}^p$ such that for each coordinate $j \in [p]$,

$$x_j = \left(1 + \frac{\lambda}{\mu} + \lambda\left|\tau_j - \mathrm{I}_{y_j<0}\right|^{-1}\right)^{-1} y_j.$$

Then:

$$\frac{\|\mathbf{x}\|_{\ell_2}^2}{\mu^2} = \sum_{j=1}^{p} \frac{y_j^2}{\mu^2\left(1 + \frac{\lambda}{\mu} + \lambda\left|\tau_j - \mathrm{I}_{y_j<0}\right|^{-1}\right)^2} = 1.$$

Consequently

$$x_j = \left(1 + \frac{\lambda}{\|\mathbf{x}\|_{\ell_2}} + \lambda\left|\tau_j - \mathrm{I}_{y_j<0}\right|^{-1}\right)^{-1} y_j.$$

and $\mathbf{x}$ satisfies the optimality conditions. This concludes the proof. ∎

Appendix C. Numerical experiments

Table 1 reports the average empirical loss (scaled by 100) along with the standard deviations. It completes Talbe 2 from the body of the text. For each dataset, the bold-face numbers are the two lowest losses. These values should be compared to the loss for $\epsilon = 0$.

References

M.A. Alvarez, L. Rosasco, and N.D. Lawrence. Kernels for Vector-Valued Functions: a Review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with Sparsity-Inducing Penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, January 2012.

Table 1: Empirical pinball loss $\times 100$ along with percentage of support vectors (the less, the better).

Data set	$\epsilon = 0$	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.5$	$\epsilon = 0.75$	$\epsilon = 1$	$\epsilon = 1.5$	$\epsilon = 2$	$\epsilon = 3$
caution	67.50 ± 12.96	**67.40** ± 13.12	**67.17** ± 12.68	67.54 ± 13.00	69.93 ± 12.51	73.24 ± 13.10	76.80 ± 12.51	83.42 ± 14.97	100.22 ± 16.63	142.19 ± 16.11
ftcollinssnow	**109.07** ± 5.88	109.12 ± 5.95	109.14 ± 6.00	109.15 ± 6.00	109.11 ± 6.30	110.39 ± 7.10	**109.05** ± 6.72	110.90 ± 6.87	109.81 ± 6.22	113.50 ± 9.29
highway	79.29 ± 17.06	78.10 ± 16.15	76.75 ± 15.42	76.66 ± 19.11	**75.09** ± 18.82	**70.94** ± 21.58	75.10 ± 21.02	97.67 ± 22.91	112.30 ± 20.10	112.09 ± 20.12
heights	91.05 ± 1.12	91.00 ± 1.20	90.98 ± 1.19	**90.98** ± 1.21	91.18 ± 1.09	91.21 ± 1.27	**90.98** ± 1.13	91.09 ± 1.02	91.51 ± 1.20	93.34 ± 1.84
sniffer	32.34 ± 5.14	**31.40** ± 4.81	32.31 ± 5.19	**31.40** ± 2.98	34.64 ± 3.85	39.84 ± 5.03	41.82 ± 4.14	52.06 ± 5.88	62.21 ± 11.77	103.76 ± 16.42
snowgeese	**49.62** ± 19.38	**50.51** ± 18.23	51.25 ± 18.64	51.08 ± 17.69	52.88 ± 15.11	53.81 ± 13.97	62.81 ± 19.66	90.15 ± 23.97	107.53 ± 23.13	94.25 ± 24.65
ufc	57.87 ± 3.09	57.90 ± 3.07	**57.78** ± 2.99	57.84 ± 3.01	**57.67** ± 2.84	57.84 ± 2.76	58.19 ± 2.92	61.04 ± 3.68	66.81 ± 4.00	86.23 ± 4.79
birthwt	99.93 ± 8.51	99.95 ± 8.53	99.93 ± 8.63	99.70 ± 8.64	**99.25** ± 8.66	100.50 ± 10.31	99.80 ± 11.29	**98.71** ± 10.04	99.56 ± 9.43	103.39 ± 8.83
crabs	8.59 ± 0.66	**8.52** ± 0.68	**8.49** ± 0.73	9.44 ± 0.57	19.94 ± 1.38	23.08 ± 1.59	31.44 ± 3.75	44.08 ± 4.64	53.45 ± 5.65	86.91 ± 9.48
GAGurine	44.30 ± 5.85	**44.26** ± 5.79	**44.25** ± 5.76	44.86 ± 6.04	46.20 ± 5.35	49.87 ± 4.88	52.88 ± 3.94	57.06 ± 3.47	65.89 ± 3.88	103.32 ± 24.62
geyser	**77.81** ± 5.36	78.15 ± 5.39	**78.12** ± 5.38	78.45 ± 5.35	78.40 ± 5.77	78.28 ± 5.88	78.54 ± 5.82	80.55 ± 6.34	85.15 ± 6.18	99.92 ± 8.65
gilgais	**32.96** ± 4.09	**33.12** ± 3.99	33.27 ± 4.11	33.42 ± 3.88	35.08 ± 3.35	36.62 ± 3.59	37.94 ± 3.68	48.17 ± 9.44	94.65 ± 4.98	104.12 ± 5.92
topo	47.49 ± 7.93	48.93 ± 7.43	48.74 ± 7.10	48.17 ± 7.01	**41.65** ± 5.60	**45.24** ± 3.53	51.19 ± 7.92	53.68 ± 8.39	58.21 ± 13.35	80.57 ± 15.18
BostonHousing	**34.54** ± 3.34	34.68 ± 3.46	34.70 ± 3.39	**34.09** ± 3.37	35.27 ± 3.02	37.65 ± 3.18	41.31 ± 3.41	55.04 ± 5.61	73.39 ± 12.35	112.22 ± 12.91
CobarOre	**0.50** ± 0.38	**5.05** ± 1.90	8.75 ± 3.44	12.47 ± 4.27	23.84 ± 6.03	35.82 ± 8.20	47.35 ± 10.94	66.15 ± 14.56	84.51 ± 17.70	106.89 ± 15.52
engel	43.57 ± 6.05	43.50 ± 6.02	**43.47** ± 6.08	**43.44** ± 5.99	57.36 ± 46.14	43.98 ± 5.37	46.31 ± 6.29	53.15 ± 5.45	69.43 ± 9.22	100.48 ± 11.63
mcycle	**63.95** ± 5.25	**63.88** ± 5.20	64.26 ± 5.99	64.90 ± 6.68	65.89 ± 5.89	67.29 ± 6.13	70.11 ± 7.65	74.78 ± 6.43	86.49 ± 6.77	109.79 ± 12.67
BigMac2003	**49.94** ± 12.85	**49.97** ± 12.84	50.00 ± 12.83	50.27 ± 13.19	51.16 ± 13.37	51.44 ± 10.57	53.63 ± 14.29	77.40 ± 24.48	106.38 ± 13.97	136.76 ± 61.70
UN3	71.27 ± 4.69	**70.94** ± 4.57	**71.03** ± 4.68	71.49 ± 5.06	71.37 ± 5.01	71.53 ± 5.90	72.68 ± 6.17	76.72 ± 6.13	84.50 ± 7.00	109.59 ± 4.71
cpus	**11.31** ± 9.32	**13.32** ± 8.95	15.57 ± 9.16	20.16 ± 8.06	25.88 ± 8.93	35.66 ± 11.61	55.27 ± 14.69	65.05 ± 9.70	65.05 ± 9.70	65.02 ± 9.65

P.L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, New York, 2013.

S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

M. Sangnier, O. Fercoq, and F. d'Alché Buc. Joint quantile regression in vector-valued RKHSs. In *Advances in Neural Information Processing Systems 29*, 2016.

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