Infinite Task Learning in RKHSs

R. Brault♣, A. Lambert†, Z. Szabó⋆, M. Sangnier♠, F. d'Alché-Buc†

♣ L2S, Centrale-Supélec. † LTCI, Télécom ParisTech. ⋆ CMAP, École Polytechnique. ♠ LIP6, Sorbonne Université.

A Kind of Multi-Output Regression

Risk minimization for function-valued regression:

- $\mathcal{X}$ input space ($\mathbb{R}^d$), Θ parameter space ($\subset \mathbb{R}$), $\mathcal{Y}$ output space ($\subset \mathbb{R}$).
- Hypothesis space $\mathcal{H} \subset \mathcal{F}(\mathcal{X}; \mathcal{F}(\Theta; \mathcal{Y}))$, *i.e.* $h(x) \in \mathcal{F}(\Theta; \mathcal{Y})$.
- Parametrized cost $v\colon \Theta \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Local loss $V(y, h(x)) := \int_\Theta v(\theta, y, h(x)(\theta)) d\mu(\theta)$.

Minimizing population risk:

$$\operatorname*{arg\,min}_{h \in \mathcal{H}} R(h) := \mathbf{E}_{X,Y}\left[V(Y, h(X))\right]. \tag{1}$$

$\Rightarrow$ Extension of Multi-Task Learning to an infinite number of tasks [1].

Two examples

Quantile Regression (QR): Given $X, Y \in \mathcal{X} \times \mathcal{Y}$ random variables, estimate the quantile function of the conditional distribution $\mathbf{P}_{Y|X}$:

$$q(x)(\theta) = \inf\left\{t \in \mathcal{Y},\ \mathbf{P}_{Y|X=x}[Y \leqslant t] \geqslant \theta\right\} \quad \forall (x, \theta) \in \mathcal{X} \times (0,1). \tag{2}$$

Pinball loss:

$$v(\theta, y, h(x)) = |\theta - \mathbb{1}_{\mathbb{R}_-}(y - h(x))||y - h(x)|. \tag{3}$$

Proposition. q *defined in (2) minimizes (1) for the pinball loss (3).*

Cost-Sensitive Classification (CSC): Support Vector Machine with asymmetric loss function

$$v(\theta, y, h(x)) = \left|\frac{\theta + 1}{2} - \mathbb{1}_{\{-1\}}(y)\right| |1 - y h(x)|_+.$$

The value of θ influences how the different classes are penalized.

Sampled Empirical Risk

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

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

Sampled empirical risk:

$$\widetilde{R}_{\mathcal{S}}(h) := \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m v(\theta_j, y_i, h(x_i)(\theta_j)).$$

Regularized problem:

$$\operatorname*{arg\,min}_{h \in \mathcal{H}} \widetilde{R}_{\mathcal{S}}(h) + \lambda \Omega(h). \tag{4}$$

Vector-Valued RKHSs

Natural extension of RKHS for modelling 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.
- $\mathcal{H}_K = \overline{\operatorname{span}}\left\{K(\cdot, x) f \mid x \in \mathcal{X},\ f \in \mathcal{H}_{k_\Theta}\right\} \cong \mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_\Theta}$.
- Hilbert norm $\frac{1}{2}\|h\|^2_{\mathcal{H}_K}$ as regularizer $\Omega(h)$.

Optimization

Proposition (Representer)**.** *If* $\forall \theta \in \Theta$, $v(\theta, \cdot, \cdot)$ *is proper lower semicontinuous with respect to its second argument, (4) 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 v + L-BFGS.

Excess Risk Guarantees

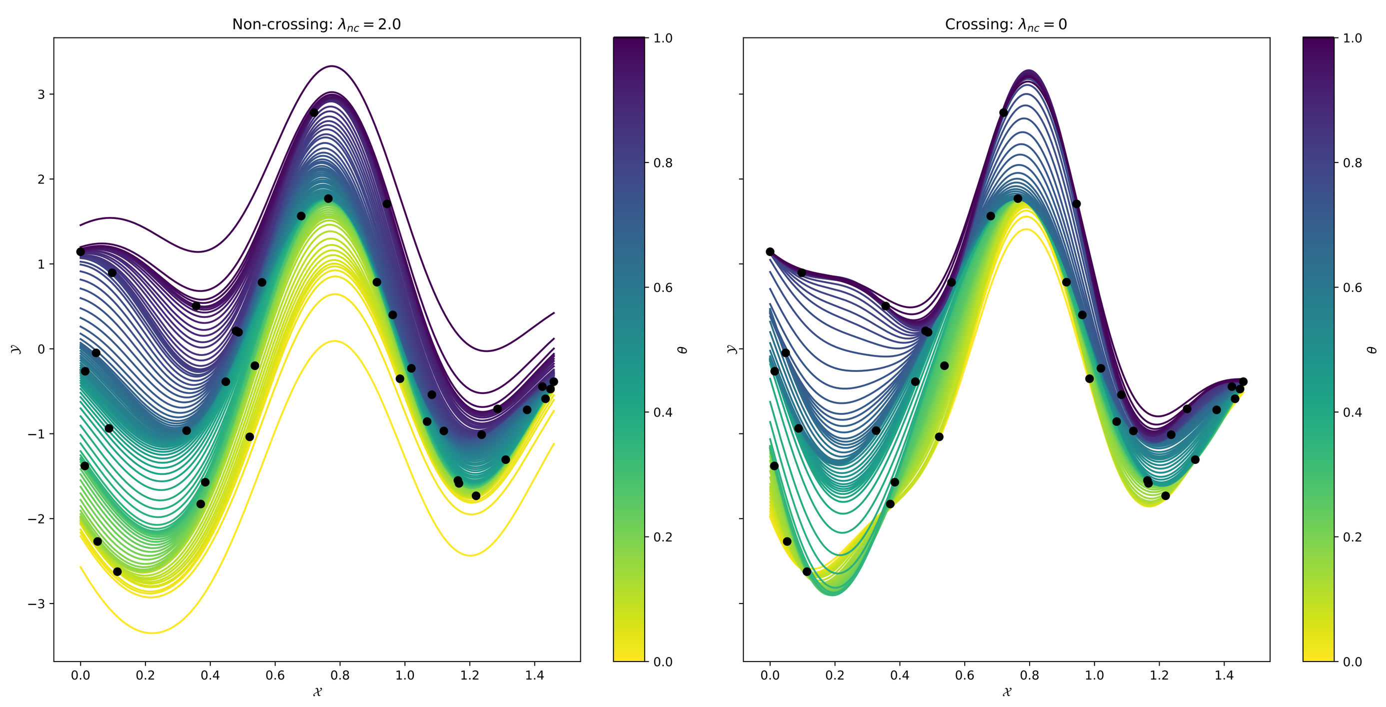
Framework of vv-RKHS allows for proper analysis [2], tradeoff n/m

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

Numerical Experiments

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

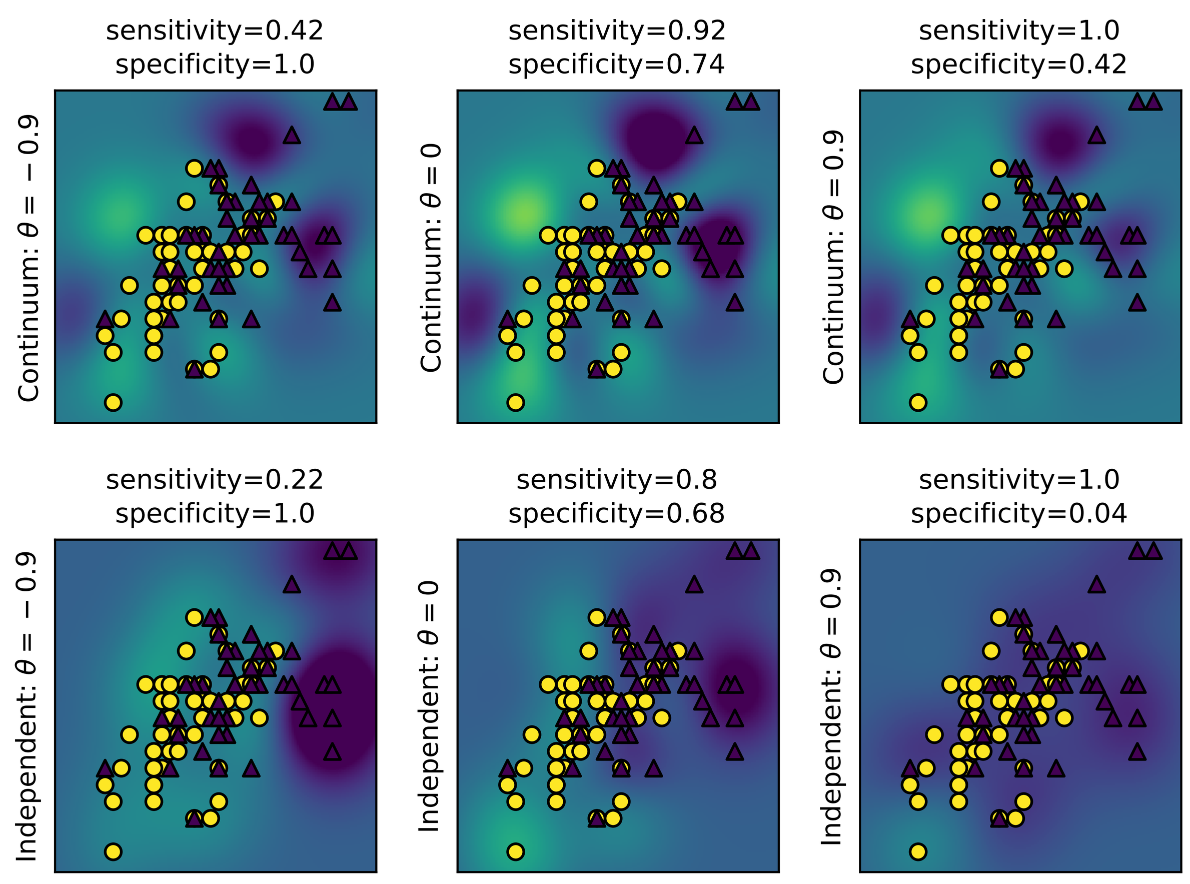
$$\widetilde{\Omega}_{nc}(h) = \frac{\lambda_{nc}}{nm}\sum_{i=1}^n \sum_{j=1}^m \left|-\frac{\partial h}{\partial \theta}(x_i)(\theta_j)\right|_+.$$



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

$\Rightarrow$ Matches state of the art [3] on 20 UCI datasets

CSC: Improved performance:



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 supported by the Labex DigiCosme (project ANR-11-LABEX-0045- DIGICOSME) and the industrial chair Machine Learning for Big Data at Télécom ParisTech.

[1] Takeuchi, Ichiro and Hongo, Tatsuya and Sugiyama, Masashi and Nakajima, Shinichi. Parametric task learning. In *NIPS*, pp. 1358–1366, 2013.

[2] 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. *JMLR*, vol 16 pp 1–54, 2015.

[3] Sangnier, Maxime and Fercoq, Olivier and d'Alché-Buc, Florence. Joint quantile regression in vector-valued RKHSs. In *NIPS*, pp. 3693–3701, 2016.