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

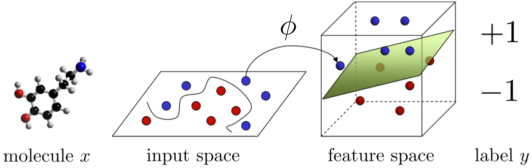
Pierre Laforgue, Alex Lambert, Luc Brogat-Motte, Florence d'Alché-Buc

LTCI, Télécom Paris, Institut Polytechnique de Paris, France

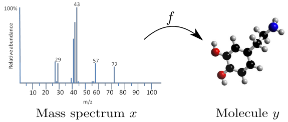
Outline

Motivation 1: structured prediction by surrogate approach

Kernel trick in the input space.



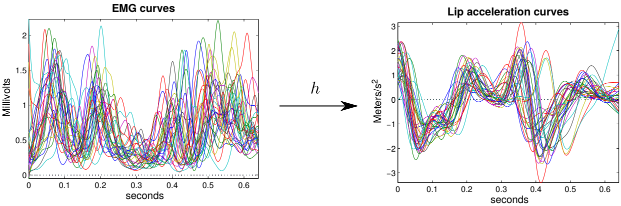
Kernel trick in the output space [Cortes '05, Geurts '06, Brouard '11, Kadri '13, Brouard '16]**, Input Output Kernel Regression (IOKR).**



$$
\begin{tikzcd}
 & & \phi(y) \in \mathcal{F}_{\mathcal{Y}} \arrow[d, dashed, bend left, "\text{argmin}"] \\
x \in \mathcal{X} \arrow[urr, "\hat{h} \in \mathcal{H}_{\mathcal{K}}"] \arrow[rr, "g"'] & & y \in \mathcal{Y} \arrow[u, "\phi"]
\end{tikzcd}
$$

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

Motivation 2: function to function regression



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

$$\min_{h \in \mathcal{H}_{\mathcal{K}}} \quad \frac{1}{2n} \sum_{i=1}^{n} \left\| y_i - h(x_i) \right\|_{L^2}^2 + \frac{\Lambda}{2} \|h\|^2$$ [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 $\mathcal{H}_{\mathcal{K}} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$, $\quad \mathcal{H}_{\mathcal{K}} = \overline{\mathrm{Span}\left\{\mathcal{K}(\cdot, x) y : x, y \in \mathcal{X} \times \mathcal{Y}\right\}}$
- **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:

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

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

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

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

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!**

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

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

Motivations

A duality theory for general OVKs

Robust losses as convolutions

Experiments

Conclusion

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:

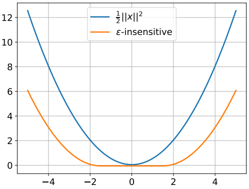
$$\ell_\epsilon(y) = (\ell \,\square\, \chi_{\mathcal{B}_\epsilon})(y) = \begin{cases} \ell(0) & \text{if } \|y\|_{\mathcal{Y}} \le \epsilon \\ \inf\limits_{\|d\|_{\mathcal{Y}} \le 1} \ell(y - \epsilon d) & \text{otherwise} \end{cases},$$

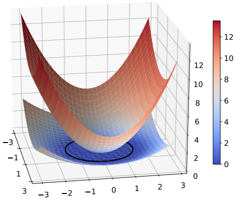
and has FL transform:

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

Interesting loss functions: sparsity and robustness

ϵ-Ridge

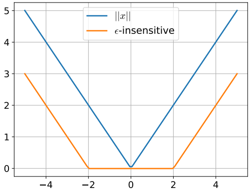


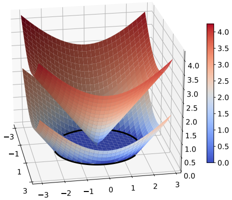


$$\frac{1}{2}\|\cdot\|^2 \,\square\, \chi_{\mathcal{B}_\epsilon}$$

(Sparsity)

ϵ-SVR

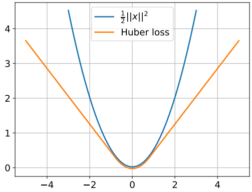


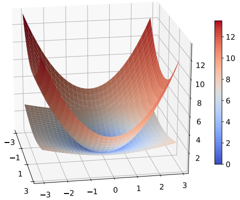


$$\|\cdot\| \,\square\, \chi_{\mathcal{B}_\epsilon}$$

(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:

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

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

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

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

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

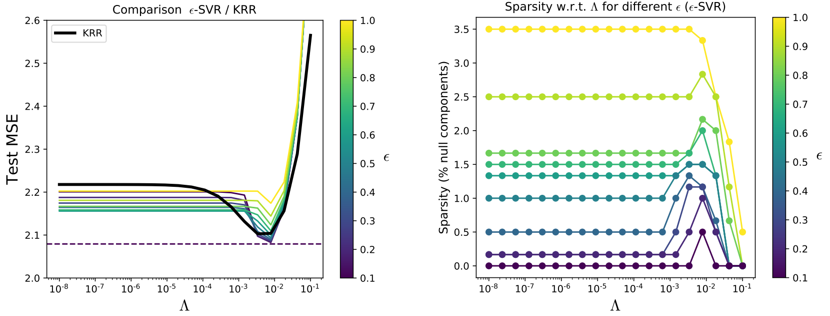


Figure 1: MSEs and sparsity w.r.t. Λ for several ϵ

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)

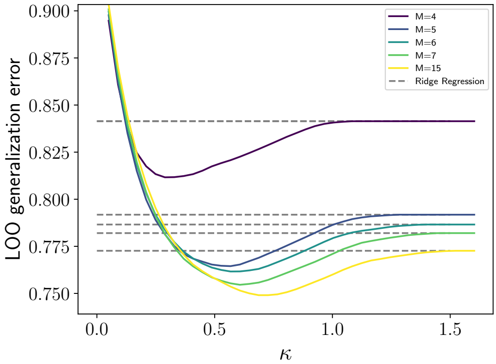


Figure 2: LOO generalization error w.r.t. κ

Outline

Motivations

A duality theory for general OVKs

Robust losses as convolutions

Experiments

Conclusion

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

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

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.

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.

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

References II

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

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.