JMLR: Workshop and Conference Proceedings 80:1–17, 2017 ACML 2017

Data sparse nonparametric regression with ϵ-insensitive losses

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

Abstract

Leveraging the celebrated support vector regression (SVR) method, we propose a unifying framework in order to deliver regression machines in reproducing kernel Hilbert spaces (RKHSs) with data sparsity. The central point is a new definition of ϵ-insensitivity, valid for many regression losses (including quantile and expectile regression) and their multivariate extensions. We show that the dual optimization problem to empirical risk minimization with ϵ-insensitivity involves a data sparse regularization. We also provide an analysis of the excess of risk as well as a randomized coordinate descent algorithm for solving the dual. Numerical experiments validate our approach.

Keywords: Quantile regression, Expectile regression, Operator-valued kernel.

1. Introduction

In supervised learning, sparsity is a common feature (Hastie et al., 2015). When it is put on the table, the first reaction is certainly to invoke feature selection (Tibshirani, 1996). In many domains such as finance and biology, feature selection aims at discovering the covariates that really explain the outcomes (hopefully in a causal manner). However, when dealing with nonparametric modeling, sparsity may also occur in the orthogonal direction: data, which support ultimately the desired estimator. Data sparsity is mainly valuable for two reasons: first, controlling the number of training data used in the model can prevent from overfitting. Second, data sparsity implies less data to store and less time needed for prediction as well as potentially less time needed for training. The main evidence of this phenomenon lies in the celebrated support vector machines (SVMs) (Boser et al., 1992; Vapnik, 2010; Mohri et al., 2012). The principle of margin maximization results in selecting only a few data points that support the whole prediction function, leaving the others sink into oblivion.

Data sparsity is particularly known for classification thanks to SVMs but is less predominant for regression. Indeed, while classifying relies on the few points close to the frontier (they are likely to be misclassified), regression is supported by points for which the outcome is far from the conditional expectation. Those points, that are likely to be poorly characterized by the prediction, gather most of the training points. Thus, by nature, regressors tend to be less data sparse than classifiers. In order to obtain the valued data sparsity for a wide class of regression problems, we consider regression relying on empirical risk minimization (Vapnik, 2010) with non-linear methods based on RKHSs for uni and multidimensional outcomes (Micchelli and Pontil, 2005).

The companion of SVM for regression, SVR (Drucker et al., 1997), is probably the most relevant representative of data sparse regression methods in the frequentist domain (in the Bayesian literature, relevance vector machines can deliver sparse regressors (Tipping, 2001)). SVR makes use of the celebrated ϵ-insensitive loss $\xi \in \mathbb{R} \mapsto \max(0, |\xi| - \epsilon)$, where $\epsilon > 0$, in order to produce data sparsity. In practice, residues close to 0 are not penalized, and consequently corresponding points are not considered in the prediction function. Later, Park and Kim (2011) extended this ϵ-insensitive loss to quantile regression (QR) (Koenker, 2005) in an *ad hoc* manner.

These approaches are in deep contrast compared to kernel ridge regression (using the loss $\xi \in \mathbb{R} \mapsto \frac{1}{2}\xi^2$), that does not benefit from data sparsity. To remedy that lack of sparsity, Lim et al. (2014) leveraged the representer theorem (Schölkopf et al., 2001) for ridge regression and penalized the coefficients associated to each training point. The same procedure was used by Zhang et al. (2016) for QR but with a sparse constraint instead of a sparse regularization. Contrarily to ϵ-insensitive losses, the latter approaches constrain the hypotheses while keeping the vanilla regression losses.

The previous works are either restricted to unidimensional losses (Drucker et al., 1997; Park and Kim, 2011; Zhang et al., 2016) or compelled to assume a finite representation of the optimal regressor (Lim et al., 2014; Zhang et al., 2016). Our contribution in this context is to provide a novel and unifying view of ϵ-insensitive losses for data sparse kernel regression, that is theoretically well founded (Section 3). In the spirit of SVR, we show that, for a large number of regression problems (uni or multivariate, quantile or traditional regression), data sparsity can be obtained by properly defining an ϵ-insensitive variant of the loss in the data-fitting term which turns to be a data sparse regularization in the dual. This framework contains SVR as a special case but also a new loss for QR (that is different from the one introduced by Park and Kim (2011)) and expectile regression (Newey and Powell, 1987), as well as their multivariate extensions (Section 4) in the context of vector-valued RKHSs. We also provide an analysis of the excess of risk (Section 5). Special attention is paid to the dual form of the optimization problem and a randomized primal-dual block coordinate descent algorithm for estimating our data sparse regressor (Section 6) is derived. Numerical experiments conclude the discussion (Section 7).

2. Framework

Notations In the whole paper, we denote vectors with bold-face letters. Moreover, $\tau \in (0,1)$ is a quantile level, $\boldsymbol{\tau} \in (0,1)^p$ is a vector of quantile levels, n and p are two positive integers (samples size and dimension), $[n]$ is the range of integers between 1 and n, $\boldsymbol{\alpha}^l \in \mathbb{R}^n$ denotes the l^{th} *row* vector of any $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_i)_{1 \le i \le n} \in (\mathbb{R}^p)^n$, $\mathrm{I}_{\cdot}$ is the indicator function (set to 1 when the condition is fulfilled and 0 otherwise), $\mathbb{1}$ denotes the all-ones vector, $\chi_{\cdot}$ is the characteristic function (set to 0 when the condition is met and to ∞ otherwise), $\preccurlyeq$ (respectively $\prec$) denotes the pointwise (respectively strict) inequality, diag is the operator mapping a vector to a diagonal matrix, $\partial\psi$ is the subdifferential of any function ψ (when it exists), prox_ψ denotes its proximal operator (Bauschke and Combettes, 2011), $\text{proj}_{\mathbb{1}}$ is the projector onto the span of $\mathbb{1}$.

General framework Let $\mathcal{X}$ be a nonempty input space and $\mathbb{R}^p$ be the output Hilbert space (for a given integer p). Let also $\mathcal{H}$ be a Hilbert space of hypotheses $h\colon \mathcal{X} \to \mathbb{R}^p$ and $\{(\mathbf{x}_i, \boldsymbol{y}_i)\}_{1\leq i\leq n}$ be a training sample of n couples $(\mathbf{x}_i, \boldsymbol{y}_i) \in \mathcal{X} \times \mathbb{R}^p$. We consider the setting of regularized empirical risk minimization for regression, based on a real-valued convex loss function $\ell\colon \mathbb{R}^p \to \mathbb{R}$. This is to provide a minimizer to the optimization problem:

$$\underset{h\in\mathcal{H},\, \boldsymbol{b}\in\mathbb{R}^p}{\text{minimize}} \ \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \frac{1}{n}\sum_{i=1}^{n} \ell\left(\boldsymbol{y}_i - (h(\mathbf{x}_i) + \boldsymbol{b})\right), \tag{P1}$$

where $\lambda > 0$ is a trade-off parameter and $\|\cdot\|_{\mathcal{H}}$ is the norm associated to $\mathcal{H}$. Problem (P1) consists in minimizing a data-fitting term $\frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\xi}_i)$, where the residues $\boldsymbol{\xi}_i = \boldsymbol{y}_i - (h(\mathbf{x}_i) + \boldsymbol{b})$ are driven close to zero. Here, the prediction function (or regressor) is $f\colon \mathbf{x} \in \mathcal{X} \mapsto h(\mathbf{x}) + \boldsymbol{b}$. It is made of a functional part $h \in \mathcal{H}$ and of an intercept $\boldsymbol{b} \in \mathbb{R}^p$ (such as what we can find in SVMs). The last component of Problem (P1), $\frac{\lambda}{2}\|\cdot\|_{\mathcal{H}}^2$, is a regularizer, which penalizes functions with high complexities, allowing for generalizing on unknown observations (Mohri et al., 2012).

Let, for any $\mathbf{x} \in \mathcal{X}$, $E_{\mathbf{x}}\colon h \in \mathcal{H} \mapsto h(\mathbf{x})$ be the evaluation map and $E_{\mathbf{x}}^*$ be its adjoint operator. A common theorem in nonparametric regression, known as the representer theorem (Schölkopf et al., 2001), states that any estimator based on (P1) is a finite expansion of local functions supported by the training examples. More formally, the functional part of any solution $(\hat{h}, \hat{\boldsymbol{b}})$ of Problem (P1) admits a representation of the form $\hat{h} = \sum_{i=1}^{n} E_{\mathbf{x}_i}^* \boldsymbol{\alpha}_i$, where $\boldsymbol{\alpha}_i \in \mathbb{R}^p$ for all $i \in [n]$.

A proof of this statement based on Karush-Kuhn-Tucker (KKT) conditions can be found in Supplementary B, but many proofs already exist for kernel methods in the scalar ($p = 1$) and the vectorial ($p > 1$) cases (Schölkopf et al., 2001; Brouard et al., 2016). The representer theorem states that $\hat{h}$ can be parametrized by a finite number of $\boldsymbol{\alpha}_i \in \mathbb{R}^p$ (which boil down to be Lagrange multipliers), such that each $\boldsymbol{\alpha}_i$ is associated to a training point $\mathbf{x}_i$. Therefore, as soon as $\boldsymbol{\alpha}_i = 0$, $\mathbf{x}_i$ does not appear in $\hat{h}$ (that is, $\mathbf{x}_i$ is not a support vector). When many $\boldsymbol{\alpha}_i = 0$, we say that $\hat{h}$ is data sparse. This feature is very attractive since it makes possible to get rid of useless training points when computing $\hat{h}(\mathbf{x})$, and speeds up the prediction task. The next section introduces a way to get such a sparsity.

3. ϵ-insensitive losses and the associated minimization problem

In the whole discussion, we will assume that the loss function ℓ is convex, lower semi-continuous and has a unique minimum at 0 with null value. Then, we develop the idea that data sparsity can be obtained by substituting the original loss function ℓ by a slightly different version ℓ_ϵ, where ℓ_ϵ is a kind of *soft-thresholding of* ℓ. More formally, let ϵ be a positive parameter and define the ϵ-insensitive the loss ℓ_ϵ by:

$$\forall \boldsymbol{\xi} \in \mathbb{R}^p\colon \quad \ell_\epsilon(\boldsymbol{\xi}) = \begin{cases} 0 & \text{if } \|\boldsymbol{\xi}\|_{\ell_2} \leq \epsilon \\ \inf\limits_{\boldsymbol{d}\in\mathbb{R}^p\,:\, \|\boldsymbol{d}\|_{\ell_2}=1} \ell\left(\boldsymbol{\xi} - \epsilon\boldsymbol{d}\right) & \text{otherwise.} \end{cases}$$

	Name	Loss $\ell(\boldsymbol{\xi})$	ϵ-insensitive loss $\ell_\epsilon(\boldsymbol{\xi})$	Conjugate loss $\ell^\star(\boldsymbol{\alpha})$
Unidimensional	Absolute	$\lvert\xi\rvert$	$\max(0, \lvert\xi\rvert - \epsilon)$	$\chi_{-1 \leq \alpha \leq 1}$
$(\xi \in \mathbb{R}, \alpha \in \mathbb{R})$	Squared	$\frac{1}{2}\xi^2$	$\frac{1}{2}\left(\max\left(0, \lvert\xi\rvert - \epsilon\right)\right)^2$	$\frac{1}{2}\alpha^2$
	Pinball	$\left(\tau - \mathrm{I}_{\xi<0}\right)\xi$	$\begin{cases} 0 & \text{if } \lvert\xi\rvert \leq \epsilon \\ \tau(\xi - \epsilon) & \text{if } \xi \geq \epsilon \\ (\tau - 1)(\xi + \epsilon) & \text{if } \xi \leq -\epsilon \end{cases}$	$\chi_{\tau - 1 \leq \alpha \leq \tau}$
	Squared pinball	$\frac{1}{2}\left\lvert\tau - \mathrm{I}_{\xi<0}\right\rvert \xi^2$	$\begin{cases} 0 & \text{if } \lvert\xi\rvert \leq \epsilon \\ \frac{\tau}{2}(\xi - \epsilon)^2 & \text{if } \xi \geq \epsilon \\ \frac{1-\tau}{2}(\xi + \epsilon)^2 & \text{if } \xi \leq -\epsilon \end{cases}$	$\frac{1}{2}\left\lvert\tau - \mathrm{I}_{\alpha<0}\right\rvert^{-1}\alpha^2$
Multidimensional	ℓ_1-norm	$\lVert\boldsymbol{\xi}\rVert_{\ell_1}$	No closed-form found	$\chi_{-\mathbb{1} \preccurlyeq \boldsymbol{\alpha} \preccurlyeq \mathbb{1}}$
$(\boldsymbol{\xi} \in \mathbb{R}^p, \boldsymbol{\alpha} \in \mathbb{R}^p)$	ℓ_2-norm	$\frac{1}{2}\lVert\boldsymbol{\xi}\rVert_{\ell_2}^2$	$\frac{1}{2}\left\lVert\boldsymbol{\xi} - \min(\lVert\boldsymbol{\xi}\rVert_{\ell_2}, \epsilon)\frac{\boldsymbol{\xi}}{\lVert\boldsymbol{\xi}\rVert_{\ell_2}}\right\rVert_{\ell_2}^2$	$\frac{1}{2}\lVert\boldsymbol{\alpha}\rVert_{\ell_2}^2$
	Multiple pinball	$\sum_{j=1}^p\left(\tau_j - \mathrm{I}_{\xi_j<0}\right)\xi_j$	No closed-form found	$\chi_{\boldsymbol{\tau} - \mathbb{1} \preccurlyeq \boldsymbol{\alpha} \preccurlyeq \boldsymbol{\tau}}$
	Multiple squared pinball	$\frac{1}{2}\sum_{j=1}^p\left\lvert\tau_j - \mathrm{I}_{\xi_j<0}\right\rvert \xi_j^2$	No closed-form found	$\frac{1}{2}\sum_{j=1}^p\left\lvert\tau_j - \mathrm{I}_{\alpha_j<0}\right\rvert^{-1}\alpha_j^2$

Table 1: Examples of ϵ-insensitive losses.

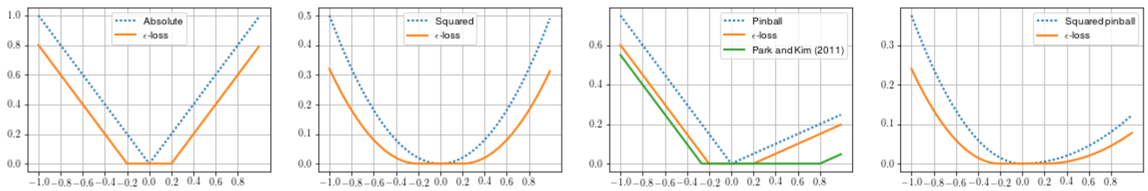

Figure 1: Examples of unidimensional ϵ-losses ($\epsilon = 0.2$, $\tau = 0.25$).

Put another way, the new loss value associated to a residue $\boldsymbol{\xi}$ is set to 0 if the magnitude of the residue is sufficiently small, or to the smallest loss value in a neighborhood centered at the original residue. Let us remark that, by the previous assumptions, ℓ_ϵ is convex (see Supplementary B). To illustrate this definition, Table 1 as well as Figure 1 provide several examples of ϵ-insensitive losses, particularly for regression, QR and expectile regression. Before developing further these examples in the next section, we explain why using this ϵ-insensitive loss helps in getting data sparsity for the regressor.

Proposition 1 (Dual optimization problem) *A dual to the learning problem*

$$\underset{h \in \mathcal{H},\, \boldsymbol{b} \in \mathbb{R}^p}{\text{minimize}} \ \frac{\lambda}{2}\lVert h\rVert_{\mathcal{H}}^2 + \frac{1}{n}\sum_{i=1}^n \ell_\epsilon\left(\boldsymbol{y}_i - (h(\mathbf{x}_i) + \boldsymbol{b})\right) \tag{P2}$$

is

$$\begin{aligned} &\underset{\forall i \in [n],\, \boldsymbol{\alpha}_i \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^n \ell^\star(\boldsymbol{\alpha}_i) + \frac{1}{2\lambda n^2}\sum_{i,j=1}^n \left\langle \boldsymbol{\alpha}_i, E_{\mathbf{x}_i}E_{\mathbf{x}_j}^*\boldsymbol{\alpha}_j\right\rangle_{\ell_2} - \frac{1}{n}\sum_{i=1}^n \langle\boldsymbol{\alpha}_i, \boldsymbol{y}_i\rangle_{\ell_2} + \frac{\epsilon}{n}\sum_{i=1}^n \lVert\boldsymbol{\alpha}_i\rVert_{\ell_2} \\ &\text{s.t.} \qquad \sum_{i=1}^n \boldsymbol{\alpha}_i = 0, \end{aligned} \tag{P3}$$

where $\ell^\star \colon \boldsymbol{\alpha} \in \mathbb{R}^p \mapsto \sup_{\boldsymbol{\xi} \in \mathbb{R}^p} \langle\boldsymbol{\alpha}, \boldsymbol{\xi}\rangle_{\ell_2} - \ell(\boldsymbol{\xi})$ is the Fenchel-Legendre transform of ℓ. In addition, let $(\hat{\boldsymbol{\alpha}}_i)_{i\in[n]}$ be solutions to Problem (P3). Then, the function $\hat{h} = \frac{1}{\lambda n}\sum_{i=1}^n E_{\mathbf{x}_i}^*\hat{\boldsymbol{\alpha}}_i$ ($\hat{h} \in \mathcal{H}$) is solution to Problem (P2) for a given intercept $\hat{\boldsymbol{b}} \in \mathbb{R}^p$.

We see that turning ℓ into an ϵ-insensitive loss leads to blending a sparsity regularization in the learning process: the ℓ_1/ℓ_2-norm $\sum_{i=1}^n \|\boldsymbol{\alpha}_i\|_{\ell_2}$, which is known to induce sparsity in the vectors $\boldsymbol{\alpha}_i$ (Bach et al., 2012). Even though a proof of this proposition is proposed in Supplementary B, we can grasp this phenomenon by remarking that ℓ_ϵ turns out to be an infimal convolution (Bauschke and Combettes, 2011), noted $\square$, of two functions. Thus, its Fenchel-Legendre transform is the sum of Fenchel-Legendre transforms of each contribution (Bauschke and Combettes, 2011):

$$\ell_\epsilon^\star = \left(\ell \,\square\, \chi_{\|\cdot\|_{\ell_2} \leq \epsilon}\right)^\star = \ell^\star + \left(\chi_{\|\cdot\|_{\ell_2} \leq \epsilon}\right)^\star = \ell^\star + \epsilon \|\cdot\|_{\ell_2}.$$

The next section discusses several applications of ϵ-insensitive losses.

4. Examples of ϵ-insensitive losses

To illustrate the link between ϵ-insensitive losses and data sparsity, we focus on RKHSs $\mathcal{H}$, based either on a scalar-valued kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Steinwart and Christmann, 2008) or on a matrix-valued kernel (MVK) $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{p \times p}$ (Micchelli and Pontil, 2005). Therefore, the base functions $E_{\mathbf{x}}^* \boldsymbol{\alpha}_i$, appearing in Proposition 1 become $\mathbf{x}' \in \mathcal{X} \mapsto \alpha_i k(\mathbf{x}', \mathbf{x}) \in \mathbb{R}$ ($\alpha_i \in \mathbb{R}$) and $\mathbf{x}' \in \mathcal{X} \mapsto K(\mathbf{x}', \mathbf{x}) \boldsymbol{\alpha}_i \in \mathbb{R}^p$ ($\alpha_i \in \mathbb{R}^p$), respectively for the uni and multidimensional case.

4.1. Least absolute deviation

The notion of ϵ-insensitive loss is very well illustrated in SVR: let us study the problem of least absolute deviation, that is, solving (P1), where there is a single output ($h \colon \mathcal{X} \to \mathbb{R}$, $y_i \in \mathbb{R}, b \in \mathbb{R}, \alpha_i \in \mathbb{R}$) and the loss to be minimized is $\ell(\xi) = |\xi|$. In this situation, it is easy to check that ℓ_ϵ boils down to: $\ell_\epsilon(\xi) = \max(0, |\xi| - \epsilon)$, which is the well known ϵ-insensitive loss considered in SVR (Hastie et al., 2009).

Now, since $\ell^\star(\alpha) = 0$ when $-1 \leq \alpha \leq 1$ and $+\infty$ otherwise, Problem (P3) becomes (up to normalizing the objective function by n):

$$\begin{aligned} &\underset{\forall i \in [n],\, \alpha_i \in \mathbb{R}}{\text{minimize}} \; \frac{1}{2\lambda n} \sum_{i,j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^n \alpha_i y_i + \epsilon \sum_{i=1}^n |\alpha_i| \\ &\text{s. t.} \quad \forall i \in [n], -1 \leq \alpha_i \leq 1, \quad \sum_{i=1}^n \alpha_i = 0. \end{aligned}$$

Using that $|\alpha| = \inf_{\substack{\alpha^+, \alpha^- \geq 0 \\ \alpha = \alpha^+ - \alpha^-}} \{\alpha^+ + \alpha^-\}$, we can introduce auxiliary variables $\alpha_i^+ \geq 0$ and $\alpha_i^- \geq 0$, decouple α_i in $\alpha_i^+ - \alpha_i^-$ and replace $|\alpha_i|$ by $\alpha_i^+ + \alpha_i^-$. This turns the previous optimization problem into:

$$\begin{aligned} &\underset{\forall i \in [n],\, \alpha_i^+, \alpha_i^- \in \mathbb{R}}{\text{minimize}} \; \frac{1}{2\lambda n} \sum_{i,j=1}^n (\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-) k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^n (\alpha_i^+ - \alpha_i^-) y_i + \epsilon \sum_{i=1}^n (\alpha_i^+ + \alpha_i^-) \\ &\text{s. t.} \quad \forall i \in [n], 0 \leq \alpha_i^+ \leq 1, \quad 0 \leq \alpha_i^- \leq 1, \quad \sum_{i=1}^n \alpha_i^+ - \alpha_i^- = 0, \end{aligned}$$

which is the well known dual of SVR. In this perspective, SVR is a data sparse version of least absolute deviation, for which the sparsity inducing norm appears in the dual.

4.2. Least mean squares

Besides SVR, Kernel Ridge Regression is also well known among kernel methods. In particular, for the general case of multivariate regression ($\boldsymbol{y}_i \in \mathbb{R}^p$), we aim at minimizing the squared loss $\ell(\cdot) = \frac{1}{2} \|\cdot\|_{\ell_2}^2$. Referring to Table 1, Problem (P3) then becomes:

$$\begin{aligned} \underset{\forall i \in [n],\, \boldsymbol{\alpha}_i \in \mathbb{R}^p}{\text{minimize}} \quad & \frac{1}{2} \sum_{i,j=1}^{n} \left\langle \boldsymbol{\alpha}_i, \left(\mathrm{I}_{i=j} \boldsymbol{I}_p + \frac{1}{\lambda n} K(\mathbf{x}_i, \mathbf{x}_j) \right) \boldsymbol{\alpha}_j \right\rangle_{\ell_2} - \sum_{i=1}^{n} \langle \boldsymbol{\alpha}_i, \boldsymbol{y}_i \rangle_{\ell_2} + \epsilon \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_{\ell_2} \\ \text{s.t.} \quad & \sum_{i=1}^{n} \boldsymbol{\alpha}_i = 0, \end{aligned} \tag{P4}$$

where $\boldsymbol{I}_p$ is the p-dimensional identity matrix. This is similar to the dual of kernel ridge regression but with an extra ℓ_1/ℓ_2-norm on Lagrange multipliers.

As far as we know, data sparsity for multivariate kernel ridge regression has been first introduced by Lim et al. (2014). In this work, the authors assume first that the optimal predictor $\hat{h}$ has a finite representation $\hat{h} = \frac{1}{\lambda n} \sum_{i=1}^{n} K(\cdot, \mathbf{x}_i)\boldsymbol{c}_i$ without relying on a representer theorem and regularize the vectors of weights $\boldsymbol{c}_i$ associated to each point thanks to an ℓ_1/ℓ_2-norm. They also dismiss the intercept $\boldsymbol{b}$ for simplicity. The learning problem, as presented in (Lim et al., 2014) (with normalization chosen at purpose), is:

$$\begin{aligned} \underset{\substack{h \in \mathcal{H}, \\ \forall i \in [n],\, \boldsymbol{c}_i \in \mathbb{R}^p}}{\text{minimize}} \quad & \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \frac{1}{2n} \sum_{i=1}^{n} \|\boldsymbol{y}_i - h(\mathbf{x}_i)\|_{\ell_2}^2 + \frac{\epsilon}{\lambda n^2} \sum_{i=1}^{n} \|\boldsymbol{c}_i\|_{\ell_2} \\ \text{s.t.} \quad & h = \frac{1}{\lambda n} \sum_{i=1}^{n} K(\cdot, \mathbf{x}_i)\boldsymbol{c}_i. \end{aligned}$$

For the sake of readability, we write $\underline{\boldsymbol{c}} = (\boldsymbol{c}_1^\top, \dots, \boldsymbol{c}_n^\top)^\top$ the vector of all weights and $\|\underline{\boldsymbol{c}}\|_{\ell_1/\ell_2} = \sum_{i=1}^{n} \|\boldsymbol{c}_i\|_{\ell_2}$. Let also $\underline{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{1 \le i,j \le n}$ be the symmetric positive semi-definite (PSD) kernel matrix. Fermat's rule indicates that a solution $\hat{\underline{\boldsymbol{c}}}$ of the previous learning problem must verify $\underline{K} \left(\boldsymbol{y} - \frac{1}{\lambda n} \underline{K} \hat{\underline{\boldsymbol{c}}} - \hat{\underline{\boldsymbol{c}}} \right) \in \partial(\epsilon \|\cdot\|_{\ell_1/\ell_2})(\hat{\underline{\boldsymbol{c}}})$.

To compare Lim et al.'s approach to the framework proposed in this paper, let us dismiss the intercept $\boldsymbol{b}$, which means omitting the dual constraint $\sum_{i=1}^{n} \boldsymbol{\alpha}_i = 0$ in (P4), and let $\hat{\underline{\boldsymbol{\alpha}}} = (\hat{\boldsymbol{\alpha}}_1^\top, \dots, \hat{\boldsymbol{\alpha}}_n^\top)^\top$ be a solution to (P4). Then Fermat's rule gives: $\left(\boldsymbol{y} - \frac{1}{\lambda n} \underline{K} \hat{\underline{\boldsymbol{\alpha}}} - \hat{\underline{\boldsymbol{\alpha}}} \right) \in \partial(\epsilon \|\cdot\|_{\ell_1/\ell_2})(\hat{\underline{\boldsymbol{\alpha}}})$. It appears that both approaches induce data sparsity thanks to a structured norm and are equivalent up to a normalization by $\underline{K}$.

4.3. Quantile regression

A slight generalization of least absolute deviation consists in changing the slope of the absolute loss $\ell(\cdot) = |\cdot|$, asymmetrically around 0. For a quantile level $\tau \in (0, 1)$, the so

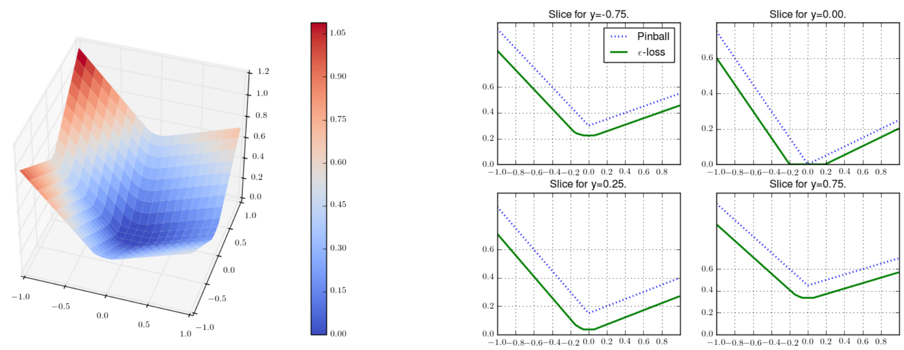

Figure 2: Multiple ϵ-pinball loss and slices of it ($\epsilon = 0.2$, $\tau = (0.25, 0.6)$).

called *pinball loss* is defined by:

$$\forall \xi \in \mathbb{R}, \quad \rho_\tau(\xi) = (\tau - \mathrm{I}_{\xi<0})\,\xi = \max\left(\tau\xi, (\tau-1)\xi\right).$$

Such a loss is used to estimate a conditional quantile of the output random variable, instead of a conditional median such as in standard regression (Koenker, 2005).

In order to improve prediction, some works propose to estimate and to predict simultaneously several quantiles, which is referred to as *joint* QR (Takeuchi et al., 2013; Sangnier et al., 2016). In this context, close to multivariate regression, the output space is $\mathbb{R}^p$ and the loss is $\ell(\boldsymbol{\xi}) = \sum_{j=1}^{p} \rho_{\tau_j}(\xi_j)$, where $\boldsymbol{\tau}$ is a vector of p quantile levels $\tau_j \in (0,1)$. Even though we do not have a closed-form expression for the corresponding ϵ-loss, Figure 2 provides graphical representations of ℓ_ϵ.

Besides the loss described above, the multiple QR framework comes with $\boldsymbol{y}_i \in \mathbb{R}^p$, $\boldsymbol{\alpha}_i \in \mathbb{R}^p$, $\boldsymbol{b} \in \mathbb{R}^p$ and $h\colon \mathcal{X} \to \mathbb{R}^p$, such that the j^{th} component of the prediction value, $f_j(\mathbf{x}) = h_j(\mathbf{x}) + b_j$, estimates the τ_j-conditional quantile of the output random variable.

In a manner very similar to least absolute deviation, the Fenchel-Legendre transformation of the loss of interest is $\ell^\star(\boldsymbol{\alpha}) = 0$ when $\boldsymbol{\tau} - \mathbb{1} \preccurlyeq \boldsymbol{\alpha} \preccurlyeq \boldsymbol{\tau}$ and $+\infty$ otherwise. Therefore, Problem (P3) becomes:

$$\begin{aligned}
&\underset{\forall i \in [n],\, \boldsymbol{\alpha}_i \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2\lambda n} \sum_{i,j=1}^{n} \langle \boldsymbol{\alpha}_i, K(\mathbf{x}_i, \mathbf{x}_j)\boldsymbol{\alpha}_j \rangle_{\ell_2} - \sum_{i=1}^{n} \langle \boldsymbol{\alpha}_i, \boldsymbol{y}_i \rangle_{\ell_2} + \epsilon \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_{\ell_2} \\
&\text{s.t.} \qquad \forall i \in [n], \boldsymbol{\tau} - \mathbb{1} \preccurlyeq \boldsymbol{\alpha}_i \preccurlyeq \boldsymbol{\tau}, \quad \sum_{i=1}^{n} \boldsymbol{\alpha}_i = 0,
\end{aligned} \tag{P5}$$

where the sparsity inducing term is the ℓ_1/ℓ_2-norm $\sum_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_{\ell_2}$ (Bach et al., 2012).

Very recently, Park and Kim (2011) also proposed an ϵ-insensitive loss for *single* QR (that is, in the unidimensional case, when estimating only a single τ-conditional quantile), based on a generalization of the one used for SVR. The proposition made by Park and Kim (2011) is:

$$\ell_\epsilon^{\mathrm{PK}}(\xi) = \max\left(0, \rho_\tau(\xi) - \epsilon\right) = \begin{cases} 0 & \text{if } \rho_\tau(\xi) \leq \epsilon \\ \rho_\tau(\xi) - \epsilon & \text{otherwise.} \end{cases}$$

That is different from the ϵ-loss proposed in this paper, as illustrated in Figure 1. As a comparison, the single QR dual in (Park and Kim, 2011) is similar to (P5) but with $\max\left(\frac{\alpha_i}{\tau}, \frac{\alpha_i}{\tau-1}\right)$ instead of $|\alpha_i|$. Let us remark that, because of the asymmetry, it is difficult to get a sparsity intuition for this modified ℓ_1-penalization.

The advantage of our approach over (Park and Kim, 2011) is to provide a unifying view of ϵ-losses in the multidimensional setting, which matches SVR as a special case. Moreover, the possibility to extend Park and Kim's loss to joint QR is unclear and could not guarantee data sparsity: as explained in Remark 1 (Supplementary A), this goal is achieved thanks to the ℓ_1/ℓ_2-norm in the dual, which is equivalent to our proposition of ϵ-loss.

4.4. Expectile regression

As for QR, it is worth considering weighting the squared loss asymmetrically around 0. The resulting framework is called *expectile regression* (Newey and Powell, 1987), by analogy with quantiles and conditional expectation estimation with the squared loss. Despite the fact that the expectile regression estimator has no statistical interpretation, it can be seen as a smooth (and less robust) version of a conditional quantile. This explains that this topic, at least in the unidimensional context, attracts a growing interest in the learning community (Farooq and Steinwart, 2015; Yang et al., 2015; Farooq and Steinwart, 2017).

Following the guiding principle of QR (Sangnier et al., 2016), we can imagine learning simultaneously several expectiles in a *joint expectile regression* framework. As far as we know, this multivariate setting did not appear in the literature yet. Therefore, we consider the multiple squared pinball loss:

$$\forall \xi \in \mathbb{R}, \quad \ell(\boldsymbol{\xi}) = \sum_{j=1}^{p} \psi_{\tau_j}(\xi_j), \quad \text{with } \psi_\tau(\xi) = \frac{1}{2}\left|\tau - \mathrm{I}_{\xi<0}\right| \xi^2.$$

Similarly to the squared loss, the Fenchel-Legendre transformation of ℓ is $\ell^\star \colon \boldsymbol{\alpha} \in \mathbb{R}^p \mapsto \frac{1}{2}\boldsymbol{\alpha}^\top \Delta(\boldsymbol{\alpha})\boldsymbol{\alpha}$, where $\Delta(\boldsymbol{\alpha})$ is the $p \times p$ diagonal matrix whose entries are $\left(\left|\tau_j - \mathrm{I}_{\boldsymbol{\alpha}_j<0}\right|^{-1}\right)_{1\le j\le p}$. Therefore, a dual optimization problem of sparse joint expectile regression is:

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

We see that the only difference compared to multivariate regression is the anisotropic ridge $\Delta(\boldsymbol{\alpha}_i)$. Moreover, compared to other kernelized approaches for expectiles estimation such as (Farooq and Steinwart, 2015; Yang et al., 2015), this setting goes beyond both in its multivariate feature as well as its data sparsity.

5. Statistical guarantees

In this section, we focus on the excess of risk for learning with the proposed ϵ-insensitive loss. As it is usual in statistical learning, we consider the constrained version of Problem (P2), which is legitimated by convexity of ℓ_ϵ (Tikhonov and Arsenin, 1977):

$$\underset{\substack{h \in \mathcal{H}\,:\,\|h\|_{\mathcal{H}} \leq c \\ \boldsymbol{b} \in \mathbb{R}^p}}{\text{minimize}} \; \frac{1}{n} \sum_{i=1}^{n} \ell_\epsilon \left(\boldsymbol{y}_i - (h(\mathbf{x}_i) + \boldsymbol{b}) \right), \tag{P6}$$

where c is a positive parameter linked to λ. This change of optimization problem is motivated by the fact that any solution to (P2) is also solution to (P6) for a well chosen parameter c.

Let (X, Y) be the couple of random variables of interest, following an unknown joint distribution. We assume being provided with n independent and identically distributed (*iid*) copies of (X, Y), denoted $(X_i, Y_i)_{1 \leq i \leq n}$. Let us now fix an intercept $\boldsymbol{b} \in \mathbb{R}^p$. $\mathcal{F} = \{h(\cdot) + \boldsymbol{b} : h \in \mathcal{H}, \|h\|_{\mathcal{H}} \leq c\}$ is a class of hypotheses. We denote $\mathcal{R}_n(\mathcal{F})$ the Rademacher average of the class $\mathcal{F}$ (Bartlett and Mendelson, 2002). Let $f^\dagger \in \arg\min_{f \in \mathcal{F}} \mathbb{E}\left[\ell\left(Y - f(X)\right)\right]$ be the target function, where the expectation is computed jointly on X and Y. For instance, in the context of single QR, if we assume that the conditional quantile lives in $\mathcal{F}$, then the target $f^\dagger$ can be this conditional quantile (up to uniqueness). Moreover, let $\hat{f}_\epsilon \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell_\epsilon \left(Y_i - f(X_i)\right)$ be the empirical estimator of $f^\dagger$.

Theorem 2 (Generalization) *Let us assume that the loss ℓ is L-Lipschitz ($L > 0$) and that residues are bounded: $\forall f \in \mathcal{F}, \|Y - f(X)\|_{\ell_2} \leq M$ almost surely (where $M > 0$). Then, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the random sample $(X_i, Y_i)_{1 \leq i \leq n}$:*

$$\mathbb{E}\left[\ell\left(Y - \hat{f}_\epsilon(X)\right)\right] - \mathbb{E}\left[\ell\left(Y - f^\dagger(X)\right)\right] \leq 2\sqrt{2} L \mathcal{R}_n(\mathcal{F}) + 2LM\sqrt{\frac{\log(2/\delta)}{2n}} + L\epsilon.$$

A proof of this theorem is given in Supplementary B. The bound is similar to what we usually observe (Boucheron et al., 2005; Mohri et al., 2012), but suffers from an extra term, which is linear in ϵ. The latter embodies the bias induced by soft-thresholding the original loss ℓ to induce data sparsity.

6. A primal-dual training algorithm

6.1. Description

For traditional kernel methods (real or vector-valued ones), pairwise coordinate descent as well as deterministic and randomized coordinate descents are popular and efficient training algorithms (Platt, 1999; Shalev-Shwartz and Zhang, 2013; Minh et al., 2016). However, few of current algorithms are able to handle multiple non-differentiable contributions in the objective value (such as the ones introduced by $\ell_\epsilon^\star$) and multiple linear constraints (coming from considering an intercept in our regressor, which is mandatory for QR (Takeuchi et al., 2006; Sangnier et al., 2016)). For these reasons, we propose to use a randomized primal-dual coordinate descent (PDCD) technique, introduced by Fercoq and Bianchi (2015) and utterly workable for the problem at hand. Moreover, PDCD has been proved favorably competitive with several state-of-the-art approaches (Fercoq and Bianchi, 2015).

The learning problem we are interested in is:

$$\underset{\forall i \in [n],\, \boldsymbol{\alpha}_i \in \mathbb{R}^p}{\text{minimize}} \quad \underbrace{s(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n) + \overbrace{\sum_{i=1}^{n} \ell^\star(\boldsymbol{\alpha}_i)}}_{\text{differentiable}} \underbrace{\quad + \epsilon \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_{\ell_2}}_{\text{not differentiable}} + \underbrace{\chi_{\sum_{i=1}^n \boldsymbol{\alpha}_i = 0}}_{\text{not differentiable}}, \tag{P7}$$

where $s(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n) = \frac{1}{2\lambda n} \sum_{i,j=1}^{n} \langle \boldsymbol{\alpha}_i, K(\mathbf{x}_i, \mathbf{x}_j)\boldsymbol{\alpha}_j \rangle_{\ell_2} - \frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{\alpha}_i, \boldsymbol{y}_i \rangle_{\ell_2}$ is a quadratic (differentiable) function. Depending on the kind of regression (see in Table 1), the mapping $\ell^\star$ may be differentiable or not. In any case, the objective function in (P7) can be written as the summation of three components: one differentiable and two non-differentiable. Given this three-term decomposition, PDCD (see Algorithm 1) dualizes the second non-differentiable component and deploys a randomized block coordinate descent, in which each iteration involves the proximal operators of the first non-differential function and of the Fenchel-Legendre transformation of the second one. We can see in Algorithm 1 that PDCD uses dual variables $\boldsymbol{\theta} \in (\mathbb{R}^p)^n$ (which are updated during the descent) and has two sets of parameters $\boldsymbol{\nu} \in \mathbb{R}^n$ and $\boldsymbol{\mu} \in \mathbb{R}^n$, that verify $\forall i \in [n]$: $\mu_i < \frac{1}{\lambda_{\max,i} + \nu_i}$, where $\lambda_{\max,i}$ is the largest eigenvalue of $K(\mathbf{x}_i, \mathbf{x}_i)$. In practice, we keep the same parameters as in (Fercoq and Bianchi, 2015): $\nu_i = 10\lambda_{\max,i}$ and μ_i equal to 0.95 times the bound.

Algorithm 1 Primal-Dual Coordinate Descent.

Initialize $\boldsymbol{\alpha}_i, \boldsymbol{\theta}_i \in \mathbb{R}^p$ $(\forall i \in [n])$.
repeat
  Choose $i \in [n]$ uniformly at random.
  Set $\overline{\boldsymbol{\theta}}^l \leftarrow \operatorname{proj}_{\mathbb{1}} \left(\boldsymbol{\theta}^l + \operatorname{diag}(\boldsymbol{\nu})\boldsymbol{\alpha}^l \right)$ for all $l \in [p]$.
  Set $\boldsymbol{d}_i \leftarrow \nabla_{\boldsymbol{\alpha}_i} s(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n) + 2\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i$.
  Set $\overline{\boldsymbol{\alpha}}_i \leftarrow \operatorname{prox}_{\mu_i(\epsilon\|\cdot\|_{\ell_2} + \ell^\star)}(\boldsymbol{\alpha}_i - \mu_i \boldsymbol{d}_i)$.
  Update coordinate i: $\boldsymbol{\alpha}_i \leftarrow \overline{\boldsymbol{\alpha}}_i$, $\boldsymbol{\theta}_i \leftarrow \overline{\boldsymbol{\theta}}_i$,
  and keep other coordinates unchanged.
until convergence

PDCD is a workable algorithm for non-differentiable $\ell^\star$ as soon as the proximal operator $\operatorname{prox}_{\epsilon\|\cdot\|_{\ell_2} + \ell^\star}$ can be computed (when $\ell^\star$ is differentiable, it can be moved to s and there is only $\operatorname{prox}_{\epsilon\|\cdot\|_{\ell_2}}$, which is easy to compute (Bach et al., 2012)). For the situations of interest in this paper, non-differentiable $\ell^\star$ appear for multiple QR and take the form of the characteristic function of box constraints (see Table 1). Nevertheless, as far as we know, computing the proximal operator of the sum of the ℓ_2-norm and box constraints has not been done yet. Therefore, the following section provides a way to compute it for box constraints intersecting both the negative and the positive orthants (proofs are in Supplementary B.4).

6.2. Proximal operator for multiple quantile regression

Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors from $\mathbb{R}^p$ with positive entries, and $\lambda > 0$. From now on, let us denote $[\cdot]_{-\boldsymbol{a}}^{\boldsymbol{b}} : \mathbb{R}^p \to \mathbb{R}^p$ the *clip operator*, defined by:

$$\forall \boldsymbol{y} \in \mathbb{R}^p, \forall j \in [p]: \quad \left([\boldsymbol{y}]_{-\boldsymbol{a}}^{\boldsymbol{b}}\right)_j = \begin{cases} y_j & \text{if } -a_j < y_j < b_j \\ b_j & \text{if } y_j \geq b_j \\ -a_j & \text{if } y_j \leq -a_j. \end{cases}$$

Lemma 3 *Let $\boldsymbol{y} \in \mathbb{R}^p$ such that $\|\boldsymbol{y}\|_{\ell_2} \geq \lambda$ be a fixed vector and $\mu \in [0, 1]$. The equation*

$$\left(1 + \frac{\lambda}{\left\|[\mu \boldsymbol{y}]_{-\boldsymbol{a}}^{\boldsymbol{b}}\right\|_{\ell_2}}\right)\mu = 1 \tag{1}$$

is well defined and has a solution $\mu \in [0, 1]$.

Proposition 4

$$\forall \boldsymbol{y} \in \mathbb{R}^p, \quad \operatorname{prox}_{\lambda\|\cdot\|_{\ell_2} + \chi_{-\boldsymbol{a} \preccurlyeq \cdot \preccurlyeq \boldsymbol{b}}}(\boldsymbol{y}) = [\mu \boldsymbol{y}]_{-\boldsymbol{a}}^{\boldsymbol{b}},$$

where $\mu = 0$ if $\|\boldsymbol{y}\|_{\ell_2} \leq \lambda$ and solution to Equation (1) otherwise.

Proposition 4 states that $\operatorname{prox}_{\lambda\|\cdot\|_{\ell_2} + \chi_{-\boldsymbol{a} \preccurlyeq \cdot \preccurlyeq \boldsymbol{b}}}$ can be computed as the composition of a clip and a scaling operator, for which the scaling factor can be easily obtained by a bisection of a Newton-Raphson method. As a straightforward consequence of Proposition 4, we can express the proximal operator of the sum of an ℓ_1/ℓ_2-norm and box constraints as the concatenation of $\operatorname{prox}_{\lambda\|\cdot\|_{\ell_2} + \chi_{-\boldsymbol{a} \preccurlyeq \cdot \preccurlyeq \boldsymbol{b}}}$ (see Supplementary B for a formal statement), which involves as many scaling factors as groups (n in this case).

6.3. Active set for multiple quantile regression

Learning sparse models gains a lot in identifying early the *active points* and optimizing only over them. This is a way to speed up learning, that is particularly topical (Ndiaye et al., 2015; Shibagaki et al., 2016; Ndiaye et al., 2016). *Active points* may have different meanings depending on the context (sparse regression, SVR, SVM) but the unifying concept is to detect optimizing variables for which the optimal value cannot be figured out beforehand.

In the context of QR (but this also holds true for SVR and SVM (Shibagaki et al., 2016)), *active points* correspond to every data point $\mathbf{x}_i$ for which the dual vector $\boldsymbol{\alpha}_i$ is neither null nor on the border of the box constraint. These active points can be identified thanks to optimality conditions.

Let $\hat{f} = \hat{h} + \hat{\boldsymbol{b}}$ be an optimal solution to (P2) and $(\hat{\boldsymbol{\alpha}}_i)_{1 \leq i \leq n}$ be optimal dual vectors to (P3). In the general setting, primal feasibility and stationarity from KKT conditions yield:

$$\forall i \in [n]: \quad \boldsymbol{y}_i - \hat{f}(\mathbf{x}_i) \in \partial\left(\ell_\epsilon^\star\right)(\hat{\boldsymbol{\alpha}}_i), \tag{2}$$

where, in the case of QR, $\ell_\epsilon^\star = \epsilon \left\|\cdot\right\|_{\ell_2} + \chi_{\boldsymbol{\tau}-\mathbb{1} \preccurlyeq \cdot \preccurlyeq \boldsymbol{\tau}}$. Therefore, $\forall \boldsymbol{\alpha} \in \mathbb{R}^p$:

$$\partial\left(\ell_\epsilon^\star\right)(\boldsymbol{\alpha}) = \begin{cases} \left\{\boldsymbol{d} : \boldsymbol{d} \in \mathbb{R}^p, \left\|\boldsymbol{d}\right\|_{\ell_2} \leq \epsilon\right\} & \text{if } \boldsymbol{\alpha} = 0 \\ \frac{\epsilon}{\left\|\boldsymbol{\alpha}\right\|_{\ell_2}} \boldsymbol{\alpha} & \text{if } \boldsymbol{\alpha} \neq 0, \boldsymbol{\tau} - \mathbb{1} \prec \boldsymbol{\alpha} \prec \boldsymbol{\tau} \\ \left\{\boldsymbol{d} \in \partial\left(\chi_{\boldsymbol{\tau}-\mathbb{1} \preccurlyeq \cdot \preccurlyeq \boldsymbol{\tau}}\right)(\boldsymbol{\alpha}) : \left\langle \boldsymbol{d}, \frac{\boldsymbol{\alpha}}{\left\|\boldsymbol{\alpha}\right\|_{\ell_2}} \right\rangle_{\ell_2} \geq \epsilon \right\} & \text{otherwise.} \end{cases}$$

Consequently, optimality condition in Equation (2) indicates that:

1. $\left\|\boldsymbol{y}_i - \hat{f}(\mathbf{x}_i)\right\|_{\ell_2} < \epsilon \implies \hat{\boldsymbol{\alpha}}_i = 0$;
2. $\left\|\boldsymbol{y}_i - \hat{f}(\mathbf{x}_i)\right\|_{\ell_2} > \epsilon \implies \forall j \in [p], (\hat{\boldsymbol{\alpha}}_i)_j = \tau_j$ or $(\hat{\boldsymbol{\alpha}}_i)_j = \tau_j - 1$.

Therefore, for each situation (1. or 2.), if both conditions are fulfilled for the current estimates f and $\boldsymbol{\alpha}_i$, the corresponding dual vector $\boldsymbol{\alpha}_i$ is put aside (at least temporally) and is not updated until optimality conditions are violated. In section 7.1, we will show that this strategy dramatically speed up the learning process when ϵ is large enough.

7. Numerical experiments

This section presents two numerical experiments in the context of multiple QR with data sparsity. The first experiment deals with training time. We compare an implementation of Algorithm 1 with an off-the-shelf solver and study the impact of the active set strategy. The second experiment analyses the effect of ϵ on quantile prediction and on the number of support vectors (data points $\mathbf{x}_i$ for which $\boldsymbol{\alpha}_i \neq 0$).

Following Sangnier et al. (2016), we use a matrix-valued kernel of the form $K(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}')\boldsymbol{B}$, where $\boldsymbol{B} = (\exp(-\gamma(\tau_i - \tau_j)^2))_{1 \leq i,j \leq p}$ and $k(\mathbf{x}, \mathbf{x}') = \exp(-\left\|\mathbf{x} - \mathbf{x}'\right\|_{\ell_2}^2 / 2\sigma^2)$. In the first experiment (simulated data), $\gamma = 1$ and $\sigma^2 = 1$, while in the second one (real data) γ is set to 0.1 and σ is the 0.7-quantile of the pairwise distances of the training data $\{\mathbf{x}_i\}_{1 \leq i \leq n}$. Quantile levels of interest are $\boldsymbol{\tau} = (0.25, 0.5, 0.75)$.

7.1. Training time

We aim at comparing the execution time of four approaches for solving (P5): i) Algorithm 1 with active set (*PDCD with AS*); ii) Algorithm 1 without active set (*PDCD without AS*); iii) an off-the-shelf solver based on an interior-point method for quadratic cone programming (*CVXOPT (CONEQP)*) and iv) quadratic programming (*CVXOPT (QP)*) (Anderson et al., 2012). For the last approach, we leverage a variational formulation of the sparsity constraint (Bach et al., 2012): $\sum_{i=1}^n \left\|\boldsymbol{\alpha}_i\right\|_{\ell_2} = \inf_{\forall i \in [n], \mu_i \in \mathbb{R}_+} \frac{1}{2} \sum_{i=1}^n \left(\frac{1}{\mu_i} \left\|\boldsymbol{\alpha}_i\right\|_{\ell_2}^2 + \mu_i\right)$, and alternate minimization with respect to $(\boldsymbol{\alpha}_i)_{1 \leq i \leq n}$ and $(\mu_i)_{1 \leq i \leq n}$. By convexity and differentiability of the objective, alternate minimization converges to an optimal solution (Rockafellar, 1970).

We fix $1/(\lambda n) = 10^2$ and we use a synthetic dataset for which $X \in [0, 1.5]$. The target Y is computed as a sine curve at 1 Hz modulated by a sine envelope at 1/3 Hz and mean 1. Moreover, this pattern is distorted with a random Gaussian noise with mean 0 and a linearly decreasing standard deviation from 1.2 at $X = 0$ to 0.2 at $X = 1.5$.

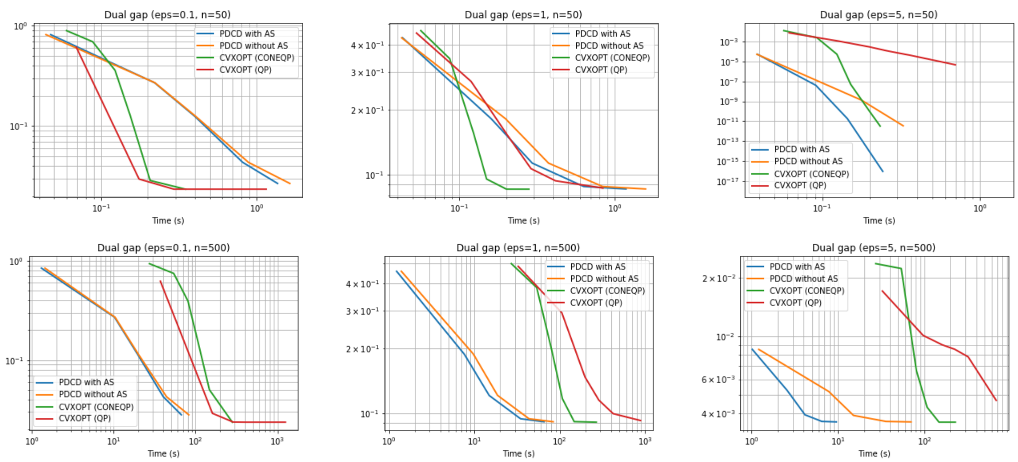

Figure 3: Training time (synthetic dataset).

Figure 3 depicts the dual gap achieved by the competitors with respect to the CPU time in different configurations ($\epsilon \in \{0.1, 1, 5\}$ and sample size $n \in \{50, 500\}$). These curves are obtained by increasing the maximal number of iterations of each solver and recording primal and dual objective values attained along with the CPU time. Given a configuration (ϵ and n fixed), all dual gaps are computed with respect to the lowest primal obtained. This comparison procedure is motivated by the fact that for multiple QR, we cannot compute the exact duality gap (we cannot compute the primal loss ℓ_ϵ), but only over-estimate it (see Remark 3 in Supplementary A). This explains why plots in Figure 3 do not necessarily converge to 0 (we conjecture that when a method faces no progress, the exact dual gap is almost null). Overall, the curves portray the behavior of the dual objective function rather than the exact duality gap.

For small-scale data and small values of ϵ, CVXOPT is more efficient than our approach, but its supremacy caves in when increasing ϵ and/or the sample size n. The impact of the problem size n on the comparison between interior-point methods (such as CVXOPT) and first-order ones (such as PDCD) is known for a long time, so this first finding is comforting. Also, increasing the sparsity regularizer ϵ requires more and more calls to the alternate procedure CVXOPT (QP), so the behavior with respect to ϵ was expected.

Analyzing the active-set strategy, we remark that it is never a drawback but becomes a real asset for hard sparsity constraints. This comes from many $\hat{\boldsymbol{\alpha}}_i$ to be null and thus put aside. Yet, it appears that, contrarily to the unidimensional case, only a few $\boldsymbol{\alpha}_i$ are on the border of the box constraint. Therefore, the active-set strategy provides only a limited (but real) gain compared to the vanilla version of PDCD for small values of ϵ.

7.2. Data sparsity

This section investigates the actual data sparsity introduced by the proposed ϵ-insensitive loss, along with its impact on the generalization error. For this purpose, we consider the datasets used in (Takeuchi et al., 2006; Zhang et al., 2016), coming from the UCI repository and three R packages: quantreg, alr3 and MASS. The sample sizes n vary from

Table 2: Empirical (test) pinball loss ×100 (percentage of support vectors in parentheses).

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 (100)	**67.40** (99)	**67.17** (95)	67.54 (87)	69.93 (65)	73.24 (49)	76.80 (39)	83.42 (24)	100.22 (17)	142.19 (6)
ftcollinssnow	**109.07** (100)	109.12 (100)	109.14 (100)	109.15 (100)	109.11 (100)	110.39 (98)	**109.05** (82)	110.90 (53)	109.81 (34)	113.50 (8)
highway	79.29 (100)	78.10 (97)	76.75 (97)	76.66 (91)	**75.09** (64)	**70.94** (51)	75.10 (37)	97.67 (31)	112.30 (0)	112.09 (1)
heights	91.05 (100)	91.00 (100)	90.98 (100)	**90.98** (100)	91.18 (99)	91.21 (86)	**90.98** (67)	91.09 (40)	91.51 (23)	93.34 (5)
sniffer	32.34 (100)	**31.40** (99)	32.31 (97)	**31.40** (84)	34.64 (43)	39.84 (24)	41.82 (12)	52.06 (6)	62.21 (4)	103.76 (3)
snowgeese	**49.62** (100)	**50.51** (98)	51.25 (87)	51.08 (70)	52.88 (37)	53.81 (23)	62.81 (17)	90.15 (15)	107.53 (14)	94.25 (1)
ufc	57.87 (100)	57.90 (100)	**57.78** (100)	57.84 (100)	**57.67** (78)	57.84 (52)	58.19 (35)	61.04 (15)	66.81 (5)	86.23 (2)
birthwt	99.93 (100)	99.95 (100)	99.93 (100)	99.70 (100)	**99.25** (100)	100.50 (87)	99.80 (67)	**98.71** (43)	99.56 (24)	103.39 (6)
crabs	8.59 (100)	**8.52** (91)	**8.49** (55)	9.44 (16)	19.94 (6)	23.08 (2)	31.44 (3)	44.08 (2)	53.45 (1)	86.91 (2)
GAGurine	44.30 (100)	**44.26** (99)	**44.25** (99)	44.86 (87)	46.20 (50)	49.87 (33)	52.88 (22)	57.06 (12)	65.89 (6)	103.32 (2)
geyser	**77.81** (100)	78.15 (100)	**78.12** (100)	78.45 (100)	78.40 (92)	78.28 (78)	78.54 (59)	80.55 (32)	85.15 (16)	99.92 (1)
gilgais	**32.96** (100)	**33.12** (99)	33.27 (96)	33.42 (81)	35.08 (43)	36.62 (25)	37.94 (14)	48.17 (5)	94.65 (7)	104.12 (0)
topo	47.49 (100)	48.93 (100)	48.74 (98)	48.17 (94)	**41.65** (57)	**45.24** (38)	51.19 (26)	53.68 (16)	58.21 (12)	80.57 (6)
BostonHousing	**34.54** (100)	34.68 (99)	34.70 (97)	**34.09** (80)	35.27 (35)	37.65 (20)	41.31 (13)	55.04 (7)	73.39 (7)	112.22 (12)
CobarOre	**0.50** (100)	**5.05** (38)	8.75 (36)	12.47 (26)	23.84 (18)	35.82 (17)	47.35 (15)	66.15 (14)	84.51 (12)	106.89 (6)
engel	43.57 (100)	43.50 (100)	**43.47** (99)	**43.44** (89)	57.36 (39)	43.98 (19)	46.31 (11)	53.15 (4)	69.43 (5)	100.48 (0)
mcycle	**63.95** (100)	**63.88** (99)	64.26 (99)	64.90 (98)	65.89 (88)	67.29 (70)	70.11 (51)	74.78 (26)	86.49 (14)	109.79 (2)
BigMac2003	**49.94** (100)	**49.97** (98)	50.00 (96)	50.27 (85)	51.16 (56)	51.44 (36)	53.63 (28)	77.40 (18)	106.38 (14)	136.76 (4)
UN3	71.27 (100)	**70.94** (100)	**71.03** (100)	71.49 (99)	71.37 (87)	71.53 (65)	72.68 (50)	76.72 (27)	84.50 (13)	109.59 (0)
cpus	**11.31** (100)	**13.32** (28)	15.57 (21)	20.16 (15)	25.88 (8)	35.66 (6)	55.27 (6)	65.05 (0)	65.05 (0)	65.02 (0)

38 (CobarOre) to 1375 (heights) and the numbers of explanatory variables vary from 1 (6 sets) to 12 (BostonHousing). The datasets are standardized coordinate-wise to have zero mean and unit variance. In addition, the generalization error is estimated by the mean over 10 trials of the empirical loss $\frac{1}{m}\sum_{i=1}^{m} \ell(\boldsymbol{y}_i^{\text{test}} - f(\mathbf{x}_i^{\text{test}}))$ computed on a test set. For each trial, the whole dataset is randomly split in a train and a test set with ratio 0.7-0.3. The parameter $1/(\lambda n)$ is chosen by a 3-fold cross-validation (minimizing the pinball loss) on a logarithmic scale between 10^{-3} and 10^{3} (10 values). At last, a training point $\mathbf{x}_i$ is considered a support vector if $\|\boldsymbol{\alpha}_i\|_{\ell_2}/p > 10^{-3}$.

Table 2 reports the average empirical test loss (scaled by 100) along with the percentage of support vectors (standard deviations appear in Supplementary C). For each dataset, the bold-face numbers are the two lowest losses, to be compared to the loss for $\epsilon = 0$. We observe first that increasing ϵ does promote data sparsity. Second, the empirical loss tends to increase with ϵ, as expected, but we do not lose much by requiring sparsity (we can even gain a little).

8. Conclusion

This paper introduces a novel and unifying framework concerning ϵ-insensitive losses, which offers a variety of opportunities for traditional, quantile and expectile regression, for uni and multivariate outcomes. Estimators are based on vector-valued RKHSs (Micchelli and Pontil, 2005) and benefit from theoretical guarantees concerning generalization as well as an efficient learning algorithm. Future directions of research include improving further the learning procedure and deriving sharper generalization bounds.

References

M.S. Anderson, J. Dahl, and L. Vandenberghe. CVXOPT: A Python package for convex optimization, version 1.1.5., 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.

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.

H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

B.E. Boser, I.M. Guyon, and V.N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Conference on Learning Theory*, 1992.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: a Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

C. Brouard, M. Szafranski, and F. d'Alché Buc. Input Output Kernel Regression: Supervised and Semi-Supervised Structured Output Prediction with Operator-Valued Kernels. *Journal of Machine Learning Research*, 17(176):1–48, 2016.

H. Drucker, C.J.C. Burges, L. Kaufman, A.J. Smola, and V. Vapnik. Support Vector Regression Machines. In *Advances in Neural Information Processing Systems 9*, 1997.

M. Farooq and I. Steinwart. An SVM-like Approach for Expectile Regression. *arXiv:1507.03887 [stat]*, 2015.

M. Farooq and I. Steinwart. Learning Rates for Kernel-Based Expectile Regression. *arXiv:1702.07552 [cs, stat]*, 2017.

O. Fercoq and P. Bianchi. A Coordinate Descent Primal-Dual Algorithm with Large Step Size and Possibly Non Separable Functions. *arXiv:1508.04625 [math]*, 2015.

T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Taylor & Francis, 2015.

R. Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, New York, 2005.

N. Lim, F. d'Alché Buc, C. Auliac, and G. Michailidis. Operator-valued kernel-based vector autoregressive models for network inference. *Machine Learning*, 99(3):489–513, 2014.

C.A. Micchelli and M.A. Pontil. On Learning Vector-Valued Functions. *Neural Computation*, 17:177–204, 2005.

H.Q. Minh, L. Bazzani, and V. Murino. A Unifying Framework in Vector-valued Reproducing Kernel Hilbert Spaces for Manifold Regularization and Co-Regularized Multi-view Learning. *Journal of Machine Learning Research*, 17(25):1–72, 2016.

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

E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. GAP Safe screening rules for sparse multi-task and multi-class models. In *Advances in Neural Information Processing Systems 28*, 2015.

E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. GAP Safe Screening Rules for Sparse-Group Lasso. In *Advances in Neural Information Processing Systems 29*, 2016.

W.K. Newey and J.L. Powell. Asymmetric Least Squares Estimation and Testing. *Econometrica*, 55(4):819–847, 1987.

J. Park and J. Kim. Quantile regression with an epsilon-insensitive loss in a reproducing kernel Hilbert space. *Statistics & Probability Letters*, 81(1):62–70, 2011.

J.C. Platt. *Fast training of support vector machines using sequential minimal optimization*. Advances in Kernel Methods. MIT Press, Cambridge, MA, USA, 1999.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

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.

B. Schölkopf, R. Herbrich, and A.J. Smola. A Generalized Representer Theorem. In *Computational Learning Theory*, pages 416–426. Springer, Berlin, Heidelberg, 2001.

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

A. Shibagaki, M. Karasuyama, K. Hatano, and I. Takeuchi. Simultaneous Safe Screening of Features and Samples in Doubly Sparse Modeling. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 2016.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, NY, 2008.

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

I. Takeuchi, T. Hongo, M. Sugiyama, and S. Nakajima. Parametric Task Learning. In *Advances in Neural Information Processing Systems 26*, pages 1358–1366, 2013.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

A.N. Tikhonov and V.Y. Arsenin. *Solutions of ill-posed problems*. Winston, Washington, DC, 1977.

M.E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

V. Vapnik. *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer, 2010.

Y. Yang, T. Zhang, and H. Zou. Flexible Expectile Regression in Reproducing Kernel Hilbert Space. *arXiv:1508.05987 [stat]*, 2015.

C. Zhang, Y. Liu, and Y. Wu. On Quantile Regression in Reproducing Kernel Hilbert Spaces with the Data Sparsity Constraint. *Journal of Machine Learning Research*, 17(1): 1374–1418, 2016.

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:*

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

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

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

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:

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

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:

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

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: $\ell^\star \colon \boldsymbol{\alpha} \in \mathbb{R}^p \mapsto \sup_{\boldsymbol{\xi}\in\mathbb{R}^p} \langle \boldsymbol{\alpha}, \boldsymbol{\xi}\rangle_{\ell_2} - \ell(\boldsymbol{\xi})$. Thus, it remains to compute:

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

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:

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

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 $\inf_{\forall i\in[n],\, \mu_i\in\mathbb{R}_+} \frac{1}{2}\sum_{i=1}^n \left(\frac{1}{\mu_i}\|\boldsymbol{\alpha}_i\|_{\ell_2}^2 + \mu_i\right) =$

$\sum_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_{\ell_2}$ (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.