Functional Output Regression with Infimal Convolution: Exploring the Huber and -insensitive Losses

Alex Lambert, Dimitri Bouche, Zoltan Szabo, Florence d'Alché-Buc

Functional Output Regression with Infimal Convolution: Exploring the Huber and ϵ-insensitive Losses

Alex Lambert [1 2] **Dimitri Bouche** [1] **Zoltán Szabó** [3] **Florence d'Alché-Buc** [1]

Abstract

The focus of the paper is functional output regression (FOR) with convoluted losses. While most existing work consider the square loss setting, we leverage extensions of the Huber and the ϵ-insensitive loss (induced by infimal convolution) and propose a flexible framework capable of handling various forms of outliers and sparsity in the FOR family. We derive computationally tractable algorithms relying on duality to tackle the resulting tasks in the context of vector-valued reproducing kernel Hilbert spaces. The efficiency of the approach is demonstrated and contrasted with the classical squared loss setting on both synthetic and real-world benchmarks.

1. Introduction

Functional data analysis (FDA, Ramsay & Silverman 1997; Wang et al. 2016) has attracted a growing attention in the field of machine learning and statistics, with applications for instance in biomedical signal processing (Ullah & Finch, 2013), epidemiology monitoring and climate science (Ramsay & Silverman, 2007). The key assumption is that we have access to densely-measured observations, in which case functional data description becomes the most natural and adequate. An important subfield of FDA is functional output regression (FOR) which focuses on regression problems where the output variable is a function. There are numerous ways to tackle the FOR problem family. The simplest approach is to assume linear dependence between the inputs and the outputs (Morris, 2015). However, in order to cope with more complex dependencies, various nonlinear approaches have been designed. In nonparametric statistics, Ferraty et al. (2011) proposed a Banach-valued Nadaraya-Watson estimator. The flexibility of kernel methods (Steinwart & Christmann, 2008) and the richness of the associated reproducing kernel Hilbert spaces (RKHSs; Micchelli et al. 2006) have proven to be particularly useful in the area, with works involving tri-variate regression problem (Reimherr et al., 2018), and approximated kernel ridge regression (KRR) using orthonormal bases (Oliva et al., 2015). In the operator-valued kernel (Pedrick, 1957; Carmeli et al., 2010) literature, examples include function-valued KRR with double representer theorem (Lian, 2007), solvers based on the discretization of the loss function (Kadri et al., 2010), purely functional methods relying on approximate inversion of integral operators (Kadri et al., 2016) or techniques relying on finite-dimensional coefficients of the functional outputs in a dictionary basis (Bouche et al., 2021).

Most of these works employ the square loss which induces an estimate of the conditional expectation of the functional outputs given the input data. However defective sensors or malicious attacks can lead to erroneous or contaminated measurements (Hubert et al., 2015), resulting in local or global functional outliers. The square loss is expected to be badly affected in those cases and considering alternative losses is a natural way to obtain reliable and robust prediction systems. For scalar-valued outputs, the Huber loss (Huber, 1964) and the ϵ-insensitive loss (Lee et al., 2005) are particularly popular and well-suited to construct outlier-robust estimators. In the FDA setting, robustness has been investigated using Bayesian methods (Zhu et al., 2011), trading the mean for the median (Cadre, 2001), using bounded loss functions (Maronna & Yohai, 2013), or leveraging principal component analysis (Kalogridis & Van Aelst, 2019).

In the operator-valued kernel literature, ϵ-insensitive losses for vector-valued regression have been proposed by Sangnier et al. (2017) for finite-dimensional outputs. The use of convex optimization tools such as the infimal convolution operator and parametric duality leads to efficient solvers and provides sparse estimators. This idea is exploited by Laforgue et al. (2020) where a generalization of this approach to infinite-dimensional outputs encompassing both the Huber and ϵ-insensitive losses is developed.

In this paper, we extend the families of losses considered by Laforgue et al. (2020) by leveraging specific p-norms in

[1]LTCI, Telecom Paris, IP Paris, France [2]ESAT, KU Leuven, Belgium [3]Department of Statistics, London School of Economics, United Kingdom. Correspondence to: Alex Lambert <alex.lambert@kuleuven.be>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

functional spaces to handle various forms of outliers (with Huber loss) and sparsity (with ϵ-insensitive losses). We study the properties of their Fenchel-Legendre conjugates, and derive the associated dual optimization problems, which require suitable representations and approximations adapted to each situation to be manageable computationally. We propose tractable optimization algorithms for $p \in \{1, 2\}$ in the Huber loss scenario, and for $p \in \{2, +\infty\}$ with the ϵ-insensitive family. Finally, we provide an empirical study of the proposed algorithms over synthetic and real functional datasets.

The paper is structured as follows. After introducing the general problem in Section 2, we focus in Section 3 on a generalized family of Huber losses and propose loss-specific tractable optimization schemes, before turning to the family of ϵ-insensitive losses in Section 4. We illustrate the benefits of the approach on several benchmarks in Section 5. Proofs are deferred to the supplement.

2. Problem Formulation

In this section, we introduce the general setting of FOR in the context of vv-RKHSs, chosen for their modeling flexibility. To benefit from duality principles, we focus on losses that can be expressed as infimal convolutions in the functional output space.

Notations: Let $\mathcal{X}$ be an input set, $\Theta \subset \mathbb{R}$ a compact set endowed with a Borel probability measure μ, $\mathcal{Y} := L^2[\Theta, \mu]$ the space of square μ-integrable real-valued functions. For $p \in [1, +\infty[$ and $f \in \mathcal{Y}$, let $\|f\|_p = \left[\int_\Theta |f(\theta)|^p \mathrm{d}\mu(\theta)\right]^{\frac{1}{p}} \in [0, +\infty]$; $\|\cdot\|_\infty$ refers to the essential supremum. In both cases, the norm is allowed to take infinite value ($\|\cdot\|_p : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$).[1] Two numbers p and $q \in [1, +\infty]$ are said to be conjugate exponents if $\frac{1}{p} + \frac{1}{q} = 1$, with the classical $0 = \frac{1}{\infty}$ convention. The ball in $\mathcal{Y}$ of radius $\epsilon > 0$ and center 0 w.r.t. $\|\cdot\|_p$ is denoted by $\mathcal{B}_\epsilon^p$. The space of bounded linear operators over $\mathcal{Y}$ is $\mathcal{L}(\mathcal{Y})$. An operator-valued kernel (OVK) $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is a mapping such that $\sum_{i=1}^n \sum_{j=1}^n \langle K(x_i, x_j) y_i, y_j \rangle_\mathcal{Y} \geq 0$ for all $(x_i, y_i)_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ and positive integer n. An OVK K gives rise to a space of functions from $\mathcal{X}$ to $\mathcal{Y}$ called vector-valued RKHS (vv-RKHS); it is defined as $\mathcal{H}_K := \overline{\mathrm{Span}}\{K(\cdot, x)y : (x, y) \in \mathcal{X} \times \mathcal{Y}\}$, where $\mathrm{Span}(\cdot)$ denotes the linear hull of its argument, $\bar{\cdot}$ stands for closure, and $K(\cdot, x)y$ is the function $x' \in \mathcal{X} \mapsto K(x', x)y \in \mathcal{Y}$ while keeping $x \in \mathcal{X}$ fixed. The *Fenchel-Legendre conjugate* of a function $f : \mathcal{Y} \to [-\infty, +\infty]$ is defined as $f^\star(z) := \sup_{y \in \mathcal{Y}} \langle z, y \rangle_\mathcal{Y} - f(y)$ where $z \in \mathcal{Y}$. Given a convex set $\mathcal{C} \subset \mathcal{Y}$, $\iota_\mathcal{C}(\cdot)$ is its indicator function ($\iota_\mathcal{C}(x) = 0$ if $x \in \mathcal{C}$, and $\iota_\mathcal{C}(x) = +\infty$ otherwise), and $\mathrm{Proj}_\mathcal{C}(\cdot)$ is the orthogonal projection on $\mathcal{C}$ when $\mathcal{C}$ is also closed. Given two functions $f, g : \mathcal{Y} \to]-\infty, +\infty]$, their *infimal convolution* is defined as $f \square g\ (y) = \inf_{y' \in \mathcal{Y}} f(y - y') + g(y')$ for $y \in \mathcal{Y}$ and the *proximal operator* of f (when f is convex, lower semi-continuous) is $\mathrm{prox}_f(y) = \arg\min_{y' \in \mathcal{Y}} \frac{1}{2} \|y - y'\|_\mathcal{Y}^2 + f(y')$ for all $y \in \mathcal{Y}$. For a positive integer n, let $[n] = \{1, \ldots, n\}$. For $p \in [1, \infty[$, the p-norm of a vector $\mathbf{v} \in \mathbb{R}^m$ is denoted by $\|\mathbf{v}\|_p = \left(\sum_{j=1}^m |v_j|^p\right)^{\frac{1}{p}}$, and $\|\mathbf{v}\|_\infty = \max_{j \in [m]} |v_j|$. Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $p, s \in [1, +\infty]$, $\|\mathbf{A}\|_{p,s}$ is the s-norm of the p-norms of the rows of $\mathbf{A}$. The positive part of $x \in \mathbb{R}$ is denoted by $|x|_+ = \max(x, 0)$.

Next we introduce the FOR problem in vv-RKHSs. Recall that $\mathcal{X}$ is a set and $\mathcal{Y} = L^2[\Theta, \mu]$, the latter capturing the functional outputs. Assume that we have i.i.d. samples $(x_i, y_i)_{i \in [n]}$ from a random variable $(\mathsf{X}, \mathsf{Y}) \in \mathcal{X} \times \mathcal{Y}$. Given a proper, convex lower-semicontinuous loss function $L : \mathcal{Y} \to \mathbb{R}$ and a regularization parameter $\lambda > 0$, we consider the *regularized empirical risk minimization* problem

$$\inf_{h \in \mathcal{H}_K} \frac{1}{n} \sum_{i \in [n]} L(y_i - h(x_i)) + \frac{\lambda}{2} \|h\|_{\mathcal{H}_K}^2 , \qquad (1)$$

where $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is a *decomposable* OVK of the form $K = k_\mathcal{X} T_{k_\Theta}$. Here $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_\Theta : \Theta \times \Theta \to \mathbb{R}$ are continuous real-valued kernels, and $T_{k_\Theta} \in \mathcal{L}(\mathcal{Y})$ is the integral operator associated to k_Θ, defined for all $f \in \mathcal{Y}$ as $(T_{k_\Theta} f)(\theta) := \int_\Theta k(\theta, \theta') f(\theta) \mathrm{d}\mu(\theta')$ where $\theta \in \Theta$. Similarly to the scalar case (Wahba, 1990), the minimizer of Problem 1 enjoys a representer theorem (Micchelli & Pontil, 2005) and writes as

$$\hat{h} = \frac{1}{\lambda n} \sum_{i \in [n]} k_\mathcal{X}(\cdot, x_i) T_{k_\Theta} \hat{\alpha}_i$$

for some coefficients $\{\hat{\alpha}_i\}_{i \in [n]} \subset \mathcal{Y}$. However, the functional nature of these parameters renders the problem extremely challenging, with quite few existing solutions. Particularly, even in the case of the square loss

$$L(f) = \frac{1}{2} \|f\|_\mathcal{Y}^2 = \frac{1}{2} \int_\Theta f(\theta)^2 \mathrm{d}\mu(\theta),$$

the value of $\{\hat{\alpha}_i\}_{i \in [n]}$ can not be computed in closed form, and some level of approximation is required. For instance in Lian (2007), L is approximated as a discrete sum, allowing for a double application of the representer theorem and yielding tractable models. In Kadri et al. (2016), the integral operator is traded for a finite rank approximation based on its eigendecomposition, providing a computable closed-form expression for the coefficients. Aiming at robustness, Laforgue et al. (2020) propose a Huber loss based on infimal

[1] This assumption is natural in convex optimization which is designed to handle functions taking infinite values.

convolution, yet limited to a narrow choice of kernels k_Θ. Moreover, the lack of flexibility in the definition of the loss prevents the resulting estimators from being robust to a large variety of outliers.

The **goal** of this work is (i) to widen the scope of the FOR problem (Problem 1) by considering losses capable of handling different forms of outliers and sparsity, and (ii) to design efficient optimization schemes for the resulting tasks. The proposed two loss families are based on infimal convolution and can be written in the form

$$\frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \square\, g \tag{2}$$

for appropriately chosen functions $g\colon \mathcal{Y} \to]-\infty, +\infty]$. The key property which allows one to handle these convoluted losses from an optimization perspective is the fact that

$$\left(\frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \square\, g\right)^\star = \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 + g^\star,$$

as it makes the associated Problem 1 amenable for dual approaches. The losses with the proposed dedicated optimization schemes are detailed in Section 3 and Section 4, respectively.

The starting point for working with convoluted losses in vv-RKHSs is the following lemma.

Lemma 2.1 (Dualization for convoluted losses; Laforgue et al. 2020)**.** *Let L be a loss function defined as $L = \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \square\, g$ for some $g\colon \mathcal{Y} \to]-\infty, +\infty]$. The solution of Problem 1 is given by*

$$\hat{h} = \frac{1}{\lambda n}\sum_{i\in[n]} k_{\mathcal{X}}(\cdot, x_i) T_{k_\Theta}\hat{\alpha}_i, \tag{3}$$

with $(\hat{\alpha}_i)_{i\in[n]} \in \mathcal{Y}^n$ being the solution of the dual task

$$\begin{aligned}\inf_{(\alpha_i)_{i\in[n]}\in\mathcal{Y}^n} &\sum_{i\in[n]}\left[\frac{1}{2}\|\alpha_i\|_{\mathcal{Y}}^2 - \langle\alpha_i, y_i\rangle_{\mathcal{Y}} + g^\star(\alpha_i)\right] \\ &+ \frac{1}{2\lambda n}\sum_{i,j\in[n]} k_{\mathcal{X}}(x_i, x_j)\,\langle\alpha_i, T_{k_\Theta}\alpha_j\rangle_{\mathcal{Y}}\,.\end{aligned} \tag{4}$$

Solving Problem 4 in the general case for various g and k_Θ raises multiple **challenges** which have to be handled simultaneously. Problem 4 is often referred to as a *composite* optimization problem, with a differentiable term consisting of a quadratic part added to a non-differentiable term induced by $g^\star$. The first challenge is to be able to compute the proximal operator associated to $g^\star$. The second and third difficulties arise from the fact that the dual variables are functions ($\alpha_i \in \mathcal{Y}$) and hence managing them computationally requires specific care. Particularly, evaluation of $\langle\alpha_i, T_{k_\Theta}\alpha_j\rangle_{\mathcal{Y}}$ can be non-trivial. In addition, one has to design a finite-dimensional description of the dual variables that is compatible with T_{k_Θ} and the proximal operator of $g^\star$. The primary focus and technical contribution of the paper is to handle these challenges, after which proximal gradient descent optimization can be applied. We detail our proposed solution in the next section.

3. Learning with $\mathcal{Y}$-Huber Losses

In this section, we propose a generalized Huber loss on $\mathcal{Y}$ based on infimal convolution, followed by an efficient dual optimization approach to solve the corresponding Problem 1. This loss (as illustrated in Section 5) shows robustness against different kind of outliers. Our proposed loss on $\mathcal{Y}$ relies on functional p-norms where $p \in [1, +\infty]$.

Definition 3.1 ($\mathcal{Y}$-Huber loss)**.** Let $\kappa > 0$ and $p \in [1, +\infty]$. We define the Huber loss with parameters (κ, p) as

$$H_\kappa^p := \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \square\, \kappa\|\cdot\|_p\,.$$

Notice that in the specific case of $\mathcal{Y} = \mathbb{R}$, H_κ^p reduces to the classical Huber loss on the real line for arbitrary p. Our following result describes the behavior of H_κ^p.

Proposition 3.2. *Let $\kappa > 0$, $p \in [1, +\infty]$, and q the conjugate exponent of p. Then for all $f \in \mathcal{Y}$,*

$$H_\kappa^p(f) = \frac{1}{2}\left\|\mathrm{Proj}_{\mathcal{B}_\kappa^q}(f)\right\|_{\mathcal{Y}}^2 + \kappa\left\|f - \mathrm{Proj}_{\mathcal{B}_\kappa^q}(f)\right\|_p.$$

Remark: For general p, the value of $H_\kappa^p(f)$ can not be computed straighforwardly due to the complexity of the projection on $\mathcal{B}_\kappa^q$. As we show however using a dual approach Problem 1 is still computationally manageable. For $p = 2$, one gets back the loss investigated by Laforgue et al. (2020).

The following proposition is a key result of this work that allows to leverage p-norms as suitable candidates for g in (2). It extends to $\mathcal{Y}$ the well-known finite-dimensional case that can *e.g.* be found in Bauschke et al. (2011).

Proposition 3.3. *Let $p, q \in [1, +\infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\|\cdot\|_p^\star = \iota_{\mathcal{B}_1^q}(\cdot).$$

Our next result provides the dual of Problem 1, and shows the impact of the parameters (p, κ).

Proposition 3.4 (Dual Huber)**.** *Let $\kappa > 0$, $p \in [1, +\infty]$, and $\frac{1}{p} + \frac{1}{q} = 1$. The dual of Problem 1 with loss H_κ^p writes*

$$\begin{aligned}\inf_{(\alpha_i)_{i=1}^n\in\mathcal{Y}^n} &\sum_{i=1}^n\left[\frac{1}{2}\|\alpha_i\|_{\mathcal{Y}}^2 - \langle\alpha_i, y_i\rangle_{\mathcal{Y}}\right] \\ &+ \frac{1}{2\lambda n}\sum_{i=1}^n\sum_{j=1}^n k_{\mathcal{X}}(x_i, x_j)\,\langle\alpha_i, T_{k_\Theta}\alpha_j\rangle_{\mathcal{Y}} \\ &\text{s.t } \|\alpha_i\|_q \le \kappa, \quad i \in [n].\end{aligned} \tag{5}$$

Remarks:

- Influence of κ and p: The difference between using the square loss and the Huber loss H_κ^p lies in the constraint on the q-norm of the dual variables. The parameter p influences the *shape* of the ball via the dual exponent q ($\|\alpha_i\|_q$) defining the admissible region for dual variables, and κ determines its *size* ($\|\alpha_i\|_q \leq \kappa$). As κ grows, the constraint becomes void and we recover the solution of the classical ridge regression problem. In Appendix C, we explore how different choices of p can affect the sensitivity of the loss to two different types of outliers.
- Partially observed data: The observed data $(y_i)_{i\in[n]}$ enter into Problem 5 only via their scalar product with the dual variables $(\alpha_i)_{i\in[n]}$. However in real life scenarii one never fully observe the y_i functions and these inner products are to be estimated. One can instead assume access to a sampling at some locations $(\theta_j)_{j=1}^m$ which can be used to approximate the inner products (see Section 3.1).

Let us now recall the **challenges** to be tackled to solve Problem 5. Firstly, as $\mathcal{Y}$ is infinite-dimensional no finite parameterization of the dual variables can be assumed a priori. Secondly, even computing the different terms of the objective function is non-trivial. Indeed, computing the quadratic term corresponding to the regularization is not straightforward as it involves the terms $\langle \alpha_i, T_{k_\Theta}\alpha_j\rangle_\mathcal{Y}$, which require the knowledge of the action of the integral operator T_{k_Θ}. The third difficulty comes from handling the constraints. Gradient-based optimization algorithms will require the projection of the dual variables on the feasible set $\mathcal{B}_\kappa^q \subset \mathcal{Y}$, which can be intractable to evaluate for some choices of q depending on the chosen representation.

The next proposition ensures the tractability of the projection step for specific choices of q.

Proposition 3.5 (Projection on $\mathcal{B}_\kappa^q$)**.** *Let $\kappa > 0$. The projection on $\mathcal{B}_\kappa^q$ is tractable for $q = 2$ and $q = \infty$ and can be expressed for all $(\alpha, \theta) \in \mathcal{Y} \times \Theta$ as*

$$\mathrm{Proj}_{\mathcal{B}_\kappa^2}(\alpha) = \min\left(1, \frac{\kappa}{\|\alpha\|_2}\right)\alpha, \tag{6}$$

$$\left(\mathrm{Proj}_{\mathcal{B}_\kappa^\infty}(\alpha)\right)(\theta) = \mathrm{sign}\left(\alpha(\theta)\right)\min\left(\kappa, |\alpha(\theta)|\right). \tag{7}$$

The projection operator for $p = q = 2$ in Equation (6) simply consists of a multiplication by a scalar involving the 2-norm of the dual variable. In the $p = 1$ case (i.e., $q = \infty$), the projection Equation (7) involves a *pointwise* projection. A suitable representation must guarantee the feasibility of this projection, which requires a pointwise control over the dual variables.

In order to solve Problem 5, we propose to use two different representations. In Section 3.1, we advocate representing the dual variables by linear splines and approximating the action of T_{k_Θ} by Monte-Carlo (MC) sampling. Splines allow pointwise control of the dual variable which is well-suited to both $p = 1$ and $p = 2$. Our alternative approach (elaborated in Section 3.2) relies on a finite-rank approximation of T_{k_Θ} using its eigendecomposition. This method is applicable when $p = 2$ with the complementary advantage of performing dimensionality reduction.

3.1. The Linear Spline Based Approach

In this section we introduce a linear spline based representation for the dual variables to tackle the challenges outlined.

Linear splines represent an easy-to-handle function class which provides pointwise control over the dual variables as they are encoded by their evaluations at some knots. While the class lacks smoothness, the dual variables are smoothed out by T_{k_Θ} in the estimator expression from Equation (3). Indeed, given that the kernel k_Θ is $2s$-times continuously differentiable, the RKHS $\mathcal{H}_{k_\Theta}$ (where T_{k_Θ} maps) consists of s-times continuously differentiable functions (Zhou, 2008), a desirable property in many settings making linear splines good candidates for modeling the dual variables.

A linear spline is a piecewise linear curve which can be encoded by a set of ordered locations or anchor points $(\theta_j)_{j\in[m]} \in \Theta^m$, and by a vector of size m corresponding to the evaluation of the spline at these points. We choose the anchors to be distributed i.i.d. according to μ; in practice, we often take the locations to be those of the available sampling of the observed data $(y_i)_{i\in[n]} \in \mathcal{Y}^n$.

Fixing the anchors, the n dual variables are encoded by the matrix of evaluations $\mathbf{A} = [\mathbf{a}_i]_{i\in[n]} = [\alpha_i(\theta_j)]_{i\in[n],j\in[m]} \in \mathbb{R}^{n\times m}$ with $\mathbf{a}_i$ being the i^{th} row of $\mathbf{A}$. The action of T_{k_Θ} on a function $\alpha \in \mathcal{Y}$ is then approximated using MC approximation as

$$T_{k_\Theta}\alpha \approx \frac{1}{m}\sum_{j\in[m]} k_\Theta(\cdot, \theta_j)\alpha(\theta_j),$$

resulting in the estimator

$$h(x)(\theta) = \frac{1}{\lambda nm}\sum_{i\in[n]} k_\mathcal{X}(x, x_i)\sum_{j\in[m]} a_{ij}k_\Theta(\theta, \theta_j).$$

Using this $\mathbf{A}$ parameterization of h, the different terms in Problem 5 are approximated as follows.

- **Squared norm of the dual variables**: We approximate the squared norm of the dual variables using MC sampling with locations $(\theta_j)_{j\in[m]}$:

$$\sum_{i\in[n]} \frac{1}{2}\|\alpha_i\|_\mathcal{Y}^2 \approx \frac{1}{2m}\mathrm{Tr}\left(\mathbf{A}^\top\mathbf{A}\right).$$

- **Scalar product with the data:** We encode the evaluations of the observed functions $(y_i)_{i\in[n]}$ at locations

Table 1: Correspondence between the quantities involved in Problem 4 depending on the representation.

	Linear splines	Eigenvectors of T_{k_Θ}
$\sum_{i\in[n]} \frac{1}{2}\|\alpha_i\|_{\mathcal{Y}}^2$	$\frac{1}{2m}\operatorname{Tr}\left(\mathbf{A}\mathbf{A}^\top\right)$	$\operatorname{Tr}\left(\frac{1}{2}\mathbf{A}\mathbf{A}^\top\right)$
$\sum_{i\in[n]}\langle\alpha_i, y_i\rangle_{\mathcal{Y}}$	$\frac{1}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{Y}^\top\right)$	$\operatorname{Tr}\left(\mathbf{A}\mathbf{R}^\top\right)$
$\sum_{i,j\in[n]} k_{\mathcal{X}}(x_i,x_j)\langle\alpha_i, T_{k_\Theta}\alpha_j\rangle_{\mathcal{Y}}$	$\frac{1}{m^2}\operatorname{Tr}\left(\mathbf{K}_{\mathcal{X}}\mathbf{A}\mathbf{K}_\Theta\mathbf{A}^\top\right)$	$\operatorname{Tr}\left(\mathbf{K}_{\mathcal{X}}\mathbf{A}\mathbf{\Delta}\mathbf{A}^\top\right)$

$(\theta_j)_{j\in[m]}$ in a matrix $\mathbf{Y} = [y_i(\theta_j)]_{i\in[n],\, j\in[m]} \in \mathbb{R}^{n\times m}$, and again use MC approximation

$$\sum_{i\in[n]}\langle\alpha_i, y_i\rangle_{\mathcal{Y}} \approx \frac{1}{m}\operatorname{Tr}\left(\mathbf{A}\mathbf{Y}^\top\right).$$

- **Regularization term:** Encoding the dual variables as linear splines hinders the exact computation of the quadratic terms $\langle\alpha_i, T_{k_\Theta}\alpha_j\rangle_{\mathcal{Y}}$, which we propose to approximate using a MC approximation of T_{k_Θ}. Letting $\mathbf{K}_{\mathcal{X}} \in \mathbb{R}^{n\times n}, \mathbf{K}_\Theta \in \mathbb{R}^{m\times m}$ be the Gram matrices respectively associated to the data $(x_i)_{i\in[n]}$ and kernel $k_{\mathcal{X}}$, and to the locations $(\theta_j)_{j\in[m]}$ and kernel k_Θ, the regularization term can be rephrased as $\frac{1}{\lambda nm^2}\operatorname{Tr}\left(\mathbf{K}_{\mathcal{X}}\mathbf{A}\mathbf{K}_\Theta\mathbf{A}^\top\right)$.
- **Constraints:** The constraints $\|\alpha_i\|_q \leq \kappa$ can be handled similarly. Particularly, let $\alpha_i \in \mathcal{S}_m$ with associated evaluations vector $\mathbf{a}_i \in \mathbb{R}^m$ and $q \in [1, +\infty[$. We trade the integral expression in the norm for an MC sum, resulting in the constraint $\|\mathbf{a}_i\|_q \leq m^{\frac{1}{q}}\kappa$. This expression also holds for $q = +\infty$, as $\|\alpha_i\|_\infty \leq \kappa$ iff. $\|\mathbf{a}_i\|_\infty \leq \kappa$.

Gathering the different terms (summarized in Table 1) yields the following relaxation of Problem 5:

$$\inf_{\mathbf{A}\in\mathbb{R}^{n\times m}} \operatorname{Tr}\left(\frac{1}{2}\mathbf{A}\mathbf{A}^\top - \mathbf{A}\mathbf{Y}^\top + \frac{1}{2\lambda nm}\mathbf{K}_{\mathcal{X}}\mathbf{A}\mathbf{K}_\Theta\mathbf{A}^\top\right) \quad s.t.\ \|\mathbf{A}\|_{q,\infty} \leq m^{\frac{1}{q}}\kappa. \tag{8}$$

Remark: The decomposable assumption on the kernel K plays a role in the regularization. It has the effect of disentangling the action of both Gram matrices $\mathbf{K}_{\mathcal{X}}$ and $\mathbf{K}_\Theta$.

We propose to tackle Problem 8 using accelerated proximal gradient descent (APGD), where the proximal step amounts to projecting the coefficients on the q-ball of radius $m^{1/q}\kappa$. The technique is summarized in Algorithm 1. The gradient stepsize γ can be computed exactly from the parameters of the problem. Indeed, for guaranteed convergence, one must set $\gamma < \frac{2}{C}$ where C is the Lipschitz constant associated to the gradient of the objective function; here $C \leq 1 + \frac{1}{\lambda n}\|\mathbf{K}_{\mathcal{X}}\|_{\text{op}}\|\mathbf{K}_\Theta\|_{\text{op}}$. The initialization can either be the null matrix $\mathbf{A}^{(0)} = \mathbf{0}_{\mathbb{R}^{n\times m}}$ or the solution of the unconstrained optimization problem which can be obtained in closed form. This solution (i) can dramatically reduce the number of iterations for small ϵ or large κ, and (ii) can be computed in $\mathcal{O}(n^3+m^3)$ time exploiting the Kronecker structure inherited from the separable kernel with a

Algorithm 1 APGD with linear splines

input : Gram matrices $\mathbf{K}_{\mathcal{X}}, \mathbf{K}_\Theta$, data matrix $\mathbf{Y}$, regularization parameter λ, loss parameters (κ, p) or (ϵ, p), gradient step γ
init : $\mathbf{A}^{(0)}, \mathbf{A}^{(-1)} = \mathbf{0} \in \mathbb{R}^{n\times m}$
for *epoch* t *from* 0 *to* $T-1$ **do**
  // gradient step
  $\mathbf{V} = \mathbf{A}^{(t)} + \frac{t-2}{t+1}\left(\mathbf{A}^{(t)} - \mathbf{A}^{(t-1)}\right)$
  $\mathbf{U} = \mathbf{V} - \gamma\left(\mathbf{V} + \frac{1}{\lambda nm}\mathbf{K}_{\mathcal{X}}\mathbf{V}\mathbf{K}_\Theta - \mathbf{Y}\right)$
  // projection step
  if $p = 2$ **then**
    for *row* $i \in [n]$ **do**
      $\mathbf{a}_i^{(t+1)} = \min\left(\frac{\sqrt{m}\kappa}{\|\mathbf{u}_i\|_2}, 1\right)\mathbf{u}_i$ // if H_κ^2
      $\mathbf{a}_i^{(t+1)} = \left|1 - \frac{\gamma\epsilon}{\sqrt{m}\|\mathbf{u}_i\|_2}\right|_+ \mathbf{u}_i$ // if ℓ_ϵ^2
  else
    for *row* $i \in [n]$ **do**
      for *column* $j \in [m]$ **do**
        $a_{ij}^{(t+1)} = \operatorname{sign}(u_{ij})\min\left(\kappa, |u_{ij}|\right)$ // if H_κ^1
        $a_{ij}^{(t+1)} = \operatorname{sign}\left(u_{ij}\right)\left||u_{ij}| - \frac{\gamma\epsilon}{m}\right|_+$ // if ℓ_ϵ^∞
return $\mathbf{A}^{(T)}$

Sylvester equation solver (Sima, 1996; Dinuzzo et al., 2011). Since the objective function in Problem 8 is the sum of two functions, one convex and differentiable with Lipschitz continuous gradient (the quadratic form) and one convex and lower semi-continuous (the indicator function of the constraint set), the optimal worst case complexity is $\mathcal{O}\left(\frac{1}{T^2}\right)$ (Beck & Teboulle, 2009). The time complexity per iteration is dominated by the computation of the matrix $\mathbf{K}_{\mathcal{X}}\mathbf{V}\mathbf{K}_\Theta$ which is $\mathcal{O}(n^2m + m^2n)$.

3.2. The Eigendecomposition Approach

In this section, we propose an alternative finite-dimensional description of the dual variables relying on an approximate eigendecomposition of T_{k_Θ} when $p = 2$. The rationale of this approach is to decrease the number of parameters needed to represent the estimator by selecting directions well-suited to T_{k_Θ}, namely the dominant r eigenvectors of T_{k_Θ}. As computing the eigendecomposition of T_{k_Θ} is generally intractable, we propose an approximation detailed in the following.

Let us consider the problem of finding a continuous eigen-

vector ψ of a sampled version of the integral operator T_{k_Θ} with eigenvalue $\delta > 0$ (Hoegaerts et al., 2005):

$$\frac{1}{m}\sum_{j\in[m]} k_\Theta(\theta,\theta_j)\psi(\theta_j) = \delta\psi(\theta) \text{ for } \forall\theta\in\Theta.$$

By evaluating it at points $(\theta_j)_{j\in[m]}$, one gets that $(\delta m, \psi(\theta_j)_{j\in[m]})$ form the eigensystem of the Gram matrix $\mathbf{K}_\Theta \in \mathbb{R}^{m\times m}$. These can be computed using for instance singular value decomposition, and by substitution one arrives at an approximated eigenbasis of dimension at most m; this can be used as a proxy instead of the true eigenvectors of T_{k_Θ}. By using the first r of these vectors for some $r < m$ we are able to lower the size of the parameterization of the model. We store in a diagonal matrix $\mathbf{\Delta} \in \mathbb{R}^{r\times r}$ the first r eigenvalues.

The problem is now parameterized by a matrix $\mathbf{A} = [\mathbf{a}_i]_{i\in[n]} \in \mathbb{R}^{n\times r}$ with each row $\mathbf{a}_i \in \mathbb{R}^r$ encoding the coefficients of the dual variable α_i on the $(\psi_j)_{j\in[r]}$ basis. The estimator then reads as

$$h(x)(\theta) = \frac{1}{\lambda n}\sum_{i\in[n],\, j\in[r]} a_{ij}\delta_j k_\mathcal{X}(x,x_i)\psi_j(\theta).$$

We store in $\mathbf{R} = [R_{ij}]_{i\in[n],j\in[r]} \in \mathbb{R}^{n\times r}$ the scalar products between the observed data and the eigenbasis: $R_{ij} = \langle y_i, \psi_j\rangle_\mathcal{Y}$. The correspondence between the different optimization terms are summarized in Table 1; the optimization reduces to

$$\inf_{\mathbf{A}\in\mathbb{R}^{n\times r}} \operatorname{Tr}\left(\frac{1}{2}\mathbf{A}\mathbf{A}^\top - \mathbf{A}\mathbf{R}^\top + \frac{1}{2\lambda n}\mathbf{K}_\mathcal{X}\mathbf{A}\mathbf{\Delta}\mathbf{A}^\top\right) \quad s.t.\ \|\mathbf{A}\|_{2,\infty} \le \kappa. \tag{9}$$

Again, one can use APGD to solve this task; the resulting computations are deferred to Algorithm 2 in the supplement.

Remark: For κ large enough, the solution reduces to that of Kadri et al. (2016). However, our approximation allows handling a wide range of kernels in contrast to the analytical knowledge imposed for the eigensystem of T_{k_Θ} by Kadri et al. (2016); Laforgue et al. (2020).

4. Learning with ϵ-insensitive Losses

In this section, we propose a generalized ϵ-insensitive version of the square loss on $\mathcal{Y}$ involving infimal convolution, and derive tractable dual optimization algorithm to solve Problem 1. This loss induces sparsity on the matrix of coefficients as illustrated in Section 5.

Definition 4.1 (ϵ-insensitive loss)**.** Let $\epsilon > 0$ and $p \in [1,+\infty]$. We define the ϵ-insensitive version of the square loss with parameters (ϵ, p) as

$$\ell_\epsilon^p := \frac{1}{2}\|\cdot\|_\mathcal{Y}^2 \,\square\, \iota_{\mathcal{B}_\epsilon^p}(\cdot).$$

When $\mathcal{Y} = \mathbb{R}$, the loss reduces to the classical ϵ-insensitive version of the square loss regardless of p. The following proposition (counterpart of Proposition 3.2) sheds light on the effect of the infimal convolution on the square loss.

Proposition 4.2. *Let $\epsilon > 0$ and $p \in [1,+\infty]$. Then for all $f \in \mathcal{Y}$,*

$$\ell_\epsilon^p(f) = \frac{1}{2}\left\|f - \operatorname{Proj}_{\mathcal{B}_\epsilon^p}(f)\right\|_\mathcal{Y}^2. \tag{10}$$

Remark: Proposition 4.2 means that $\ell_\epsilon^p(f) = 0$ when $\|f\|_p \le \epsilon$, i.e. small residuals do not contribute to the risk. For general p, $\ell_\epsilon^p(f)$ is not straightforward to compute due to the complexity of $\operatorname{Proj}_{\mathcal{B}_\epsilon^p}(f)$. As we however use a dual approach, Problem 1 can still be tackled computationally.

The next result shows how to dualize Problem 1 when the proposed ϵ-insensitive loss is used.

Proposition 4.3 (Dual ϵ-insensitive)**.** *Let $\epsilon \ge 0, p \in [1,+\infty]$, and $\frac{1}{p}+\frac{1}{q}=1$. The dual of Problem 1 writes as*

$$\inf_{(\alpha_i)_{i\in[n]}\in\mathcal{Y}^n} \sum_{i\in[n]}\left[\frac{1}{2}\|\alpha_i\|_\mathcal{Y}^2 - \langle\alpha_i, y_i\rangle_\mathcal{Y} + \epsilon\|\alpha_i\|_q\right] + \frac{1}{2\lambda n}\sum_{i\in[n],\,j\in[n]} k_\mathcal{X}(x_i,x_j)\,\langle\alpha_i, T_{k_\Theta}\alpha_j\rangle_\mathcal{Y}. \tag{11}$$

Remark (influence of ϵ and p): Compared to the square loss, ℓ_ϵ^p induces an additional term $\sum_{i\in[n]}\epsilon\|\alpha_i\|_q$ in the dual. Setting $\epsilon = 0$ recovers the square loss case.

The challenges involving the representation of the dual variables, and the computability of the different terms composing Problem 11 are similar to those evoked in Section 3. We have however traded the constraints on the q-norms of the dual variables against an additional non-smooth term. As for the Huber loss family in Section 3, we address this convex non-smooth optimization problem through the APGD algorithm. The proximal step involves the computation of $\operatorname{prox}_{\gamma\epsilon\|\cdot\|_q}$ for a suitable gradient stepsize $\gamma > 0$, which is the focus of the next proposition.

Proposition 4.4 (Proximal q-norm)**.** *Let $\epsilon > 0$. The proximal operator of $\epsilon\|\cdot\|_q$ is computable for $q = 1$ and $q = 2$, and given for all $(\alpha,\theta) \in \mathcal{Y}\times\Theta$ by*

$$\left(\operatorname{prox}_{\epsilon\|\cdot\|_1}(\alpha)\right)(\theta) = \operatorname{sign}(\alpha(\theta))\,\big||\alpha(\theta)| - \epsilon\big|_+, \tag{12}$$

$$\operatorname{prox}_{\epsilon\|\cdot\|_2}(\alpha) = \alpha\left|1 - \frac{\epsilon}{\|\alpha\|_\mathcal{Y}}\right|_+. \tag{13}$$

We recognize in Equation (12) a pointwise soft thresholding, and in Equation (13) an analogous to the block soft thresholding, both are known to promote sparsity.

To solve Problem 11 we rely on the two kinds of finite representations introduced previously. In Section 4.1, we

tackle the $p = 2$ and $p = \infty$ case with the linear splines based method from Section 3.1, before using dimensionality reduction from Section 3.2 for the case $p = 2$ in Section 4.2.

4.1. The Linear Spline Based Approach

Similarly to what was presented in Section 3.1, we use linear splines to represent the dual variables $(\alpha_i)_{i\in[n]}$ as they allow a pointwise control over the dual variable and thus give rise to computable proximal operators. Keeping the notations, the optimization boils down to

$$\inf_{\mathbf{A}\in\mathbb{R}^{n\times m}} \operatorname{Tr}\left(\frac{1}{2}\mathbf{A}\mathbf{A}^\top - \mathbf{A}\mathbf{Y}^\top + \frac{1}{2\lambda nm}\mathbf{K}_\mathcal{X}\mathbf{A}\mathbf{K}_\Theta\mathbf{A}^\top\right) + \frac{\epsilon}{m^{\frac{1}{q}}}\sum_{i\in[n]}\|\mathbf{a}_i\|_q\,.$$

We use APGD to solve it with steps detailed in Algorithm 1. When $q = 1$, the proximal operator is the soft thresholding operator, akin to promote sparsity in the dual coefficients.

4.2. The Eigendecomposition Approach

We mobilize the eigendecomposition technique from Section 3.2 to solve Problem 11 in the case $p = 2$. Using the same notation as in Problem 9, we get the following task:

$$\inf_{\mathbf{A}\in\mathbb{R}^{n\times r}} \operatorname{Tr}\left(\frac{1}{2}\mathbf{A}\mathbf{A}^\top - \mathbf{A}\mathbf{R}^\top + \frac{1}{2\lambda n}\mathbf{K}_\mathcal{X}\mathbf{A}\mathbf{\Delta}\mathbf{A}^\top\right) + \epsilon\sum_{i\in[n]}\|\mathbf{a}_i\|_2\,.$$

APGD is applied to tackle this problem; the details are deferred to Algorithm 2 in the supplement. Notice that the proximal operator in this case is the block soft thresholding operator, known to promote structured row-wise sparsity.

5. Numerical Experiments

In this section, we demonstrate the efficiency of the proposed convoluted losses. The implementation is done in Python, and is available in the form of an open source package at https://github.com/allambert/foreg.
The experiments are centered around **two key directions**:

1. The first goal is to understand the accuracy-sparsity trade-offs of the ϵ-insensitive loss as a function of the regularization λ and insensitivity parameter ϵ.
2. Our second aim is to quantify the robustness of the Huber losses w.r.t. different forms of outliers with a particular focus on global versus local ones. To gain further insight into the robustness w.r.t. corruption, we designed 3 types of functional outliers with distinct characteristics.

Our proposed losses are investigated on 3 benchmarks: a synthetic one associated to Gaussian processes, followed by two real-world ones arising in the context of neuroimaging and speech analysis. We investigate both questions on the synthetic data, and provide further insights for the first and the second question on the neuroimaging and the speech dataset, respectively.

We now detail the 3 functional outlier types used in our experiments on robustness to study the effect of local and global corruption. Local outliers affect the functions only on small portions of Θ whereas global ones contaminates them in their entirety. To corrupt the functions $(y_i)_{i\in[n]}$, we first draw a set $I \subset \{0,\dots,n\}$ of size $\lfloor \tau n \rfloor$ corresponding to the indices to contaminate; $\tau \in [0,1]$ being the proportion of contaminated functions. Then, we perform different kinds of corruption:

- **Type 1**: Let ω be the permutation defined for $j \in [|I|]$ as $\omega(I_j) = I_{j+1}$ if $j < |I|$ and $\omega(I_{|I|}) = I_1$, then for $i \in I$, the data point (x_i, y_i) is replaced by $(x_i, -y_{\omega(i)})$.
- **Type 2**: Given covariance parameters $\boldsymbol{\sigma} \in \mathbb{R}^r$ and an intensity parameter $\zeta > 0$, we draw a Gaussian process $g_c \sim \mathcal{GP}(0, k_{\sigma_c})$ for $c \in [r]$ where k_{σ_c} is the Gaussian covariance function with standard deviation σ_c. Then, for $i \in I$, we replace (x_i, y_i) with $\left(x_i, \sum_{c\in[r]} a_{ic} g_c\right)$ where the coefficients a_{ic} are drawn i.i.d. from a uniform distribution $\mathcal{U}([-0.5\zeta, 0.5\zeta])$.
- **Type 3**: For each $i \in I$, a randomly chosen fraction $\xi \in [0,1]$ of the discrete observations for y_i is replaced by random draws from a uniform distribution $\mathcal{U}([-b_{\max}, b_{\max}])$, where $b_{\max} := \max_{i,j} |y_i(\theta_j)|$. The corruptions of Type 1 and 2 are global ones whereas Type 3 is a local one. In terms of the characteristics of the different corruptions, for Type 1 the properties of the outlier functions remain close to those of the non-outlier ones, whereas with Type 2 they become completely different. Finally, for corrupted data in the hyperparameter choice using cross-validation the mean was replaced with median.

For the losses H^1_κ and ℓ^∞_ϵ we solve the problem based on the representation with linear splines (see Section 3.1 and Section 4.1 respectively); this is the only possible approach. However for the losses H^2_κ and ℓ^2_ϵ we exploit the representation using a truncated basis of approximate eigenfunctions (see Section 3.2 and Section 4.2 respectively), in doing so we reduce the computational cost compared to the linear splines approach. Concerning optimization, we deployed the APGD method (Beck & Teboulle, 2009) with backtracking line search, and adaptive restart (O'Donoghue & Candès, 2015). The initialization in APGD was carried out with the closed-form solution available for the square loss using a Sylvester solver.

Regarding the performance measure applied for evaluation, let $((y_i(\theta_{ij}))_{j\in[m_i]})_{i\in[n]}$ be the set of observed discretized functions and let $(\widehat{y}_i(\theta_{ij}))_{j\in[m_i]})_{i\in[n]}$ be an estimated set of discretized functions, where $(\theta_{ij})_{j\in[m_i]}$ denotes the observation locations for y_i. We used the mean squared error

Table 2: MSEs and sparsity on the DTI dataset

λ	METRIC	$1/2\|\cdot\|_{\mathcal{Y}}^2$	H_κ^2	H_κ^1	ℓ_ϵ^2	ℓ_ϵ^∞
10^{-5}	MSE (10^{-1})	2.5±0.19	**2.21**±0.31	**2.21**±0.31	2.41±0.26	2.5±0.23
	SPARSITY	-	-	-	27.4±17.2%	85.9±10.7%
10^{-3}	MSE (10^{-1})	**2.18**±0.27	2.23±0.32	2.21±0.32	2.2±0.29	**2.18**±0.28
	SPARSITY	-	-	-	3.4±6.9%	12.7±10.5%

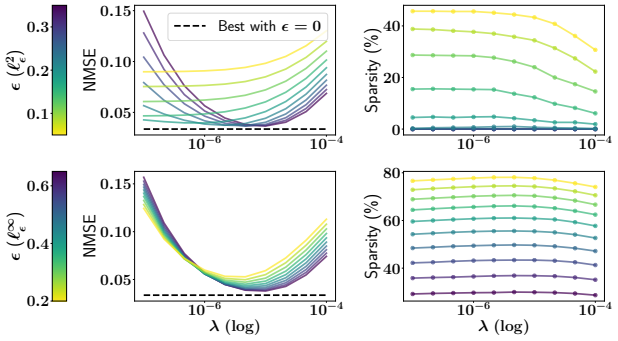


Figure 1: Interaction between regularization λ and insensitivity ϵ for the loss ℓ_ϵ^2 (1st row) and ℓ_ϵ^∞ (2nd row).

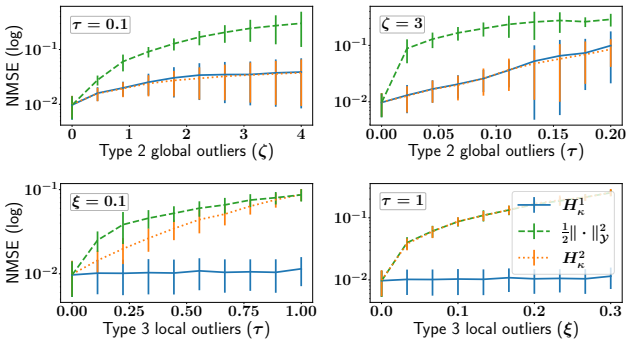


Figure 2: Robustness to different types of outliers.

defined as

$$\text{MSE} := \frac{1}{n}\sum_{i\in[n]}\sum_{j\in[m_i]}[y_i(\theta_{ij}) - \widehat{y}_i(\theta_{ij})]^2.$$

When $m_i = m$ for all i, we normalize it by m and define NMSE:$= \frac{1}{m}$ MSE.

5.1. Experiments on the synthetic dataset

The impact of the different losses are investigated in detail on a function-to-function synthetic dataset whose construction is detailed in Appendix B.2. The kernels $k_{\mathcal{X}}$ and k_Θ are chosen to be Gaussian and the experiments are averaged over 20 draws with training and testing samples of size 100.

ϵ-insensitive loss: To study the interaction between λ and ϵ and the resulting sparsity-accuracy trade-offs, we added i.i.d. Gaussian noise with standard deviation 0.5 to the observations of the output functions. The resulting MSE values are summarized in Fig. 1. For both the ℓ_ϵ^2 and the ℓ_ϵ^∞ loss, one can reduce λ, increase ϵ and get a fair amount of sparsity while making a small compromise in terms of accuracy.

Huber loss: We investigate the robustness of the Huber loss to different types of outliers while selecting both λ and κ through robust cross validation. The resulting MSE values are summarized in Fig. 2. As it can be seen in the first row of the figure, the losses H_κ^1 and H_κ^2 are significantly more robust to global outliers than the square loss, both when the outliers' intensity ζ and the proportion τ of contaminated samples vary. The second row of the figure shows that when dealing with local outliers, the closer one gets to the whole sample being contaminated ($\tau = 1$), the less robust H_κ^2 becomes. On the other hand, H_κ^1 shows remarkable robustness as it can be observed at the bottom right panel (in case of $\tau = 1$). One can interpret this phenomenon by noticing that the loss H_κ^2 can penalize less big discrepancy between functions in the $\|.\|_{\mathcal{Y}}$ norm sense, but if all samples are contaminated locally a little, the outliers are meddled in the norm and so H_κ^2 becomes inefficient.

5.2. Experiments on the DTI dataset

In our next experiment we considered the DTI benchmark[2]. The dataset contains a collection of fractional anisotropy profiles deduced from diffusion tensor imaging scans, and we take the first scans of the $n = 100$ multiple sclerosis patients. The profiles are given along two tracts, the corpus callosum and the right cortiospinal. The goal is to predict the latter function from the former, which can be framed as a function-to-function regression problem. When some functions admit missing observations, we fill in the gaps by linear interpolation, and later use the MSE as metric. We use a Gaussian kernel for $k_{\mathcal{X}}$ and a Laplacian one for k_Θ and average over 10 runs with $n_{\text{train}} = 70$ and $n_{\text{test}} = 30$.

Similarly to our experiences gained on the synthetic dataset, a compromise can be made between the two parameters λ and ϵ to get increased sparsity, as can be observed in Table 2.

[2]This dataset was collected at the Johns Hopkins University and the Kennedy-Krieger Institute.

Table 3: MSEs on speech data

VT	$1/2\|\cdot\|_{\mathbb{y}}^2$	H_κ^2	H_κ^1	ℓ_ϵ^2	ℓ_ϵ^∞
LP	**6.58**±0.62	6.59±0.62	6.59±0.64	**6.58**±0.62	**6.58**±0.62
LA	4.65±0.55	4.65±0.55	4.66±0.55	**4.64**±0.55	**4.64**±0.55
TBCL	**4.26**±0.46	**4.26**±0.46	4.27±0.46	**4.26**±0.46	**4.26**±0.46
TBCD	4.67±0.37	4.68±0.38	4.7±0.38	**4.67**±0.38	**4.67**±0.38
VEL	**2.94**±0.5	**2.94**±0.5	2.95±0.5	**2.94**±0.5	**2.94**±0.5
GLO	**7.25**±0.65	7.26±0.65	**7.25**±0.64	**7.25**±0.65	**7.25**±0.65
TTCL	3.76±0.21	3.76±0.21	3.74±0.2	**3.73**±0.21	**3.73**±0.21
TTCD	5.93±0.34	5.94±0.34	5.93±0.35	**5.92**±0.34	**5.92**±0.34

Table 4: MSEs on contaminated speech data

VT	TYPE 1 OUTLIERS ($\tau = 0.1$)			TYPE 3 OUTLIERS ($\tau = 0.1, \xi = 0.1$)		
	$1/2\|\cdot\|_{\mathbb{y}}^2$	H_κ^2	H_κ^1	$1/2\|\cdot\|_{\mathbb{y}}^2$	H_κ^2	H_κ^1
LP	9.4±0.75	9.4±0.66	**9.19**±0.79	7.53±0.58	7.62±0.59	**7.0**± 0.59
LA	5.72±0.76	5.63±0.71	**5.52**±0.69	**5.06**±0.6	5.11±0.6	5.09±0.55
TBCL	6.71±0.96	6.14±0.97	**5.98**±0.93	5.06±0.51	5.16±0.48	**4.72**± 0.54
TBCD	**5.8**±0.41	5.86±0.44	5.83±0.44	5.18±0.4	5.26±0.41	**5.08**± 0.4
VEL	4.37±0.56	3.76±0.62	**3.76**±0.59	3.52±0.57	3.54±0.58	**3.41**± 0.57
GLO	9.61±0.87	**9.51**±0.86	9.53±0.84	7.94±0.61	8.02±0.61	**7.76**± 0.61
TTCL	15.06±2.22	9.51±0.63	**9.48**±0.6	**5.89**±0.43	5.91±0.45	6.62±0.66
TTCD	8.15±0.48	**7.96**±0.49	8.02±0.51	6.63±0.44	6.74±0.42	**6.36**± 0.39

Moreover, we highlight that even for optimal regularization with respect to the square loss $\lambda = 10^{-3}$, one gets a fair amount of sparsity while getting the same score with ℓ_ϵ^∞ and a very small difference with ℓ_ϵ^2.

5.3. Speech data

In this section, we focus on a speech inversion problem (Mitra et al., 2009). Particularly, our goal is to predict a vocal tract (VT) configuration that likely produced a speech signal (Richmond, 2002). This benchmark encompasses $n = 413$ synthetically pronounced words to which 8 VT functions are associated: LA, LP, TTCD, TTCL, TBCD, TBCL, VEL, GLO. This is then a time-series–to–function regression problem. We predict the VT functions separately in eight subproblems.

Since the words are of varying length, we use the MSE as metric and extend symmetrically the signals to match the longest word for in training. We encode the input sounds through 13 mel-frequency cepstral coefficients (MFCC) and normalize the VT functions' values to the range [-1, 1]. We average over 10 train-test splits taking $n_{\text{train}} = 250$ and $n_{\text{test}} = 163$. Finally we take an integral Gaussian kernel on the standardized MFCCs (see Appendix B for further details) as $k_{\mathbb{X}}$ and a Laplace kernel as k_Θ.

We first compare all the losses on untainted data in Table 3. Then to evaluate the robustness of the Huber losses, we ran experiments on contaminated data with two configurations. In the first case, we added Type 1 (global) outliers with $\tau = 0.1$ and in the second one, we added Type 3 (local) outliers with $\tau = 0.1$ and $\xi = 0.1$. The results are displayed in Table 4. In the contaminated setting, one gets results similar to ones obtained on the synthetic dataset. The loss H_κ^1 works especially well for local outliers whereas the loss H_κ^2 is robust only to global outliers.

6. Conclusion

In this paper we introduced generalized families of loss functions based on infimal convolution and p-norms for functional output regression. The resulting optimization problems were handled using duality principles. Future work could focus on extending these techniques to a wider choice of p using iterative techniques for the proximal steps.

Acknowledgements

AL, DB, and FdB received funding from the *Télécom Paris Research and Testing Chair on Data Science and Artificial Intelligence for Digitalized Industry and Service*. AL also obtained additional funding from the *ERC Advanced Grant E-DUALITY (787960)*.

References

Bauschke, H. H., Combettes, P. L., et al. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Bouche, D., Clausel, M., Roueff, F., and d'Alché Buc, F. Nonlinear functional output regression: A dictionary approach. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 235–243, 2021.

Cadre, B. Convergent estimators for the l1-median of Banach valued random variable. *Statistics: A Journal of Theoretical and Applied Statistics*, 35(4):509–521, 2001.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

Dinuzzo, F., Ong, C. S., Gehler, P., and Pillonetto, G. Learning output kernels with block coordinate descent. In *International Conference on Machine Learning (ICML)*, pp. 49–56, 2011.

Ferraty, F., Laksaci, A., Tadj, A., Vieu, P., et al. Kernel regression with functional response. *Electronic Journal of Statistics*, 5:159–171, 2011.

Hoegaerts, L., Suykens, J. A., Vandewalle, J., and De Moor, B. Subset based least squares subspace regression in RKHS. *Neurocomputing*, 63:293–323, 2005.

Huber, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pp. 73–101, 1964.

Hubert, M., Rousseeuw, P. J., and Segaert, P. Multivariate functional outlier detection. *Statistical Methods and Applications*, 24:177–202, 2015.

Kadri, H., Duflos, E., Preux, P., Canu, S., and Davy, M. Nonlinear functional regression: a functional RKHS approach. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 374–380, 2010.

Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., and Audiffren, J. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20):1–54, 2016.

Kalogridis, I. and Van Aelst, S. Robust functional regression based on principal components. *Journal of Multivariate Analysis*, 173:393–415, 2019.

Laforgue, P., Lambert, A., Brogat-Motte, L., and d'Alché Buc, F. Duality in RKHSs with infinite dimensional outputs: Application to robust losses. In *International Conference on Machine Learning (ICML)*, pp. 5598–5607, 2020.

Lee, Y.-J., Hsieh, W.-F., and Huang, C.-M. epsilon-SSVR: A smooth support vector machine for epsilon-insensitive regression. *IEEE Transactions on Knowledge & Data Engineering*, (5):678–685, 2005.

Lian, H. Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *Canadian Journal of Statistics*, 35(4):597–606, 2007.

Maronna, R. A. and Yohai, V. J. Robust functional linear regression based on splines. *Computational Statistics & Data Analysis*, 65:46–55, 2013.

Micchelli, C. and Pontil, M. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

Micchelli, C., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

Mitra, V., Ozbek, Y., Nam, H., Zhou, X., and Espy-Wilson, C. Y. From acoustics to vocal tract time functions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4497–4500, 2009.

Moreau, J. J. Proximité et dualité dans un espace hilbertien. Technical report, 1965. (https://hal.archives-ouvertes.fr/hal-01740635).

Morris, J. S. Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.

Oliva, J., Neiswanger, W., Póczos, B., Xing, E., Trac, H., Ho, S., and Schneider, J. Fast function to function regression. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 717–725, 2015.

O'Donoghue, B. and Candès, E. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732, 2015.

Parikh, N. and Boyd, S. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

Pedrick, G. Theory of reproducing kernels for Hilbert spaces of vector-valued functions. Technical report, University of Kansas, Department of Mathematics, 1957.

Ramsay, J. and Silverman, B. Functional data analysis, 1997.

Ramsay, J. O. and Silverman, B. W. *Applied functional data analysis: methods and case studies*. Springer, 2007.

Reimherr, M., Sriperumbudur, B., Taoufik, B., et al. Optimal prediction for additive function-on-function regression. *Electronic Journal of Statistics*, 12(2):4571–4601, 2018.

Richmond, K. *Estimating Articulatory Parameters from the Acoustic Speech Signal*. PhD thesis, The Center for Speech Technology Research, Edinburgh University, 2002.

Sangnier, M., Fercoq, O., and d'Alché-Buc, F. Data sparse nonparametric regression with ϵ-insensitive losses. In *Asian Conference on Machine Learning (ACML)*, pp. 192–207, 2017.

Sima, V. *Algorithms for Linear-Quadratic Optimization*. Chapman and Hall/CRC, 1996.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Ullah, S. and Finch, C. F. Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13(1):1–12, 2013.

Wahba, G. *Spline Models for Observational Data*. SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics, 1990.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

Zhou, D.-X. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220:456–463, 2008.

Zhu, H., Brown, P. J., and Morris, J. S. Robust, adaptive functional regression in functional mixed model framework. *Journal of the American Statistical Association*, 106(495):1167–1179, 2011.

The supplement is structured as follows. We present the proofs of our results in Appendix A. Appendix B complements the main part of the paper by providing additional algorithmic and experimental details. Finally, Appendix C includes additional plots and insights about the loss functions.

A. Proofs

In this section, we present the proofs of our results. In Appendix A.1 we recall some definitions from convex optimization used throughout the proofs, with focus on Fenchel-Legendre conjugation and proximal operators, followed by the proofs themselves (Appendix A.2-A.8).

A.1. Reminder on Convex Optimization

Recall that $\mathcal{Y} := L^2[\Theta, \mu]$ where $\Theta \subset \mathbb{R}$ is a compact set endowed with a probability measure μ.

Definition A.1 (Proper, convex, lower semi-continuous functions)**.** We denote by $\Gamma_0(\mathcal{Y})$ the set of functions $J\colon \mathcal{Y} \to]-\infty, +\infty]$ that are

1. *proper*: $\mathrm{dom}(J) := \{f \in \mathcal{Y} \,:\, J(f) < +\infty\} \neq \varnothing$,
2. *convex*: $J(tf + (1-t)g) \leq tJ(f) + (1-t)J(g)$ for $\forall f, g \in \mathcal{Y}, \forall t \in [0,1]$, and
3. *lower semicontinuous*: $\underline{\lim}_{g \to f} J(g) \geq J(f)$ for $\forall f \in \mathcal{Y}$, where $\underline{\lim}$ denotes limit inferior.

Definition A.2. The *Fenchel-Legendre conjugate* of a function $J\colon \mathcal{Y} \to [-\infty, +\infty]$ is defined as

$$J^\star(f) := \sup_{g \in \mathcal{Y}} \langle f, g \rangle_{\mathcal{Y}} - J(g), \quad f \in \mathcal{Y}.$$

The Fenchel-Legendre conjugate of a function J is always convex. It is also involutive on $\Gamma_0(\mathcal{Y})$, meaning that $(J^\star)^\star = J$ for any $J \in \Gamma_0(\mathcal{Y})$. We gather in Table 5 examples and properties of Fenchel-Legendre conjugates.

We now introduce the infimal convolution operator following Bauschke et al. (2011).

Definition A.3 (Infimal convolution)**.** The *infimal convolution* of two functions $L, J\colon \mathcal{Y} \to]-\infty, +\infty]$ is

$$L \,\square\, J\colon \begin{pmatrix} \mathcal{Y} \to [-\infty, +\infty] \\ f \mapsto \inf\limits_{g \in \mathcal{Y}} L(f-g) + J(g) \end{pmatrix}.$$

One key property of the infimal convolution operator is that it behaves nicely under Fenchel-Legendre conjugation, as it is detailed in the following proposition.

Lemma A.4 (Bauschke et al. 2011, Proposition 13.24)**.** *Let $L, J\colon \mathcal{Y} \to]-\infty, +\infty]$. Then*

$$(L \,\square\, J)^\star = L^\star + J^\star.$$

We now define the proximal operator, used as a replacement for the classical gradient step in the presence of non-differentiable objective functions.

Definition A.5 (Proximal operator, Moreau 1965)**.** The *proximal operator* (or proximal map) is defined as

$$\mathrm{prox}_J(f) := \operatorname*{arg\,min}_{g \in \mathcal{Y}} \; J(g) + \frac{1}{2} \|f - g\|_{\mathcal{Y}}^2, \text{ for } (J, f) \in \Gamma_0(\mathcal{Y}) \times \mathcal{Y}. \tag{14}$$

One advantage of working with functions in $\Gamma_0(\mathcal{Y})$ is that the proximal operator is always well-defined. Its computation is doable for various losses thanks to the following lemma.

Lemma A.6 (Moreau decomposition, Moreau 1965)**.** *Let $J \in \Gamma_0(\mathcal{Y})$ and $\gamma > 0$. Then*

$$\mathrm{Id} = \mathrm{prox}_{\gamma J}(\cdot) + \gamma \, \mathrm{prox}_{J^\star/\gamma}(\cdot/\gamma), \tag{15}$$

where Id *stands for the identity operator.*

Table 5: Properties of Fenchel-Legendre conjugate, for any $J, L \colon \mathcal{Y} \to [-\infty, +\infty]$ and $p, q \in [1, +\infty]$ s.t. $\frac{1}{p} + \frac{1}{q} = 1$.

Function	Fenchel-Legendre conjugate
$\frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2$	$\frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2$
$\|\cdot\|_p$	$\iota_{\mathcal{B}_1^q}$
ϵJ	$\epsilon J^\star(\frac{\cdot}{\epsilon})$ for all $\epsilon > 0$
$J(\cdot - y)$	$J^\star + \langle \cdot, y \rangle_{\mathcal{Y}}$ for all $y \in \mathcal{Y}$
$L \,\square\, J$	$L^\star + J^\star$

We remind the reader that we want to solve

$$\inf_{h \in \mathcal{H}_K} \frac{1}{n} \sum_{i \in [n]} L(y_i - h(x_i)) + \frac{\lambda}{2} \|h\|_{\mathcal{H}_K}^2 , \tag{16}$$

where $K \colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is a *decomposable* OVK of the form $K = k_{\mathcal{X}} T_{k_\Theta}$. Here $k_{\mathcal{X}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_\Theta \colon \Theta \times \Theta \to \mathbb{R}$ are continuous real-valued kernels, and $T_{k_\Theta} \in \mathcal{L}(\mathcal{Y})$ is the integral operator associated to k_Θ, defined for all $\alpha \in \mathcal{Y}$ by

$$(T_{k_\Theta} \alpha)(\theta) = \int_\Theta k(\theta, \theta') \alpha(\theta) \mathrm{d}\mu(\theta') \text{ for all } \theta \in \Theta.$$

We also remind the reader to the dual of Problem 16 when the loss writes as an infimal convolution.

Lemma A.7 (Dualization for convoluted losses, Laforgue et al. 2020)**.** *Let L be a loss function defined as $L = \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2 \,\square\, g$ for some $g \colon \mathcal{Y} \to]-\infty, +\infty]$. Then the solution of Problem 16 is given by*

$$\hat{h} = \frac{1}{\lambda n} \sum_{i \in [n]} k_{\mathcal{X}}(\cdot, x_i) T_{k_\Theta} \hat{\alpha}_i, \tag{17}$$

with $(\hat{\alpha}_i)_{i \in [n]} \in \mathcal{Y}^n$ being the solution of the dual task

$$\inf_{(\alpha_i)_{i \in [n]} \in \mathcal{Y}^n} \sum_{i \in [n]} \left[\frac{1}{2} \|\alpha_i\|_{\mathcal{Y}}^2 - \langle \alpha_i, y_i \rangle_{\mathcal{Y}} + g^\star(\alpha_i) \right] + \frac{1}{2\lambda n} \sum_{i,j \in [n]} k_{\mathcal{X}}(x_i, x_j) \langle \alpha_i, T_{k_\Theta} \alpha_j \rangle_{\mathcal{Y}} . \tag{18}$$

A.2. Proof of Proposition 3.3

Before going through the proof, let us recall Hölder's inequality.

Lemma A.8 (Hölder's inequality)**.** *Let $p, q \in [1, +\infty]$ be conjugate exponents, in other words $\frac{1}{p} + \frac{1}{q} = 1$. Let Θ be a measurable space enriched with probability measure μ. Then for any $f, g : \Theta \to \mathbb{R}$ measurable functions one has*

$$\int_\Theta |f(\theta) g(\theta)| \, \mathrm{d}\mu(\theta) \leq \|f\|_p \|g\|_q .$$

Moreover, if $p \in]1, +\infty[$, $f \in L^p[\Theta, \mu]$ and $g \in L^q[\Theta, \mu]$, then equality is attained if and only if $|f|^p$ and $|g|^p$ are linearly dependent in $L^1[\Theta, \mu]$.

We now introduce a lemma useful to the proof of Proposition 3.3.

Lemma A.9. *Let $p, q \in]1, +\infty[$ be conjugate exponents and $f \in \mathcal{Y}$ such that $1 < \|f\|_q < +\infty$. Then there exist $h \in \mathcal{Y}$ and $C > 0$ such that*

$$\langle f, h \rangle_{\mathcal{Y}} - \|h\|_p \geq C.$$

Moreover, one can choose h such that whenever $f(\theta) = 0$, $h(\theta) = 0$ also holds.

Proof of Lemma A.9 Let $p, q \in]1, +\infty[$ be conjugate exponents and $f \in \mathcal{Y}$ such that $1 < \|f\|_q < +\infty$. We know that Hölder's inequality becomes an equality if and only if $|f|^q$ and $|g|^p$ are linearly dependent in $L^1[\Theta, \mu]$. To that end, let $g\colon \Theta \to \mathbb{R}$ be defined as

$$g(\theta) = \operatorname{sign}(f(\theta))\, |f(\theta)|^{\frac{q}{p}} \quad \text{where } \theta \in \Theta. \tag{19}$$

It is to be noted that g does not necessarily belong to $\mathcal{Y}$, yet it belongs to $L^p[\Theta, \mu]$. By construction, we have

$$\int_\Theta f(\theta) g(\theta) \mathrm{d}\mu(\theta) = \|f\|_q \|g\|_p . \tag{20}$$

We consider a sequence $(g_n)_{n\in\mathbb{N}} \in \mathcal{Y}^{\mathbb{N}}$ such that $g_n(\theta) = \operatorname{sign}(g(\theta)) \min(|g(\theta)|, n)$ with $(n, \theta) \in \mathbb{N} \times \Theta$. As $|g_n(\theta)| \leq n$ for all $(n, \theta) \in \mathbb{N} \times \Theta$ and μ is a probability measure, the functions g_n belong to $\mathcal{Y}$. Since (i) $g_n(\theta) \xrightarrow{n\to\infty} g(\theta)$ for all $\theta \in \Theta$ and (ii) $|g_n(\theta)| \leq |g(\theta)|$ for any $n \in \mathbb{N}$ holds μ-almost everywhere, the dominated convergence theorem in $L^p[\Theta, \mu]$ ensures that $\|g - g_n\|_p \xrightarrow{n\to\infty} 0$. Consequently, it holds that for all $n \in \mathbb{N}$,

$$\left| \int_\Theta f(\theta) g(\theta) \mathrm{d}\mu(\theta) - \int_\Theta f(\theta) g_n(\theta) \mathrm{d}\mu(\theta) \right| = \left| \int_\Theta f(\theta)[g(\theta) - g_n(\theta)] \mathrm{d}\mu(\theta) \right| \overset{(a)}{\leq} \int_\Theta |f(\theta)|\, |g(\theta) - g_n(\theta)|\, \mathrm{d}\mu(\theta)$$
$$\overset{(b)}{\leq} \|f\|_q \|g - g_n\|_p .$$

In (a) we used that the absolute value of the integral can be upper bounded by the integral of the absolute value, in (b) the Hölder's inequality was invoked. Thus by $\|g - g_n\|_p \xrightarrow{n\to\infty} 0$ and $\|f\|_q < +\infty$, this means that $\langle f, g_n\rangle_{\mathcal{Y}} \xrightarrow{n\to\infty} \int_\Theta f(\theta) g(\theta) \mathrm{d}\mu(\theta) \overset{(20)}{=} \|f\|_q \|g\|_p$, and for all $\epsilon > 0$, there exist $N \in \mathbb{N}$ such that for all $n \geq N$, $\langle f, g_n\rangle_{\mathcal{Y}} > (\|f\|_q - \epsilon) \|g\|_p$. In particular for $\epsilon = \frac{\|f\|_q - 1}{2} > 0$, we have $\langle f, g_N\rangle_{\mathcal{Y}} > \frac{1 + \|f\|_q}{2} \|g\|_p$. Then,

$$\langle f, g_N\rangle_{\mathcal{Y}} - \|g_N\|_p \overset{(c)}{\geq} \langle f, g_N\rangle_{\mathcal{Y}} - \|g\|_p \overset{(d)}{\geq} \frac{1 + \|f\|_q}{2} \|g\|_p - \|g\|_p \geq \underbrace{\frac{\|f\|_q - 1}{2} \|g\|_p}_{>0} .$$

In (c) we used that $\|g_N\|_p \leq \|g\|_p$, (d) is implied by $\langle f, g_N\rangle_{\mathcal{Y}} > \frac{1+\|f\|_q}{2} \|g\|_p$. Taking $h = g_N$ and $C = \frac{\|f\|_q - 1}{2} \|g\|_p$ yields the announced result, by noticing that (19) shows that $f(\theta) = 0$ also implies $h(\theta) = g_N(\theta) = g(\theta) = 0$. □

We are now ready to prove Proposition 3.3, which is the building block for dualizing optimization problems resulting from the generalized Huber and ϵ-insensitive losses whose definition can be found respectively in Definition 3.1 and Definition 4.1. The proposition is an extension of the well-studied finite-dimensional case to the space $\mathcal{Y}$.

Proposition (3.3). *Let $p, q \in [1, +\infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\|\cdot\|_p^\star = \iota_{\mathcal{B}_1^q}(\cdot). \tag{21}$$

Proof The proof is structured as follows. We first consider the case of $p = 1$, followed by $p \in]1, +\infty]$, and $p = +\infty$. The reasoning in all cases rely heavily on Hölder's inequality. Throughout the proof it is assumed that $f \in \mathcal{Y}$.

Case $p = 1$: The reasoning goes as follows: we show that $\|f\|_\infty \leq 1$ implies $\|\cdot\|_1^\star(f) = 0$, and $\|f\|_\infty > 1$ gives $\|\cdot\|_1^\star(f) = +\infty$, which allow one to conclude that $\|\cdot\|_1^\star = \iota_{\mathcal{B}_1^\infty}(\cdot)$.

- **When** $\|f\|_\infty \leq 1$: Exploiting Hölder's inequality, it holds that

  $$\langle f, g\rangle_{\mathcal{Y}} \leq \|f\|_\infty \|g\|_1 \quad \text{for all } g \in \mathcal{Y}.$$

  Since $\|f\|_\infty \leq 1$, this implies that

  $$\langle f, g\rangle_{\mathcal{Y}} - \|g\|_1 \leq 0 \quad \text{for all } g \in \mathcal{Y}.$$

  The supremum being attained for $g = 0$, we conclude that $\|\cdot\|_1^\star(f) = 0$.

- **When** $\|f\|_\infty > 1$: Let $A = \left\{\theta \in \Theta \,:\, |f(\theta)| > \frac{1+\|f\|_\infty}{2}\right\}$. By the definition of the essential supremum, $\mu(A) > 0$. We define $g\colon \Theta \to \mathbb{R}$ to be the function: $g(\theta) = \operatorname{sign}(f(\theta))$ if $\theta \in A$ and 0 otherwise. Since g is bounded, $g \in \mathcal{Y}$. Denoting by $t > 0$ a running parameter, it holds that

$$\begin{aligned}\langle f, tg\rangle_{\mathcal{Y}} - \|tg\|_1 &\overset{(a)}{=} \langle f, tg\rangle_{\mathcal{Y}} - t\mu(A) = t\int_\Theta f(\theta)g(\theta)\mathrm{d}\mu(\theta) - t\mu(A) \overset{(a)}{=} t\int_A |f(\theta)|\,\mathrm{d}\mu(\theta) - t\mu(A)\\ &\overset{(b)}{\geq} t\mu(A)\frac{1+\|f\|_\infty}{2} - t\mu(A) = t\underbrace{\mu(A)\frac{\|f\|_\infty - 1}{2}}_{>0} \xrightarrow{t\to\infty} +\infty.\end{aligned}$$

  In (a) we used the definition of g, (b) is implied by the fact that $|f(\theta)| > \frac{1+\|f\|_\infty}{2}$ for all $\theta \in A$. Thus $\|\cdot\|_1^\star(f) = +\infty$, which concludes the proof.

Case $p \in]1, +\infty[$**:** The reasoning proceeds as follows: we show that (i) $\|f\|_q \leq 1$ implies $\|\cdot\|_p^\star(f) = 0$, (ii) $1 < \|f\|_q < +\infty$ gives $\|\cdot\|_p^\star(f) = +\infty$, and (iii) $\|f\|_q = +\infty$ results in $\|\cdot\|_p^\star(f) = +\infty$. This allows us to conclude that $\|\cdot\|_p^\star = \iota_{\mathcal{B}_1^q}(\cdot)$.

- **When** $\|f\|_q \leq 1$: By Hölder's inequality, it holds that

$$\langle f, g\rangle_{\mathcal{Y}} \leq \|f\|_q \|g\|_p \text{ for all } g \in \mathcal{Y}.$$

  Exploiting $\|f\|_q \leq 1$, we get that

$$\langle f, g\rangle_{\mathcal{Y}} - \|g\|_p \leq 0 \text{ for all } g \in \mathcal{Y}.$$

  The supremum being reached for $g = 0$; we conclude that $\|\cdot\|_p^\star(f) = 0$.

- **When** $1 < \|f\|_q < +\infty$: According to Lemma A.9, there exist $g \in \mathcal{Y}$ and $C > 0$ such that

$$\langle f, g\rangle_{\mathcal{Y}} - \|g\|_p \geq C.$$

  Denoting by $t > 0$ a running parameter, one arrives at

$$\langle f, tg\rangle_{\mathcal{Y}} - \|tg\|_p \geq tC \xrightarrow{t\to\infty} +\infty.$$

  This shows that $\|\cdot\|_p^\star(f) = +\infty$.

- **When** $\|f\|_q = +\infty$: We consider the sequence of functions $(f_n)_{n\in\mathbb{N}}$ defined as $f_n(\theta) = f(\theta)$ if $|f(\theta)| \leq n$ and $f_n(\theta) = 0$ otherwise, where $(n, \theta) \in \mathbb{N} \times \Theta$. Each f_n is bounded, thus belongs to $L^q[\Theta, \mu]$, and the monotone convergence theorem applied to the functions $|f_n|^q$ states that $\|f_n\|_q \xrightarrow{n\to\infty} \|f\|_q = +\infty$. Thus, there exists $N \in \mathbb{N}$ such that $\|f_N\|_q > 1$. We can then apply Lemma A.9 to get $g \in \mathcal{Y}$ and $C > 0$ such that

$$\langle f_N, g\rangle_{\mathcal{Y}} - \|g\|_p \geq C.$$

  According to Lemma A.9, $g(\theta) = 0$ whenever $f_N(\theta) = 0$, which ensures that

$$\langle f, g\rangle_{\mathcal{Y}} = \langle f_N, g\rangle_{\mathcal{Y}}.$$

  Taking a running parameter $t > 0$, this means that

$$\langle f_N, tg\rangle_{\mathcal{Y}} - \|tg\|_p = \langle f, tg\rangle_{\mathcal{Y}} - \|tg\|_p \geq tC \xrightarrow{t\to\infty} +\infty,$$

  which shows that $\|\cdot\|^\star(f) = +\infty$.

Case $p = +\infty$: The reasoning goes as follows: we show that $\|f\|_1 \leq 1$ implies $\|\cdot\|_\infty^\star(f) = 0$, and that $\|f\|_1 > 1$ gives $\|\cdot\|_\infty^\star(f) = +\infty$, which allows one to conclude that $\|\cdot\|_\infty^\star = \iota_{\mathcal{B}_1^1}(\cdot)$.

- **When** $\|f\|_1 \leq 1$: By applying Hölder's inequality we get that $\langle f, g\rangle_{\mathcal{Y}} \leq \|f\|_1 \|g\|_\infty$ for all $g \in \mathcal{Y}$. Using the condition that $\|f\|_1 \leq 1$, this means that $\langle f, g\rangle_{\mathcal{Y}} - \|g\|_\infty \leq 0$ for all $g \in \mathcal{Y}$. Since the supremum is reached for $g = 0$, we get that $\|\cdot\|_\infty^\star(f) = 0$.
- **When** $\|f\|_1 > 1$: Let $g\colon \theta \mapsto \operatorname{sign}(f(\theta))$. Since g is bounded by 1, it belongs to $\mathcal{Y}$, and $\langle f, g\rangle_{\mathcal{Y}} = \|f\|_1$. Running a free parameter $t > 0$, this means that $\langle f, tg\rangle_{\mathcal{Y}} - t\|g\|_\infty = t\underbrace{(\|f\|_1 - 1)}_{>0} \xrightarrow{t\to\infty} +\infty$ which implies that $\|\cdot\|_\infty^\star(f) = +\infty$.

□

A.3. Proof of Proposition 3.2

Proposition (3.2). *Let $\kappa > 0$, $p \in [1, +\infty]$, and q the dual exponent of p (i.e., $\frac{1}{p} + \frac{1}{q} = 1$). Then for all $f \in \mathcal{Y}$,*

$$H_\kappa^p(f) = \begin{cases} \frac{1}{2}\|f\|_{\mathcal{Y}}^2 & \text{if } \|f\|_q \leq \kappa \\ \frac{1}{2}\left\|\operatorname{Proj}_{\mathcal{B}_\kappa^q}(f)\right\|_{\mathcal{Y}}^2 + \kappa\left\|f - \operatorname{Proj}_{\mathcal{B}_\kappa^q}(f)\right\|_p & \text{otherwise.} \end{cases}.$$

Proof Let us introduce the notation $R(g) = \frac{1}{2}\|f - g\|_{\mathcal{Y}}^2 + \kappa\|g\|_p$ where $f \in \mathcal{Y}$, $g \in \mathcal{Y}$. Then

$$H_\kappa^p(f) \overset{(a)}{=} \inf_{g\in\mathcal{Y}} R(g) \overset{(b)}{=} R(\operatorname{prox}_{\kappa\|\cdot\|_p}(f)) \overset{(c)}{=} \frac{1}{2}\left\|\operatorname{Proj}_{\mathcal{B}_\kappa^q}(f)\right\|_{\mathcal{Y}}^2 + \kappa\left\|f - \operatorname{Proj}_{\mathcal{B}_\kappa^q}(f)\right\|_p, \tag{22}$$

where (a) follows from the definition of the infimal convolution, (b) is implied by that of the proximal operator using that $\kappa\|\cdot\|_p \in \Gamma_0(\mathcal{Y})$. (c) is a consequence of the Moreau decomposition (Lemma A.6) as

$$\operatorname{prox}_{\kappa\|\cdot\|_p}(f) = f - \operatorname{prox}_{(\kappa\|\cdot\|_p)^\star}(f) \overset{(d)}{=} f - \operatorname{prox}_{\iota_{\mathcal{B}_\kappa^q}}(f) \overset{(e)}{=} f - \operatorname{Proj}_{\mathcal{B}_\kappa^q}(f), \tag{23}$$

where in (d) and (e) we used that

$$\left(\kappa\|\cdot\|_p\right)^\star \overset{(f)}{=} \iota_{\mathcal{B}_\kappa^q}(\cdot) \text{ with } \frac{1}{p} + \frac{1}{q} = 1, \tag{24}$$

$$\operatorname{prox}_{\iota_{\mathcal{B}_\kappa^q}} \overset{(g)}{=} \operatorname{Proj}_{\mathcal{B}_\kappa^q}. \tag{25}$$

(f) follows from the facts listed in the 3rd and the 2nd line of Table 5:

$$\left(\kappa\|\cdot\|_p\right)^\star = \kappa\left(\|\cdot\|_p\right)^\star(\cdot/\kappa) = \kappa\iota_{\mathcal{B}_1^q}(\cdot/\kappa) = \iota_{\mathcal{B}_\kappa^q}(\cdot).$$

(g) is implied by $\iota_{\mathcal{B}_\kappa^q} = \iota_{\mathcal{B}_1^q}(\cdot/\kappa)$, the precomposition rule of proximal operators ($\operatorname{prox}_{f(\alpha\cdot)} = \frac{1}{\alpha}\operatorname{prox}_{\alpha^2 f}(\alpha\cdot)$ holding for any $\alpha > 0$; see (2.2) in (Parikh & Boyd, 2014)), and $\operatorname{prox}_{\iota_{\mathcal{B}_1^q}} = \operatorname{Proj}_{\mathcal{B}_1^q}$:

$$\operatorname{prox}_{\iota_{\mathcal{B}_\kappa^q}} = \operatorname{prox}_{\iota_{\mathcal{B}_1^q}(\cdot/\kappa)} = \kappa\operatorname{prox}_{\frac{1}{\kappa^2}\iota_{\mathcal{B}_1^q}}(\cdot/\kappa) = \kappa\operatorname{prox}_{\iota_{\mathcal{B}_1^q}}(\cdot/\kappa) = \kappa\operatorname{Proj}_{\mathcal{B}_1^q}(\cdot/\kappa) = \operatorname{Proj}_{\mathcal{B}_\kappa^q}.$$

Finally we note that when $f = \operatorname{Proj}_{\mathcal{B}_\kappa^q}(f)$ ($\Leftrightarrow f \in \mathcal{B}_\kappa^q \Leftrightarrow \|f\|_q \leq \kappa$), (22) simplifies to $\frac{1}{2}\|f\|_{\mathcal{Y}}^2$. □

A.4. Proof of Proposition 3.4

Proposition (3.4). *Let $p, \kappa \in [1, +\infty]\times]0, +\infty[$ and $\frac{1}{p} + \frac{1}{q} = 1$. Then the dual of Problem 16 with loss H_κ^p writes as*

$$\inf_{(\alpha_i)_{i\in[n]}\in\mathcal{Y}^n} \sum_{i\in[n]} \frac{1}{2}\|\alpha_i\|_{\mathcal{Y}}^2 - \langle\alpha_i, y_i\rangle_{\mathcal{Y}} + \frac{1}{2\lambda n}\sum_{i,j\in[n]} k_{\mathcal{X}}(x_i, x_j)\langle\alpha_i, T_{k_\Theta}\alpha_j\rangle_{\mathcal{Y}} \text{ s.t. } \|\alpha_i\|_q \leq \kappa \text{ for } \forall i \in [n]. \tag{26}$$

Proof Applying Lemma A.7 and (24) give the result. □

A.5. Proof of Proposition 3.5

Proposition (3.5). *Let $\kappa > 0$. The projection on $\mathcal{B}_\kappa^q$ is tractable for $q = 2$ and $q = \infty$ and can be expressed for all $(\alpha, \theta) \in \mathcal{Y} \times \Theta$ as*

$$\mathrm{Proj}_{\mathcal{B}_\kappa^2}(\alpha) = \min\left(1, \frac{\kappa}{\|\alpha\|_2}\right)\alpha, \quad \text{if } \alpha \neq 0 \tag{27}$$

$$\left(\mathrm{Proj}_{\mathcal{B}_\kappa^\infty}(\alpha)\right)(\theta) = \mathrm{sign}\left(\alpha(\theta)\right)\min\left(\kappa, |\alpha(\theta)|\right). \tag{28}$$

Proof The projection on the 2-ball of radius κ is similar to the finite-dimensional case ($\mathcal{Y} = \mathbb{R}^d$) for which Equation (27) is well-known.

We now turn to the case of $q = \infty$. Let $\alpha \in \mathcal{Y}$. By definition,

$$\mathrm{Proj}_{\mathcal{B}_\kappa^\infty}(\alpha) = \operatorname*{arg\,min}_{y \in \mathcal{Y}} \frac{1}{2}\int_\Theta \left[\alpha(\theta) - y(\theta)\right]^2 \mathrm{d}\mu(\theta) + \iota_{\mathcal{B}_\kappa^\infty}(y). \tag{29}$$

Since $\alpha \in \mathcal{Y}$, the function g defined as

$$g(\theta) = \mathrm{sign}\left(\alpha(\theta)\right)\min\left(\kappa, |\alpha(\theta)|\right), \quad \theta \in \Theta,$$

is measurable, it is in $\mathcal{Y}$, and one can verify easily that it is the solution of Equation (29); it corresponds to taking the pointwise projection of α on the segment $[-\kappa, \kappa]$. □

A.6. Proof of Proposition 4.2

Proposition (4.2). *Let $\epsilon > 0$ and $p \in [1, +\infty]$. Then $\ell_\epsilon^p(f) = \frac{1}{2}\left\|f - \mathrm{Proj}_{\mathcal{B}_\epsilon^p}(f)\right\|_{\mathcal{Y}}^2$ for all $f \in \mathcal{Y}$.*

Proof Let $R(g) = \frac{1}{2}\|f - g\|_{\mathcal{Y}}^2 + \iota_{\mathcal{B}_\epsilon^p}(g)$ where $f \in \mathcal{Y}$ and $g \in \mathcal{Y}$. Then

$$\ell_\epsilon^p(f) \overset{(a)}{=} \inf_{g \in \mathcal{Y}} R(g) \overset{(b)}{=} R\left(\mathrm{prox}_{\iota_{\mathcal{B}_\epsilon^p}}(f)\right) \overset{(c)}{=} R\left(\mathrm{Proj}_{\mathcal{B}_\epsilon^p}(f)\right) \overset{(d)}{=} \frac{1}{2}\left\|f - \mathrm{Proj}_{\mathcal{B}_\epsilon^p}(f)\right\|_{\mathcal{Y}}^2,$$

where (a) follows from the definition of the infimal convolution, (b) is implied by that of the proximal operator and by $\iota_{\mathcal{B}_\epsilon^p} \in \Gamma_0(\mathcal{Y})$, (c) is the consequence of $\mathrm{prox}_{\iota_{\mathcal{B}_\epsilon^p}}(f) = \mathrm{Proj}_{\mathcal{B}_\epsilon^p}(f)$ implied by (25), in (d) the definition of R was applied. □

A.7. Proof of Proposition 4.3

Proposition (4.3). *Let $(p, \epsilon) \in [1, +\infty] \times]0, +\infty[$, and $\frac{1}{p} + \frac{1}{q} = 1$. Then, the dual of Problem 16 writes as*

$$\inf_{(\alpha_i)_{i\in[n]} \in \mathcal{Y}^n} \sum_{i \in [n]} \left[\frac{1}{2}\|\alpha_i\|_{\mathcal{Y}}^2 - \langle \alpha_i, y_i \rangle_{\mathcal{Y}} + \epsilon \|\alpha_i\|_q\right] + \frac{1}{2\lambda n} \sum_{i,j \in [n]} k_{\mathcal{X}}(x_i, x_j)\left\langle \alpha_i, T_{k_\Theta}\alpha_j \right\rangle_{\mathcal{Y}}. \tag{30}$$

Proof Applying Lemma A.7 with

$$\left(\iota_{\mathcal{B}_\epsilon^p}(\cdot)\right)^\star = \epsilon \|\cdot\|_q \tag{31}$$

gives the result. (31) is the consequence of (24) and the involutive property of the Fenchel-Legendre conjugate. □

A.8. Proof of Proposition 4.4

Proposition (4.4). *Let $\epsilon > 0$. The proximal operator of $\epsilon\|\cdot\|_q$ is computable for $q = 1$ and $q = 2$, and given for all $(\alpha, \theta) \in \mathcal{Y} \times \Theta$ by*

$$\left(\mathrm{prox}_{\epsilon\|\cdot\|_1}(\alpha)\right)(\theta) = \mathrm{sign}\left(\alpha(\theta)\right)\left||\alpha(\theta)| - \epsilon\right|_+, \tag{32}$$

$$\mathrm{prox}_{\epsilon\|\cdot\|_2}(\alpha) = \alpha\left|1 - \frac{\epsilon}{\|\alpha\|_{\mathcal{Y}}}\right|_+ \quad \text{if } \alpha \neq 0. \tag{33}$$

Algorithm 2 APGD with eigenbasis representation

input : Gram matrix $\mathbf{K}_{\mathcal{X}}$, matrix of eigenvalues $\mathbf{\Delta}$, data scalar product matrix $\mathbf{R}$, regularization parameter λ, loss parameters $(\kappa, 2)$ or $(\epsilon, 2)$, gradient step γ

init : $\mathbf{A}^{(0)}, \mathbf{A}^{(-1)} = \mathbf{0} \in \mathbb{R}^{n \times r}$

for *epoch* t *from* 0 *to* $T-1$ **do**

  // gradient step

  $\mathbf{V} = \mathbf{A}^{(t)} + \frac{t-2}{t+1}\left(\mathbf{A}^{(t)} - \mathbf{A}^{(t-1)}\right)$

  $\mathbf{U} = \mathbf{V} - \gamma\left(\mathbf{V} + \frac{1}{\lambda n}\mathbf{K}_{\mathcal{X}}\mathbf{V}\mathbf{\Delta} - \mathbf{R}\right)$

  // proximal step

  for *row* $i \in [n]$ **do**

    $\mathbf{a}_i^{(t+1)} = \min\left(\frac{\kappa}{\|\mathbf{u}_i\|_2}, 1\right)\mathbf{u}_i$ // if H_κ^2

    $\mathbf{a}_i^{(t+1)} = \left|1 - \frac{\gamma\epsilon}{\|\mathbf{u}_i\|_2}\right|_+ \mathbf{u}_i$ // if ℓ_ϵ^2

return $\mathbf{A}^{(T)}$

Proof By (23) we know that

$$\operatorname{prox}_{\epsilon\|\cdot\|_q}(f) = f - \operatorname{Proj}_{\mathcal{B}_\epsilon^p}(f). \tag{34}$$

The projection operator is known from Proposition 3.5 in the case of $p = 2$ and $p = \infty$, which allows to express the proximal operator of the q-norm for $q = 2$ and $q = 1$ and by substituting Equation (27) and Equation (28) into Equation (34). □

B. Additional Details

In this section we present additional algorithmic details as well as complement the numerical experiments presented in the main document.

B.1. Algorithmic Details

Algorithm 2 fully describes how to learn models with the representation relying on the eigendecomposition of the integral operator developed in Section 3.2 and Section 4.2.

B.2. Synthetic Data

Below we detail the generation process of the synthetic dataset (Section B.2.1), we expose in full detail the parameters used in the experiments (Section B.2.2; see Fig. 1 and 2 in the main paper), and we provide additional illustration for the interaction between the Huber loss' κ and the regularization parameter λ (Section B.2.3).

B.2.1. Generation Process

Given covariance parameters $(\boldsymbol{\sigma}^{\text{in}}, \boldsymbol{\sigma}^{\text{out}}) \in \mathbb{R}^r \times \mathbb{R}^r$ for $c \in [r]$ we draw and fix Gaussian processes $g_c^{\text{in}} \sim \mathcal{GP}(0, k_{\sigma_c^{\text{in}}})$ and $g_c^{\text{out}} \sim \mathcal{GP}(0, k_{\sigma_c^{\text{out}}})$. We then generate n samples as $\left(\sum_{c\in[r]} u_{ic} g_c^{\text{in}}, \sum_{c\in[r]} u_{ic} g_c^{\text{out}}\right)_{i\in[n]}$, where the coefficients u_{ic} are drawn i.i.d. according to a uniform distribution $\mathcal{U}([-0.5, 0.5])$. In the experiments, we take $r = 4$ and set $\boldsymbol{\sigma}^{\text{in}} = \boldsymbol{\sigma}^{\text{out}} = (0.05, 0.1, 0.5, 0.7)$. We show input and output functions drawn in this manner in the first and second row of Fig. 3. In the bottom row we display outliers of Type 2 with $\boldsymbol{\sigma} = (0.01, 0.05, 1, 4)$ and intensity $\zeta = 2$. For the contaminated indices i in I we add the corresponding outlier to the function y_i

B.2.2. Experimental Details

We provide here the full details of the parameters used for the experiments on the toy dataset. For all experiments, we fix the parameter ρ^{in} of the input Gaussian kernel $k_{\mathcal{X}} : (x_1, x_2) \longmapsto \exp\left(-\rho\|x_0 - x_1\|_{\mathcal{X}}^2\right)$ to $\rho^{\text{in}} = 0.01$ and that of the output Gaussian kernel to $\rho^{\text{out}} = 100$. Indeed, since we are only given discrete observations for the input functions as well, we use the available observations to approximate the norms in the above kernels. For the experiments on robustness which results are displayed in Fig. 2 of the main paper, we select via cross-validation the regularization parameter λ and the κ parameters of the Huber loss, considering values in a geometric grid of size 10 ranging from 10^{-6} to 10^{-3} for λ and values

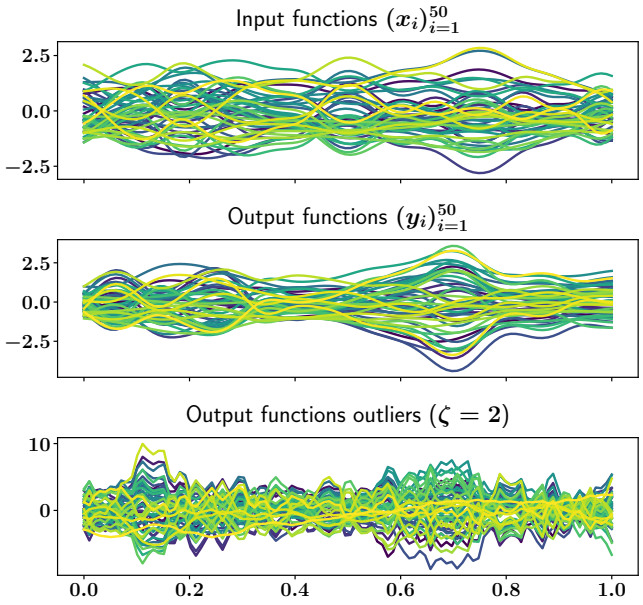


Figure 3: Examples from the toy dataset and corresponding type 2 outliers.

in a geometric grid of size 25 ranging from 10^{-3} to 10^{-1} for κ.

B.2.3. Additional Illustrations

To highlight the interaction between the regularization parameter λ and the parameter κ of the Huber loss, we plot the NMSE values for various values of λ and κ using the toy dataset corrupted with the two main types of contamination used in the main paper, Type 2 and Type 3 outliers. The results are displayed in Fig. 4 and confirm that by making κ and λ vary, when the data are corrupted, we can always find a configuration that is significantly more robust than the square loss. In accordance with one's expectation, when dealing with local outliers (Fig. 4b), the loss H_κ^1 is much more efficient than the loss H_κ^2. However, when dealing with global outliers (Fig. 4a), the two losses perform equally well.

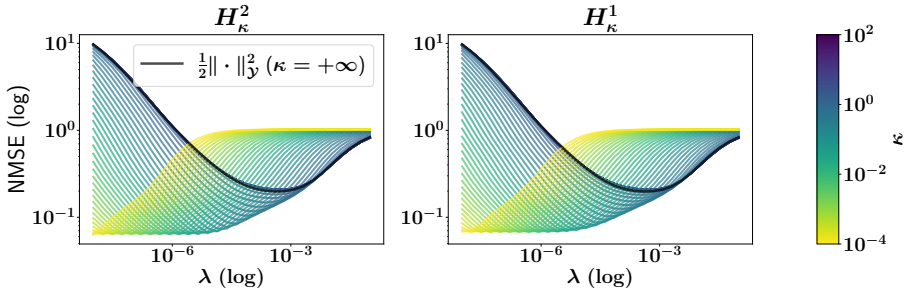
B.3. DTI Data

In this section we provide details regarding the experiments on the DTI dataset. For this dataset, we use a Gaussian kernel as input kernel and a Laplace kernel as output kernel, for the first we fix its parameter to $\rho^{\text{in}} = 1.25$, and for the second, defined as $k_\Theta : (\theta_1, \theta_2) \longmapsto \exp(-\rho^{\text{out}} \|x_0 - x_1\|_{\mathcal{X}})$, we fix its parameter to $\rho^{\text{out}} = 10$. We consider two values of λ, the first one ($\lambda = 10^{-5}$) is chosen too small for the square loss to highlight the additional sparsity-inducing regularization possibilities offered by the ϵ-insensitive loss through the parameter ϵ, while the second one ($\lambda = 10^{-3}$) corresponds to a near-optimal value for the square loss. We do cross-validate the parameters of the losses. For the loss ℓ_ϵ^2 we consider values of ϵ in a geometric grid of size 50 ranging from 10^{-3} to 10^{-1}, while for the loss ℓ_ϵ^∞, we search in a geometric grid of the same size, however ranging from 10^{-3} to $10^{-0.5}$. For the Huber losses H_κ^1 and H_κ^2, we search for κ using a geometric grid of size 50 ranging this time from 10^{-4} to 10^{-1}.

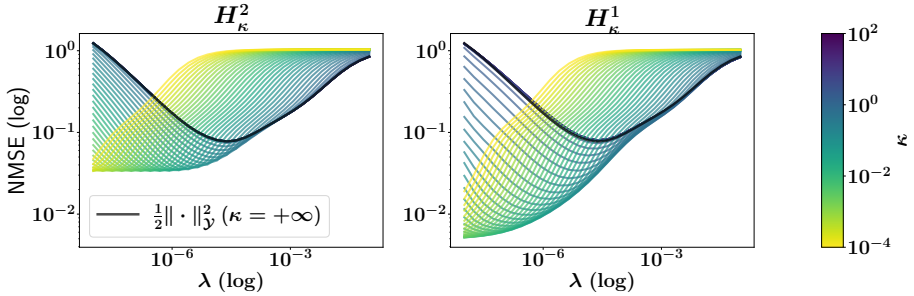
B.4. Speech Data

This section is dedicated to additional details about the experiments carried out on the speech benchmark.

Input kernel: As highlighted in the main paper, we encode the input sounds through 13 mel-frequency cepstral coefficients (MFCC). To deal with this particular input data type we used the following kernel. Let $((x_{ij}^{(v)})_{v\in[13]})_{i\in[n],j\in[m]}$ be the transformed input data where v serves as an index for the MFCC number. The number of locations m is the same for all $i \in [n]$ since we extend the signals to match the longest one to be able to train the models. We then center and reduce each MFCC using all samples and sampling locations to compute the mean and standard deviation; let $((\tilde{x}_{ij}^{(v)})_{v\in[13]})_{i\in[n],j\in[m]}$



(a) Type 2 outliers with $\tau = 0.2$ and $\zeta = 2$



(b) Type 3 outliers with $\tau = 0.3$ and $\xi = 0.3$

Figure 4: NMSE as a function of λ and Huber losses' κ with two types of outliers.

be the resulting standardized input data. Denoting by $\tilde{x}_i$ the element $((\tilde{x}_{ij}^{(v)})_{v\in[13]})_{j\in[m]}$, we then use the following kernel:

$$k_{\mathcal{X}} : (\tilde{x}_0, \tilde{x}_1) \longmapsto \frac{1}{m}\sum_{j\in[m]} \exp\left(-\rho^{\text{in}} \sum_{v\in[13]} \left(\tilde{x}_{0j}^{(v)} - \tilde{x}_{1j}^{(v)}\right)^2\right).$$

Experimental details: For all the experiments (with or without corruption), we select the parameter of the input kernel ρ^{in}, the regularization parameter and the parameters of the losses using cross-validation. We fix the parameter of the Laplace output kernel to $\rho^{\text{out}} = 10$. However, to reduce the computational burden, we perform the selection of the parameter ρ^{in} only for the square loss, and then take the corresponding values for the other losses. For this parameter values in a geometric grid of size 15 ranging from 10^{-2} to $10^{-0.5}$ are considered. For λ, the search space is a geometric grid of size 10 ranging from 10^{-10} to 10^{-6}. Finally, for the ϵ-insensitive loss, values of ϵ in a geometric grid of size 80 ranging from 10^{-5} to 10^{-1} are considered, while for the Huber losses we search for κ in a geometric grid of size 100 ranging from 10^{-7} to 1.

C. Illustration of Loss Functions

In this section we illustrate the differences between our proposed convoluted losses in several ways. In Section C.1 we study empirically how the choice of p affects the sensitivity of the Huber loss H^p_κ to different kind of outliers. In Section C.2, we plot some of our proposed losses when they are defined either on $\mathbb{R}$ or $\mathbb{R}^2$.

C.1. Discussion on the Choice of p for H^p_κ

As it is highlighted in the main paper, solving Problem (8) for $p \notin \{1, 2\}$ is unpractical since it involves the computation of n projections on a q-ball at each APGD iteration. Performing such projection is feasible (it is a convex optimization problem) but it has to be done in an iterative way. In our case, to run APGD with such inner iterations turns out to be too time consuming. However we still can approximately calculate the losses H^p_κ using Proposition 3.2 for any p (computing the involved projection iteratively). We thus propose to leverage this possibility to study empirically the sensitivity of the Huber losses H^p_κ to global and local outliers, for different values of p.

The impact of the outliers on the solution of a regularized empirical risk minimization problem is partly determined by the

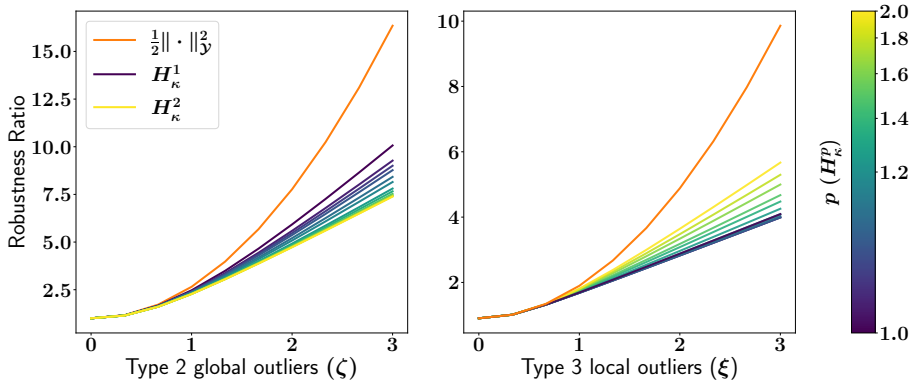


Figure 5: Sensitivity of H^p_κ to outliers for various $p \in [1, 2]$

contribution of the outliers to the data-fitting term relatively to the contribution of the normal observations. In order to investigate this aspect, we study and define next a quantity which we call Robustness Ratio.

Let $(e_i)_{i\in[n]} \in (L^2[\Theta, \mu])^n$ be a set of functional residuals and let $(\tilde{e}_i)_{i\in[n]}$ be the same functional residuals but contaminated with outliers. In practice, we have to choose a probability distribution to draw the functions $(e_i)_{i\in[n]}$ from, and an outlier distribution to corrupt those. For the functions $(e_i)_{i\in[n]}$ we use our synthetic data generation process (see Section B.2.1), and for the outliers, we consider the same type 2 and type 3 outliers as in the experiments in Section 5.1 from the main paper. We then define the Robustness Ratio as

$$\text{Robustness Ratio} := \inf_{\kappa\geq 0} \frac{1}{n} \sum_{i\in[n]} \frac{H^p_\kappa(\tilde{e}_i)}{H^p_\kappa(e_i)}.$$

The best value of this quantity is 1; it means that the loss is not affected at all by the outliers, but it is indeed not possible to reach such value. In practice, we restrain our study to $p \in [1, 2]$. For each p we reduce the search for κ to different empirical quantiles of the q-norms of the uncorrupted functions $(e_i)_{i\in[n]}$, where q is the dual exponent of p. It makes sense to do so since κ corresponds to a q-norm threshold which separates observations considered to be outliers from those deemed normal (see Proposition 3.2). We consider the $\{0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$-th such empirical quantiles. For each p, we compute the robustness ratio for κ equal to each of those quantiles, and then for each level of corruption, we select the value which minimizes the ratio. This indeed corresponds to an ideal setting, since in practice, we never have access to the uncorrupted data and we can never optimize κ in this way. Thus the robustness ratio reflects more of a general robustness property of the loss in an optimal setting.

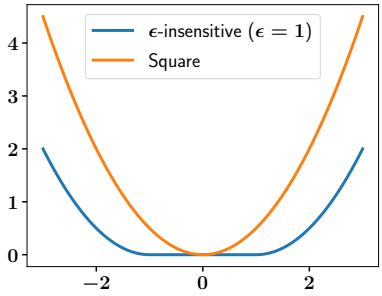
In accordance with one's expectation, when the data is contaminated with global outliers (left panel of Fig. 5), it is better to choose $p = 2$ whereas when the contamination is local (right panel of Fig. 5), $p = 1$ is almost the best choice; even though it seems that choosing p slightly bigger than 1 could be a tad better. Even though, we highlight that this analysis based on the Robustness Ratio has its limits; indeed we do not take into account the interplay between the data-fitting term and the regularization term which takes place during optimization. This certainly explains why we found the losses H^1_κ and H^2_κ to perform equally well in practice whereas based only on the Robustness Ratio analysis (left panel of Fig. 5) we would have said otherwise. The findings in presence of local outliers (right panel of Fig . 5) are nevertheless coherent with what we observed in practice for the losses H^1_κ and H^2_κ in our experiments.

C.2. Loss Examples in 1d and 2d

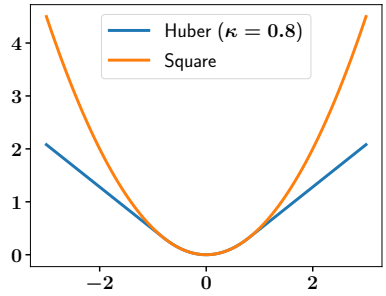
In this section, we plot several of the proposed losses when they are defined on $\mathbb{R}$ and $\mathbb{R}^2$. In Fig. 6, we compare the Huber (Fig. 6b) and the ϵ-insensitive (Fig. 6a) losses with the square loss when they are defined on $\mathbb{R}$.

Then in Fig. 7 we highlight the influence of p on the shape of the ϵ-insensitive loss ℓ^p_ϵ defined on $\mathbb{R}^2$. We set $\epsilon = 1$ and consider values of $p \in \{1.01, 1.5, 2, 3, 5, +\infty\}$. We display $\ell^{1.01}_\epsilon$ in Fig. 7a, $\ell^{1.5}_\epsilon$ in Fig. 7b, ℓ^2_ϵ in Fig. 7c, ℓ^3_ϵ in Fig. 7d, ℓ^5_ϵ in Fig. 7e and ℓ^∞_ϵ in Fig. 7f.

Finally, in Fig. 8 we underline the influence that the parameter p has on our proposed Huber losses when it is defined on $\mathbb{R}^2$;



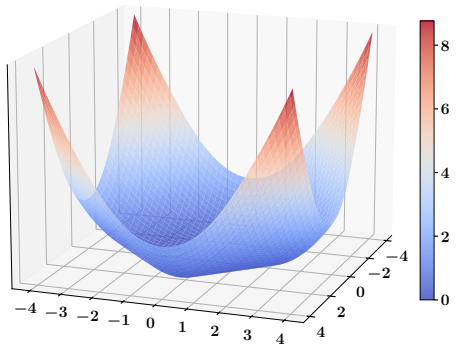
(a) ϵ-insensitive ($\epsilon = 1$) and square loss



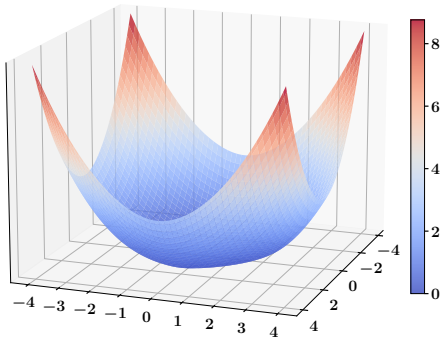
(b) Huber loss ($\kappa = 0.8$) and square loss

Figure 6: Illustrations of the different losses defined on $\mathbb{R}$.

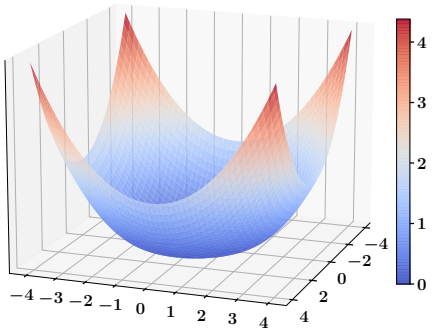
we take $\kappa = 0.8$ and we display H_κ^2 in Fig. 8a, $H_\kappa^{1.5}$ in Fig. 8b, $H_\kappa^{1.25}$ in Fig. 8c and H_κ^1 in Fig. 8d.



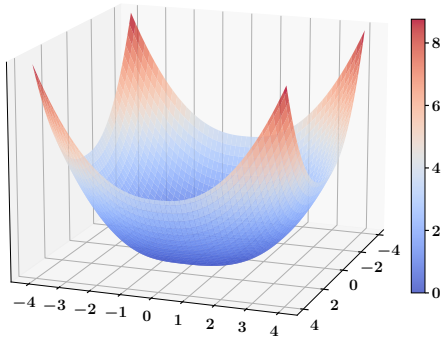
(a) $\ell_\epsilon^{1.01}$ $(\epsilon = 1)$



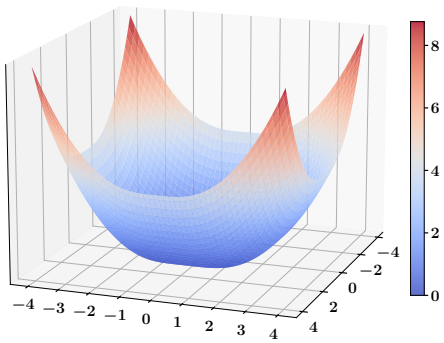
(b) $\ell_\epsilon^{1.5}$ $(\epsilon = 1)$



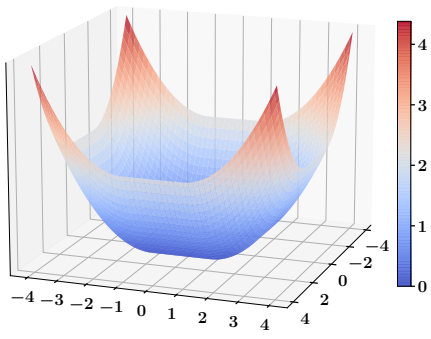
(c) ℓ_ϵ^{2} $(\epsilon = 1)$



(d) ℓ_ϵ^{3} $(\epsilon = 1)$

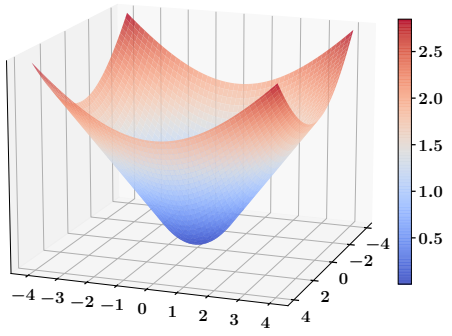


(e) ℓ_ϵ^{5} $(\epsilon = 1)$

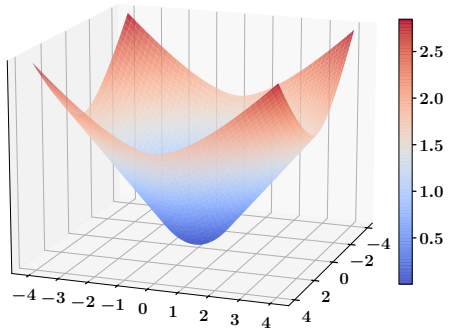


(f) ℓ_ϵ^{∞} $(\epsilon = 1)$

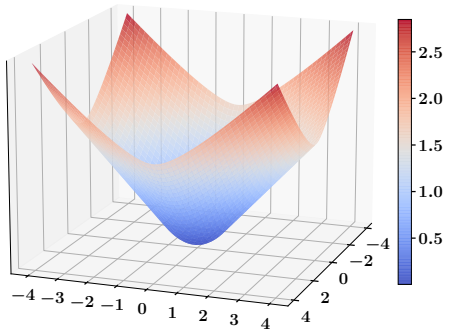
Figure 7: Examples of the proposed ϵ-insensitive losses defined on $\mathbb{R}^2$ for different values of p.



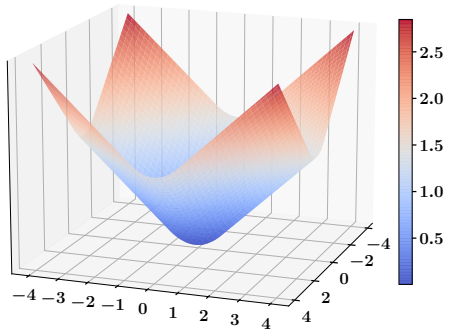
(a) H_κ^2 ($\kappa = 0.8$)



(b) $H_\kappa^{1.5}$ ($\kappa = 0.8$)



(c) $H_\kappa^{1.25}$ ($\kappa = 0.8$)



(d) H_κ^1 ($\kappa = 0.8$)

Figure 8: Examples of the proposed Huber losses defined on $\mathbb{R}^2$ for different values of p.