Infinite Task Learning

In RKHSs

PASADENA

Presented by:

R. Brault (Thales).

Joint Work with:

A. Lambert (LTCI), Z. Szabó (CMAP), M. Sangnier (LPSM), F. d'Alché-Buc (LTCI)

Acknowledgement:

A. Tenenhaus (L2S)

Friday, the 15^{th} February 2019

https://arxiv.org/pdf/1805.08809.pdf

Motivations

E.g. Quantile Regression

Let X and Y be two random variables taking values in $\mathcal{X}$ and $\mathbb{R}$.

Objective:

Learn the quantile $\theta \in (0,\ 1)$:

$$q_\theta(x) = \inf\ \{y \in \mathbb{R}, P(\{Y \leq y | X = x\}) = \theta\},$$

from i.i.d. training copies:

$$\mathcal{S} := ((X_i, Y_i))_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X, Y).$$

Method
[koenker, 1978; Takeuchi, 2006]:

Minimize the "pinball loss" in the function h:

$$q_\theta = \arg\min_h R(h) := \arg\min_h E\left[\max(\theta(Y - h(X)), (1-\theta)(h(X) - Y))\right].$$

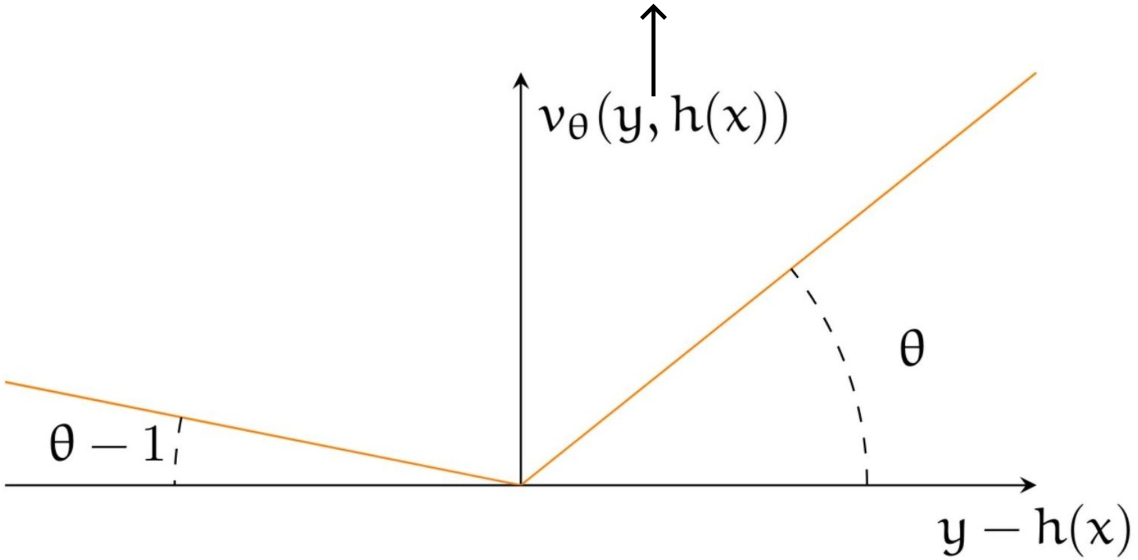
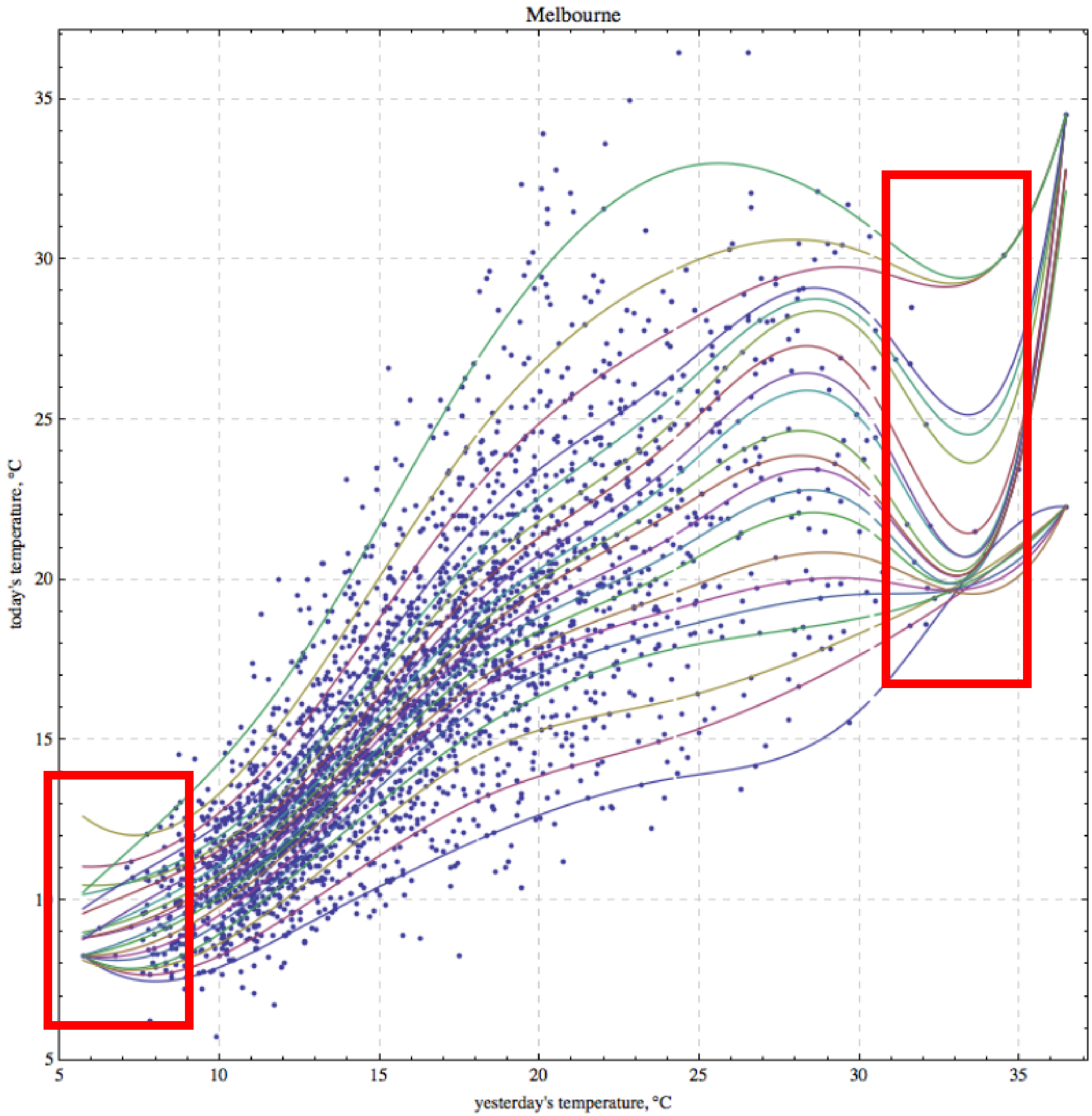


Figure S.4: Pinball loss for $\theta = 0.8$.

$$\arg\min_h R_{\mathcal{S}}(h) = \arg\min_h \frac{1}{n}\sum_{i=1}^{n} \max(\theta(Y_i - h(X_i)), (1-\theta)(h(X_i) - Y_i)) + \lambda\Omega(h).$$



https://mathematicaforprediction.wordpress.com/category/quantile-regression/page/2/

Drawbacks:

- Not adapted to the structure of the problem,
- No way to recover other non-learned quantiles,
- Inefficient.

Setting

Learn problem with risk depending on hyperparameters for **all** hyperparameters values.

Related work: [Takeuchi, 2013; Sangnier et al. 2016, Glazer et al 2013].

- **Quantile level,**
- **Density level sets,**
- **Error sensitivity in classification.**

Framework

A Functional Approach

Idea

Learn Function-Valued Functions:

'input $\mapsto$ (hyperparameter $\mapsto$ output)'

In a nutshell '$x \mapsto (\theta \mapsto y)$'.

Given an input '$x \in \mathcal{X}$' the model returns a function, with suitable properties, that predict an output '$y \in \mathbb{R}$' from an hyperparameter '$\theta \in \Theta$'.

The Infinite Task Learning Framework

Remember:

$$\arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} v_\theta(h(x_i), y_i) + \lambda\Omega(h)$$

e.g. pinball loss

In Infinite task learning:

$$\arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \int_\Theta v_\theta(h(x_i)(\theta), y_i) d\mu(\theta) + \lambda\Omega(h)$$

Let $V(h(x), y) = \int_\Theta v_\theta(h(x)(\theta), y) d\mu(\theta)$.

Finite Sample Properties

How Do We Learn in Practice?

Estimating The Integral Term

Replace $V(h(x), y) = \int_\Theta v_\theta(h(x)(\theta), y) d\mu(\theta)$

with $\tilde{V}(h(x), y) = \sum_{j=1}^{m} w_j v_{\theta_j}(h(x)(\theta_j), y)$.

- $\theta_j' s$ cannot depend on h,
- Quasi Monte-Carlo: low discrepency sequences have error rate of $\mathcal{O}(m^{-1}\log(m))$,
- No overkill in precision.

Handling Function-Valued Functions

The model h lives in a Vector-Valued RKHS [Pedrick, 1957]

- Hilbert space of functions with values in a Hilbert space.
- Regularity property (continuous evaluation functional) and inner product [Carmeli et al. 2006].

Take two scalar-valued kernels $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_\Theta : \Theta \times \Theta \to \mathbb{R}$.
Construct

$$K : \begin{cases} \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H}_{K_\Theta}) \\ (x, z) \mapsto k_{\mathcal{X}}(x, z) I_{\mathcal{H}_{k_\Theta}} \end{cases}$$

Then $\mathcal{H}_K \simeq \mathcal{H}_{k_{\mathcal{X}} \times k_\Theta} = \overline{\operatorname{span}} \left\{ k_{\mathcal{X}}(\cdot, x) k_\Theta(\cdot, \theta) \quad \middle| \quad \forall (x, \theta) \in \mathcal{X} \times \Theta \right\}$.

A Representer Theorem

$$h^* = \arg\min_{h \in \mathcal{H}_K} \sum_{i=1}^{n} \tilde{V}(h(x_i), y_i) + \lambda ||h||^2_{\mathcal{H}_K}, \text{ with } \lambda > 0. \qquad (1)$$

Representer Theorem:

Assume that the local loss function v_θ is convex, lower semicontinous. Then the solution for (1) exists, is unique an verifies for all $(x, \theta) \in \mathbb{R}^d \times \Theta$,

$$h^*(x)(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} k_{\mathcal{X}}(x, x_i) k_\Theta(\theta, \theta_j), \qquad (2)$$

for some $\alpha_{ij} \in \mathbb{R}^{n \times m}$.

Plug back (2) in (1). We solved with L-BFGS.

Smoothing or Keskar and Wächter (2017).

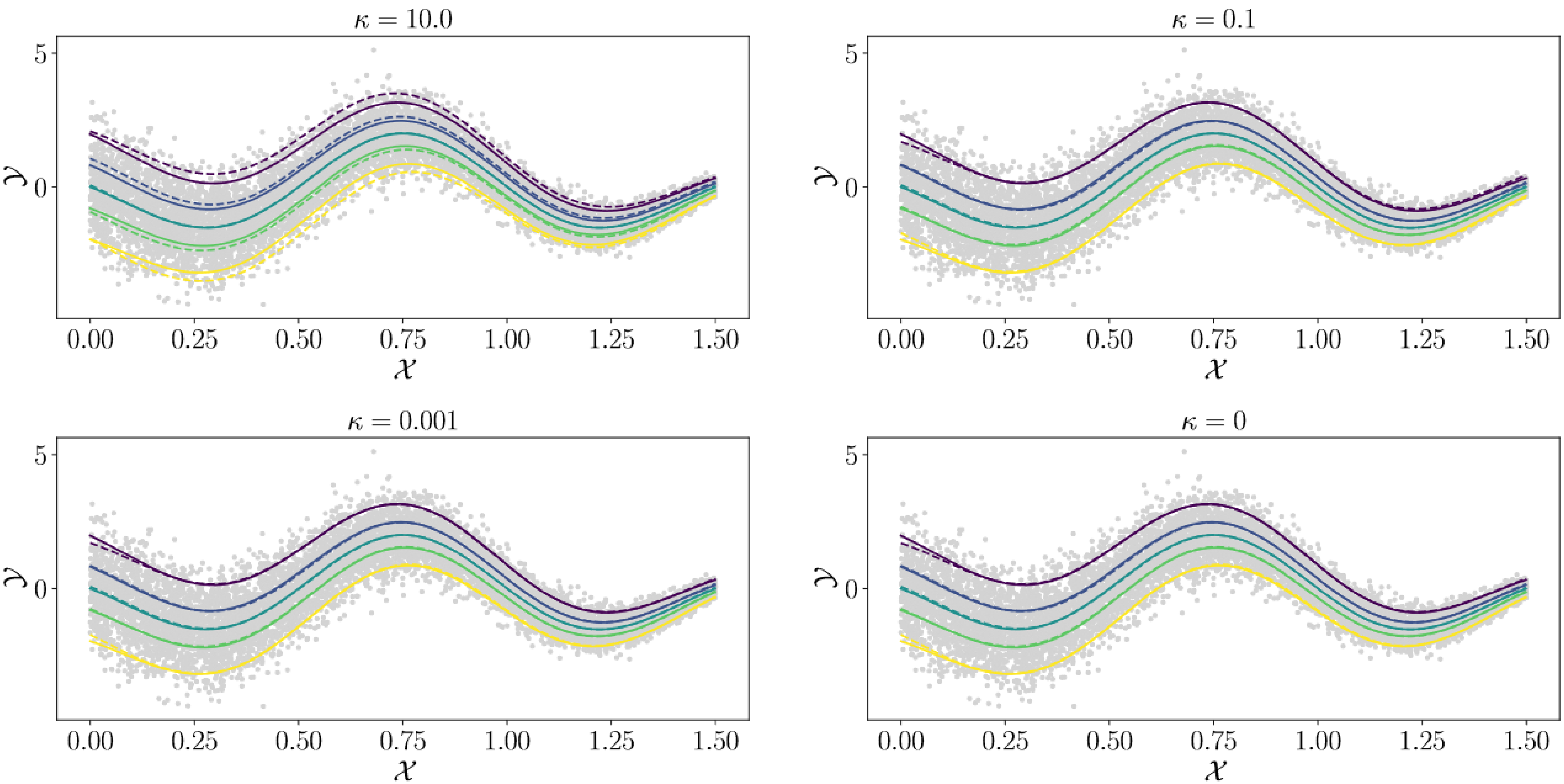


Figure S.5: Impact of the Huber loss smoothing of the pinball loss for differents values of κ.

Numerical Study

Quantile Regression

&

Cost Sensitive Classification

Quantile Regression: Crossing Penalty

$$\Omega_{\mathrm{nc}}(h) = \lambda_{\mathrm{nc}} \int_{\mathcal{X}} \int_{\Theta} max \left(-\frac{\partial h}{\partial \theta}(X)(\theta), 0 \right) d\mu(\theta) dP(X)$$

We have a representer theorem.

Zhou, D.-X. (2008)

Perspectives:

Christian Agrell (2019): https://arxiv.org/pdf/1901.03134.pdf

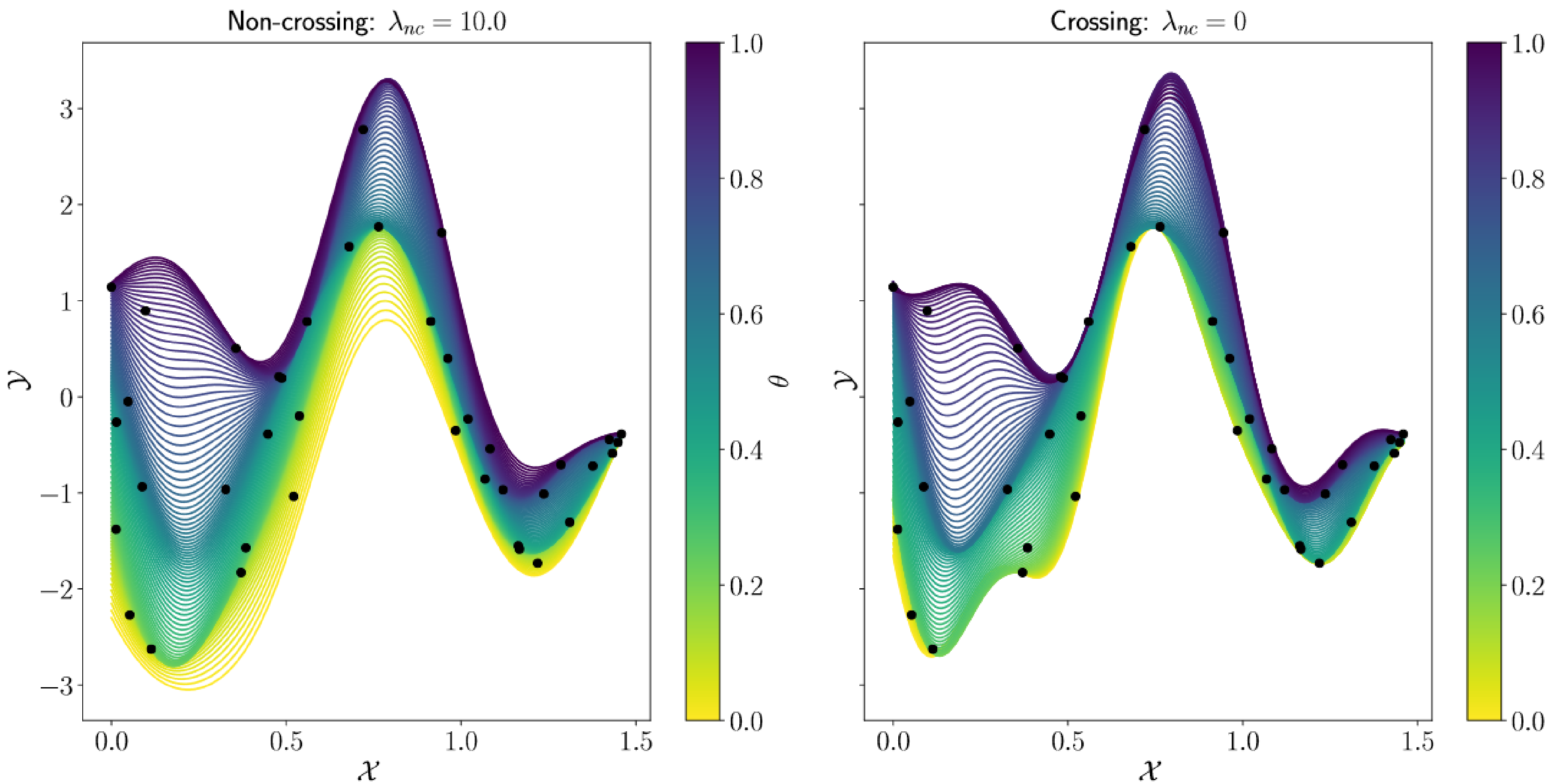


Figure 1: Impact of crossing penalty on toy data. Left plot: strong non-crossing penalty ($\lambda_{\text{nc}} = 10$). Right plot: no non-crossing penalty ($\lambda_{\text{nc}} = 0$). The plots show 100 quantiles of the continuum learned, linearly spaced between 0 (blue) and 1 (red). Notice that the non-crossing penalty does not provide crossings to occur in the regions where there is no points to enforce the penalty (e. g. $x \in [0.13, 0.35]$). This phenomenon is alleviated by the regularity of the model.

Quantile Regression: Real Data

	JQR				IND				∞-QR	
	(PINBALL	P.-VAL.)	(CROSS	P.-VAL.)	(PINBALL	P.-VAL.)	(CROSS	P.-VAL.)	PINBALL	CROSS
COBARORE	159 ± 24	$9 \cdot 10^{-01}$	0.1 ± 0.4	$6 \cdot 10^{-01}$	150 ± 21	$2 \cdot 10^{-01}$	0.3 ± 0.8	$7 \cdot 10^{-01}$	165 ± 36	2.0 ± 6.0
ENGEL	175 ± 555	$6 \cdot 10^{-01}$	0.0 ± 0.2	$1 \cdot 10^{+00}$	63 ± 53	$8 \cdot 10^{-01}$	4.0 ± 12.8	$8 \cdot 10^{-01}$	47 ± 6	0.0 ± 0.1
BOSTONHOUSING	49 ± 4	$8 \cdot 10^{-01}$	0.7 ± 0.7	$2 \cdot 10^{-01}$	49 ± 4	$8 \cdot 10^{-01}$	$\mathbf{1.3 \pm 1.2}$	$1 \cdot 10^{-05}$	49 ± 4	0.3 ± 0.5
CAUTION	88 ± 17	$6 \cdot 10^{-01}$	0.1 ± 0.2	$6 \cdot 10^{-01}$	89 ± 19	$4 \cdot 10^{-01}$	$\mathbf{0.3 \pm 0.4}$	$2 \cdot 10^{-04}$	85 ± 16	0.0 ± 0.1
FTCOLLINSSNOW	154 ± 16	$8 \cdot 10^{-01}$	0.0 ± 0.0	$6 \cdot 10^{-01}$	155 ± 13	$9 \cdot 10^{-01}$	0.2 ± 0.9	$8 \cdot 10^{-01}$	156 ± 17	0.1 ± 0.6
HIGHWAY	103 ± 19	$4 \cdot 10^{-01}$	0.8 ± 1.4	$2 \cdot 10^{-02}$	99 ± 20	$9 \cdot 10^{-01}$	$\mathbf{6.2 \pm 4.1}$	$1 \cdot 10^{-07}$	105 ± 36	0.1 ± 0.4
HEIGHTS	127 ± 3	$1 \cdot 10^{+00}$	0.0 ± 0.0	$1 \cdot 10^{+00}$	127 ± 3	$9 \cdot 10^{-01}$	0.0 ± 0.0	$1 \cdot 10^{+00}$	127 ± 3	0.0 ± 0.0
SNIFFER	43 ± 6	$8 \cdot 10^{-01}$	0.1 ± 0.3	$2 \cdot 10^{-01}$	44 ± 5	$7 \cdot 10^{-01}$	$\mathbf{1.4 \pm 1.2}$	$6 \cdot 10^{-07}$	44 ± 7	0.1 ± 0.1
SNOWGEESE	55 ± 20	$7 \cdot 10^{-01}$	0.3 ± 0.8	$3 \cdot 10^{-01}$	53 ± 18	$6 \cdot 10^{-01}$	0.4 ± 1.0	$5 \cdot 10^{-02}$	57 ± 20	0.2 ± 0.6
UFC	81 ± 5	$6 \cdot 10^{-01}$	$\mathbf{0.0 \pm 0.0}$	$4 \cdot 10^{-04}$	82 ± 5	$7 \cdot 10^{-01}$	$\mathbf{1.0 \pm 1.4}$	$2 \cdot 10^{-04}$	82 ± 4	0.1 ± 0.3
BIGMAC2003	80 ± 21	$7 \cdot 10^{-01}$	$\mathbf{1.4 \pm 2.1}$	$4 \cdot 10^{-04}$	74 ± 24	$9 \cdot 10^{-02}$	$\mathbf{0.9 \pm 1.1}$	$7 \cdot 10^{-05}$	84 ± 24	0.2 ± 0.4
UN3	98 ± 9	$8 \cdot 10^{-01}$	0.0 ± 0.0	$1 \cdot 10^{-01}$	99 ± 9	$1 \cdot 10^{+00}$	$\mathbf{1.2 \pm 1.0}$	$1 \cdot 10^{-05}$	99 ± 10	0.1 ± 0.4
BIRTHWT	141 ± 13	$1 \cdot 10^{+00}$	0.0 ± 0.0	$6 \cdot 10^{-01}$	140 ± 12	$9 \cdot 10^{-01}$	0.1 ± 0.2	$7 \cdot 10^{-02}$	141 ± 12	0.0 ± 0.0
CRABS	$\mathbf{11 \pm 1}$	$4 \cdot 10^{-05}$	0.0 ± 0.0	$8 \cdot 10^{-01}$	$\mathbf{11 \pm 1}$	$2 \cdot 10^{-04}$	$\mathbf{0.0 \pm 0.0}$	$2 \cdot 10^{-05}$	13 ± 3	0.0 ± 0.0
GAGURINE	61 ± 7	$4 \cdot 10^{-01}$	0.0 ± 0.1	$3 \cdot 10^{-03}$	62 ± 7	$5 \cdot 10^{-01}$	$\mathbf{0.1 \pm 0.2}$	$4 \cdot 10^{-04}$	62 ± 7	0.0 ± 0.0
GEYSER	105 ± 7	$9 \cdot 10^{-01}$	0.1 ± 0.3	$9 \cdot 10^{-01}$	105 ± 6	$9 \cdot 10^{-01}$	0.2 ± 0.3	$6 \cdot 10^{-01}$	104 ± 6	0.1 ± 0.2
GILGAIS	51 ± 6	$5 \cdot 10^{-01}$	0.1 ± 0.1	$1 \cdot 10^{-01}$	49 ± 6	$6 \cdot 10^{-01}$	$\mathbf{1.1 \pm 0.7}$	$2 \cdot 10^{-05}$	49 ± 7	0.3 ± 0.3
TOPO	69 ± 18	$1 \cdot 10^{+00}$	0.1 ± 0.5	$1 \cdot 10^{+00}$	71 ± 20	$1 \cdot 10^{+00}$	$\mathbf{1.7 \pm 1.4}$	$3 \cdot 10^{-07}$	70 ± 17	0.0 ± 0.0
MCYCLE	66 ± 9	$9 \cdot 10^{-01}$	0.2 ± 0.3	$7 \cdot 10^{-03}$	66 ± 8	$9 \cdot 10^{-01}$	$\mathbf{0.3 \pm 0.3}$	$7 \cdot 10^{-06}$	65 ± 9	0.0 ± 0.1
CPUS	$\mathbf{7 \pm 4}$	$2 \cdot 10^{-04}$	$\mathbf{0.7 \pm 1.0}$	$5 \cdot 10^{-04}$	$\mathbf{7 \pm 5}$	$3 \cdot 10^{-04}$	$\mathbf{1.2 \pm 0.8}$	$6 \cdot 10^{-08}$	16 ± 10	0.0 ± 0.0

Table 1: Quantile Regression on 20 UCI datasets. Reported: 100×value of the pinball loss, 100×crossing loss (smaller is better). p.-val.: outcome of the Mann-Whitney-Wilcoxon test of JQR vs. ∞-QR and Independent vs. ∞-QR. Boldface: significant values.

Cost Sensitive Classification

$$V(h(x), y) = \int_{[-1,1]} \left| \frac{\theta+1}{2} - 1_{\{-1\}}(y) \right| \max(1 - h(x)(\theta)y, 0) d\mu(\theta)$$

DATASET	METHOD	$\theta = -0.9$		$\theta = 0$		$\theta = +0.9$	
		SENSITIVITY	SPECIFICITY	SENSITIVITY	SPECIFICITY	SENSITIVITY	SPECIFICITY
TWO-MOONS	IND	0.3 ± 0.05	0.99 ± 0.01	0.83 ± 0.03	0.86 ± 0.03	0.99 ± 0	0.32 ± 0.06
	∞-CSC	0.32 ± 0.05	0.99 ± 0.01	0.84 ± 0.03	0.87 ± 0.03	1 ± 0	0.36 ± 0.04
CIRCLES	IND	0 ± 0	1 ± 0	0.82 ± 0.02	0.84 ± 0.03	1 ± 0	0 ± 0
	∞-CSC	$\boxed{0.15 \pm 0.05}$	1 ± 0	0.82 ± 0.02	0.84 ± 0.03	1 ± 0	$\boxed{0.12 \pm 0.05}$
IRIS	IND	0.88 ± 0.08	0.94 ± 0.06	0.94 ± 0.05	0.92 ± 0.06	0.97 ± 0.05	0.87 ± 0.06
	∞-CSC	0.89 ± 0.08	0.94 ± 0.05	0.94 ± 0.06	0.92 ± 0.05	0.97 ± 0.04	0.90 ± 0.05
TOY	IND	0.51 ± 0.06	0.98 ± 0.01	0.83 ± 0.03	0.86 ± 0.03	0.97 ± 0.01	0.49 ± 0.07
	∞-CSC	$\boxed{0.63 \pm 0.04}$	0.96 ± 0.01	0.83 ± 0.03	0.85 ± 0.03	0.95 ± 0.02	$\boxed{0.61 \pm 0.04}$

Table 2: ∞-CSC vs Independent (IND)-CSC. Higher is better.

Statistical Study

Generalization Error

β-Stability

[Kadri et al., 2015; Bousquet et al., 2002]

Generalization Bound:

Let h^* be the unique solution of the Quantile Regression or Cost Sensitive Classification problem with Quasi Monte-Carlo sampling. Under mild assumptions it holds,

$$R(h^*) \leq \tilde{R}_{\mathcal{S}}(h^*) + \mathcal{O}_{P_{(X,Y)}}\left(\frac{1}{\sqrt{n}}\right) + \mathcal{O}\left(\frac{\log(m)}{m}\right). \qquad (3)$$

- Requires bounded random variable in Quantile Regression,
- Indicates the potential tradeoff between 'n' and 'm',
- Mild assumptions on the kernel.

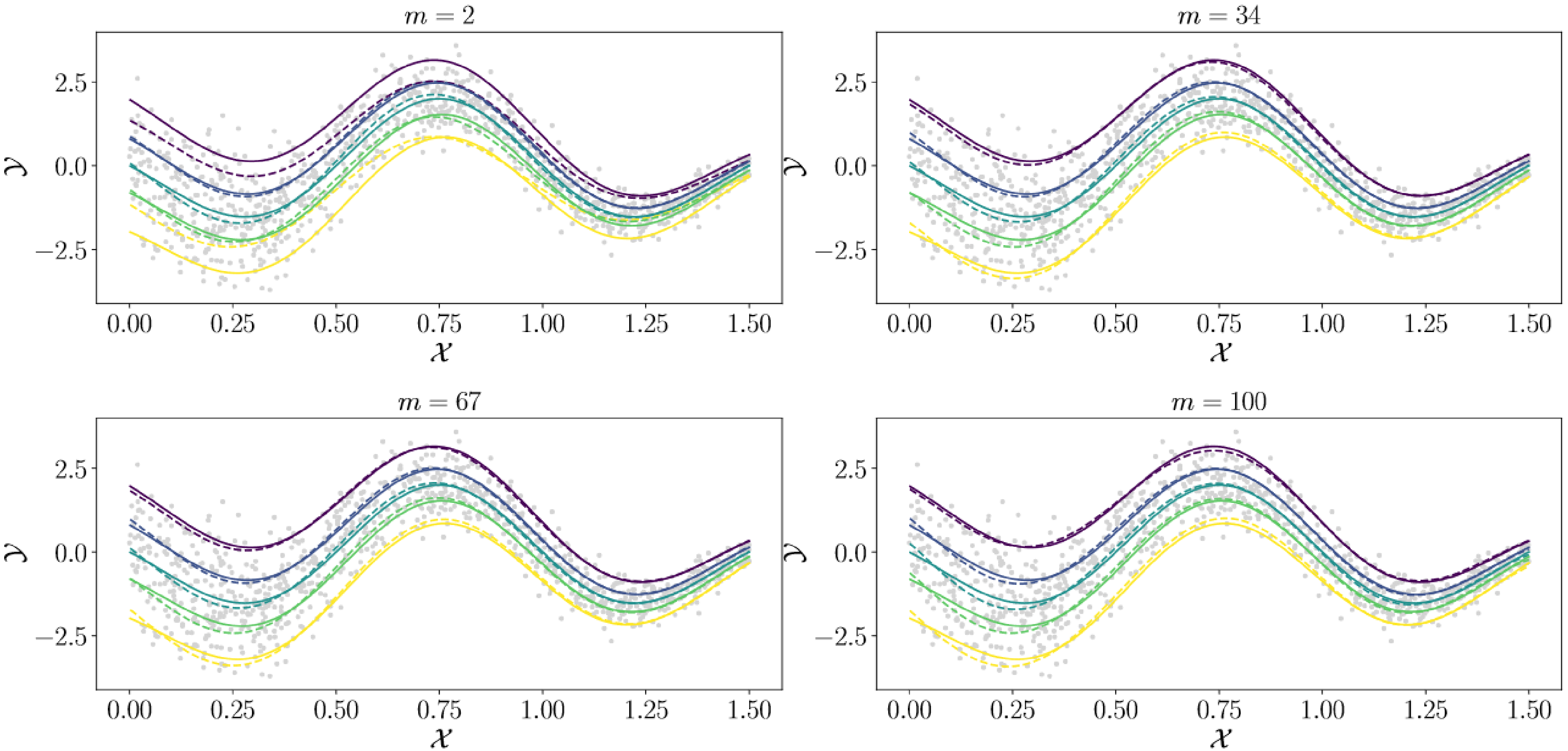


Figure S.4: Impact of the number of hyperparameters sampled.

Extensions

Unsupervised Tasks

The One-Class SVM [Schölkopf, 2000]

Given $(x_i)_{i=1}^n \overset{\text{i. i. d.}}{\sim} X$ and $\theta \in (0, 1)$, minimize for $(h, t) \in \mathcal{H}_{k_\mathcal{X}} \times \mathbb{R}$

$$J(h, t) = \frac{1}{n} \sum_{i=1}^{n} \frac{max(0, t - h(x_i))}{\theta} - t + ||h||^2_{\mathcal{H}_{K_\mathcal{X}}}.$$

Decision function

$$d(x) = 1_{\mathbb{R}_+}(h(x) - t)$$

θ-property:

The decision function should separate the training data into two subsets (normal / abnormal) with proportion θ of abnormal.

The ∞-OCSVM

Given $(x_i)_{i=1}^N \overset{\text{i. i. d.}}{\sim} X$ and $\theta \in (0,\, 1)$, minimize for $(h,\, t) \in \mathcal{H}_{k_\mathcal{X}} \times \mathbb{R}$

$$J(h,\, t) = \frac{1}{n} \sum_{i=1}^{n} \int_\Theta \frac{max(0, t - h(x_i)(\theta))}{\theta} - t(\theta) + ||h(\cdot)(\theta)||^2_{\mathcal{H}_{K_\mathcal{X}}} d\mu(\theta). \quad (4)$$

New regularizer

Again one can use Quasi Monte-Carlo or quadrature rules to approximate the integral.

A New Representer Theorem

(weak) Representer Theorem:

The solution for (4) exists, is unique an verifies for all $(x,\, \theta) \in \mathbb{R}^d \times (0, 1)$,

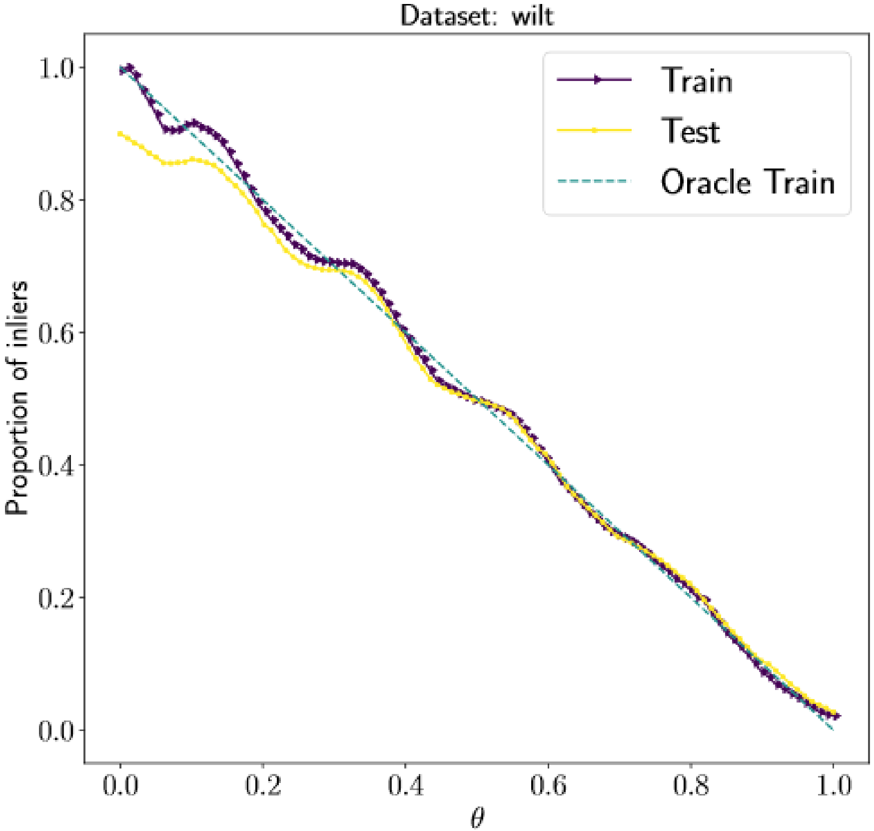
$$h^*(x)(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} k_{\mathcal{X}}(x,\, x_i) k_{\Theta}(\theta,\, \theta_j), \qquad (5)$$

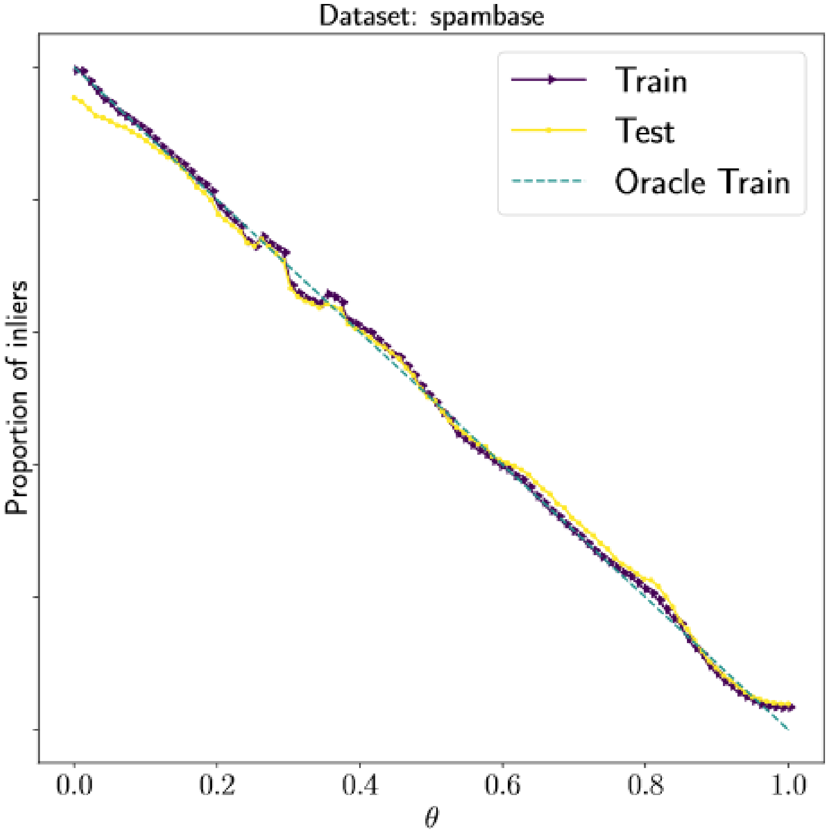
$$t^*(\theta) = \sum_{j=1}^{m} \beta_j k_b(\theta,\, \theta_j),$$

for some $\alpha_{ij} \in \mathbb{R}^{n \times m}$ and $\beta_j \in \mathbb{R}^m$.

- Weaker regularizer, coercivity is not trivial,
- Convex problem with $(n+1)m$ parameters sloved again with L-BFGS.

Numerical Illustrations





- The θ-property is approximately respected

	LOSS	PENALTY
QUANTILE	$\int_{[0,1]} \left\|\theta - \mathbb{1}_{\mathbb{R}_-}(y - h_x(\theta))\right\| \|y - h_x(\theta)\| d\mu(\theta)$	$\lambda_{nc} \int_{[0,1]} \left\|-\frac{dh_x}{d\theta}(\theta)\right\|_+ d\mu(\theta) + \frac{\lambda}{2}\\|h\\|^2_{\mathcal{H}_K}$
M-QUANTILE (SMOOTH)	$\int_{[0,1]} \left\|\theta - \mathbb{1}_{\mathbb{R}_-}(y - h_x(\theta))\right\| \psi_1^\kappa(y - h_x(\theta))\, d\mu(\theta)$	$\lambda_{nc} \int_{(0,1)} \psi_+^\kappa\left(-\frac{dh_x}{d\theta}(\theta)\right) d\mu(\theta) + \frac{\lambda}{2}\\|h\\|^2_{\mathcal{H}_K}$
EXPECTILES (SMOOTH)	$\int_{[0,1]} \left\|\theta - \mathbb{1}_{\mathbb{R}_-}(y - h_x(\theta))\right\| (y - h_x(\theta))^2\, d\mu(\theta)$	$\lambda_{nc} \int_{(0,1)} \left\|-\frac{dh_x}{d\theta}(\theta)\right\|_+^2 d\mu(\theta) + \frac{\lambda}{2}\\|h\\|^2_{\mathcal{H}_K}$
COST-SENSITIVE	$\int_{[-1,1]} \left\|\frac{\theta+1}{2} - \mathbb{1}_{\{-1\}}(y)\right\| \|1 - yh_x(\theta)\|_+ d\mu(\theta)$	$\frac{\lambda}{2}\\|h\\|^2_{\mathcal{H}_K}$
COST-SENSITIVE (SMOOTH)	$\int_{[-1,1]} \left\|\frac{\theta+1}{2} - \mathbb{1}_{\{-1\}}(y)\right\| \psi_+^\kappa(1 - yh_x(\theta))\, d\mu(\theta)$	$\frac{\lambda}{2}\\|h\\|^2_{\mathcal{H}_K}$
LEVEL-SET	$\int_{[\epsilon,1]} -t(\theta) + \frac{1}{\theta}\|t(\theta) - h_x(\theta)\|_+ d\mu(\theta)$	$\frac{1}{2}\int_{[\epsilon,1]} \\|h(\cdot)(\theta)\\|^2_{\mathcal{H}_{k_\mathcal{X}}} d\mu(\theta) + \frac{\lambda}{2}\\|t\\|^2_{\mathcal{H}_{k_b}}$

Table S.3: Examples for objective (8). ψ_1^κ, ψ_+^κ: κ-smoothed absolute value and positive part. $h_x(\theta) := h(x)(\theta)$.

Conclusion

Wrap-Up

- New flexible setting functional multitask (multioutput),
- Recover some settings as limit cases [Sangnier et al. 2016, Glazer et al. 2013],
- New representer theorems and statistical guarantees,
- Compares well to the state of the art.
- https://bitbucket.org/RomainBrault/itl/
- https://arxiv.org/pdf/1805.08809.pdf

Conclusion

Future Directions

Investigate:

- Further algorithmic and statistical guarantees,
- Efficient solvers,
- New regularization term $\int_{\Theta} ||h(\cdot)(\theta)||^2 d\mu(\theta)$,
- Other algorithms (LASSO, SVR, ...),
- Scaling up with Random Fourier Features [Brault et al., 2016],
- Deep architectures?