Deep Kernel Representation Learning for Complex Data and Reliability Issues

Pierre LAFORGUE, *PhD Defense*
LTCI, Télécom Paris, France

Jury:

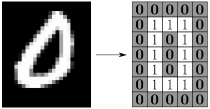
d'ALCHÉ-BUC Florence	Supervisor
BONALD Thomas	President
CLÉMENÇON Stephan	co-Supervisor
KADRI Hachem	Examiner
LUGOSI Gábor	Reviewer
MAIRAL Julien	Examiner
VERT Jean-Philippe	Reviewer

Motivation: need for structured data representations

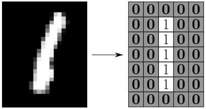
Goal of ML: infer from a set of examples, the relationship between some explanatory variables x, and a target output y

A representation: set of features characterizing the observations

Ex 1:
digit recognition (MNIST)

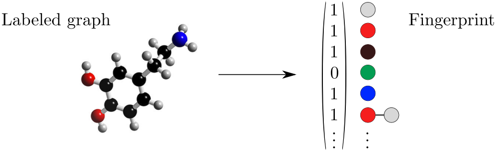


0	0	0	0	0
0	1	1	1	0
0	1	0	1	0
0	1	0	1	0
0	1	0	1	0
0	1	1	1	0
0	0	0	0	0



0	0	0	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	0	0	0

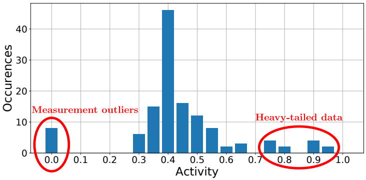
Ex 2:
molecule activity prediction



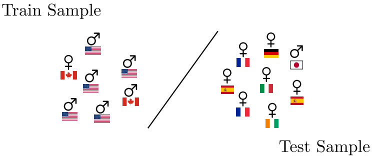
How to (automatically) learn structured data representations?

Motivation: need for reliable procedures

Empirical Risk Minimization: minimize the average error on train data



Ordinary Least Squares fail, need for more robust loss functions and/or mean estimators



Importance Sampling may only correct on the space covered by the training observations

How to adapt to data with outliers and/or biased?

Outline for today

Empirical Risk Minimization (ERM), formally:

$$\min_{h \text{ measurable}} \mathbb{E}_P\Big[\ell(h(X), Y)\Big] \quad \rightarrow \quad \min_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(h(x_i), y_i\Big)$$

Part I: Deep kernel architectures for complex data

Part II: Robust losses for RKHSs with infinite dimensional outputs

Part III: Reliable learning through Median-of-Means approaches

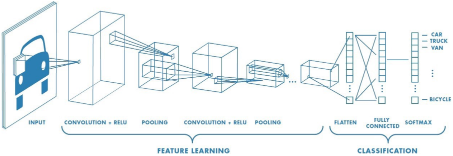
Backup: Statistical learning from biased training samples

Part I:
Deep kernel architectures for complex data

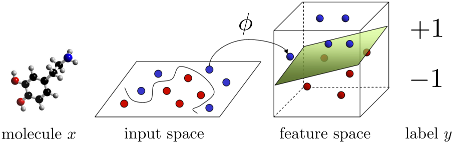
$$\min_{h \text{ measurable}} \mathbb{E}_P\Big[\ell(h(X), Y)\Big] \quad \rightarrow \quad \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\Big(h(x_i), y_i\Big)$$

Two opposite representation learning paradigms

Deep Learning: representations learned along with the training, key to the success [Erhan et al., 2009]



Kernel Methods: linear method after embedding through feature map ϕ, choice of kernel $\Longleftrightarrow$ choice of representation



Question: Is it possible to combine both approaches [Mairal et al., 2014]?

Autoencoders (AEs)

- **Idea:** compress and reconstruct inputs by a Neural Net (NN)
- Base mapping: $f : \mathbb{R}^d \to \mathbb{R}^p$ such that $f_{\boldsymbol{W},\boldsymbol{b}}(x) = \sigma(\boldsymbol{W}x + \boldsymbol{b})$
- Hour-glass shaped network, reconstruction criterion:

$$\min_{\boldsymbol{W},\boldsymbol{b},\boldsymbol{W'},\boldsymbol{b'}} \left\|x - f_{\boldsymbol{W'},\boldsymbol{b'}} \circ f_{\boldsymbol{W},\boldsymbol{b}}(x)\right\|^2 \qquad \text{(self-supervised)}$$

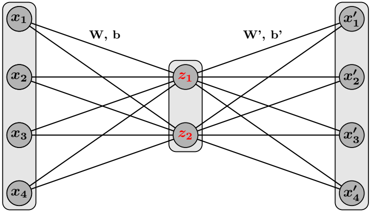


Fig. 1: A 1 hidden layer autoencoder

Autoencoders: uses

- Data compression, link to Principal Component Analysis (PCA) [Bourlard and Kamp, 1988, Hinton and Salakhutdinov, 2006]
- Pre-training of neural networks [Bengio et al., 2007]
- Denoising [Vincent et al., 2010]
- **For non-vectorial data?**

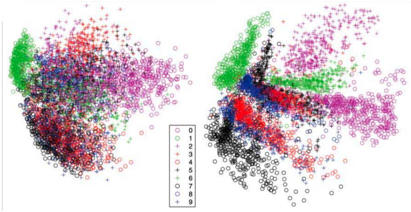


Fig. 2: PCA/AE on MNIST (reproduced from HS '06)

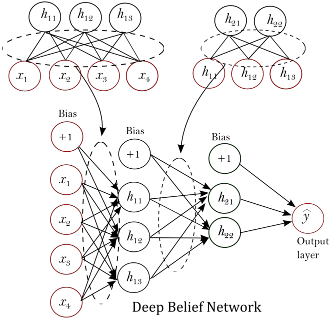


Fig. 3: Pre-training of bigger network through AEs

Scalar kernel methods [Schölkopf et al., 2004]

- feature map $\phi\colon \mathcal{X} \to \mathcal{H}_k$ associated to scalar kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}_k} = k(x, x')$

- Replace x with $\phi(x)$ and use linear methods. Ridge regression:

$$\min_{\beta \in \mathbb{R}^p} \sum_i (y_i - \langle x_i, \beta \rangle_{\mathbb{R}^p})^2 + 2n\lambda \|\beta\|^2_{\mathbb{R}^p}$$

$$\min_{\omega \in \mathcal{H}_k} \sum_i (y_i - \langle \phi(x_i), \omega \rangle_{\mathcal{H}_k})^2 + 2n\lambda \|\omega\|^2_{\mathcal{H}_k}$$

- In an autoencoder? **Need for Hilbert-valued functions!**

$$\min_{f_l \in \mathrm{NN}_{\mathrm{em}}} \quad \frac{1}{n} \sum_{i=1}^{n} \left\| \phi(x_i) - f_L \circ \ldots \circ f_1(\phi(x_i)) \right\|^2_{\mathcal{H}_k}$$

Operator-valued kernel methods [Carmeli et al., 2006]

Generalization of scalar kernel methods to output Hilbert spaces:

- $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ $\qquad$ $\mathcal{K} \colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$
- $k(x', x) = k(x, x')$ $\qquad$ $\mathcal{K}(x', x) = \mathcal{K}(x, x')^*$
- $\sum_{i,j} \alpha_i k(x_i, x_j) \alpha_j \geq 0$ $\qquad$ $\sum_{i,j} \langle \alpha_i, \mathcal{K}(x_i, x_j) \alpha_j \rangle_{\mathcal{Y}} \geq 0$
- $\mathcal{H}_k = \overline{\mathrm{Span}\left\{k(\cdot, x)\right\}} \subset \mathbb{R}^{\mathcal{X}}$ $\qquad$ $\mathcal{H}_{\mathcal{K}} = \overline{\mathrm{Span}\left\{\mathcal{K}(\cdot, x) y\right\}} \subset \mathcal{Y}^{\mathcal{X}}$

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

$$\begin{tikzcd}
\phi(x) \in \mathcal{F}_{\mathcal{X}} \arrow[rd, "\hat{h} \in \mathcal{H}_k"] & \\
x \in \mathcal{X} \arrow[u, "\phi"] \arrow[r, "g"'] & y \in \mathbb{R}
\end{tikzcd}$$

$$\begin{tikzcd}
& \phi(z) \in \mathcal{F}_{\mathcal{Z}} = \mathcal{Y} \arrow[d, dashed, bend left, "\mathrm{argmin}"] \\
x \in \mathcal{X} \arrow[ru, "\hat{h} \in \mathcal{H}_{\mathcal{K}}"] \arrow[r, "g"'] & z \in \mathcal{Z} \arrow[u, "\phi"]
\end{tikzcd}$$

How to learn in vector-valued RKHSs? OVK ridge regression

For $\{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ with $\mathcal{Y}$ a Hilbert space, we want to solve:

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

Representer Theorem [Micchelli and Pontil, 2005]:

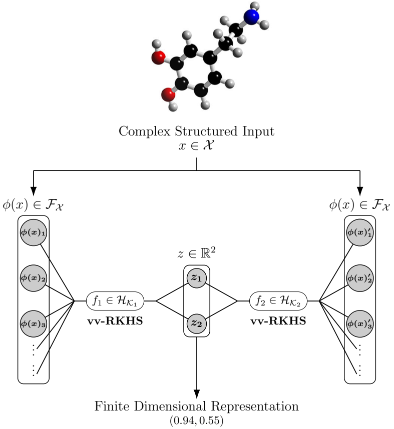
$$\exists (\hat{\alpha}_i)_{i=1}^n \in \mathcal{Y}^n \quad \text{s.t.} \quad \hat{h}(x) = \sum_{i=1}^n \mathcal{K}(\cdot, x_i) \hat{\alpha}_i, \quad \text{and differentiating gives:}$$

$$\begin{cases} \sum_{i=1}^n \left(\mathcal{K}(x_1, x_i) + \Lambda n \delta_{1i} \mathbf{I}_\mathcal{Y} \right) \hat{\alpha}_i = y_1, \\ \qquad \qquad \dots \\ \sum_{i=1}^n \left(\mathcal{K}(x_n, x_i) + \Lambda n \delta_{ni} \mathbf{I}_\mathcal{Y} \right) \hat{\alpha}_i = y_n. \end{cases}$$

If $\mathcal{K}(x, x') = k(x, x') \mathbf{I}_\mathcal{Y}$, then **closed form solution**:

$$\hat{\alpha}_i = \textstyle\sum_j A_{ij} y_j \quad \text{with} \quad A = (K + \Lambda n \mathbf{I}_n)^{-1}$$

The Kernel Autoencoder [Laforgue et al., 2019a]


Complex Structured Input
$x \in \mathcal{X}$
$\phi(x) \in \mathcal{F}_\mathcal{X}$
$\phi(x)_1$
$\phi(x)_2$
$\phi(x)_3$
$f_1 \in \mathcal{H}_{\mathcal{K}_1}$
vv-RKHS
$z \in \mathbb{R}^2$
z_1
z_2
$f_2 \in \mathcal{H}_{\mathcal{K}_2}$
vv-RKHS
$\phi(x) \in \mathcal{F}_\mathcal{X}$
$\phi(x)'_1$
$\phi(x)'_2$
$\phi(x)'_3$
Finite Dimensional Representation
$(0.94, 0.55)$

$$\mathbf{K^2AE:} \min_{f_l \in \text{vv-RKHS}} \frac{1}{n} \sum_{i=1}^{n} \left\| \phi(x_i) - f_L \circ \ldots \circ f_1(\phi(x_i)) \right\|^2_{\mathcal{F}_\mathcal{X}} + \sum_{l=1}^{L} \lambda_l \|f_l\|^2_{\mathcal{H}_l}$$

Connection to kernel Principal Component Analysis (PCA)

2-layer K^{2}AE with linear kernels, internal layer of size p, and no penalization. Let $((\sigma_1, u_1) \ldots, (\sigma_p, u_p))$ denote the largest eigen values/vectors of K_{in}. It holds:

K^{2}AE output: $\left(\sqrt{\sigma_1} u_1, \ldots, \sqrt{\sigma_p} u_p\right) \in \mathbb{R}^{n \times p}$

KPCA output: $(\sigma_1 u_1, \ldots, \sigma_p u_p) \in \mathbb{R}^{n \times p}$

Proof: $X \in \mathbb{R}^{n \times d}, \;\; Y = XX^\top A \in \mathbb{R}^{n \times p}, \;\; Z = YY^\top B.$

The objective writes $\min_{A,B} \; \|X - Z\|^2_{\mathrm{Fro}}$ and Eckart-Young gives:

$$Z^* = U_d \; \overline{\Sigma}_p \; V_d{}^\top \quad \text{with} \quad X = U_d \; \overline{\Sigma}_d \; V_d{}^\top.$$

Sufficient: $A = U_p \; \overline{\Sigma}_p{}^{-\frac{3}{2}} \in \mathbb{R}^{n \times p} \qquad B = U_d \; V_d{}^\top \in \mathbb{R}^{n \times d}.$

Extends to $X \in \mathcal{L}(\mathcal{Y}, \mathbb{R}^n)$ as SVD exists for compact operators.

A composite representer theorem [Laforgue et al., 2019a]

How to train the Kernel Autoencoder?

$$\min_{f_l \in \text{vv-RKHS}} \quad \frac{1}{n}\sum_{i=1}^{n} \left\| \phi(x_i) - f_L \circ \ldots \circ f_1(\phi(x_i)) \right\|_{\mathcal{F}_\mathcal{X}}^2 + \sum_{l=1}^{L} \lambda_l \|f_l\|_{\mathcal{H}_l}^2$$

For $l \leq L$, $\mathcal{X}_l$ Hilbert, $\mathcal{X}_0 = \mathcal{X}_L = \mathcal{F}_\mathcal{X}$, $\mathcal{K}_l \colon \mathcal{X}_{l-1} \times \mathcal{X}_{l-1} \to \mathcal{L}(\mathcal{X}_l)$.

For all $L_0 \leq L$, there is $(\hat{\alpha}_{1,1}, \ldots, \hat{\alpha}_{1,n}, \ldots, \hat{\alpha}_{L_0,1}, \ldots, \hat{\alpha}_{L_0,n}) \in \mathcal{X}_1^n \times \ldots \times \mathcal{X}_{L_0}^n$, such that for all $l \leq L$ it holds:

$$\hat{f}_l(\cdot) = \sum_{i=1}^{n} \mathcal{K}_l\left(\ \cdot \ , {x_i}^{(l-1)}\right) \hat{\alpha}_{l,i},$$

with the notation for all $i \leq n$:

$$x_i^{(l)} = f_l \circ \ldots \circ f_1(x_i) \quad \text{and} \quad x_i^{(0)} = x_i.$$

Optimization algorithm

How to train the Kernel Autoencoder?

$$\min_{f_l \in \text{vv-RKHS}} \quad \frac{1}{n}\sum_{i=1}^{n} \left\| \phi(x_i) - f_L \circ \ldots \circ f_1(\phi(x_i)) \right\|_{\mathcal{F}_{\mathcal{X}}}^2 + \sum_{l=1}^{L} \lambda_l \|f_l\|_{\mathcal{H}_l}^2$$

- Last layer's infinite dimensional coefficients makes it impossible to perform Gradient Descent directly

- Yet, gradient can propagate through last layer ($[N_L]_{ij} = \langle \alpha_{L,i}, \alpha_{L,j} \rangle$):
$$\sum_{i,i'=1}^{n} [N_l]_{ii'} \left(\nabla^{(1)} k_l \left({x_i}^{(l-1)}, {x_{i'}}^{(l-1)} \right) \right)^\top \mathbf{Jac}_{{x_i}^{(l-1)}}(\alpha_{l_0,i_0})$$

- If inner layers fixed and $\mathcal{K}_L = k_L \mathbf{I}_{\mathcal{X}_0}$, closed-form solution for N_L

Alternate descent: gradient steps and OVKRR resolution

Application to molecule activity prediction

KAE representations are useful for posterior supervised tasks

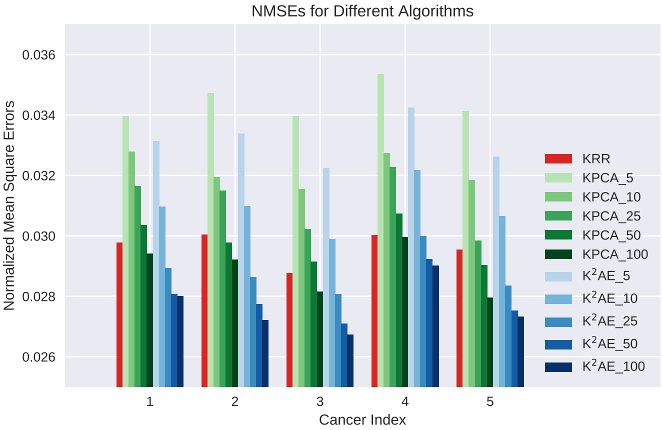


Fig. 4: Performance of the different strategies on 5 cancers (NCI dataset)

Part II: Robust losses for RKHSs with infinite dimensional outputs

$$\min_{h \text{ measurable}} \mathbb{E}_P\Big[\ell(h(X), Y)\Big] \quad \rightarrow \quad \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\Big(h(x_i), y_i\Big)$$

Infinite dimensional outputs in machine learning

Kernel Autoencoder [Laforgue et al., 2019a].

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

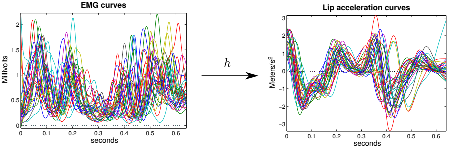
Structured prediction by ridge-IOKR [Brouard et al., 2016].

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

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

$$g(x) = \operatorname*{argmin}_{z \in \mathcal{Z}} \ \left\| \phi(z) - \hat{h}(x) \right\|_{\mathcal{F}_\mathcal{Z}}$$

Function to function regression [Kadri et al., 2016].



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

$$\min_{h \in \mathcal{H}_\mathcal{K}} \ \frac{1}{2n} \sum_{i=1}^{n} \left\| y_i - h(x_i) \right\|_{L^2}^2 + \frac{\Lambda}{2} \|h\|^2$$

Purpose of this part

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?

Learning in vector-valued RKHSs (reminder)

For $\{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ with $\mathcal{Y}$ a Hilbert space, we want to solve:

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

Representer Theorem [Micchelli and Pontil, 2005]:

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

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

Applying duality

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

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

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

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

Applying duality

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

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

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

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

The double representer theorem [Laforgue et al., 2020]

Assume that OVK $\mathcal{K}$ and loss ℓ satisfy the appropriate assumptions (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 $\hat{\Omega} = [\hat{\omega}_{ij}] \in \mathbb{R}^{n \times n}$ 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 and K_{kl}^Y if $\mathcal{K}$ is decomposable.

How to apply the duality approach?

Infimal convolution and Fenchel-Legendre transforms

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

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

Relation to FL transform:

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

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

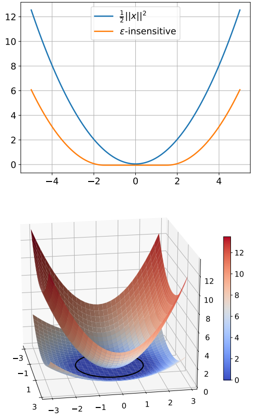
$$\ell_\epsilon(y) = (\ell \,\Box\, \chi_{\mathcal{B}_\epsilon})(y) = \begin{cases} \ell(0) & \text{if } \|y\|_{\mathcal{Y}} \le \epsilon \\ \inf\limits_{\|d\|_{\mathcal{Y}} \le 1} \ell(y - \epsilon d) & \text{otherwise} \end{cases},$$

and has FL transform:

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

Interesting loss functions: sparsity and robustness

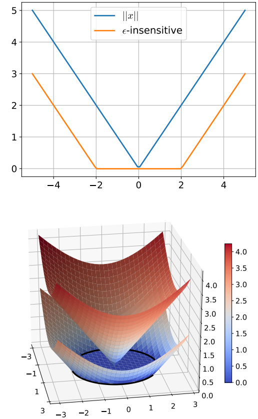
ϵ-Ridge



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

(Sparsity)

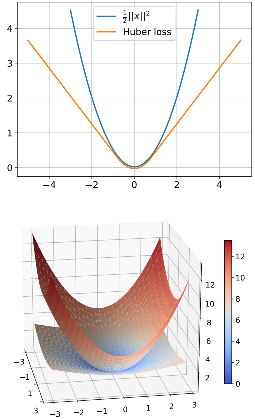
ϵ-SVR



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

(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:

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

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

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

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 $A^\top B = V$.

Application to 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

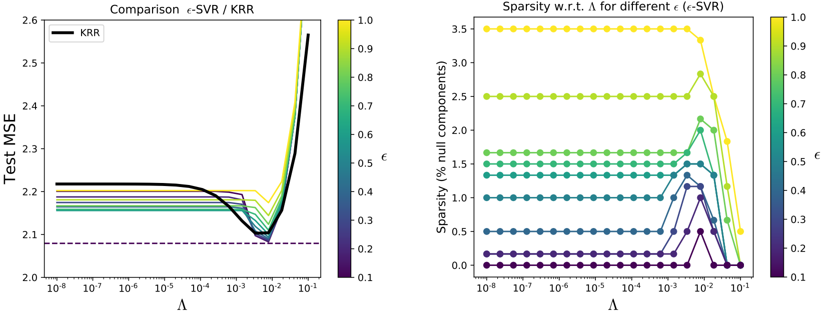


Fig. 5: MSEs and sparsity w.r.t. Λ for several ϵ

Part III:
Reliable learning through Median-of-Means approaches

$$\min_{h \text{ measurable}} \mathbb{E}_P\left[\ell(h(X), Y)\right] \quad \rightarrow \quad \min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(h(x_i), y_i\Big)$$

Preliminaries

Sample $\mathcal{S}_n = \{Z_1, \dots, Z_n\} \sim Z$ i.i.d. such that $\mathbb{E}[Z] = \theta$

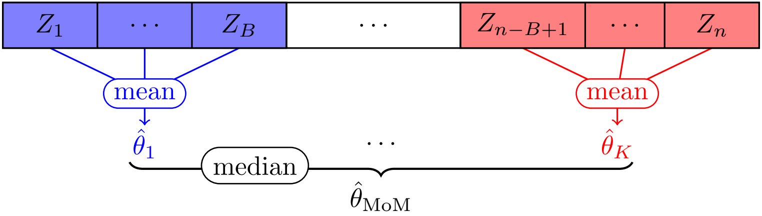
- $\hat{\theta}_{\mathsf{avg}} = \frac{1}{n} \sum_{i=1}^{n} Z_i$
- $\hat{\theta}_{\mathsf{med}} = Z_{\sigma(\frac{n+1}{2})}$, with $Z_{\sigma(1)} \leq \ldots \leq Z_{\sigma(n)}$
- Deviation Probabilities [Catoni, 2012]: $\mathbb{P}\left\{|\hat{\theta} - \theta| > t\right\}$.
- If Z is **bounded** (see Hoeffding's Inequality) or sub-Gaussian:

$$\mathbb{P}\left\{\left|\hat{\theta}_{\mathsf{avg}} - \theta\right| > \sigma\sqrt{\frac{2\ln(2/\delta)}{n}}\right\} \leq \delta.$$

Do estimators exist with same guarantees under weaker assumptions?

How to use them to perform (robust) learning?

The Median-of-Means

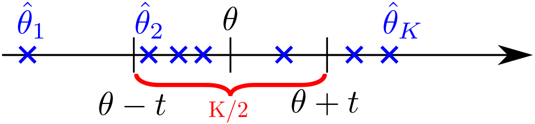


$Z_1, \ldots, Z_n$ i.i.d. realizations of r.v. Z s.t. $\mathbb{E}[Z] = \theta$, $\mathrm{Var}(Z) = \sigma^2$.

$\forall \delta \in [e^{1-\frac{2n}{9}}, 1[$, for $K = \left\lceil \frac{9}{2} \ln(1/\delta) \right\rceil$ it holds [Devroye et al., 2016]:

$$\mathbb{P}\left\{ \left| \hat{\theta}_{\mathrm{MoM}} - \theta \right| > 3\sqrt{6}\sigma \sqrt{\frac{1 + \ln(1/\delta)}{n}} \right\} \leq \delta.$$

Proof



$$\hat{\theta}_k = \frac{1}{B}\sum_{i\in B_k} Z_i, \qquad \hat{I}_{k,t} = \mathbb{I}\left\{|\hat{\theta}_k - \theta| > t\right\}, \qquad \hat{p}_t = \mathbb{E}[\hat{I}_{1,t}] = \mathbb{P}\left\{|\hat{\theta}_1 - \theta| > t\right\}$$

$$\begin{aligned}\mathbb{P}\left\{\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right| > t\right\} &\leq \mathbb{P}\left\{\sum_{k=1}^{K}\hat{I}_{k,t} \geq \frac{K}{2}\right\} \leq \mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^{K}(\hat{I}_{k,t} - p_t) \geq \frac{1}{2} - \frac{\sigma^2}{Bt^2}\right\}, \\ &\leq \exp\left(-2K\left(\frac{1}{2} - \frac{\sigma^2}{Bt^2}\right)^2\right), \\ &\leq \delta \text{ for } K = \frac{9\ln(1/\delta)}{2} \text{ and } \frac{\sigma^2}{Bt^2} = \frac{1}{6} \Leftrightarrow t = 3\sqrt{3}\sigma\sqrt{\frac{\ln(1/\delta)}{n}}.\end{aligned}$$

U-statistics & pairwise learning

Estimator of $\mathbb{E}[h(Z, Z')]$ with minimal variance, defined from an i.i.d. sample $Z_1, \ldots, Z_n$ as:

$$U_n(h) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h(Z_i, Z_j).$$

Ex: the empirical variance when $h(z, z') = \frac{(z-z')^2}{2}$.

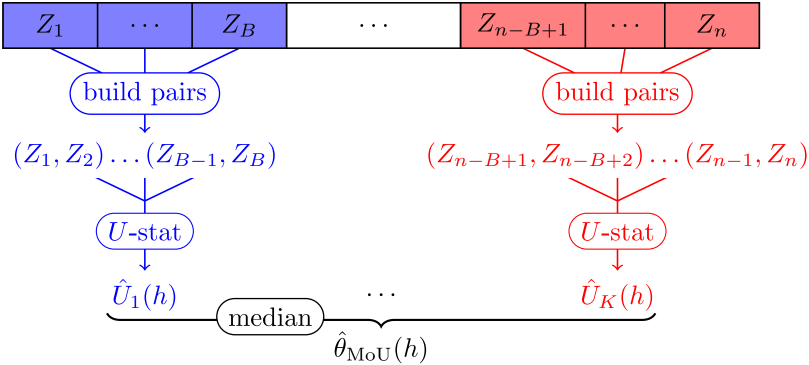
Encountered *e.g.* in **pairwise ranking** and **metric learning**:

$$\widehat{\mathcal{R}}_n(r) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{I}\left\{r(X_i, X_j) \cdot (Y_i - Y_j) \leq 0\right\}.$$

$$\widehat{\mathcal{R}}_n(d) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{I}\left\{Y_{ij} \cdot (d(X_i, X_j) - \epsilon) > 0\right\}.$$

How to extend MoM to U-statistics?

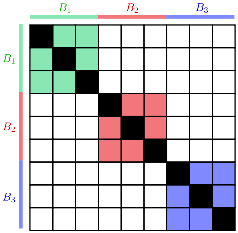
The Median-of-U-statistics



w.p. $1-\delta$, $\quad \left|\hat{\theta}_{\mathrm{MoU}}(h) - \theta(h)\right| \leq C_1(h)\sqrt{\frac{1+\ln(1/\delta)}{n}} + C_2(h)\,\frac{1+\ln(1/\delta)}{n}$

Why randomization?

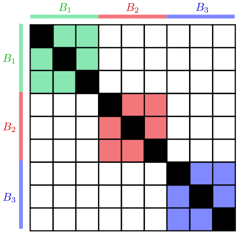
Partition



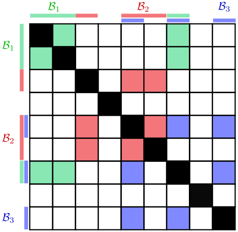
Build all possible blocks
[Joly and Lugosi, 2016]

Why randomization?

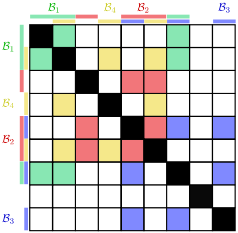
Partition



Random Blocks

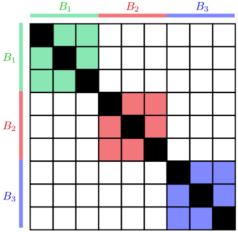


Build all possible blocks
[Joly and Lugosi, 2016]

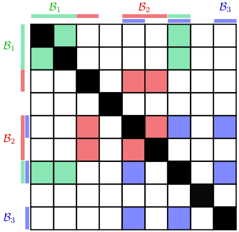


Why randomization?

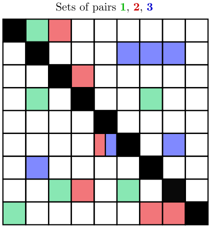
Partition



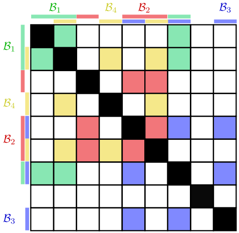
Random Blocks

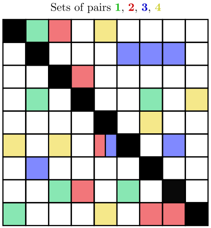


Random Pairs



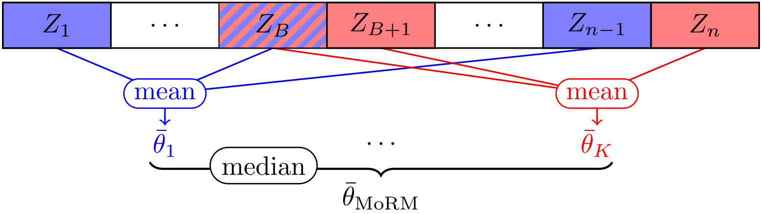
Build all possible blocks
[Joly and Lugosi, 2016]





Randomization allows for a better exploration

The Median-of-Randomized-Means [Laforgue et al., 2019b]



With blocks formed by SWoR, $\forall\ \tau \in]0, 1/2[,\ \forall\ \delta \in [2e^{-\frac{8\tau^2 n}{9}}, 1[$, set

$K := \left\lceil \frac{\ln(2/\delta)}{2(1/2-\tau)^2} \right\rceil$, and $B := \left\lfloor \frac{8\tau^2 n}{9\ln(2/\delta)} \right\rfloor$, it holds:

$$\mathbb{P}\left\{ \left| \bar{\theta}_{\mathrm{MoRM}} - \theta \right| > \frac{3\sqrt{3}\ \sigma}{2\ \tau^{3/2}} \sqrt{\frac{\ln(2/\delta)}{n}} \right\} \leq \delta.$$

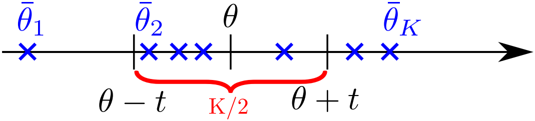
Proof

Random block $\mathcal{B}_k$ characterized by random vector $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, \ldots, \epsilon_{k,n}) \in \{0,1\}^n$ i.i.d. uniformly over $\Lambda_{n,B} = \left\{\boldsymbol{\epsilon} \in \{0,1\}^n : \mathbf{1}^\top \boldsymbol{\epsilon} = B\right\}$, of cardinality $\binom{n}{B}$.

$$\bar{\theta}_k = \frac{1}{B} \sum_{i \in \mathcal{B}_k} Z_i, \qquad \bar{I}_{\boldsymbol{\epsilon}_k,t} = \mathbb{I}\{|\bar{\theta}_k - \theta| > t\}, \qquad \bar{p}_t = \mathbb{E}[\bar{I}_{\boldsymbol{\epsilon}_k,t}] = \mathbb{P}\left\{|\bar{\theta}_1 - \theta| > t\right\}$$

$$\bar{U}_{n,t} = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{1}{K}\sum_{k=1}^{K} \bar{I}_{\boldsymbol{\epsilon}_k,t} \middle| \mathcal{S}_n\right] = \frac{1}{\binom{n}{B}} \sum_{\boldsymbol{\epsilon} \in \Lambda(n,B)} \bar{I}_{\boldsymbol{\epsilon},t} = \frac{1}{\binom{n}{B}} \sum_{I} \mathbb{I}\left\{\left|\frac{1}{B}\sum_{j=1}^{B} X_{I_j} - \theta\right| > t\right\}$$

$$\mathbb{P}\left\{\left|\bar{\theta}_{\text{MoRM}} - \theta\right| > t\right\} \leq \mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^{K} \bar{I}_{\boldsymbol{\epsilon}_k,t} \qquad\qquad\qquad \geq \frac{1}{2} \qquad\qquad\qquad \right\},$$



Proof

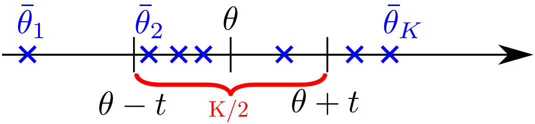
Random block $\mathcal{B}_k$ characterized by random vector $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, \dots, \epsilon_{k,n}) \in \{0,1\}^n$ i.i.d. uniformly over $\Lambda_{n,B} = \left\{\boldsymbol{\epsilon} \in \{0,1\}^n : \mathbf{1}^\top \boldsymbol{\epsilon} = B\right\}$, of cardinality $\binom{n}{B}$.

$$\bar\theta_k = \frac{1}{B}\sum_{i\in\mathcal{B}_k} Z_i, \qquad \bar I_{\boldsymbol\epsilon_k,t} = \mathbb{I}\{|\bar\theta_k-\theta|>t\}, \qquad \bar p_t = \mathbb{E}[\bar I_{\boldsymbol\epsilon_k,t}] = \mathbb{P}\left\{|\bar\theta_1-\theta|>t\right\}$$

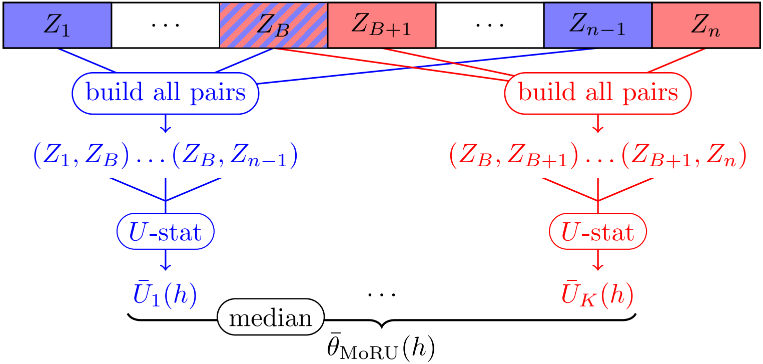
$$\bar U_{n,t} = \mathbb{E}_{\boldsymbol\epsilon}\left[\frac1K\sum_{k=1}^K \bar I_{\boldsymbol\epsilon_k,t}\middle|\mathcal{S}_n\right] = \frac{1}{\binom{n}{B}}\sum_{\boldsymbol\epsilon\in\Lambda(n,B)} \bar I_{\boldsymbol\epsilon,t} = \frac{1}{\binom{n}{B}}\sum_{I} \mathbb{I}\left\{\left|\frac1B\sum_{j=1}^B X_{I_j}-\theta\right|>t\right\}$$

$$\mathbb{P}\left\{\left|\bar\theta_{\text{MoRM}}-\theta\right|>t\right\} \le \mathbb{P}\left\{\frac1K\sum_{k=1}^K \bar I_{\boldsymbol\epsilon_k,t} - \bar U_{n,t} + \bar U_{n,t} - \bar p_t \ge \frac12 - \bar p_t + \tau - \tau\right\},$$

$$\le \exp\left(-2K\left(\frac12-\tau\right)^2\right) + \exp\left(-2\frac{n}{B}\left(\tau-\frac{\sigma^2}{Bt^2}\right)^2\right).$$



The Median-of-Randomized-U-statistics [Laforgue et al., 2019b]



$$\text{w.p.a.l. } 1-\delta, \quad \left|\bar{\theta}_{\mathrm{MoRU}} - \theta(h)\right| \leq C_1(h,\tau)\sqrt{\frac{\ln(2/\delta)}{n}} + C_2(h,\tau)\ \frac{\ln(2/\delta)}{n}$$

The tournament procedure [Lugosi and Mendelson, 2016]

We want $g^* \in \underset{g \in \mathcal{G}}{\operatorname{argmin}} \mathcal{R}(g) = \mathbb{E}[(g(X) - Y)^2]$. For any pair $(g, g') \in \mathcal{G}^2$:

1) Compute the MoM estimate of $\|g - g'\|_{L_1}$

$$\Phi_{\mathcal{S}}(g, g') = \text{median}\left(\hat{\mathbb{E}}_1|g - g'|, \ldots, \hat{\mathbb{E}}_K|g - g'|\right).$$

2) If it is *large enough*, compute the *match*

$$\Psi_{\mathcal{S}'}(g, g') = \text{median}\Big(\hat{\mathbb{E}}_1[(g(X) - Y)^2 - (g'(X) - Y)^2], \ldots, \\ \hat{\mathbb{E}}_K[(g(X) - Y)^2 - (g'(X) - Y)^2]\Big).$$

$\hat{g}$ winning all its matches verifies w.p.a.l. $1 - \exp(c_0 n \min\{1, r^2\})$

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g^*) \leq cr.$$

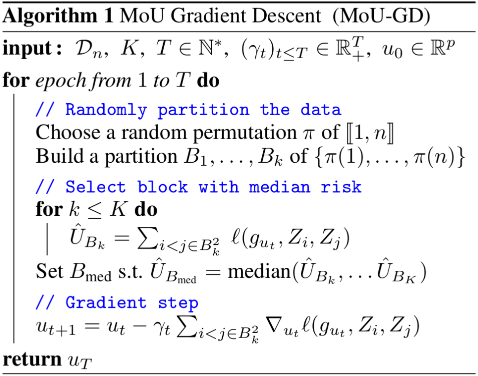
Can be extended to pairwise learning thanks to MoU

The MoM Gradient Descent [Lecué et al., 2018]

If $\mathcal{G}$ is parametric, want to compute the minimizer of:

$$\mathrm{MoM}[\ell(g_u, Z)] = \mathrm{median}\left(\hat{\mathbb{E}}_1[\ell(g_u, Z)], \ldots, \hat{\mathbb{E}}_K[\ell(g_u, Z)]\right)$$

Idea: find the block with median risk, and use it as mini-batch

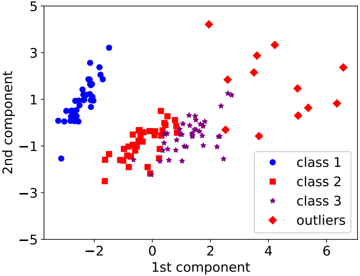


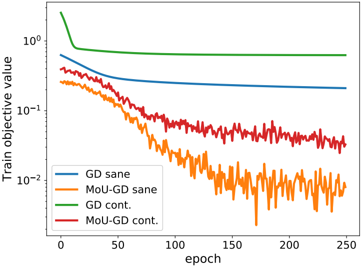
```
Algorithm 1 MoU Gradient Descent (MoU-GD)
input: $\mathcal{D}_n,\ K,\ T \in \mathbb{N}^*,\ (\gamma_t)_{t \leq T} \in \mathbb{R}_+^T,\ u_0 \in \mathbb{R}^p$
for epoch from 1 to $T$ do
    // Randomly partition the data
    Choose a random permutation $\pi$ of $[\![1, n]\!]$
    Build a partition $B_1, \ldots, B_k$ of $\{\pi(1), \ldots, \pi(n)\}$
    // Select block with median risk
    for $k \leq K$ do
        $\hat{U}_{B_k} = \sum_{i<j \in B_k^2} \ell(g_{u_t}, Z_i, Z_j)$
    Set $B_{\text{med}}$ s.t. $\hat{U}_{B_{\text{med}}} = \text{median}(\hat{U}_{B_k}, \ldots \hat{U}_{B_K})$
    // Gradient step
    $u_{t+1} = u_t - \gamma_t \sum_{i<j \in B_k^2} \nabla_{u_t} \ell(g_{u_t}, Z_i, Z_j)$
return $u_T$
```

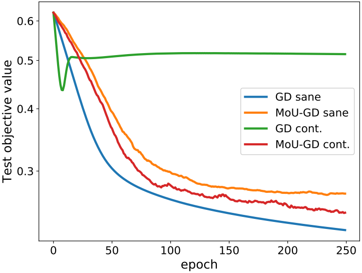
MoU Gradient Descent for metric learning

We want to minimize for $M \in S_q^+(\mathbb{R})$:

$$\frac{2}{n(n-1)} \sum_{i<j} \max\Big(0, 1+y_{ij}(d_M^2(x_i, x_j)-2)\Big)$$







Conclusion

Conclusion

$$\min_{h \text{ measurable}} \mathbb{E}_P\Big[\ell(h(X), Y)\Big] \quad \rightarrow \quad \min_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(h(x_i), y_i\Big)$$

- New hypothesis set for RL inspired from deep and kernel
- Link with Kernel PCA, optimization based on composite RT
- Allows to autoencode any type of data, empirical success on molecules

Conclusion

$$\min_{h \text{ measurable}} \mathbb{E}_P\Big[\ell(h(X), Y)\Big] \quad \rightarrow \quad \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\Big(h(x_i), y_i\Big)$$

- New hypothesis set for RL inspired from deep and kernel
- Link with Kernel PCA, optimization based on composite RT
- Allows to autoencode any type of data, empirical success on molecules

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

Conclusion

$$\min_{h \text{ measurable}} \mathbb{E}_P\Big[\ell(h(X), Y)\Big] \quad \rightarrow \quad \min_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(h(x_i), y_i\Big)$$

- New hypothesis set for RL inspired from deep and kernel
- Link with Kernel PCA, optimization based on composite RT
- Allows to autoencode any type of data, empirical success on molecules

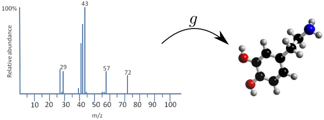
- Double RT: coefficients linear combinations of the outputs
- Allows to cope with many losses (ϵ, Huber) and kernels
- Empirical improvements on surrogate tasks

- Extension of MoM to randomized blocks and/or U-statistics
- Extension of MoM tournaments and MoM-GD to pairwise learning
- Remarkable empirical resistance to the presence of outliers

Perspectives

From K^{2}AE to deep IOKR.

- fully supervised scheme
- benefits of a hybrid architecture?
- learning the output embeddings?



Mass spectrum x Molecule z

Y's invariance: the good characterization for $\mathcal{K}$?

- what if we relax the hypothesis?
- case of integral losses: $\ell(h(x), y) = \int \ell_\theta[h(x)(\theta), y(\theta)]d\theta$

Among the numerous MoM possibilities.

- a partial representer theorem?
- concentration in presence of outliers?

Remerciements

- **PhD supervisors:** Florence d'Alché-Buc, Stephan Clémençon
- **Co-authors:** Alex Lambert, Luc Brogat-Motte, Patrice Bertail
- **Thank you:** Olivier Fercoq

- *Autoencoding any data through kernel autoencoders*
  with S. Clémençon and F. d'Alché-Buc, AISTATS 2019
- *On medians-of-randomized-(pairwise)-means*
  with S. Clémençon and P. Bertail, ICML 2019
- *Duality in RKHSs with infinite dimensional outputs: application to robust losses*
  with A. Lambert, L. Brogat-Motte and F. d'Alché-Buc, ICML 2020
- *On statistical learning from biased training samples*
  with S. Clémençon, Submitted

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.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007).
Greedy layer-wise training of deep networks.
In *Advances in neural information processing systems*, pages 153–160.

References II

- Bourlard, H. and Kamp, Y. (1988).
Auto-association by multilayer perceptrons and singular value decomposition.
Biological cybernetics, 59(4):291–294.

- 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 kernels.
Journal of Machine Learning Research, 17:176:1–176:48.

References III

Carmeli, C., De Vito, E., and Toigo, A. (2006).
Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem.
Analysis and Applications, 4(04):377–408.

Catoni, O. (2012).
Challenging the empirical mean and empirical variance: a deviation study.
In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré.

Devroye, L., Lerasle, M., Lugosi, G., Oliveira, R. I., et al. (2016).
Sub-gaussian mean estimators.
The Annals of Statistics, 44(6):2695–2725.

References IV

- Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009).
Visualizing higher-layer features of a deep network.
University of Montreal, 1341(3):1.

- Gill, R., Vardi, Y., and Wellner, J. (1988).
Large sample theory of empirical distributions in biased sampling models.
The Annals of Statistics, 16(3):1069–1112.

- Hinton, G. E. and Salakhutdinov, R. R. (2006).
Reducing the dimensionality of data with neural networks.
science, 313(5786):504–507.

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

References V

Joly, E. and Lugosi, G. (2016).
Robust estimation of u-statistics.
Stochastic Processes and their Applications, 126(12):3760–3773.

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.

References VI

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

- Laforgue, P., Clemencon, S., and Bertail, P. (2019b).
  On medians of (Randomized) pairwise means.
  In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1272–1281, Long Beach, California, USA. PMLR.

- Laforgue, P., Lambert, A., Motte, L., and d'Alché Buc, F. (2020).
  Duality in rkhss with infinite dimensional outputs: Application to robust losses.
  arXiv preprint arXiv:1910.04621.

References VII

Lecué, G., Lerasle, M., and Mathieu, T. (2018).
Robust classification via mom minimization.
arXiv preprint arXiv:1808.03106.

Lugosi, G. and Mendelson, S. (2016).
Risk minimization by median-of-means tournaments.
arXiv preprint arXiv:1608.00757.

Mairal, J., Koniusz, P., Harchaoui, Z., and Schmid, C. (2014).
Convolutional kernel networks.
In *Advances in neural information processing systems*, pages 2627–2635.

References VIII

Maurer, A. (2014).
A chain rule for the expected suprema of gaussian processes.
In *Algorithmic Learning Theory: 25th International Conference, ALT 2014, Bled, Slovenia, October 8-10, 2014, Proceedings*, volume 8776, page 245. Springer.

Maurer, A. (2016).
A vector-contraction inequality for rademacher complexities.
In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer.

Maurer, A. and Pontil, M. (2016).
Bounds for vector-valued function estimation.
arXiv preprint arXiv:1606.01487.

References IX

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.

Schölkopf, B., Tsuda, K., and Vert, J.-P. (2004).
***Support vector machine applications in computational biology*.**
MIT press.

References X

- Vardi, Y. (1985).
  Empirical distributions in selection bias models.
  Ann. Statist., 13:178–203.

- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010).
  Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.
  J. Mach. Learn. Res., 11:3371–3408.

Appendices *Kernel Autoencoder*

Functional spaces similarity

- Neural mapping f_{NN} parametrized by a matrix $A \in \mathbb{R}^{p \times d}$ with rows $(a_j)_{j \leq p}$, and and activation function σ
- Kernel mapping f_{OVK} from decomposable OVK $\mathcal{K} = k\mathbf{I}_p$, associated to the (scalar) feature map ϕ_k

$$f_{\mathrm{NN}}(x) = \begin{pmatrix} \sigma\left(\langle a_1, x\rangle\right) \\ \vdots \\ \sigma\left(\langle a_p, x\rangle\right) \end{pmatrix} \qquad f_{\mathrm{OVK}}(x) = \begin{pmatrix} f^1(x) = \left\langle f^1, \phi_k(x)\right\rangle \\ \vdots \\ f^p(x) = \langle f^p, \phi_k(x)\rangle \end{pmatrix}$$

Only differ on the order in which linear/nonlinear mappings are used (and on their nature)

Disentangling concentric circles

More complex layers enhance the learned representations

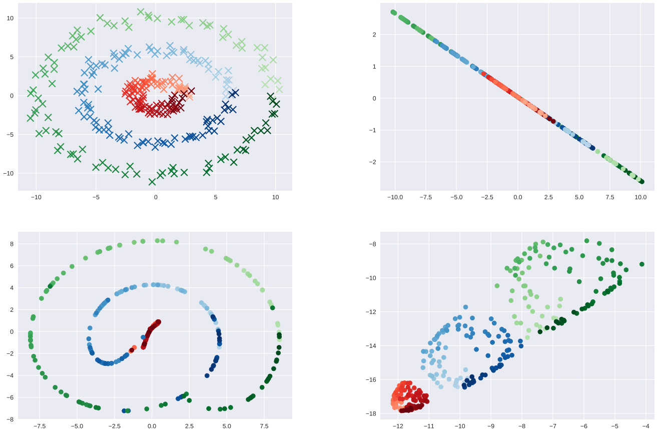


Fig. 6: KAE performance on concentric circles

KAE generalization bound

2-layer KAE on data bounded in norm by M, with:

- internal layer of size p
- encoder $f \in \mathcal{H}_1$ such that $\|f\| \leq s$
- decoder $g \in \mathcal{H}_2$ such that $\|g\| \leq t$, with Lipschitz constant L

Then it holds:

$$\epsilon(\hat{g}_n \circ \hat{f}_n) - \epsilon^* \leq C_0 L M s t \sqrt{\frac{Kp}{n}} + 24 M^2 \sqrt{\frac{\log(2)/\delta}{2n}}.$$

with $\epsilon(g \circ f) = \mathbb{E}_X \|X - g \circ f(X)\|^2_{\mathcal{X}_0}$

Proof

Based on vector-valued Rademacher average:

$$\widehat{\mathscr{R}}_n(\mathcal{C}(S)) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{\sigma}_i, h(x_i) \rangle_H \right].$$

With $\mathcal{H}_{s,t} \subset \mathcal{F}(\mathcal{X}_0, \mathcal{X}_0) = \mathcal{H}_{1,s} \circ \mathcal{H}_{2,t}$, ℓ the squared norm on $\mathcal{X}_0$, it holds:

$$\widehat{\mathscr{R}}_n\Big((\ell \circ (\mathrm{id} - \mathcal{H}_{s,t}))(S) \Big) \leq 2\sqrt{2}M \, \widehat{\mathscr{R}}_n\Big((\mathrm{id} - \mathcal{H}_{s,t})(S) \Big),$$

$$\leq 2\sqrt{2}M \, \widehat{\mathscr{R}}_n\Big(\mathcal{H}_{s,t}(S) \Big) \leq 2\sqrt{\pi}M \, \widehat{\mathscr{G}}_n\Big(\mathcal{H}_{s,t}(S) \Big).$$

$$\widehat{\mathscr{G}}_n\Big(\mathcal{H}_{s,t}(S) \Big) \leq C_1 \; L\Big(\mathcal{H}_{2,t}, \mathcal{H}_{1,s}(S) \Big) \, \widehat{\mathscr{G}}_n\Big(\mathcal{H}_{1,s}(S) \Big)$$

$$+ \frac{C_2}{n} \; R\Big(\mathcal{H}_{2,t}, \mathcal{H}_{1,s}(S) \Big) \; D\Big(\mathcal{H}_{1,s}(S) \Big) + \frac{1}{n} \; G\Big(\mathcal{H}_{2,t}(0) \Big).$$

using [Maurer, 2016, Maurer, 2014] in the spirit of [Maurer and Pontil, 2016]

Appendices *Duality in vv-RKHSs*

On the invariance assumption

With $\mathbf{Y} = \mathrm{Span}\{y_j,\ j \leq n\}$, 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

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

- $\ell_i(y) = f(\langle y, z_i \rangle)$, $z_i \in Y$ 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])

Proof of the Double Representer Theorem

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:

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

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

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

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

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

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

Application to robust function-to-function regression

- Predict lip acceleration from EMG signals [Kadri et al., 2016]
- Dataset augmented with outliers, model learned with Huber loss
- Improvement for every output size m

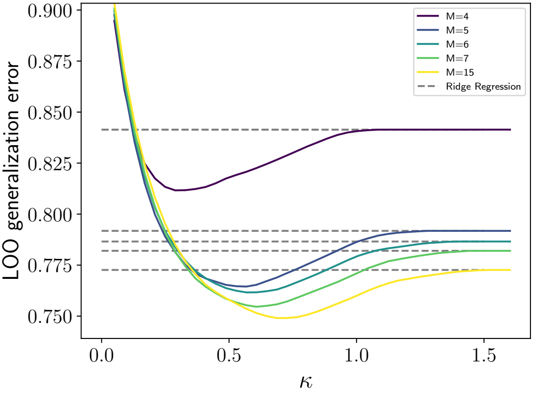


Fig. 7: LOO generalization error w.r.t. κ

Application to kernel autoencoding

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

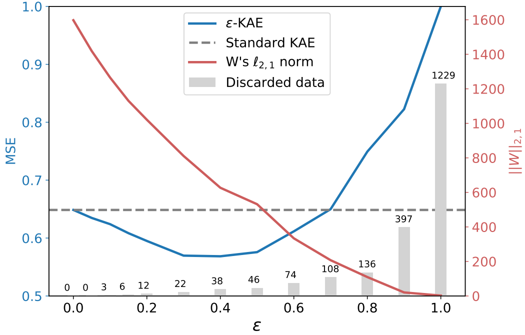


Fig. 8: Performances of ϵ-insensitive Kernel Autoencoder

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

Appendices *Learning with Sample Bias*

Empirical Risk Minimization (ERM)

General goal of supervised machine learning:
From a r.v. $Z = (X, Y)$, and a loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, find:

$$h^* = \underset{h \text{ measurable}}{\operatorname{argmin}} \; R(h) = \mathbb{E}_P \left[\ell(h(X), Y) \right].$$

Empirical Risk Minimization (ERM):

- P is unknown (and the set of measurable functions too large)
- sample $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d}{\sim} P$, hypothesis set $\mathcal{H}$

$$\hat{h}_n = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i) = \mathbb{E}_{\hat{P}_n} \left[\ell(h(X), Y) \right],$$

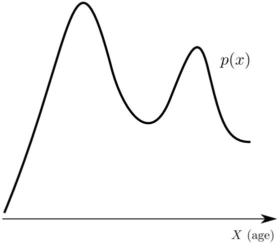
with $\hat{P}_n = \frac{1}{n} \sum_i \delta_{Z_i}$, and $Z_i = (X_i, Y_i)$. It holds $\hat{P}_n \underset{n \to +\infty}{\to} P$.

Importance Sampling (IS)

What if the data is not drawn from P?

Sample $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d}{\sim} Q$ such that $\frac{dQ}{dP}(z) = \frac{q(z)}{p(z)}$.

Now $\frac{1}{n}\sum_i \delta_{Z_i} = \hat{Q}_n \underset{n \to +\infty}{\to} Q$.



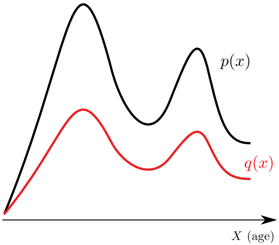
Importance Sampling (IS)

What if the data is not drawn from P?

Sample $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d}{\sim} Q$ such that $\frac{dQ}{dP}(z) = \frac{q(z)}{p(z)}$.

Now $\frac{1}{n}\sum_i \delta_{Z_i} = \hat{Q}_n \underset{n \to +\infty}{\to} Q$.

$q(x)/p(x) = 1/2.$



$$\min_{h \in \mathcal{H}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i) \cdot \frac{p(Z_i)}{q(Z_i)}$$

$$\min_{h \in \mathcal{H}} \; \mathbb{E}_{\hat{Q}_n} \left[\ell(h(X), Y) \cdot \frac{p(Z)}{q(Z)} \right]$$

$$\downarrow$$

$$\mathbb{E}_{Q} \left[\ell(h(X), Y) \cdot \frac{p(Z)}{q(Z)} \right] = \mathbb{E}_P \left[\ell(h(X), Y) \right]$$

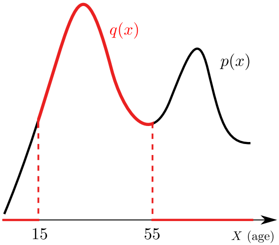
Importance Sampling (IS)

What if the data is not drawn from P?

Sample $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d}{\sim} Q$ such that $\frac{dQ}{dP}(z) = \frac{q(z)}{p(z)}$.

Now $\frac{1}{n}\sum_i \delta_{Z_i} = \hat{Q}_n \underset{n \to +\infty}{\to} Q$.

$\color{red}{q(x)}/p(x) = \mathbb{I}\{15 \leq x \leq 55\}$.



$$\min_{h \in \mathcal{H}} \ \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i) \cdot \frac{p(Z_i)}{\color{red}{q}(Z_i)}$$

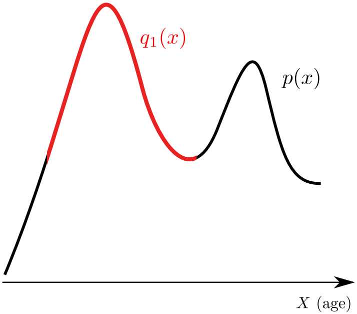
$$\min_{h \in \mathcal{H}} \ \mathbb{E}_{\hat{Q}_n}\left[\ell(h(X), Y) \cdot \frac{p(Z)}{\color{red}{q}(Z)}\right]$$

not possible!

$$\downarrow$$

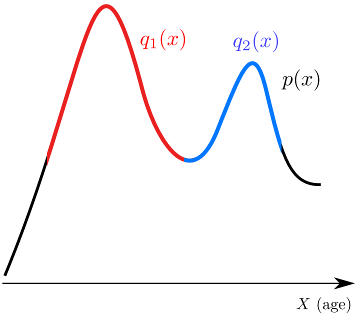
$$\mathbb{E}_Q\left[\ell(h(X), Y) \cdot \frac{p(Z)}{\color{red}{q}(Z)}\right] = \mathbb{E}_P\left[\ell(h(X), Y)\right]$$

Adding samples



$$q_1(x)/p(x) = \mathbb{I}\{15 \leq x \leq 55\}$$

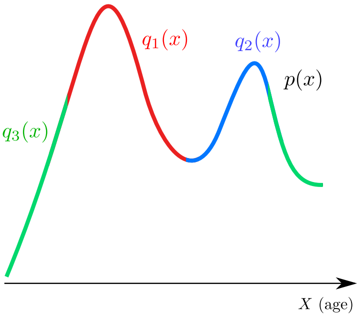
Adding samples



$q_1(x)/p(x) = \mathbb{I}\{15 \leq x \leq 55\}$

$q_2(x)/p(x) = \mathbb{I}\{50 \leq x \leq 70\}$

Adding samples



$q_1(x)/p(x) = \mathbb{I}\{15 \leq x \leq 55\}$

$q_2(x)/p(x) = \mathbb{I}\{50 \leq x \leq 70\}$

$q_3(x)/p(x) = \mathbb{I}\{x \leq 20\} + \mathbb{I}\{x \geq 60\}$

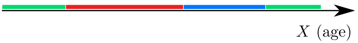
We need: $\bigcup_{k=1}^{K} \textsc{Supp}(q_k) = \textsc{Supp}(p)$.

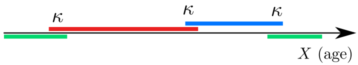
Sample-wise IS doe not work because of samples proportions.

Setting and assumptions

- K independent i.i.d. samples $\mathcal{D}_k = \{Z_{k,1}, \ \ldots, \ Z_{k,n_k}\}$
- $n = \sum_k n_k$, $\hat{\lambda}_k = n_k/n$ for $k \leq K$
- sample k drawn according to Q_k such that $\frac{dQ_k}{dP}(z) = \frac{\omega_k(z)}{\Omega_k}$
- The $\Omega_k = \mathbb{E}_P[\omega_k(Z)] = \int_{\mathcal{Z}} \omega_k(z) P(dz)$ are unknown.

- $\exists C, \underline{\lambda}, \lambda_1, \ldots, \lambda_K > 0, \quad |\lambda_k - \hat{\lambda}_k| \leq \frac{C}{\sqrt{n}}$ and $\underline{\lambda} \leq \hat{\lambda}_k$.
- The graph G_κ is connected.
- $\exists \xi > 0, \ \forall k \leq K, \quad \Omega_k \geq \xi$.
- $\exists m, M > 0, \quad m \leq \inf_z \max_{k \leq K} \ \omega_k(z)$ and $\sup_z \max_{k \leq K} \ \omega_k(z) \leq M$.





Building an unbiased estimate of P (1/2)

Without considering the bias issue:

$$\hat{Q}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Z_i} = \sum_{k=1}^{K}\frac{n_k}{n}\frac{1}{n_k}\sum_{i\in\mathcal{D}_k}\delta_{Z_i} \;\rightarrow\; \sum_{k=1}^{K}\lambda_k Q_k \neq P.$$

But it holds:

$$dQ_k = \frac{\omega_k}{\Omega_k}dP, \qquad \sum_k \hat{\lambda}_k dQ_k = \sum_k \frac{\hat{\lambda}_k\omega_k}{\Omega_k}dP$$

$$\boxed{dP = \left(\sum_k \frac{\hat{\lambda}_k\omega_k}{\Omega_k}\right)^{-1}\sum_k \hat{\lambda}_k dQ_k} \qquad (1)$$

We only need to estimate the Ω_k's.

Building an unbiased estimate of P (2/2)

It holds:

$$\Omega_k = \int \omega_k dP = \int \left(\sum_k \frac{\lambda_k \omega_k}{\Omega_k} \right)^{-1} \sum_k \lambda_k \omega_k dQ_k.$$

$\hat{\boldsymbol{\Omega}}$ solution to the system:

$$\forall k \leq K, \qquad \hat{H}_k(\boldsymbol{\Omega}) - 1 = 0,$$

with $\hat{H}_k(\boldsymbol{\Omega}) = \int \left(\sum_k \frac{\hat{\lambda}_k \omega_k}{\Omega_k} \right)^{-1} \sum_k \hat{\lambda}_k \omega_k d\hat{Q}_k.$

The final estimate is obtained by plugging $\hat{\boldsymbol{\Omega}}$ in Equation (1).

Non-asymptotic guarantees

Debiasing procedure due to [Vardi, 1985, Gill et al., 1988], but only asymptotic results.

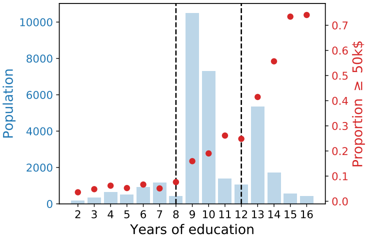
With $\hat{P}_n = \left(\sum_k \frac{\hat{\lambda}_k \omega_k}{\hat{\Omega}_k}\right)^{-1} \sum_k \hat{\lambda}_k d\hat{Q}_k$, there exists $(\pi_i)_{i \leq n}$ such that:

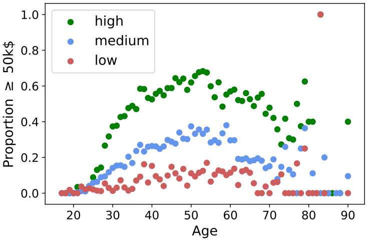
$$\mathbb{E}_{\hat{P}_n}\left[\ell(h(X), Y)\right] = \sum_{i=1}^{n} \pi_i \cdot \ell(h(X_i), Y_i), \tag{2}$$

and $\hat{h}_n$ minimizer of Equation (2) satisfies with probability $1 - \delta$:

$$R(\hat{h}_n) - R(h^*) \leq C_1 \sqrt{\frac{K^3}{n}} + C_2 \sqrt{\frac{K \log n}{n}} + C_3 \sqrt{\frac{K \log 1/\delta}{n}}.$$

Experiments on the *Adult* dataset





Dataset of size $6,000$: 98% from 13+ years of education, 2% unbiased. Scores:

	LogReg	RF
ERM	63.95 ± 1.37	42.73 ± 3.36
debiased ERM	**79.77** ± **1.72**	**43.58** ± **4.77**
unbiased sample	77.75 ± 2.27	22.16 ± 6.18