Infinite Task Learning with Vector-Valued RKHSs

R. Brault♣, A. Lambert†, Z. Szabo⋆, M. Sangnier♠, F. d'Alché-Buc†
♣ L2S, Centrale-Supelec. † LTCI, Télécom ParisTech. ⋆ CMAP, Ecole Polytechnique. ♠ LIP6, Sorbonne Université.

Parameterized Learning Tasks

Some learning tasks involve loss functions that depend on a hyperparameter. Examples: two parameterized tasks Quantile Regression (QR) and Cost-Sensitive Classification (CSC) with the following notations.

- $\mathcal{X}$ input data space ($\mathbb{R}^d$), Θ parameter space ($\mathbb{R}$), $\mathcal{Y}$ output space ($\mathbb{R}$).
- Hypothesis space $\mathcal{H} \subset \mathcal{F}(\mathcal{X}; \mathcal{Y})$
- Parameterized cost $\nu\colon \Theta \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

QR: Given $X, Y \in \mathcal{X} \times \mathcal{Y}$ random variables, estimate the θ-quantile of the conditional distribution $\mathbf{P}_{Y|X}$:

$$q(x)(\theta) = \inf\left\{t \in \mathcal{Y},\ \mathbf{P}_{Y|X=x}[Y \leqslant t] \geqslant \theta\right\} \quad \forall (x, \theta) \in \mathcal{X} \times (0,1). \tag{1}$$

Pinball loss: $$\nu(\theta, y, h(x)) = |\theta - \mathbb{1}_{\mathbb{R}_-}(y - h(x))||y - h(x)|. \tag{2}$$

CSC: Support Vector Machine with asymetric loss function

$$\nu(\theta, y, h(x)) = \left|\theta - \mathbb{1}_{\{-1\}}(y)\right||1 - yh(x)|_+. \tag{3}$$

Usual approach: Empirical risk minimization in some well chosen $\mathcal{H}$ for a given value of θ. For several values, turn to Multi-Task learning [1].

Learning an infinite number of tasks

We propose to jointly solve parameterized tasks for an infinite number of values of the hyperparameter θ using function-valued regression:

- Hypothesis space $\mathcal{H} \subset \mathcal{F}(\mathcal{X}; \mathcal{F}(\Theta; \mathcal{Y}))$ *i.e* $h(x) \in \mathcal{F}(\Theta; \mathcal{Y})$.
- Local loss $V(y, h(x)) := \int_\Theta \nu(\theta, y, h(x)(\theta)) d\mu(\theta)$.

Minimizing population risk:

$$\arg\min_{h \in \mathcal{H}} \ R(h) := \mathbf{E}_{X,Y}\left[V(Y, h(X))\right]. \tag{4}$$

Proposition. q *defined in (1) minimizes (4) for the pinball loss (2).*

$\Rightarrow$ Extension of Multi-Task Learning to an infinite number of tasks [2].

Sampled Empirical Risk

Approximate expectation over $\mathbf{P}_{X,Y}$ and $\int_\Theta$

- $(x_i, y_i)_{i=1}^n$ *i.i.d* $\sim \mathbf{P}_{X,Y}$
- $(\theta_j)_{j=1}^m \sim \mu$ (Quasi-Monte Carlo)

Sampled empirical risk: $$\widetilde{R}_S(h) := \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m \nu(\theta_j, y_i, h(x_i)(\theta_j)).$$

Regularized problem: $$\arg\min_{h \in \mathcal{H}} \widetilde{R}_S(h) + \lambda\Omega(h). \tag{5}$$

Vector-Valued RKHSs

Natural extension of RKHSs for modelizing outputs in any Hilbert space.

- $k_\mathcal{X}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_\Theta\colon \Theta \times \Theta \to \mathbb{R}$ two scalar valued kernels.
- Operator-valued kernel $K(x, z) = k_\mathcal{X}(x, z) I_{\mathcal{H}_{k_\Theta}}$ associated to $\mathcal{H}_K$ a space of function-valued functions.
- $\mathcal{H}_K = \overline{\text{span}}\ \{ K(\cdot, x)f \mid \ x \in \mathcal{X},\ \ f \in \mathcal{H}_{k_\Theta} \} \cong \mathcal{H}_{k_\mathcal{X}} \otimes \mathcal{H}_{k_\Theta}$
- Hilbert norm $\|h\|^2_{\mathcal{H}_K}$ as regularizer $\Omega(h)$

Optimization

Proposition (Representer). *If* $\forall \theta \in \Theta$, $\nu(\theta, \cdot, \cdot)$ *is proper lower semicontinuous with respect to its second argument, (5) has a unique solution* $h^* \in \mathcal{H}_K$, *and* $\exists\ (\alpha_{ij})_{i,j=1}^{n,m} \in \mathbb{R}^{n \times m}$ *such that* $\forall (x, \theta) \in \mathcal{X} \times \Theta$

$$h^*(x)(\theta) = \sum_{i=1}^n \sum_{j=1}^m \alpha_{ij} k_\mathcal{X}(x, x_i) k_\Theta(\theta, \theta_j).$$

- Solution shaped by $k_\mathcal{X}$ and k_Θ (gaussian, laplacian, ...)
- Infinite dimensional problem $\Rightarrow$ size $n \cdot m$
- In practice, solved via smoothing ν + L-BFGS.

Excess Risk Guarantees

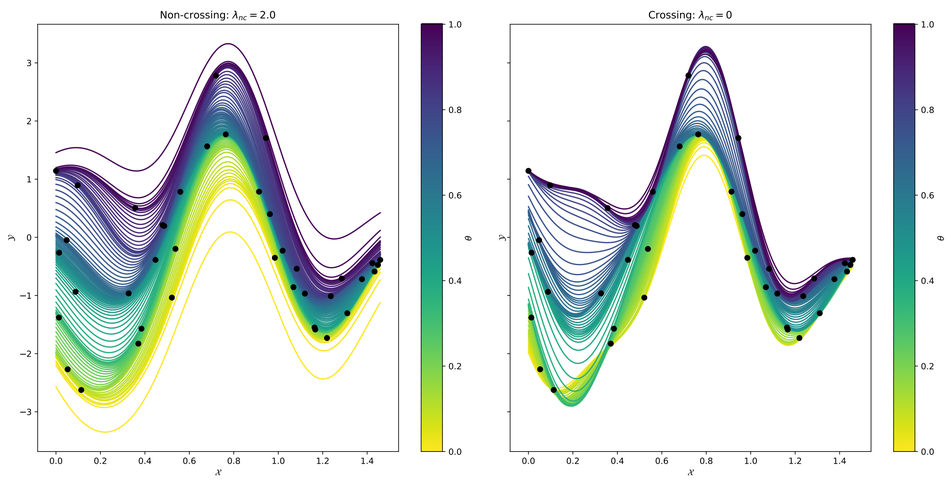
Framework of vv-RKHS allows for proper analysis [3], tradeoff n/m

$$R(h^*) \leqslant \widetilde{R}_S(h^*) + \mathcal{O}_{\mathbf{P}_{X,Y}}\left(\frac{1}{\sqrt{n}}\right) + \mathcal{O}\left(\frac{\log(m)}{m}\right).$$

Numerical Experiments

QR: Continuous model $\Rightarrow$ new non-crossing constraint:

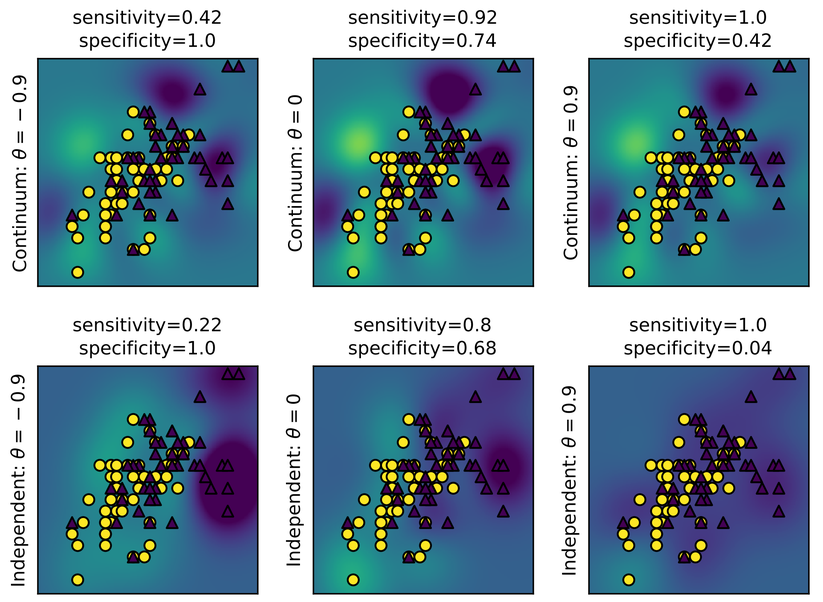
$$\widetilde{\Omega}_{nc}(h) = \frac{\lambda_{nc}}{nm}\sum_{i=1}^n \sum_{j=1}^m \left| -\frac{\partial h}{\partial \theta}(x_i)(\theta_j) \right|_+.$$



Left plot: strong non-crossing penalty ($\lambda_{nc} = 2$). Right plot: no non-crossing penalty ($\lambda_{nc} = 0$). The plots show 100 quantiles of the continuum learned, linearly spaced between 0 (blue) and 1 (red).

$\Rightarrow$ Matches state of the art on 20 UCI datasets

CSC: Improved performances:



Iris dataset. Top: infinite learning; bottom: independent learning for $\theta \in \{-0.9, 0, 0.9\}$.

Code available: https://bitbucket.org/RomainBrault/itl/

This work was funded by the chair Machine Learning for Big Data of Télécom ParisTech, the Digicosme labex, and the Data Science Initiative.

[1] Sangnier, Maxime and Fercoq, Olivier and d'Alché-Buc, Florence. Joint quantile regression in vector-valued RKHSs. In *NIPS*, pp. 3693–3701, 2016.
[2] Takeuchi, Ichiro and Hongo, Tatsuya and Sugiyama, Masashi and Nakajima, Shinichi. Parametric task learning. In *NIPS*, pp. 1358–1366, 2013.
[3] Kadri, Hachem and Duflos, Emmanuel and Preux, Philippe and Canu, Stéphane and Rakotomamonjy, Alain and Audiffren, Julien. Operator-valued kernels for learning from functional response data. *J. Mach. Learn. Res.*, vol 16 pp 1–54, 2015.