The Supplementary Material is organized as follows. Appendix A collects the technical proofs of the core article's results. Appendix B provides illustrations of the main loss functions considered (ϵ-insensitive Ridge and SVR, κ-Huber) in 1 and 2 dimensions. Appendix C gathers additional details about the experimental protocols and the code furnished.

A. Technical Proofs

A.1. Proof of Theorem 3

First, notice that the primal problem

$$\min_{h\in\mathcal{H}_\mathcal{K}} \quad \frac{1}{n}\sum_{i=1}^{n}\ell(h(x_i),y_i)+\frac{\Lambda}{2}\|h\|^2_{\mathcal{H}_\mathcal{K}}$$

can be rewritten

$$\begin{aligned}\min_{h\in\mathcal{H}_\mathcal{K}} \quad & \sum_{i=1}^{n}\ell_i(u_i)+\frac{\Lambda n}{2}\|h\|^2_{\mathcal{H}_\mathcal{K}},\\ \text{s.t.} \quad & u_i = h(x_i) \quad \forall i\leqslant n.\end{aligned}$$

Therefore, with the notation $\boldsymbol{u}=(u_i)_{i\leqslant n}$ and $\boldsymbol{\alpha}=(\alpha_i)_{i\leqslant n}$, the Lagrangian writes

$$\begin{aligned}\mathscr{L}(h,\boldsymbol{u},\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\ell_i(u_i)+\frac{\Lambda n}{2}\|h\|^2_{\mathcal{H}_\mathcal{K}}+\sum_{i=1}^{n}\langle\alpha_i,u_i-h(x_i)\rangle_\mathcal{Y}\,,\\ &= \sum_{i=1}^{n}\ell_i(u_i)+\frac{\Lambda n}{2}\|h\|^2_{\mathcal{H}_\mathcal{K}}+\sum_{i=1}^{n}\langle\alpha_i,u_i\rangle_\mathcal{Y}-\sum_{i=1}^{n}\langle\mathcal{K}(\cdot,x_i)\alpha_i,h\rangle_{\mathcal{H}_\mathcal{K}}\,.\end{aligned}$$

Differentiating with respect to h and using the definition of the Fenchel-Legendre transform, one gets

$$\begin{aligned}g(\boldsymbol{\alpha}) &= \inf_{h\in\mathcal{H}_\mathcal{K},\boldsymbol{u}\in\mathcal{Y}^n}\mathscr{L}(h,\boldsymbol{u},\boldsymbol{\alpha}),\\ &= \sum_{i=1}^{n}\inf_{u_i\in\mathcal{Y}}\left\{\ell_i(u_i)+\langle\alpha_i,u_i\rangle_\mathcal{Y}\right\}+\inf_{h\in\mathcal{H}_\mathcal{K}}\left\{\frac{\Lambda n}{2}\|h\|^2_{\mathcal{H}_\mathcal{K}}-\sum_{i=1}^{n}\langle\mathcal{K}(\cdot,x_i)\alpha_i,h\rangle_{\mathcal{H}_\mathcal{K}}\right\},\\ &= \sum_{i=1}^{n}-\ell_i^\star(-\alpha_i)-\frac{1}{2\Lambda n}\sum_{i,j=1}^{n}\langle\alpha_i,\mathcal{K}(x_i,x_j)\alpha_j\rangle_\mathcal{Y}\,,\end{aligned}$$

together with the equality $\hat{h}=\frac{1}{\Lambda n}\sum_{i=1}^{n}\mathcal{K}(\cdot,x_i)\alpha_i$. The conclusion follows immediately. □

A.2. Proof of Theorem 4

As a reminder, our goal is to compute the solutions to the following problem:

$$\hat{h}\in\operatorname*{argmin}_{h\in\mathcal{H}_\mathcal{K}}\ \frac{1}{n}\sum_{i=1}^{n}\ell(h(x_i),y_i)+\frac{\Lambda}{2}\|h\|^2_{\mathcal{H}_\mathcal{K}}.$$

Using Theorem 3, one gets that $\hat{h}=\frac{1}{\Lambda n}\sum_{i=1}^{n}\mathcal{K}(\cdot,x_i)\hat{\alpha}_i$, with the $(\hat{\alpha}_i)_{i\leqslant n}$ satisfying:

$$(\hat{\alpha}_i)_{i=1}^n\in\operatorname*{argmin}_{(\alpha_i)_{i=1}^n\in\mathcal{Y}^n}\ \sum_{i=1}^{n}\ell_i^\star(-\alpha_i)+\frac{1}{2\Lambda n}\sum_{i,j=1}^{n}\langle\alpha_i,\mathcal{K}(x_i,x_j)\alpha_j\rangle_\mathcal{Y}\,.$$

However, this optimization problem cannot be solved in a straightforward manner, as $\mathcal{Y}$ is in general infinite dimensional. Nevertheless, it is possible to bypass this difficulty by noticing that the optimal $(\hat{\alpha}_i)_{i\leqslant n}$ actually lie in $\mathbf{Y}^n$. To show this, we decompose each coefficient as $\hat{\alpha}_i=\alpha_i^{\mathbf{Y}}+\alpha_i^{\perp}$, with $(\alpha_i^{\mathbf{Y}})_{i\leqslant n},(\alpha_i^{\perp})_{i\leqslant n}\in\mathbf{Y}^n\times{\mathbf{Y}^\perp}^n$. Then, noticing that non-null $(\alpha_i^{\perp})_{i\leqslant n}$ necessarily increase the objective, we can conclude that the optimal $(\hat{\alpha}_i)_{i\leqslant n}$ have no components among $\mathbf{Y}^\perp$, or equivalently pertain to $\mathbf{Y}$. Indeed, by virtue of Assumptions 1 and 3, it holds:

$$\sum_{i=1}^{n} \ell_i^\star(-\alpha_i^{\mathbf{Y}}) + \frac{1}{2\Lambda n} \sum_{i,j=1}^{n} \left\langle \alpha_i^{\mathbf{Y}}, \mathcal{K}(x_i,x_j)\alpha_j^{\mathbf{Y}} \right\rangle_{\mathcal{Y}} \leqslant \sum_{i=1}^{n} \ell_i^\star(-\alpha_i^{\mathbf{Y}} - \alpha_i^{\perp}) + \frac{1}{2\Lambda n} \sum_{i,j=1}^{n} \left\langle \alpha_i^{\mathbf{Y}} + \alpha_i^{\perp}, \mathcal{K}(x_i,x_j)(\alpha_j^{\mathbf{Y}} + \alpha_j^{\perp}) \right\rangle_{\mathcal{Y}}.$$

If the inequality about $\ell_i^\star$ follows directly Assumption 1, that about $\mathcal{K}(x_i,x_j)$ can be obtained by Assumption 3 as follows:

$$\begin{aligned}
&\sum_{i,j=1}^{n} \left\langle \alpha_i^{\mathbf{Y}} + \alpha_i^{\perp}, \mathcal{K}(x_i,x_j)(\alpha_j^{\mathbf{Y}} + \alpha_j^{\perp}) \right\rangle_{\mathcal{Y}} \\
&\qquad = \sum_{i,j=1}^{n} \left\langle \alpha_i^{\mathbf{Y}}, \mathcal{K}(x_i,x_j)\alpha_j^{\mathbf{Y}} \right\rangle_{\mathcal{Y}} + 2\sum_{i,j=1}^{n} \left\langle \alpha_i^{\perp}, \mathcal{K}(x_i,x_j)\alpha_j^{\mathbf{Y}} \right\rangle_{\mathcal{Y}} + \sum_{i,j=1}^{n} \left\langle \alpha_i^{\perp}, \mathcal{K}(x_i,x_j)\alpha_j^{\perp} \right\rangle_{\mathcal{Y}}, \\
&\qquad = \sum_{i,j=1}^{n} \left\langle \alpha_i^{\mathbf{Y}}, \mathcal{K}(x_i,x_j)\alpha_j^{\mathbf{Y}} \right\rangle_{\mathcal{Y}} + \sum_{i,j=1}^{n} \left\langle \alpha_i^{\perp}, \mathcal{K}(x_i,x_j)\alpha_j^{\perp} \right\rangle_{\mathcal{Y}}, \\
&\qquad \geqslant \sum_{i,j=1}^{n} \left\langle \alpha_i^{\mathbf{Y}}, \mathcal{K}(x_i,x_j)\alpha_j^{\mathbf{Y}} \right\rangle_{\mathcal{Y}},
\end{aligned}$$

where we have used successively Assumption 3 and the positiveness of $\mathcal{K}$. So there exists $\Omega = [\omega_{ij}]_{1\leqslant i,j\leqslant n} \in \mathbb{R}^{n\times n}$ such that for all $i \leqslant n$, $\hat{\alpha}_i = \sum_j \omega_{ij}\, y_j$. This proof technique is very similar in spirit to that of the Representer Theorem, and yields an analogous result, the reduction of the search space to a smaller vector space, as discussed at length in the main text. The dual optimization problem thus rewrites:

$$\begin{aligned}
&\min_{\Omega\in\mathbb{R}^{n\times n}} \sum_{i=1}^{n} \ell_i^\star\left(-\sum_{j=1}^{n} \omega_{ij}\, y_j\right) + \frac{1}{2\Lambda n} \sum_{i,j=1}^{n} \left\langle \sum_{k=1}^{n} \omega_{ik}\, y_k, \mathcal{K}(x_i,x_j) \sum_{l=1}^{n} \omega_{jl}\, y_l \right\rangle_{\mathcal{Y}} \\
&\min_{\Omega\in\mathbb{R}^{n\times n}} \sum_{i=1}^{n} L_i\left((\omega_{ij})_{j\leqslant n}, K^Y\right) + \frac{1}{2\Lambda n} \sum_{i,j,k,l=1}^{n} \omega_{ik}\,\omega_{jl} \left\langle y_k, \sum_{t=1}^{T} k_t(x_i,x_j) A_t y_l \right\rangle_{\mathcal{Y}}, \\
&\min_{\Omega\in\mathbb{R}^{n\times n}} \sum_{i=1}^{n} L_i\left(\Omega_{i:}, K^Y\right) + \frac{1}{2\Lambda n} \mathbf{Tr}\left(\tilde{M}^\top(\Omega\otimes\Omega)\right),
\end{aligned} \tag{12}$$

with M the $n\times n\times n\times n$ tensor such that $M_{ijkl} = \langle y_k, \mathcal{K}(x_i,x_j) y_l\rangle_{\mathcal{Y}}$, and $\tilde{M}$ its rewriting as a $n^2\times n^2$ block matrix such that its (i,j) block is the $n\times n$ matrix with elements $\tilde{M}_{st}^{(i,j)} = \langle y_j, \mathcal{K}(x_i,x_s) y_t\rangle_{\mathcal{Y}}$.

The second term is quadratic in Ω, and consequently convex. As for the L_i's, they are basically rewritings of the Fenchel-Legendre transforms $\ell_i^\star$'s that ensure the computability of the problem (they only depend on K^Y, which is known). Regarding their convexity, we know by definition that the $\ell_i^\star$'s are convex. Composing by a linear function preserving the convexity, we know that each L_i is convex with respect to $\Omega_{i:}$, and therefore with respect to Ω.

Thus, we have first converted the infinite dimensional primal problem in $\mathcal{H}_\mathcal{K}$ into an infinite dimensional dual problem in $\mathcal{Y}^n$, which in turn is reduced to a convex optimization procedure over $\mathbb{R}^{n\times n}$, that only involves computable quantities.

If $\mathcal{K}$ satisfies Assumption 4, the tensor M simplifies to

$$M_{ijkl} = \langle y_k, \mathcal{K}(x_i,x_j) y_l\rangle_{\mathcal{Y}} = \sum_{t=1}^{T} k_t(x_i,x_j)\langle y_k, A_t y_l\rangle_{\mathcal{Y}} = \sum_{t=1}^{T} [K_t^X]_{ij}[K_t^Y]_{kl},$$

and the problem rewrites

$$\min_{\Omega\in\mathbb{R}^{n\times n}} \sum_{i=1}^{n} L_i\left(\Omega_{i:}, K^Y\right) + \frac{1}{2\Lambda n} \sum_{t=1}^{T} \mathbf{Tr}\left(K_t^X \Omega K_t^Y \Omega^\top\right).$$

□

Remark 7. *The second term of Problem (12) can be easily optimized. Indeed, let $\tilde{M}$ be a block matrix such that $\tilde{M}_{st}^{(i,j)} = \tilde{M}_{ij}^{(s,t)}$ for all $i,j,s,t \leqslant n$. Notice that $\tilde{M}$ as defined earlier satisfies this condition as a direct consequence of the OVK symmetry property. Then it holds:*

$$\frac{\partial \mathbf{Tr}\left(\tilde{M}^\top(\Omega\otimes\Omega)\right)}{\partial \omega_{st}} = 2\mathbf{Tr}\left(\tilde{M}^{(s,t)\top}\Omega\right).$$

Indeed, notice that $\mathbf{Tr}\left(\tilde{M}^\top(\Omega\otimes\Omega)\right) = \sum_{i,j=1}^n \omega_{ij}\mathbf{Tr}\left(\tilde{M}^{(i,j)\top}\Omega\right)$ and use the symmetry assumption. In the particular case of a decomposable kernel, it holds that $\tilde{M}^{(i,j)} = K^X_{i:}K^{Y\top}_{j:}$ so that

$$\frac{\partial \mathbf{Tr}\left(\tilde{M}^\top(\Omega\otimes\Omega)\right)}{\partial \omega_{st}} = 2\mathbf{Tr}\left(\tilde{M}^{(s,t)\top}\Omega\right) = 2\sum_{i,j=1}^n \left[K^X_{s:}K^{Y\top}_{t:}\right]_{ij}\omega_{ij} = 2\sum_{ij=1}^n K^X_{si}K^Y_{tj}\omega_{ij} = 2\left[K^X\Omega K^Y\right]_{st},$$

and one recovers the gradients established in Equation (15).

A.3. Proof of Proposition 1

The proof technique is the same for all losses: first explicit the FL transforms $\ell_i^\star$, then use simple arguments to verify Assumptions 1 and 2. For instance, any increasing function of $\|\alpha\|$ automatically satisfy the assumptions.

- Assume that ℓ is such that there is $f:\mathbb{R}\to\mathbb{R}$ convex, $\forall i\leqslant n, \exists z_i\in Y$, $\ell_i(y) = f(\langle y, z_i\rangle)$. Then $\ell_i^\star:\mathcal{Y}\to\mathbb{R}$ writes $\ell_i^\star(\alpha) = \sup_{y\in\mathcal{Y}}\langle\alpha,y\rangle - f(\langle y,z_i\rangle)$. If α is not collinear to z_i, this quantity is obviously $+\infty$. Otherwise, assume that $\alpha=\lambda z_i$. The FL transform rewrites: $\ell_i^\star(\alpha) = \sup_t \lambda t - f(t) = f^\star(\lambda) = f^\star(\pm\|\alpha\|/\|z_i\|)$. Finally, $\ell_i^\star(\alpha) = \chi_{\text{span}(z_i)}(\alpha) + f^\star\left(\pm\frac{\|\alpha\|}{\|z_i\|}\right)$. If $\alpha\notin Y$, then *a fortiori* $\alpha\notin\text{span}(z_i)$, so $\ell_i^\star(\alpha^Y+\alpha^\perp) = +\infty \geqslant \ell_i^\star(\alpha^Y)$ for all $(\alpha^Y,\alpha^\perp)\in Y\times Y^\perp$. For all $i\leqslant n$, $\ell_i^\star$ satisfy Assumption 1. As for Assumption 2, if $\alpha = \sum_{i=1}^n c_i y_i$, then $\chi_{\text{span}(z_i)}(\alpha)$ only depends on the $(c_i)_{i\leqslant n}$ Indeed, assume that $z_i\in Y$ writes $\sum_j b_j y_j$. Then $\chi_{\text{span}(z_i)}(\alpha)$ is equal to 0 if there exists $\lambda\in\mathbb{R}$ such that $c_j=\lambda b_j$ for all $j\leqslant n$, and to $+\infty$ otherwise. The second term of $\ell_i^\star$ depending only on $\|\alpha\|$, it directly satisfies Assumption 2. This concludes the proof.
- Assume that ℓ is such that there is $f:\mathbb{R}_+\to\mathbb{R}$ convex increasing, with $\frac{f'(t)}{t}$ continuous over $\mathbb{R}_+$, $\ell(y) = f(\|y\|)$. Although this loss may seem useless at the first sight since ℓ does not depend on y_i, it should not be forgotten that the composition with $y\mapsto y-y_i$ does not affect the validation of Assumptions 1 and 2 (see below). One has: $\ell^\star(\alpha) = \sup_{y\in\mathcal{Y}}\langle\alpha,y\rangle - f(\|y\|)$. Differentiating w.r.t. y, one gets: $\alpha = \frac{f'(\|y\|)}{\|y\|}y$, which is always well define as $t\mapsto\frac{f'(t)}{t}$ is continuous over $\mathbb{R}_+$. Reverting the equality, it holds: $y = \frac{f'^{-1}(\|\alpha\|)}{\|\alpha\|}\alpha$, and $\ell^\star(\alpha) = \|\alpha\|f'^{-1}(\|\alpha\|) - f\circ f'^{-1}(\|\alpha\|)$. This expression depending only on $\|\alpha\|$, Assumption 2 is automatically satisfied. Let us now investigate the monotonicity of $\ell^\star$ w.r.t. $\|\alpha\|$. Let $g:\mathbb{R}_+\to\mathbb{R}$ such that $g(t) = tf'^{-1}(t) - f\circ f'^{-1}(t)$. Then $g'(t) = f'^{-1}(t)\geqslant 0$. Indeed, as $f':\mathbb{R}_+\to\mathbb{R}_+$ is always positive due to the monotonicity of f, so is f'^{-1}. This final remark guarantees that $\ell^\star$ is increasing with $\|\alpha\|$. It is then direct that $\ell^\star$ fulfills Assumption 1.
- Assume that $\ell(y) = \lambda\|y\|$. It holds $\ell^\star(\alpha) = \chi_{\mathcal{B}_\lambda}(\alpha)$. So $\ell^\star$ is increasing w.r.t. $\|\alpha\|$: it fulfills Assumptions 1 and 2.
- Assume that $\ell(y) = \chi_{\mathcal{B}_\lambda}(y)$. It holds $\ell^\star(\alpha) = \lambda\|\alpha\|$. The monotonicity argument also applies.
- Assume that $\ell(y) = \lambda\|y\|\log(\|y\|)$. It can be shown that $\ell^\star(\alpha) = \lambda e^{\frac{\|\alpha\|}{\lambda}-1}$. The same argument as above applies.
- Assume that $\ell(y) = \lambda(\exp(\|y\|)-1)$. It can be shown that $\ell^\star(\alpha) = \mathbb{I}\{\|\alpha\|\geqslant\lambda\}\cdot\left(\|\alpha\|\log\left(\frac{\|\alpha\|}{\lambda e}\right)+\lambda\right)$. Again, the FL transform is an increasing function of $\|\alpha\|$: it satisfies Assumptions 1 and 2.
- Assume that $\ell_i(y) = f(y-y_i)$, with f such that $f^\star$ fulfills Assumptions 1 and 2. Then $\ell_i^\star(\alpha) = \sup_{y\in\mathcal{Y}}\langle\alpha,y\rangle - f(y-y_i) = f^\star(\alpha) + \langle\alpha,y_i\rangle$. If $f^\star$ satisfies Assumptions 1 and 2, then so does $\ell_i^\star$. This remark is very important, as it gives more sense to loss function based on $\|y\|$ only, since they can be applied to $y-y_i$ now.
- Assume that there exists f,g satisfying Assumptions 1 and 2 such that $\ell_i(y) = (f\,\square\, g)(y)$, where $\square$ denotes the infimal convolution, *i.e.* $(f\,\square\, g)(y) = \inf_x f(x)+g(y-x)$. Standard arguments about FL transforms state that $(f\,\square\, g)^\star = f^\star + g^\star$, so that if both f and g satisfy Assumptions 1 and 2, so does $f\,\square\, g$. This last example allows to deal with ϵ-insensitive losses for instance (convolution of a loss and $\chi_{\mathcal{B}_\epsilon}$), the Huber loss (convolution of $\|.\|$ and $\|.\|^2$), or more generally all Moreau envelopes (convolution of a loss and $\frac{1}{2}\|.\|^2$).

□

A.4. Proof of Theorem 5

The proof of Theorem 5 is straightforward: since the dual space $\widetilde{\mathcal{Y}}_m$ is of finite dimension m, the dual variable can be written as a linear combination of the $\{\psi_j\}_{j=1}^m$ to get Problem (7).

A.5. Proof of Theorem 6

A.5.1. ϵ-RIDGE – FROM PROBLEM $(P1)$ TO $(D1)$

Applying Theorem 3 together with the Fenchel-Legendre transforms detailed in the proof of Proposition 1, a dual to the ϵ-Ridge regression primal problem is:

$$\min_{(\alpha_i)_{i=1}^n \in \mathcal{Y}^n} \quad \frac{1}{2}\sum_{i=1}^n \|\alpha_i\|_{\mathcal{Y}}^2 - \sum_{i=1}^n \langle \alpha_i, y_i\rangle_{\mathcal{Y}} + \epsilon \sum_{i=1}^n \|\alpha_i\|_{\mathcal{Y}} + \frac{1}{2\Lambda n}\sum_{i,j=1}^n \langle \alpha_i, \mathcal{K}(x_i,x_j)\alpha_j\rangle_{\mathcal{Y}},$$

$$\min_{(\alpha_i)_{i=1}^n \in \mathcal{Y}^n} \quad \frac{1}{2}\sum_{i,j=1}^n \left\langle \alpha_i, \left(\delta_{ij}\mathbf{I}_{\mathcal{Y}} + \frac{1}{\Lambda n}\mathcal{K}(x_i,x_j)\right)\alpha_j\right\rangle_{\mathcal{Y}} - \sum_{i=1}^n \langle \alpha_i, y_i\rangle_{\mathcal{Y}} + \epsilon\sum_{i=1}^n \|\alpha_i\|_{\mathcal{Y}}.$$

By virtue of Theorem 4, we known that the optimal $(\alpha_i)_{i=1}^n \in \mathcal{Y}^n$ are in $\mathbf{Y}^n$. After the reparametrization $\alpha_i = \sum_j \omega_{ij}\, y_j$, the problem rewrites:

$$\min_{\Omega \in \mathbb{R}^{n\times n}} \quad \frac{1}{2}\mathbf{Tr}\left(\tilde{K}\Omega K^Y \Omega^\top\right) - \mathbf{Tr}\left(K^Y\Omega\right) + \epsilon\sum_{i=1}^n \sqrt{[\Omega K^Y\Omega^\top]_{ii}}, \tag{13}$$

with Ω, $\tilde{K}$, K^Y the $n\times n$ matrices such that $[\Omega]_{ij} = \omega_{ij}$, $\tilde{K} = \frac{1}{\Lambda n}K^X + \mathbf{I}_n$, and $[K^Y]_{ij} = \langle y_i, y_j\rangle_{\mathcal{Y}}$.

Now, let $K^Y = U\Sigma U^\top = \left(U\Sigma^{1/2}\right)\left(U\Sigma^{1/2}\right)^\top = VV^\top$ be the SVD of K^Y, and let $W = \Omega V$. Notice that K^Y is positive semi-definite, and can be made positive definite if necessary, so that V is full rank, and optimizing with respect to W is strictly equivalent to minimizing with respect to Ω. With this change of variable, Problem (13) rewrites:

$$\min_{W\in\mathbb{R}^{n\times n}} \quad \frac{1}{2}\mathbf{Tr}\left(\tilde{K}WW^\top\right) - \mathbf{Tr}\left(V^\top W\right) + \epsilon\|W\|_{2,1}, \tag{14}$$

with $\|W\|_{2,1} = \sum_i \|W_{i:}\|_2$ the row-wise $\ell_{2,1}$ mixed norm of matrix W. With $\tilde{K} = A^\top A$ the SVD of $\tilde{K}$, and B such that $A^\top B = V$, one can add the constant term $\frac{1}{2}\mathbf{Tr}(A^{\top -1}VV^\top A^{-1}) = \frac{1}{2}\mathbf{Tr}(BB^\top)$ to the objective without changing Problem (14). One finally gets the Multi-Task Lasso problem:

$$\min_{W\in\mathbb{R}^{n\times n}} \quad \frac{1}{2}\|AW - B\|_{\text{Fro}}^2 + \epsilon\|W\|_{2,1}.$$

□

We also emphasize that we recover the solution to the standard Ridge regression when $\epsilon = 0$. Indeed, coming back to Problem (13) and differentiating with respect to Ω, one gets:

$$\tilde{K}\hat{\Omega}K^Y - K^Y = 0 \iff \hat{\Omega} = \tilde{K}^{-1},$$

which is exactly the standard kernel Ridge regression solution, see *e.g.* Brouard et al. (2016b).

Furthermore, notice that when $\mathcal{K}$ is not identity decomposable, but only satisfies Assumption 4, then Problem (14) cannot be factorized that easily. Nonetheless, it admits a simple resolution, as detailed in the following lines. After the Ω reparametrization, the problem writes

$$\min_{\Omega\in\mathbb{R}^{n\times n}} \quad \frac{1}{2}\mathbf{Tr}(\Omega K^Y\Omega^\top) - \mathbf{Tr}\left(K^Y\Omega\right) + \epsilon\sum_{i=1}^n \sqrt{[\Omega K^Y\Omega^\top]_{i,i}} + \frac{1}{2\Lambda n}\sum_{t=1}^T \mathbf{Tr}(K_t^X\Omega K_t^Y\Omega^\top),$$

$$\min_{W\in\mathbb{R}^{n\times n}} \quad \frac{1}{2}\mathbf{Tr}(WW^\top) + \frac{1}{2\Lambda n}\sum_{t=1}^T \mathbf{Tr}(K_t^X W\tilde{K}_t^Y W^\top) - \mathbf{Tr}\left(V^\top W\right) + \epsilon\|W\|_{2,1},$$

with $K^Y = VV^\top$, $W = \Omega V$, $\tilde{K}_t^Y = V^{-1}K_t^Y(V^\top)^{-1}$. Due to the different quadratic terms, this problem cannot be summed up as a Multi-Task Lasso like before. However, it may still be solved, *e.g.* by proximal gradient descent. Indeed, the gradient of the smooth term (*i.e.* all but the $\ell_{2,1}$ mixed norm) reads

$$W + \frac{1}{\Lambda n}\sum_{t=1}^{T} K_t^X W \tilde{K}_t^Y - V, \tag{15}$$

while the proximal operator of the $\ell_{2,1}$ mixed norm is

$$\operatorname{prox}_{\epsilon\,\|\cdot\|_{2,1}}(W) = \begin{pmatrix} | \\ \operatorname{prox}_{\epsilon\,\|\cdot\|_2}(W_{i:}) \\ | \end{pmatrix} = \begin{pmatrix} | \\ \left(1 - \frac{\epsilon}{\|W_{i:}\|_2}\right)_+ W_{i:} \\ | \end{pmatrix} = \begin{pmatrix} | \\ \mathrm{BST}(W_{i:}, \epsilon) \\ | \end{pmatrix}.$$

Hence, even in the more involved case of an OVK satisfying only Assumption 4, we have designed an efficient algorithm to compute the solutions to the dual problem.

A.5.2. κ-HUBER – FROM PROBLEM $(P2)$ TO $(D2)$

Basic manipulations give the Fenchel-Legendre transforms of the Huber loss:

$$\begin{aligned}\left(y \mapsto \ell_{H,\kappa}(y - y_i)\right)^\star(\alpha) &= \left(\kappa\|\cdot\|_{\mathcal{Y}} \,\square\, \frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2\right)^\star(\alpha) + \langle \alpha, y_i\rangle_{\mathcal{Y}}\,, \\ &= \left(\kappa\|\cdot\|_{\mathcal{Y}}\right)^\star(\alpha) + \left(\frac{1}{2}\|\cdot\|_{\mathcal{Y}}^2\right)^\star(\alpha) + \langle \alpha, y_i\rangle_{\mathcal{Y}}\,, \\ &= \chi_{\mathcal{B}_\kappa}(\alpha) + \frac{1}{2}\|\alpha\|_{\mathcal{Y}}^2 + \langle \alpha, y_i\rangle_{\mathcal{Y}}\,.\end{aligned}$$

Following the same lines as for as for the ϵ-Ridge regression, the dual problem writes

$$\min_{(\alpha_i)_{i=1}^n \in \mathcal{Y}^n} \quad \frac{1}{2}\sum_{i,j=1}^{n} \left\langle \alpha_i, \left(\delta_{ij}\mathbf{I}_{\mathcal{Y}} + \frac{1}{\Lambda n}\mathcal{K}(x_i, x_j)\right)\alpha_j \right\rangle_{\mathcal{Y}} - \sum_{i=1}^{n}\langle \alpha_i, y_i\rangle_{\mathcal{Y}} + \sum_{i=1}^{n}\chi_\kappa(\|\alpha_i\|_{\mathcal{Y}}),$$

or again after the reparametrization in Ω

$$\begin{aligned}\min_{\Omega \in \mathbb{R}^{n\times n}} \quad & \frac{1}{2}\mathbf{Tr}\left(\tilde{K}\Omega K^Y \Omega^\top\right) - \mathbf{Tr}\left(K^Y\Omega\right) \\ \text{s.t.} \quad & \sqrt{[\Omega K^Y \Omega^\top]_{ii}} \leqslant \kappa \qquad \forall i \leqslant n\end{aligned}$$

The same change of variable permits to conclude. □

When $\mathcal{K}$ is not identity decomposable, but only satisfies Assumption 4, the problem rewrites

$$\begin{aligned}\min_{W \in \mathbb{R}^{n\times n}} \quad & \frac{1}{2}\mathbf{Tr}(WW^\top) + \frac{1}{2\Lambda n}\sum_{t=1}^{T}\mathbf{Tr}(K_t^X W \tilde{K}_t^Y W^\top) - \mathbf{Tr}\left(V^\top W\right), \\ \text{s.t.} \quad & \|W_{i:}\|_2 \leqslant \kappa \quad \forall i \leqslant n,\end{aligned}$$

Again, the gradient term is given by Equation (15), while the projection is similar to the identity decomposable case. The only change thus occurs in the gradient step of Algorithm 1, with a replacement by the above formula.

Notice that if κ tends to infinity, the problem is unconstrained, and one also recovers the standard Ridge regression solution.

A.5.3. ϵ-SVR – FROM PROBLEM $(P3)$ TO $(D3)$

The proof is similar to the above derivations except that the term $\sum_i \|\alpha_i\|_{\mathcal{Y}}^2$ does not appear in the dual, hence the change of matrix $\tilde{K}$. Instead, the dual problem features both the $\ell_{2,1}$ penalization and the $\ell_{2,\infty}$ constraint. □

A.6. Proof of Theorem 7

The proof is similar to Appendix A.5.2, with the finite representation coming from Theorem 5.

A.7. Proof of Theorem 10

In this section, we detail the derivation of constants in Figure 1.

A.7.1. ϵ-SVR

Using that the null function is part of the vv-RKHS, it holds

$$\frac{\Lambda}{2}\|h_{A(\mathcal{S})}\|^2_{\mathcal{H}_\mathcal{K}} \leqslant \hat{\mathcal{R}}_n(h_{A(\mathcal{S})}) \leqslant \hat{\mathcal{R}}_n(0_{\mathcal{H}_\mathcal{K}}) \leqslant M_\mathcal{Y} - \epsilon, \qquad \text{or again} \qquad \|h_{A(\mathcal{S})}\|_{\mathcal{H}_\mathcal{K}} \leqslant \sqrt{\frac{2}{\Lambda}(M_\mathcal{Y} - \epsilon)}.$$

Furthermore, the reproducing property and Assumption 8 give that for any $x \in \mathcal{X}$ and any $h \in \mathcal{H}_\mathcal{K}$ it holds

$$\|h(x)\|^2 = \left\langle \mathcal{K}(\cdot,x)\mathcal{K}(\cdot,x)^\# h, h\right\rangle_{\mathcal{H}_\mathcal{K}} \leqslant \left\|\mathcal{K}(\cdot,x)\mathcal{K}(\cdot,x)^\#\right\|_{\text{op}} \|h\|^2_{\mathcal{H}_\mathcal{K}} \leqslant \|\mathcal{K}(x,x)\|_{\text{op}} \|h\|^2_{\mathcal{H}_\mathcal{K}} \leqslant \gamma^2 \|h\|^2_{\mathcal{H}_\mathcal{K}}.$$

Therefore, one gets that for any realization $(x,y) \in \mathcal{X} \times \mathcal{Y}$ of (X,Y) it holds

$$\ell(h_{A(\mathcal{S})}(x), y) = \max(\|y - h_{A(\mathcal{S})}(x)\|_\mathcal{Y} - \epsilon, 0) \leqslant M_y - \epsilon + \|h_{A(\mathcal{S})}(x)\|_\mathcal{Y} \leqslant \sqrt{M_\mathcal{Y} - \epsilon}\left(\gamma\sqrt{\frac{2}{\Lambda}} + \sqrt{M_\mathcal{Y} - \epsilon}\right).$$

This gives M. As for C, one has

$$\ell(h_{A(\mathcal{S})}(x), y) - \ell(h_{A(\mathcal{S}^{\setminus i})}(x), y) = \max(\|y - h_{A(\mathcal{S})}(x)\|_\mathcal{Y} - \epsilon, 0) - \max(\|y - h_{A(\mathcal{S}^{\setminus i})}(x)\|_\mathcal{Y} - \epsilon, 0).$$

If both norms are smaller than ϵ, then any value of C fits. If both norms are greater than ϵ, the difference reads

$$\|y - h_{A(\mathcal{S})}(x)\|_\mathcal{Y} - \|y - h_{A(\mathcal{S}^{\setminus i})}(x)\|_\mathcal{Y} \leqslant \|h_{A(\mathcal{S})}(x) - h_{A(\mathcal{S}^{\setminus i})}(x)\|_\mathcal{Y}.$$

If only one norm is greater than ϵ (we write it only for $h_{A(\mathcal{S})}$ as it is symmetrical), the difference may be rewritten

$$\|y - h_{A(\mathcal{S})}(x)\|_\mathcal{Y} - \epsilon \leqslant \|y - h_{A(\mathcal{S})}(x)\|_\mathcal{Y} - \|y - h_{A(\mathcal{S}^{\setminus i})}(x)\|_\mathcal{Y} \leqslant \|h_{A(\mathcal{S})}(x) - h_{A(\mathcal{S}^{\setminus i})}(x)\|_\mathcal{Y}.$$

Hence we get $C = 1$.

A.7.2. ϵ-RIDGE

Using the same reasoning as for the ϵ-SVR, one has

$$\|h_{A(\mathcal{S})}\|_{\mathcal{H}_\mathcal{K}} \leqslant \sqrt{\frac{2}{\Lambda}}(M_\mathcal{Y} - \epsilon) \qquad \text{and} \qquad \|h_{A(\mathcal{S}^{\setminus i})}\|_{\mathcal{H}_\mathcal{K}} \leqslant \sqrt{\frac{2}{\Lambda}}(M_\mathcal{Y} - \epsilon). \tag{16}$$

Therefore, for any realization $(x,y) \in \mathcal{X} \times \mathcal{Y}$ of (X,Y) it holds

$$\ell(h_{A(\mathcal{S})}(x), y) = \max(\|y - h_{A(\mathcal{S})}(x)\| - \epsilon, 0)^2 \leqslant (\|y\|_\mathcal{Y} - \epsilon + \|h_{A(\mathcal{S})}(x)\|_\mathcal{Y})^2 \leqslant (M_\mathcal{Y} - \epsilon)^2\left(1 + \frac{2\sqrt{2}\gamma}{\sqrt{\Lambda}} + \frac{2\gamma^2}{\Lambda}\right).$$

As for C, one has

$$\ell(h_{A(\mathcal{S})}(x), y) - \ell(h_{A(\mathcal{S}^{\setminus i})}(x), y) = \max(\|y - h_{A(\mathcal{S})}(x)\|_\mathcal{Y} - \epsilon, 0)^2 - \max(\|y - h_{A(\mathcal{S}^{\setminus i})}(x)\|_\mathcal{Y} - \epsilon, 0)^2.$$

If both norms are smaller than ϵ, any C fits. If both are larger than ϵ, using Equation (16) the difference becomes

$$\begin{aligned}&\left(\|y - h_{A(\mathcal{S})}(x)\|_\mathcal{Y} + \|y - h_{A(\mathcal{S}^{\setminus i})}(x)\|_\mathcal{Y} - 2\epsilon\right)\left(\|y - h_{A(\mathcal{S})}(x)\|_\mathcal{Y} - \|y - h_{A(\mathcal{S}^{\setminus i})}(x)\|_\mathcal{Y}\right),\\ &\leqslant 2(M_\mathcal{Y} - \epsilon)\left(1 + \frac{\gamma\sqrt{2}}{\sqrt{\Lambda}}\right)\|h_{A(\mathcal{S})}(x) - h_{A(\mathcal{S}^{\setminus i})}(x)\|_\mathcal{Y}.\end{aligned}$$

If only one norm is greater than ϵ (again, the analysis is symmetrical), the difference may be rewritten

$$\begin{aligned}\left(\|y - h_{A(\mathcal{S})}(x)\|_{\mathcal{Y}} - \epsilon\right)^2 &\leqslant \left(\|y - h_{A(\mathcal{S})}(x)\|_{\mathcal{Y}} - \|y - h_{A(\mathcal{S}^{\setminus i})}(x)\|_{\mathcal{Y}}\right)^2 \leqslant \|h_{A(\mathcal{S})}(x) - h_{A(\mathcal{S}^{\setminus i})}(x)\|_{\mathcal{Y}}^2, \\ &\leqslant \left(\|h_{A(\mathcal{S})}(x)\|_{\mathcal{Y}} + \|h_{A(\mathcal{S}^{\setminus i})}(x)\|_{\mathcal{Y}}\right) \|h_{A(\mathcal{S})}(x) - h_{A(\mathcal{S}^{\setminus i})}(x)\|_{\mathcal{Y}}, \\ &\leqslant 2(M_{\mathcal{Y}} - \epsilon)\frac{\gamma\sqrt{2}}{\sqrt{\Lambda}}\|h_{A(\mathcal{S})}(x) - h_{A(\mathcal{S}^{\setminus i})}(x)\|_{\mathcal{Y}}.\end{aligned}$$

In every case $C = 2(M_{\mathcal{Y}} - \epsilon)\left(1 + \gamma\sqrt{2}/\sqrt{\Lambda}\right)$ works, hence the conclusion.

A.7.3. κ-HUBER

Using the same techniques, one gets

$$\|h_{A(\mathcal{S})}\|_{\mathcal{H}_{\mathcal{K}}} \leqslant \sqrt{\frac{2\kappa}{\Lambda}\left(M_{\mathcal{Y}} - \frac{\kappa}{2}\right)} \qquad \text{and} \qquad \|h_{A(\mathcal{S}^{\setminus i})}\|_{\mathcal{H}_{\mathcal{K}}} \leqslant \sqrt{\frac{2\kappa}{\Lambda}\left(M_{\mathcal{Y}} - \frac{\kappa}{2}\right)},$$

and for any realization $(x, y) \in \mathcal{X} \times \mathcal{Y}$ of (X, Y)

$$\ell(h_{A(\mathcal{S})}(x), y) \leqslant \kappa\sqrt{M_{\mathcal{Y}} - \frac{\kappa}{2}}\left(\frac{\gamma\sqrt{2\kappa}}{\sqrt{\Lambda}} + \sqrt{M_{\mathcal{Y}} - \frac{\kappa}{2}}\right).$$

If both norms are greater than κ, the difference $\ell(h_{A(\mathcal{S})}(x), y) - \ell(h_{A(\mathcal{S}^{\setminus i})}(x), y)$ writes

$$\kappa\left(\|h_{A(\mathcal{S})}(x) - y\|_{\mathcal{Y}} - \frac{\kappa}{2}\right) - \kappa\left(\|h_{A(\mathcal{S}^{\setminus i})}(x) - y\|_{\mathcal{Y}} - \frac{\kappa}{2}\right) \leqslant \kappa\|h_{A(\mathcal{S})}(x) - h_{A(\mathcal{S}^{\setminus i})}(x)\|_{\mathcal{Y}}.$$

If only one norm is greater than κ, one may upperbound the difference using the previous writing

$$\kappa\left(\|h_{A(\mathcal{S})}(x) - y\|_{\mathcal{Y}} - \frac{\kappa}{2}\right) - \frac{1}{2}\|h_{A(\mathcal{S}^{\setminus i})}(x) - y\|_{\mathcal{Y}}^2 \leqslant \kappa\left(\|h_{A(\mathcal{S})}(x) - y\|_{\mathcal{Y}} - \frac{\kappa}{2}\right) - \kappa\left(\|h_{A(\mathcal{S}^{\setminus i})}(x) - y\|_{\mathcal{Y}} - \frac{\kappa}{2}\right).$$

If both are smaller than κ, the difference becomes

$$\begin{aligned}&\frac{1}{2}\|h_{A(\mathcal{S})}(x) - y\|_{\mathcal{Y}}^2 - \frac{1}{2}\|h_{A(\mathcal{S}^{\setminus i})}(x) - y\|_{\mathcal{Y}}^2, \\ &\quad= \frac{1}{2}\left(\|h_{A(\mathcal{S})}(x) - y\|_{\mathcal{Y}} + \|h_{A(\mathcal{S}^{\setminus i})}(x) - y\|_{\mathcal{Y}}\right)\left(\|h_{A(\mathcal{S})}(x) - y\|_{\mathcal{Y}} - \|h_{A(\mathcal{S}^{\setminus i})}(x) - y\|_{\mathcal{Y}}\right), \\ &\quad\leqslant \kappa\|h_{A(\mathcal{S})}(x) - h_{A(\mathcal{S}^{\setminus i})}(x)\|_{\mathcal{Y}},\end{aligned}$$

so that $C = \kappa$.

A.8. Further Admissible Kernels for Assumption 3

In the continuation of Remark 1, we now exhibit several types of OVK that satisfy Assumption 3.

Proposition 2. *The following Operator-Valued Kernels satisfy Assumption 3:*

(i) $\forall s, t \in \mathcal{X}^2,\ \mathcal{K}(s,t) = \sum_i k_i(s,t)\, y_i \otimes y_i,$ *with k_i positive semi-definite (p.s.d.) scalar kernels for all $i \leqslant n$.*

(ii) $\forall s, t \in \mathcal{X}^2,\ \mathcal{K}(s,t) = \sum_i \mu_i\, k(s,t)\, y_i \otimes y_i,$ *with k a p.s.d. scalar kernel and $\mu_i \geqslant 0$ for all $i \leqslant n$.*

(iii) $\forall s, t \in \mathcal{X}^2,\ \mathcal{K}(s,t) = \sum_i k(s, x_i)k(t, x_i)\, y_i \otimes y_i,$

(iv) $\forall s, t \in \mathcal{X}^2,\ \mathcal{K}(s,t) = \sum_{i,j} k_{ij}(s,t)\, (y_i + y_j) \otimes (y_i + y_j),$ *with k_{ij} p.s.d. scalar kernels for all $i, j \leqslant n$.*

(v) $\forall s, t \in \mathcal{X}^2,\ \mathcal{K}(s,t) = \sum_{i,j} \mu_{ij}\, k(s,t)\, (y_i + y_j) \otimes (y_i + y_j),$ *with k a p.s.d. scalar kernel and $\mu_{ij} \geqslant 0$.*

(vi) $\forall s, t \in \mathcal{X}^2,\ \mathcal{K}(s,t) = \sum_{i,j} k(s, x_i, x_j)k(t, x_i, x_j)\, (y_i + y_j) \otimes (y_i + y_j).$

Proof.

(i) For all $(s_k, z_k)_{k\leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$, it holds: $\sum_{k,l} \langle z_k, \mathcal{K}(s_k, s_l) z_k \rangle_{\mathcal{Y}} = \sum_i \sum_{k,l} k_i(s,t) \langle z_k, y_i \rangle_{\mathcal{Y}} \langle z_l, y_i \rangle_{\mathcal{Y}}$, which is positive by the positiveness of the scalar kernels k_i's. Notice that (ii) and (iii) are then particular cases of (i).

(ii) is an application of (i), as a kernel remains p.s.d. through positive multiplication. Observe that this kernel is separable.

(iii) is also a direct application of (i), kernel $k' : s, t \mapsto k(s, x_i)k(t, x_i)$ being indeed p.s.d. for all function k and point x_i.

(iv) is proved similarly to (i). The arguments used for (ii) and (iii) also makes (v) and (vi) direct applications of (iv).

Finally, notice that for (iv), (v) and (vi), any linear combination $(\nu_i y_i + \nu_j y_j) \otimes (\nu_i y_i + \nu_j y_j)$, with $0 \leqslant \nu_i \leqslant 1$ for all $i \leqslant n$, could have been used instead of $(y_i + y_j) \otimes (y_i + y_j)$. □

B. Loss Functions Illustrations

In this section, we provide illustrations of the loss functions we used to promote sparsity and robustness. This includes ϵ-insensitive losses (Definitions 3 and 4, Figures 9 and 10) and the κ-Huber loss (Definition 5, Figure 11). First introduced for real outputs, their formulations as infimal convolutions allows for a generalization to any Hilbert space, either of finite dimension (as in Sangnier et al. (2017)) or not, which is the general case addressed in the present paper. The ϵ-insensitive loss functions promote sparsity, as reflected in the corresponding dual problems (see Theorem 6, Problems $(D1)$ and $(D3)$ therein) and the empirical results (Figures 12 and 13). On the other hand, losses whose slopes asymptotically behave as $||\cdot||_{\mathcal{Y}}$ instead of $||\cdot||^2_{\mathcal{Y}}$ (such as the κ-Huber or the ϵ-SVR loss) encourage robustness through a resistance to outliers. Indeed, under such a setting, residuals of high norm contribute less to the gradient and have a minor influence on the model output.

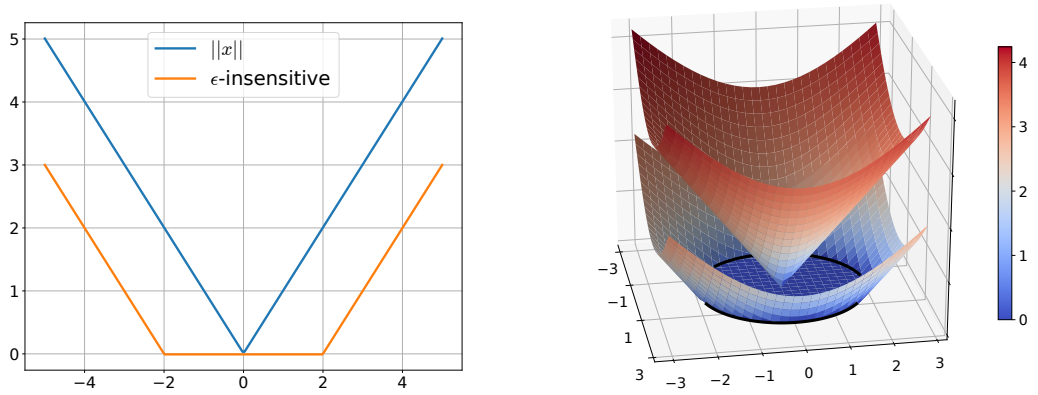

Figure 9. Standard and ϵ-insensitive versions of the SVR loss in 1 and 2 dimensions ($\epsilon = 2$).

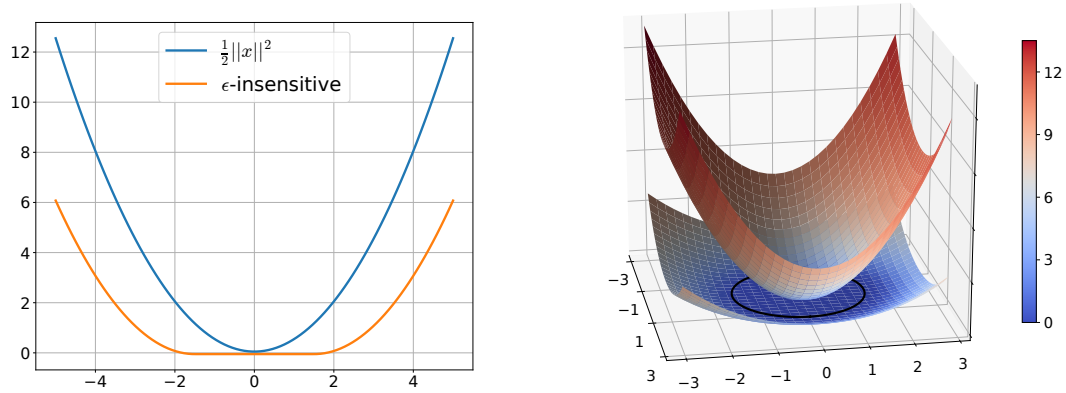

Figure 10. Standard and ϵ-insensitive versions of the square loss in 1 and 2 dimensions ($\epsilon = 1.5$).

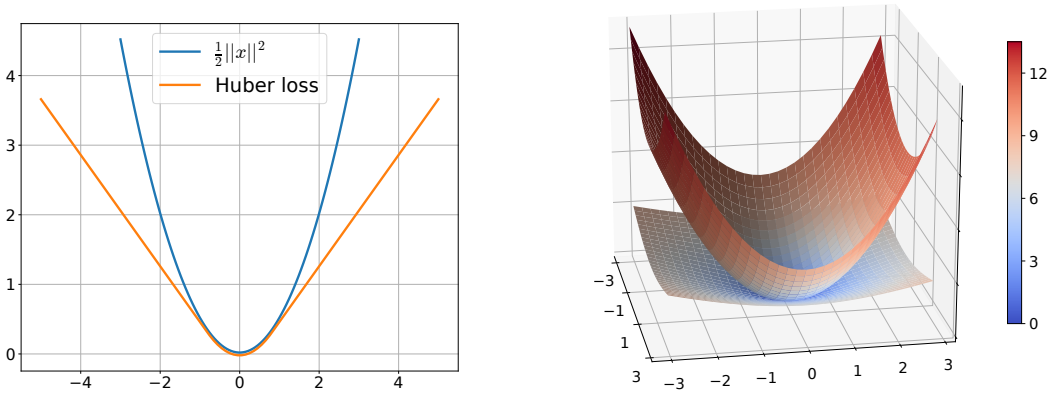

Figure 11. Standard square loss and Huber loss in 1 and 2 dimensions ($\kappa = 0.8$).

C. Numerical Experiments and Code

C.1. Provided Code

The Python code used to generate the plots and tables of the article is provided. The `README` file in the code folder contains instructions for quickly reproducing (part of) the plots. All implemented methods may be run on other datasets/problems.

C.2. Detailed Protocols

C.2.1. Structured Prediction

YEAST Dataset Description. YEAST[1] is a publicly available multi-label classification dataset used as a benchmark in several structured prediction articles. We compared our approach, with the same train/test decomposition, to those presented in Elisseeff and Weston (2002), Finley and Joachims (2008) and Belanger and McCallum (2016). The size of the training set is 1500, the test set is of size 917. The problem consists in predicting the functional classes of a gene. The inputs are micro-array expression data (representing the genes) of dimension $p = 103$. The outputs are multi-label vectors of size $d = 14$ representing the possible functional classes of the genes. The average number of labels is 4.2. These 14 functional classes correspond to the first level of a tree that structures a much bigger set of possible functional classes.

Experimental protocol: Comparison with other methods. In Figure 6, we reported the Hamming error on the test set obtained by each method. The results obtained by SSVM and SPENS are extracted from Finley and Joachims (2008) and Belanger and McCallum (2016). For our approach and its three variants (ϵ-KRR, κ-Huber, ϵ-SVR), each hyper-parameter (Λ, ϵ, or κ) has been selected by estimating the Mean Squared Error (MSE) through a 5-fold cross-validation computed on the training set. We used an input Gaussian kernel with a fixed bandwidth equal to 1.

Experimental protocol: Cross-Effect of ϵ and Λ on sparsity and MSE. In order to measure the effect of the different hyperparameters and study their interrelations, we have computed the 5-fold cross-validation MSE and sparsity/saturation for several values of Λ and ϵ/κ. The input kernel is still Gaussian with bandwidth 1. The results are plotted in Figures 3 and 4 for the ϵ-KRR, and in Figures 13 and 14 for the ϵ-SVR and κ-Huber. In Figure 4, we have measured sparsity through the number of training data which are discarded, *i.e.* not used in the finite representation of the ϵ-KRR model. The κ-Huber saturation is assessed in a similar fashion: it corresponds to the number of training data whose associated coefficient saturates the norm constraint (see Theorem 6, Problem $(D2)$ therein). Simplified versions of these graphs may be quickly reproduced using the code attached (see `README` file).

Metabolite identification dataset description. We next tested our method on a harder problem: that of metabolite identification (Brouard et al., 2016a). The goal is to predict a metabolite (small molecule) thanks to its mass spectrum. The difficulty comes from the reduced size of the training set ($n = 6974$) compared to the high dimension of the outputs ($d = 7593$). Input Output Kernel Regression (IOKR, see Brouard et al. (2016a;b)) with a Tanimoto-Gaussian kernel is state of the art on this problem.

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

Experimental protocol. We investigate the advantages of substituting the Ridge Regression for the ϵ-KRR, κ-Huber, and ϵ-SVR. Outputs are embedded in an infinite dimensional space through the use of the Tanimoto-Gaussian kernel (with bandwidth $\gamma = 0.72$). We compare the different algorithms' performances on a set of 6974 mass spectra through the top-k accuracies for $k \in \{1, 10, 20\}$. We give the average 5-fold top-k accuracies (Table 1). The 5 folds have been chosen such that a metabolite does not appear in two different folds (zero-shot learning setting).

C.2.2. Structured Representation Learning

Dataset Description. Robust structured representation learning was tested on a drug dataset, introduced in Su et al. (2010), and extracted from the NCI-Cancer database. This dataset features a set of molecules that are represented through a Gram matrix of size 2303×2303 obtained with a Tanimoto kernel. Tanimoto kernels (see Ralaivola et al. (2005) for details) are a common way to compare labeled graphs by means of a bag-of-sequences approach.

Experimental protocol: Robust KAE. We computed the mean 5-fold cross-validation Mean Squared Error. The first layer uses a linear kernel. But since inputs (and outputs) are kernelized – only the 2303×2303 Gram matrix is provided for learning –, the first layer may also be seen as a function from the associated Tanimoto-RKHS, applied to the molecules. The second layer uses a Gaussian kernel. The regularization parameters for the two layers have been fixed to $\Lambda = 1e-6$, and the inner dimension has been set to $p = 200$. In Figure 7 is plotted the MSE and the sparsity (discarded training data) for several values of ϵ in order to assess the effect of the regularization. We used an existing source code from Laforgue et al. (2019)[2], that has been adapted to our needs. The IOKR resolution part, materialized by the `compute_N_L` function therein, has been replaced by the `compute_Omega` function of the `IOKR_plus` class in the attached code.

C.2.3. Robust Function-to-Function Regression

Dataset Description. The task at hand consists in predicting lip acceleration from electromyography (EMG) signals of the corresponding muscle (Ramsay and Silverman, 2007). The dataset[3] includes 32 samples of time series obtained by recording a subject saying "say bob again", that are noted $(x_i, y_i)_{i=1}^{32}$. Each time series is of length 64. To assess the performance of our method in presence of outliers, we created 4 outliers by picking randomly some $(x_i)_{i=1}^{4}$ and adding to the dataset the samples $(x_i, -1.2 * y_i)_{i=1}^{4}$.

Experimental protocol. As the number of samples is small, one can use the Leave One Out (LOO) generalization error as a measure of the model performance. We first used it with plain Ridge Regression (Kadri et al., 2016) to select the best hyperparameter Λ. Then, we tested robustness by computing the LOO generalization error of a model output by solving Problem (9) for various κ (see Figure 8, that may also be reproduced from the attached code). For the $\{\psi_j\}_{j=1}^{m}$ we used the sine and cosine basis of $L^2([0,1])$, i.e. $\forall l \leqslant \frac{m}{2}$ and $\theta \in [0,1]$, $\psi_{2l}(\theta) = \sqrt{2}\cos(2\pi l\theta)$ and $\psi_{2l+1}(\theta) = \sqrt{2}\sin(2\pi l\theta)$. The number of basis function was set to $m = 16$, so that we get the first 8 cosines and sines of the basis. The chosen associated eigenvalues are $\lambda_{2l} = \lambda_{2l+1} = \frac{1}{(1+j)^2}$. We used as an input kernel the integral Laplacian $k_{\mathcal{X}}(x_1, x_2) = \int_0^1 \exp\left(-7|x_1(\theta) - x_2(\theta)|\right)\mathrm{d}\theta$.

C.3. Additional Figures

We now provide analogues to Figures 3 and 4 for the ϵ-SVR and κ-Huber. The ϵ-Ridge graphs are first recalled. Notice that simplified versions of these plots may be easily reproduced from the attached code.

The ϵ-KRR (Figure 12) appears as a natural regularized version of the plain KRR. For small values of Λ, the regularization effect of the ϵ induces a smaller MSE. This phenomenon is achieved for a wide range of Λ and ϵ, and coincides with an important sparsity. The counterpart is that no value of ϵ clearly allows to outperform the standard KRR for its optimal Λ. The ϵ-KRR may rather be used as an implicit regularization preventing from a cross-validation on Λ.

The ϵ-SVR (Figure 13) shares analogous characteristics for the small Λ regime. However, it further produces predictors with smaller MSE than the best KRR one. This furthermore coincides with a peak in the sparsity.

The κ-Huber (Figure 14) has a quite different behavior. When Λ tends to 0, the constraint (see Problem $(D2)$) is vacuous for all κ, and one asymptotically recovers the standard KRR. The optimal Λ now changes with κ, and better performances than the KRR for the best Λ are regularly attained.

[2] github.com/plaforgue/kae

[3] http://www.stats.ox.ac.uk/ silverma/fdacasebook/lipemg.html

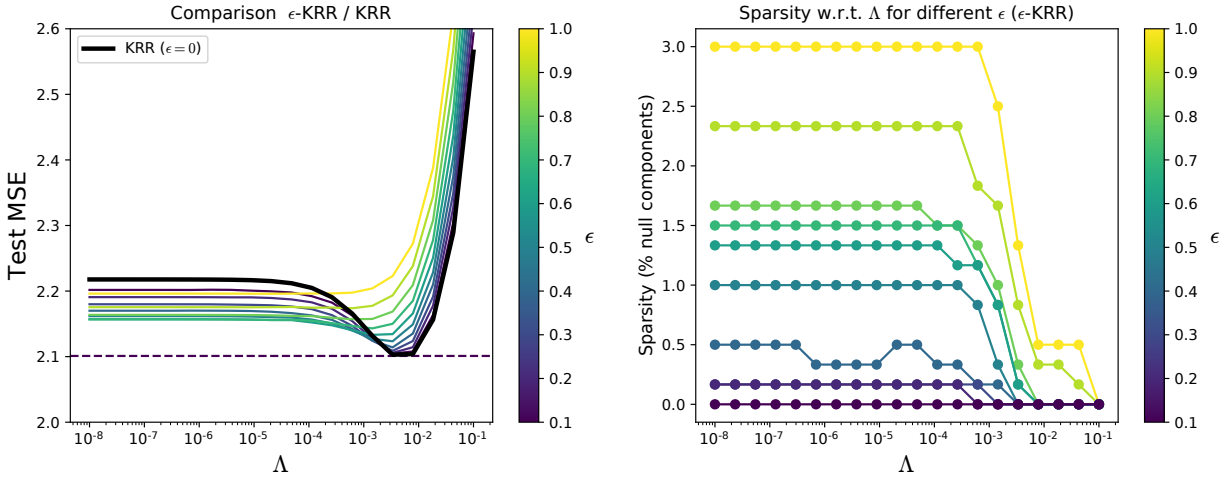

Figure 12. MSE and Sparsity w.r.t. Λ for different ϵ for the ϵ-KRR on the YEAST dataset.

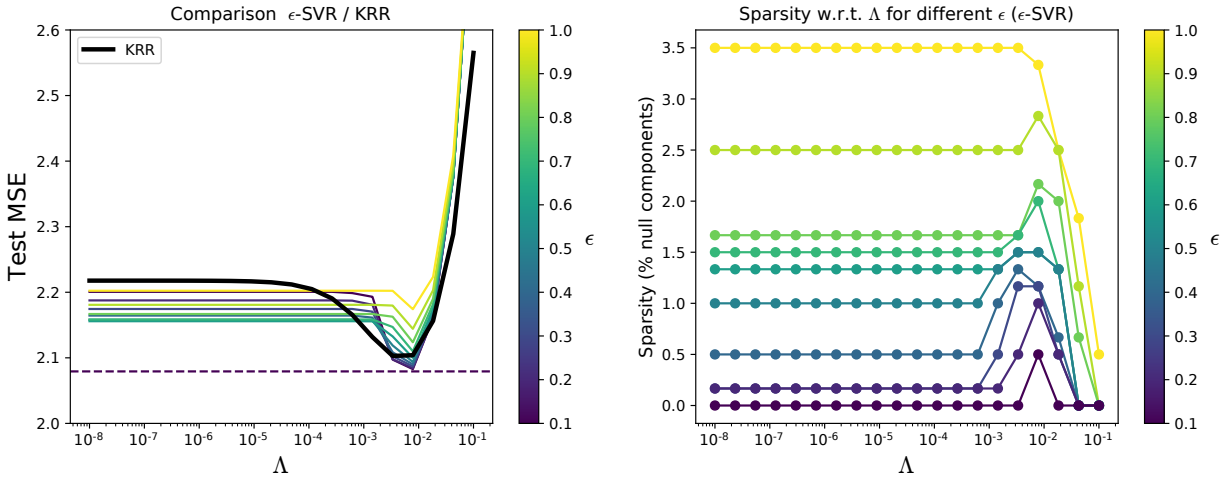

Figure 13. MSE and Sparsity w.r.t. Λ for different ϵ for the ϵ-SVR on the YEAST dataset.

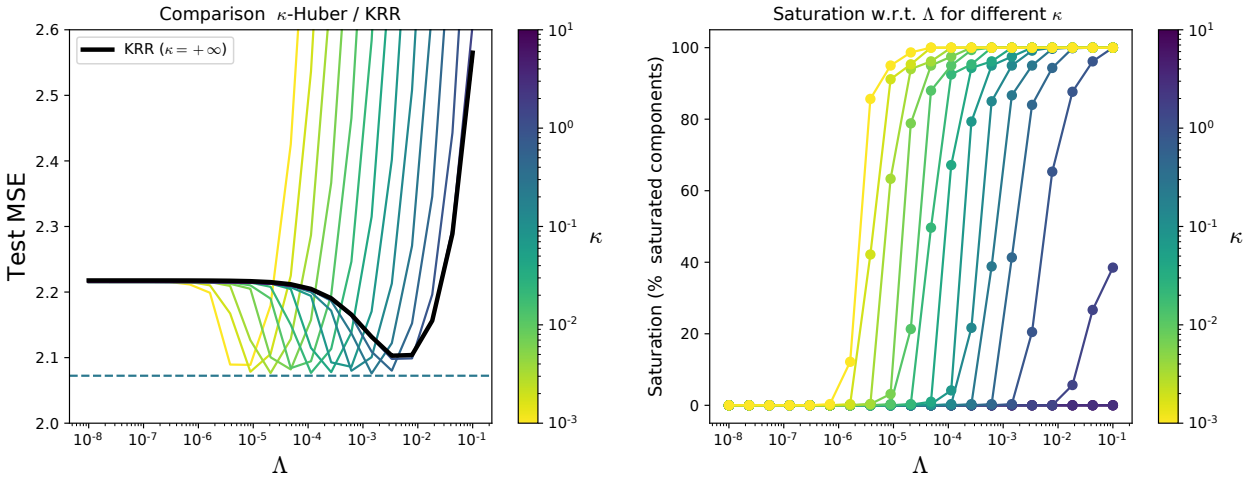

Figure 14. MSE and Saturation w.r.t. Λ for different κ for the κ-Huber on the YEAST dataset.

REFERENCES

Belanger, D. and McCallum, A. (2016). Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992.

Brouard, C., Shen, H., Dührkop, K., d'Alché Buc, F., Böcker, S., and Rousu, J. (2016a). Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12):i28–i36.

Brouard, C., Szafranski, M., and D'Alché-Buc, F. (2016b). Input output kernel regression: supervised and semi-supervised structured output prediction with operator-valued kernels. *The Journal of Machine Learning Research*, 17(1):6105–6152.

Elisseeff, A. and Weston, J. (2002). A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687.

Finley, T. and Joachims, T. (2008). Training structural svms when exact inference is intractable. In *Proceedings of the 25th international conference on Machine learning*, pages 304–311.

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

Laforgue, P., Clémençon, S., and d'Alché-Buc, F. (2019). Autoencoding any data through kernel autoencoders. In *Artificial Intelligence and Statistics*, pages 1061–1069.

Ralaivola, L., Swamidass, S. J., Saigo, H., and Baldi, P. (2005). Graph kernels for chemical informatics. *Neural networks*, 18(8):1093–1110.

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

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

Su, H., Heinonen, M., and Rousu, J. (2010). Structured output prediction of anti-cancer drug activity. In *IAPR International Conference on Pattern Recognition in Bioinformatics*, pages 38–49. Springer.