Duality in vv-RKHSs with Infinite Dimensional Outputs: Application to Robust Losses

Pierre Laforgue, Alex Lambert, Luc Brogat-Motte, Florence d'Alché-Buc

LTCI, Télécom Paris, Institut Polytechnique de Paris, France

Outline

Motivations

A duality theory for general OVKs

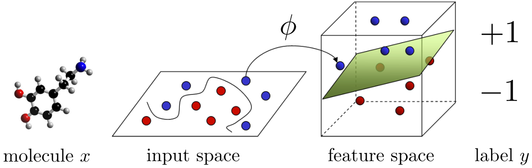
Robust losses as convolutions

Experiments

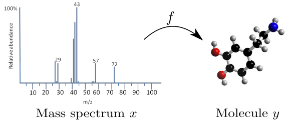
Conclusion

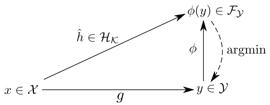
Motivation 1: structured prediction by surrogate approach

Kernel trick in the input space.



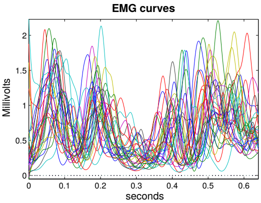
Kernel trick in the output space [Cortes '05, Geurts '06, Brouard '11, Kadri '13, Brouard '16], **Input Output Kernel Regression (IOKR).**





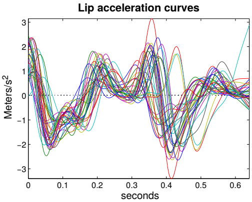
$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}_{\mathcal{K}}} \frac{1}{2n} \sum_{i=1}^{n} \left\| \phi(y_i) - h(x_i) \right\|_{\mathcal{F}_{\mathcal{Y}}}^{2} + \frac{\Lambda}{2} \|h\|_{\mathcal{H}_{\mathcal{K}}}^{2}, \qquad g(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \left\| \phi(y) - \hat{h}(x) \right\|_{\mathcal{F}_{\mathcal{Y}}}$$

Motivation 2: function to function regression



Input functions $\{x_i\}_{i \leq n}$

$h \longrightarrow$



Output functions $\{y_i\}_{i \leq n}$

$$\min_{h \in \mathcal{H}_{\mathcal{K}}} \quad \frac{1}{2n} \sum_{i=1}^{n} \left\| y_i - h(x_i) \right\|_{L^2}^2 + \frac{\Lambda}{2} \|h\|^2 \qquad \text{[Kadri et al., 2016]}$$

And many more!

e.g. *structured data autoencoding* [Laforgue et al., 2019]

$$\min_{h_1, h_2 \in \mathcal{H}_{\mathcal{K}}^1 \times \mathcal{H}_{\mathcal{K}}^2} \quad \frac{1}{2n} \sum_{i=1}^{n} \left\| \phi(x_i) - h_2 \circ h_1(\phi(x_i)) \right\|_{\mathcal{F}_{\mathcal{X}}}^2 + \Lambda \operatorname{Reg}(h_1, h_2).$$

Question: Is it possible to extend the previous approaches to different (ideally robust) loss functions?

First answer: Yes, possible extension to maximum-margin regression [Brouard et al., 2016], and ϵ-insensitive loss functions for matrix-valued kernels [Sangnier et al., 2017]

What about general Operator-Valued Kernels (OVKs)?

What about other types of loss functions?

Outline

Learning in vector-valued RKHSs (vv-RKHSs)

- $\mathcal{K}\colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$, $\quad \mathcal{K}(x, x') = \mathcal{K}(x', x)^*$, $\quad \sum_{i,j} \langle y_i, \mathcal{K}(x_i, x_j) y_j \rangle_{\mathcal{Y}} \geq 0$
- Unique vv-RKHS $\mathcal{H}_{\mathcal{K}} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$, $\quad \mathcal{H}_{\mathcal{K}} = \overline{\text{Span}\left\{\mathcal{K}(\cdot, x) y : x, y \in \mathcal{X} \times \mathcal{Y}\right\}}$
- **Ex:** decomposable OVK $\mathcal{K}(x, x') = k(x, x')A$, with k scalar, A p.s.d. on $\mathcal{Y}$

Learning in vector-valued RKHSs (vv-RKHSs)

- $\mathcal{K} \colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$, $\quad \mathcal{K}(x, x') = \mathcal{K}(x', x)^*$, $\quad \sum_{i,j} \langle y_i, \mathcal{K}(x_i, x_j) y_j \rangle_{\mathcal{Y}} \geq 0$
- Unique vv-RKHS $\mathcal{H}_\mathcal{K} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$, $\quad \mathcal{H}_\mathcal{K} = \overline{\text{Span}\left\{\mathcal{K}(\cdot, x) y : x, y \in \mathcal{X} \times \mathcal{Y}\right\}}$
- **Ex:** decomposable OVK $\mathcal{K}(x, x') = k(x, x')A$, with k scalar, A p.s.d. on $\mathcal{Y}$
- For $\{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ with $\mathcal{Y}$ a Hilbert space, we want to find:

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

Representer Theorem [Micchelli and Pontil, 2005]:

$$\exists (\hat{\alpha}_i)_{i=1}^n \in \mathcal{Y}^n \ \textbf{(infinite dimensional!)} \quad s.t. \quad \hat{h}(x) = \sum_{i=1}^n \mathcal{K}(\cdot, x_i) \hat{\alpha}_i.$$

When $\ell(\cdot, \cdot) = \frac{1}{2}\| \cdot - \cdot \|^2_{\mathcal{Y}}$, $\mathcal{K} = k \cdot \mathbf{I}_\mathcal{Y}$**:** $\hat{\alpha}_i = \sum_{j=1}^n A_{ij} y_j$, $\quad A = (K + n\Lambda \mathbf{I_n})^{-\mathbf{1}}$.

Applying duality

$$\hat{h} \in \underset{h \in \mathcal{H}_{\mathcal{K}}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \ell_i(h(x_i)) + \frac{\Lambda}{2} \|h\|_{\mathcal{H}_{\mathcal{K}}}^2 \quad \text{is given by} \quad \hat{h} = \frac{1}{\Lambda n} \sum_{i=1}^{n} \mathcal{K}(\cdot, x_i) \hat{\alpha}_i,$$

with $(\hat{\alpha}_i)_{i=1}^n \in \mathcal{Y}^n$ the solutions to the **dual problem**:

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

with $f^\star : \alpha \in \mathcal{Y} \mapsto \sup_{y \in \mathcal{Y}} \langle \alpha, y \rangle_{\mathcal{Y}} - f(y)$ the Fenchel-Legendre transform of f.

Applying duality

$$\hat{h} \in \underset{h \in \mathcal{H}_{\mathcal{K}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(h(x_i)) + \frac{\Lambda}{2} \|h\|_{\mathcal{H}_{\mathcal{K}}}^2 \quad \text{is given by} \quad \hat{h} = \frac{1}{\Lambda n} \sum_{i=1}^{n} \mathcal{K}(\cdot, x_i) \hat{\alpha}_i,$$

with $(\hat{\alpha}_i)_{i=1}^n \in \mathcal{Y}^n$ the solutions to the **dual problem**:

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

with $f^\star : \alpha \in \mathcal{Y} \mapsto \sup_{y \in \mathcal{Y}} \langle \alpha, y \rangle_{\mathcal{Y}} - f(y)$ the Fenchel-Legendre transform of f.

- **1st limitation:** the FL transform $\ell^\star$ needs to be computable ($\to$ assumption)
- **2nd limitation :** the dual variables $(\alpha_i)_{i=1}^n$ are still **infinite dimensional!**

Applying duality

$$\hat{h} \in \underset{h \in \mathcal{H}_{\mathcal{K}}}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n} \ell_i(h(x_i)) + \frac{\Lambda}{2}\|h\|^2_{\mathcal{H}_{\mathcal{K}}} \quad \text{is given by} \quad \hat{h} = \frac{1}{\Lambda n}\sum_{i=1}^{n} \mathcal{K}(\cdot, x_i)\hat{\alpha}_i,$$

with $(\hat{\alpha}_i)_{i=1}^n \in \mathcal{Y}^n$ the solutions to the **dual problem**:

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

with $f^\star : \alpha \in \mathcal{Y} \mapsto \sup_{y \in \mathcal{Y}} \langle \alpha, y \rangle_{\mathcal{Y}} - f(y)$ the Fenchel-Legendre transform of f.

- **1st limitation:** the FL transform $\ell^\star$ needs to be computable ($\rightarrow$ assumption)
- **2nd limitation :** the dual variables $(\alpha_i)_{i=1}^n$ are still **infinite dimensional!**

If $\mathbf{Y} = \text{Span}\{y_j,\ j \leq n\}$ invariant by $\mathcal{K}$, *i.e.* $\forall (x, x'), \quad y \in \mathbf{Y} \ \Rightarrow \ \mathcal{K}(x, x')y \in \mathbf{Y}$:

then $\hat{\alpha}_i \in \mathbf{Y} \quad \rightarrow \quad$ possible reparametrization: $\hat{\alpha}_i = \sum_j \hat{\omega}_{ij} y_j$

The double representer theorem (1/2)

Assume that OVK $\mathcal{K}$ and loss ℓ satisfy the appropriate assumptions (see paper for details, verified by standard kernels and losses), then

$\hat{h} = \underset{\mathcal{H}_\mathcal{K}}{\text{argmin}} \ \frac{1}{n}\sum_i \ell(h(x_i), y_i) + \frac{\Lambda}{2}\|h\|^2_{\mathcal{H}_\mathcal{K}}$ is given by

$$\hat{h} = \frac{1}{\Lambda n}\sum_{i,j=1}^{n} \mathcal{K}(\cdot, x_i)\ \hat{\omega}_{ij}\ y_j,$$

with $\hat{\Omega} = [\hat{\omega}_{ij}] \in \mathbb{R}^{n\times n}$ the solution to the **finite dimensional** problem

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

with $\tilde{M}$ the $n^2\times n^2$ matrix writing of M s.t. $M_{ijkl} = \langle y_k, \mathcal{K}(x_i, x_j)y_l\rangle_\mathcal{Y}$.

The double representer theorem (2/2)

If $\mathcal{K}$ further satisfies $\mathcal{K}(x, x') = \sum_t k_t(x, x') A_t$, then tensor M simplifies to $M_{ijkl} = \sum_t [K_t^X]_{ij} [K_t^Y]_{kl}$ and the problem rewrites

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

Rmk. Only need the n^4 tensor $\langle y_k, \mathcal{K}(x_i, x_j) y_l \rangle_{\mathcal{Y}}$ to learn OVKMs.

Simplifies to 2 n^2 matrices $K_{ij}^X K_{kl}^Y$ if $\mathcal{K}$ is decomposable.

How to apply the duality approach?

Outline

Infimal convolution and Fenchel-Legendre transforms

Infimal-convolution operator $\square$ between proper lower semicontinuous functions [Bauschke et al., 2011]:

$$(f \,\square\, g)(x) = \inf_y f(y) + g(x-y).$$

Relation to FL transform:

$$(f \,\square\, g)^\star = f^\star + g^\star$$

Ex: ϵ-insensitive losses. Let $\ell : \mathcal{Y} \to \mathbb{R}$ be a convex loss with unique minimum at 0, and $\epsilon > 0$. The ϵ-insensitive version of ℓ, denoted ℓ_ϵ, is defined by:

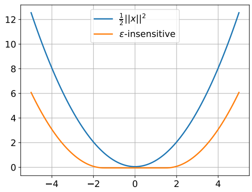
$$\ell_\epsilon(y) = (\ell \,\square\, \chi_{\mathcal{B}_\epsilon})(y) = \begin{cases} \ell(0) & \text{if } \|y\|_{\mathcal{Y}} \leq \epsilon \\ \inf_{\|d\|_{\mathcal{Y}} \leq 1} \ell(y - \epsilon d) & \text{otherwise} \end{cases},$$

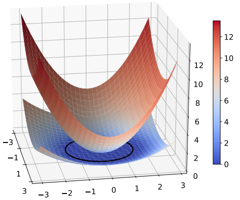
and has FL transform:

$$\ell_\epsilon^\star(y) = (\ell \,\square\, \chi_{\mathcal{B}_\epsilon})^\star(y) = \ell^\star(y) + \epsilon\|y\|.$$

Interesting loss functions: sparsity and robustness

ϵ-Ridge

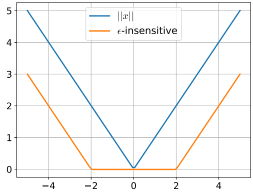


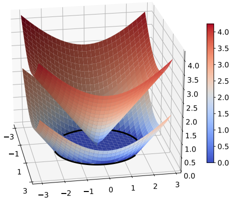


$$\frac{1}{2}\|\cdot\|^2 \,\square\, \chi_{\mathcal{B}_\epsilon}$$

(Sparsity)

ϵ-SVR

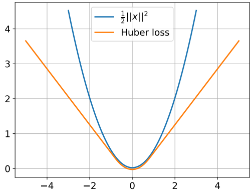


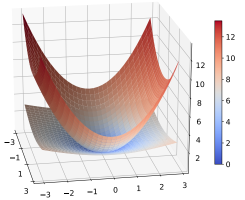


$$\|\cdot\| \,\square\, \chi_{\mathcal{B}_\epsilon}$$

(Sparsity, Robustness)

κ-Huber





$$\kappa\|\cdot\| \,\square\, \frac{1}{2}\|\cdot\|^2$$

(Robustness)

Specific dual problems

For the ϵ-ridge, ϵ-SVR and κ-Huber, it holds $\hat{\Omega} = \hat{W} V^{-1}$, with $\hat{W}$ the solution to these finite dimensional dual problems:

$$(D1) \quad \min_{W \in \mathbb{R}^{n \times n}} \quad \frac{1}{2} \|AW - B\|_{\text{Fro}}^2 + \epsilon \|W\|_{2,1},$$

$$(D2) \quad \min_{W \in \mathbb{R}^{n \times n}} \quad \frac{1}{2} \|AW - B\|_{\text{Fro}}^2 + \epsilon \|W\|_{2,1}, \quad \text{s.t.} \quad \|W\|_{2,\infty} \leq 1,$$

$$(D3) \quad \min_{W \in \mathbb{R}^{n \times n}} \quad \frac{1}{2} \|AW - B\|_{\text{Fro}}^2, \quad \text{s.t.} \quad \|W\|_{2,\infty} \leq \kappa,$$

with V, A, B such that: $VV^\top = K^Y$, $A^\top A = K^X/(\Lambda n) + \mathbf{I}_n$ (or $A^\top A = K^X/(\Lambda n)$ for the ϵ-SVR), and $A^\top B = V$.

Outline

Surrogate approaches for structured prediction

- Experiments on YEAST dataset
- Empirically, ϵ-SV-IOKR outperforms ridge-IOKR for a wide range of ϵ
- Promotes sparsity and acts as a regularizer

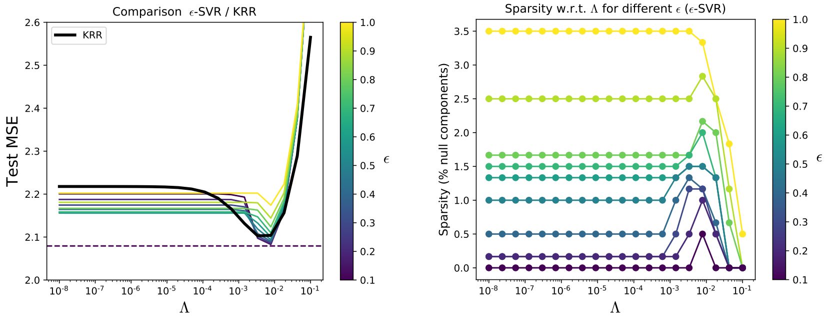


Figure 1: MSEs and sparsity w.r.t. Λ for several ϵ

Robust function-to-function regression

Task from [Kadri et al., 2016]: predict lip acceleration from EMG signals.

- Dataset augmented with outliers, model learned with Huber loss
- Improvement for every output size M (see paper for approximation)

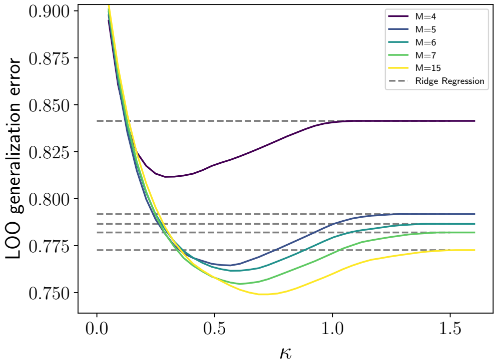


Figure 2: LOO generalization error w.r.t. κ

Outline

Conclusion

State of the art:

- OVK and vv-RKHSs tailored to infinite dimensional outputs
- RT: expansion with few information on the coefficients
- Duality: coefficients solutions to the (infinite) dual problem

Contributions:

- Double RT: coefficients linear combinations of the outputs
- Allows to cope with many losses (ϵ, Huber) and kernels
- Empirical improvements on surrogate approaches

Much more in the paper!

- Thorough algorithmic stability analysis
- What if $\mathbf{Y}$ is not invariant by $\mathcal{K}$?

References I

Audiffren, J. and Kadri, H. (2013).
Stability of multi-task kernel regression algorithms.
In *Asian Conference on Machine Learning*, pages 1–16.

Bauschke, H. H., Combettes, P. L., et al. (2011).
***Convex analysis and monotone operator theory in Hilbert spaces*, volume 408.**
Springer.

Bousquet, O. and Elisseeff, A. (2002).
Stability and generalization.
Journal of Machine Learning Research, 2(Mar):499–526.

Brouard, C., Szafranski, M., and d'Alché-Buc, F. (2016).
Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernel.
Journal of Machine Learning Research, 17:176:1–176:48.

References II

Huber, P. J. (1964).
Robust estimation of a location parameter.
The Annals of Mathematical Statistics, pages 73–101.

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, 17:20:1–20:54.

Kadri, H., Ghavamzadeh, M., and Preux, P. (2013).
A generalized kernel approach to structured output learning.
In *International Conference on Machine Learning (ICML)*, pages 471–479.

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

Micchelli, C. A. and Pontil, M. (2005).
On learning vector-valued functions.
Neural computation, 17(1):177–204.

Moreau, J. J. (1962).
Fonctions convexes duales et points proximaux dans un espace hilbertien.
Comptes rendus hebdomadaires des séances de l'Académie des sciences, 255:2897–2899.

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.

On the invariance assumption

With $\mathbf{Y} = \text{Span}\{y_j,\ j \leq n\}$, the assumption reads:

$$\forall (x, x') \in \mathcal{X}^2,\ \forall y \in \mathcal{Y}, \quad y \in \mathbf{Y} \implies \mathcal{K}(x, x')y \in \mathbf{Y}$$

- We do not need it to hold for every collection of $\{y_i\}_{i \leq n} \in \mathcal{Y}^n$
- Rather an a posteriori condition to ensure that the kernel is *aligned*
- The little we know about $\mathcal{Y}$ should be preserved through $\mathcal{K}$
- If $\mathcal{Y}$ finite dimensional, and sufficiently many outputs, then $\mathbf{Y} = \mathcal{Y}$
- Identity-decomposable kernels fit (nontrivial in infinite dimension)
- The empirical covariance kernel $\sum_i y_i \otimes y_i$ [Kadri et al., 2013] fits

Admissible kernels

- $\mathcal{K}(s,t) = \sum_i \; k_i(s,t) \; y_i \otimes y_i$,
  with k_i positive semi-definite (p.s.d.) scalar kernels for all $i \leq n$
- $\mathcal{K}(s,t) = \sum_i \; \mu_i \; k(s,t) \; y_i \otimes y_i$,
  with k a p.s.d. scalar kernel and $\mu_i \geq 0$ for all $i \leq n$
- $\mathcal{K}(s,t) = \sum_i \; k(s,x_i)k(t,x_i) \; y_i \otimes y_i$,
- $\mathcal{K}(s,t) = \sum_{i,j} \; k_{ij}(s,t) \; (y_i + y_j) \otimes (y_i + y_j)$,
  with k_{ij} p.s.d. scalar kernels for all $i, j \leq n$
- $\mathcal{K}(s,t) = \sum_{i,j} \; \mu_{ij} \; k(s,t) \; (y_i + y_j) \otimes (y_i + y_j)$,
  with k a p.s.d. scalar kernel and $\mu_{ij} \geq 0$
- $\mathcal{K}(s,t) = \sum_{i,j} \; k(s,x_i,x_j)k(t,x_i,x_j) \; (y_i + y_j) \otimes (y_i + y_j)$.

Admissible losses

$$\forall i \leq n,\ \forall (\alpha^{\mathbf{Y}}, \alpha^{\perp}) \in \mathbf{Y} \times \mathbf{Y}^{\perp}, \qquad \ell_i^\star(\alpha^{\mathbf{Y}}) \leq \ell_i^\star(\alpha^{\mathbf{Y}} + \alpha^{\perp})$$

- $\ell_i(y) = f(\langle y, z_i \rangle)$, $z_i \in Y$ and $f : \mathbb{R} \to \mathbb{R}$ convex. Maximum-margin obtained with $z_i = y_i$ and $f(t) = \max(0, 1 - t)$.

- $\ell(y) = f(\|y\|)$, $f : \mathbb{R}_+ \to \mathbb{R}$ convex increasing s.t. $t \mapsto \frac{f'(t)}{t}$ is continuous over $\mathbb{R}_+$. Includes the functions $\frac{\lambda}{\eta}\|y\|_{\mathcal{Y}}^{\eta}$ for $\eta > 1$, $\lambda > 0$.

- $\forall \lambda > 0$, with $\mathcal{B}_\lambda$ the centered ball of radius λ,
  - $\ell(y) = \lambda\|y\|$,
  - $\ell(y) = \lambda\|y\| \log(\|y\|)$,
  - $\ell(y) = \chi_{\mathcal{B}_\lambda}(y)$,
  - $\ell(y) = \lambda(\exp(\|y\|) - 1)$.

- $\ell_i(y) = f(y - y_i)$, $f^\star$ verifying the condition.

- Infimal convolution of functions verifying the condition. (ϵ-insensitive [Sangnier et al., 2017], the Huber loss [Huber, 1964], Moreau or Pasch-Hausdorff envelopes [Moreau, 1962, Bauschke et al., 2011])

Dual problem:

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

- Decompose $\hat{\alpha}_i = \alpha_i^{\mathbf{Y}} + \alpha_i^{\perp}$, with $(\alpha_i^{\mathbf{Y}})_{i \leq n}, (\alpha_i^{\perp})_{i \leq n} \in \mathbf{Y}^n \times \mathbf{Y}^{\perp\, n}$
- Assume that $\ell_i^\star(\alpha^{\mathbf{Y}}) \leq \ell_i^\star(\alpha^{\mathbf{Y}} + \alpha^{\perp})$ (satisfied if ℓ relies on $\langle \cdot, \cdot \rangle$)

Then it holds:

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

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

Approximating the dual problem if no invariance

The kernel $\mathcal{K} = k \cdot A$ is a separable OVK, with A a compact operator.

There exists an o.n.b. $(\psi_j)_{j=1}^{\infty}$ of $\mathcal{Y}$, s.t. $A = \sum_{j=1}^{\infty} \lambda_j \psi_j \otimes \psi_j$, $(\lambda_j \geq 0)$.

There exists $(\hat{\omega}_i)_{i=1}^{n} \in \ell^2(\mathbb{R})^n$ such that $\forall i \leq n,\ \hat{\alpha}_i = \sum_{j=1}^{\infty} \hat{\omega}_{ij} \psi_j$.

Denoting by $\widetilde{\mathcal{Y}}_m = \text{span}(\{\psi_j\}_{j=1}^{m})$, $S = \text{diag}(\lambda_j)_{j=1}^{m}$, solve instead:

$$\min_{(\alpha_i)_{i=1}^{n} \in \widetilde{\mathcal{Y}}_m^n} \sum_{i=1}^{n} \ell_i^{\star}(-\alpha_i) + \frac{1}{2\Lambda n} \sum_{i,j=1}^{n} \langle \alpha_i, \mathcal{K}(x_i, x_j)\alpha_j \rangle_{\mathcal{Y}} .$$

The final solution is given by: $\hat{h} = \frac{1}{\Lambda n} \sum_{i=1}^{n} \sum_{j=1}^{m} k(\cdot, x_i)\ \lambda_j\ \hat{\omega}_{ij}\ \psi_j,$

with $\hat{\Omega}$ solution to:

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

Application to kernel autoencoding

- Experiments on molecules with Tanimoto-Gaussian kernel
- Empirical improvements for wide range of ϵ
- Introduces sparsity

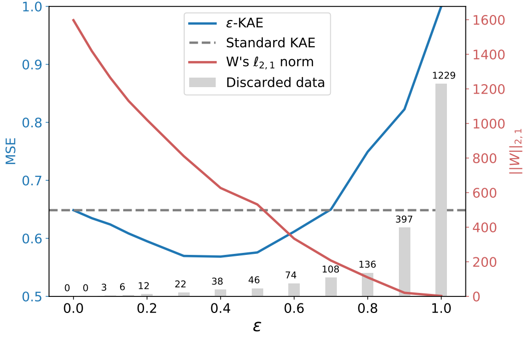


Figure 3: Performances of ϵ-insensitive Kernel Autoencoder

Algorithmic stability analysis [Bousquet and Elisseeff, 2002]

Algorithm A has stability β if for any sample $\mathcal{S}_n$, and any $i \leq n$, it holds:

$$\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |\ell(h_{A(\mathcal{S}_n)}(x), y) - \ell(h_{A(\mathcal{S}_n^{\setminus i})}(x), y)| \leq \beta$$

Let A be an algorithm with stability β and loss function bounded by M. Then, for any $n \geq 1$ and $\delta \in]0, 1[$ it holds with probability at least $1 - \delta$:

$$\mathcal{R}(h_{A(\mathcal{S}_n)}) \leq \hat{\mathcal{R}}_n(h_{A(\mathcal{S}_n)}) + 2\beta + (4n\beta + M)\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

If $\|\mathcal{K}(x, x)\|_{\text{op}} \leq \gamma^2$, and $|\ell(h_{\mathcal{S}}(x), y) - \ell(h_{\mathcal{S}^{\setminus i}}(x), y)| \leq C\|h_{\mathcal{S}}(x) - h_{\mathcal{S}^{\setminus i}}(x)\|_{\mathcal{Y}}$, then OVK algorithm has stability $\beta \leq C^2\gamma^2/(\Lambda n)$ [Audiffren and Kadri, 2013].

	M	C
ϵ-SVR	$\sqrt{M_{\mathcal{Y}} - \epsilon}\left(\frac{\sqrt{2}\gamma}{\sqrt{\Lambda}} + \sqrt{M_{\mathcal{Y}} - \epsilon}\right)$	1
ϵ-Ridge	$(M_{\mathcal{Y}} - \epsilon)^2\left(1 + \frac{2\sqrt{2}\gamma}{\sqrt{\Lambda}} + \frac{2\gamma^2}{\Lambda}\right)$	$2(M_{\mathcal{Y}} - \epsilon)\left(1 + \frac{\gamma\sqrt{2}}{\sqrt{\Lambda}}\right)$
κ-Huber	$\kappa\sqrt{M_{\mathcal{Y}} - \frac{\kappa}{2}}\left(\frac{\gamma\sqrt{2\kappa}}{\sqrt{\Lambda}} + \sqrt{M_{\mathcal{Y}} - \frac{\kappa}{2}}\right)$	κ