Regression on Probability Measures: A Simple and Consistent Algorithm

Zoltán Szabó (Gatsby Unit, UCL)

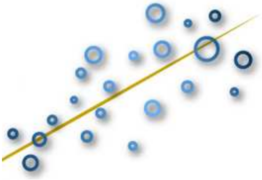
Joint work with

- Bharath K. Sriperumbudur (Department of Statistics, PSU),
- Barnabás Póczos (ML Department, CMU),
- Arthur Gretton (Gatsby Unit, UCL)

CRiSM Seminars
Department of Statistics,
University of Warwick
May 29, 2015

The task

- Samples: $\{(x_i, y_i)\}_{i=1}^{l}$. Goal: $f(x_i) \approx y_i$, find $f \in \mathcal{H}$.



- Distribution regression:
  - x_i-s are distributions,
  - available only through samples: $\{x_{i,n}\}_{n=1}^{N_i}$.
- $\Rightarrow$ Training examples: labelled *bags*.

Example: aerosol prediction from satellite images

- Bag := pixels of a multispectral satellite image over an area.
- Label of a bag := aerosol value.

- Relevance: climate research.
- Engineered methods [Wang et al., 2012]: $100 \times \text{RMSE} = 7.5 - 8.5$.
- Using distribution regression?

Wider context

- Context:
  - machine learning: multi-instance learning,
  - statistics: point estimation tasks (without analytical formula).

- Applications:
  - computer vision: image = collection of patch vectors,
  - network analysis: group of people = bag of friendship graphs,
  - natural language processing: corpus = bag of documents,
  - time-series modelling: user = set of trial time-series.

Several algorithmic approaches

1. Parametric fit: Gaussian, MOG, exp. family [Jebara et al., 2004, Wang et al., 2009, Nielsen and Nock, 2012].
2. Kernelized Gaussian measures: [Jebara et al., 2004, Zhou and Chellappa, 2006].
3. (Positive definite) kernels: [Cuturi et al., 2005, Martins et al., 2009, Hein and Bousquet, 2005].
4. Divergence measures (KL, Rényi, Tsallis): [Póczos et al., 2011].
5. Set metrics: Hausdorff metric [Edgar, 1995]; variants [Wang and Zucker, 2000, Wu et al., 2010, Zhang and Zhou, 2009, Chen and Wu, 2012].

Theoretical guarantee?

- MIL dates back to [Haussler, 1999, Gärtner et al., 2002].

- *Sensible* methods in regression: require density estimation [Póczos et al., 2013, Oliva et al., 2014, Reddi and Póczos, 2014] + assumptions:
  1. compact Euclidean domain.
  2. output $= \mathbb{R}$ ([Oliva et al., 2013] allows distribution).

Kernel, RKHS

- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ kernel on $\mathcal{D}$, if
  - $\exists \varphi : \mathcal{D} \to H$(ilbert space) feature map,
  - $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_H$ $(\forall a, b \in \mathcal{D})$.

Kernel, RKHS

- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ kernel on $\mathcal{D}$, if
  - $\exists \varphi : \mathcal{D} \to H$(ilbert space) feature map,
  - $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_H$ $(\forall a, b \in \mathcal{D})$.
- Kernel examples: $\mathcal{D} = \mathbb{R}^d$ $(p > 0,\ \theta > 0)$
  - $k(a, b) = (\langle a, b \rangle + \theta)^p$: polynomial,
  - $k(a, b) = e^{-\|a-b\|_2^2/(2\theta^2)}$: Gaussian,
  - $k(a, b) = e^{-\theta\|a-b\|_1}$: Laplacian.
- In the $H = H(k)$ RKHS $(\exists!)$: $\varphi(u) = k(\cdot, u)$.

- Euclidean space: $\mathcal{D} = \mathbb{R}^d$.
- Graphs, texts, time series, dynamical systems.

- Distributions!

Problem formulation ($Y = \mathbb{R}$)

- Given:
  - labelled bags $\hat{\mathbf{z}} = \{(\hat{x}_i, y_i)\}_{i=1}^{\ell}$,
  - i^{th} bag: $\hat{x}_i = \{x_{i,1}, \ldots, x_{i,N}\} \overset{i.i.d.}{\sim} x_i \in \mathcal{P}(\mathcal{D})$, $y_i \in \mathbb{R}$.
- Task: find a $\mathcal{P}(\mathcal{D}) \to \mathbb{R}$ mapping based on $\hat{\mathbf{z}}$.

Problem formulation ($Y = \mathbb{R}$)

- Given:
  - labelled bags $\hat{\mathbf{z}} = \{(\hat{x}_i, y_i)\}_{i=1}^{\ell}$,
  - i^{th} bag: $\hat{x}_i = \{x_{i,1}, \ldots, x_{i,N}\} \overset{i.i.d.}{\sim} x_i \in \mathcal{P}(\mathcal{D})$, $y_i \in \mathbb{R}$.
- Task: find a $\mathcal{P}(\mathcal{D}) \to \mathbb{R}$ mapping based on $\hat{\mathbf{z}}$.
- Construction: distribution embedding (μ_x)

$$\mathcal{P}(\mathcal{D}) \xrightarrow{\mu=\mu(k)} X \subseteq H = H(k)$$

Problem formulation ($Y = \mathbb{R}$)

- Given:
  - labelled bags $\hat{\mathbf{z}} = \{(\hat{x}_i, y_i)\}_{i=1}^{\ell}$,
  - i^{th} bag: $\hat{x}_i = \{x_{i,1}, \ldots, x_{i,N}\} \overset{i.i.d.}{\sim} x_i \in \mathcal{P}(\mathcal{D})$, $y_i \in \mathbb{R}$.
- Task: find a $\mathcal{P}(\mathcal{D}) \to \mathbb{R}$ mapping based on $\hat{\mathbf{z}}$.
- Construction: distribution embedding (μ_x) + ridge regression

$$\mathcal{P}(\mathcal{D}) \xrightarrow{\mu=\mu(k)} X \subseteq H = H(k) \xrightarrow{f \in \mathcal{H} = \mathcal{H}(K)} \mathbb{R}.$$

Problem formulation ($Y = \mathbb{R}$)

- Given:
  - labelled bags $\hat{\mathbf{z}} = \{(\hat{x}_i, y_i)\}_{i=1}^{\ell}$,
  - i^{th} bag: $\hat{x}_i = \{x_{i,1}, \ldots, x_{i,N}\} \overset{i.i.d.}{\sim} x_i \in \mathcal{P}(\mathcal{D})$, $y_i \in \mathbb{R}$.
- Task: find a $\mathcal{P}(\mathcal{D}) \to \mathbb{R}$ mapping based on $\hat{\mathbf{z}}$.
- Construction: distribution embedding (μ_x) + ridge regression

$$\mathcal{P}(\mathcal{D}) \xrightarrow{\mu=\mu(k)} X \subseteq H = H(k) \xrightarrow{f \in \mathcal{H} = \mathcal{H}(K)} \mathbb{R}.$$

- Our goal: risk bound compared to the regression function

$$f_\rho(\mu_x) = \int_{\mathbb{R}} y \, \mathrm{d}\rho(y|\mu_x).$$

Goal in details

- Expected risk:

$$\mathcal{R}[f] = \mathbb{E}_{(x,y)} \left| f(\mu_x) - y \right|^2.$$

- <u>Contribution</u>: analysis of the excess risk

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}) = \mathcal{R}[f_{\hat{\mathbf{z}}}^{\lambda}] - \mathcal{R}[f_{\rho}]$$

Goal in details

- Expected risk:

$$\mathcal{R}[f] = \mathbb{E}_{(x,y)} \left| f(\mu_x) - y \right|^2 .$$

- Contribution: analysis of the excess risk

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}) = \mathcal{R}[f_{\hat{\mathbf{z}}}^{\lambda}] - \mathcal{R}[f_{\rho}] \leq g(\ell, N, \lambda) \to 0 \text{ and rates,}$$

Goal in details

- Expected risk:

$$\mathcal{R}[f] = \mathbb{E}_{(x,y)} \left| f(\mu_x) - y \right|^2.$$

- <u>Contribution</u>: analysis of the excess risk

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}) = \mathcal{R}[f_{\hat{\mathbf{z}}}^{\lambda}] - \mathcal{R}[f_{\rho}] \leq g(\ell, N, \lambda) \to 0 \text{ and rates,}$$

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} \left| f(\mu_{\hat{x}_i}) - y_i \right|^2 + \lambda \left\| f \right\|_{\mathcal{H}}^2, \quad (\lambda > 0).$$

Goal in details

- Expected risk:

$$\mathcal{R}\left[f\right] = \mathbb{E}_{(x,y)}\left|f(\mu_x) - y\right|^2.$$

- Contribution: analysis of the excess risk

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}) = \mathcal{R}[f_{\hat{\mathbf{z}}}^{\lambda}] - \mathcal{R}[f_{\rho}] \leq g(\ell, N, \lambda) \to 0 \text{ and rates,}$$

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f\in\mathcal{H}} \frac{1}{\ell}\sum_{i=1}^{\ell}\left|f(\mu_{\hat{x}_i}) - y_i\right|^2 + \lambda\left\|f\right\|_{\mathcal{H}}^2, \quad (\lambda > 0).$$

- We consider two settings:
  1. well-specified case: $f_\rho \in \mathcal{H}$,
  2. misspecified case: $f_\rho \in L^2_{\rho_X} \backslash \mathcal{H}$.

Step-1: mean embedding

- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ kernel; canonical feature map: $\varphi(u) = k(\cdot, u)$.
- Mean embedding of a distribution $x, \hat{x}_i \in \mathcal{P}(\mathcal{D})$:

$$\mu_x = \int_{\mathcal{D}} k(\cdot, u) \mathrm{d}x(u) \in H(k),$$

$$\mu_{\hat{x}_i} = \int_{\mathcal{D}} k(\cdot, u) \mathrm{d}\hat{x}_i(u) = \frac{1}{N} \sum_{n=1}^{N} k(\cdot, x_{i,n}).$$

Step-1: mean embedding

- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ kernel; canonical feature map: $\varphi(u) = k(\cdot, u)$.
- Mean embedding of a distribution $x, \hat{x}_i \in \mathcal{P}(\mathcal{D})$:

$$\mu_x = \int_{\mathcal{D}} k(\cdot, u) \mathrm{d}x(u) \in H(k),$$

$$\mu_{\hat{x}_i} = \int_{\mathcal{D}} k(\cdot, u) \mathrm{d}\hat{x}_i(u) = \frac{1}{N} \sum_{n=1}^{N} k(\cdot, x_{i,n}).$$

- Linear $K \Rightarrow$ set kernel:

$$K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j}) = \left\langle \mu_{\hat{x}_i}, \mu_{\hat{x}_j} \right\rangle_H = \frac{1}{N^2} \sum_{n,m=1}^{N} k(x_{i,n}, x_{j,m}).$$

Step-1: mean embedding

- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ kernel; canonical feature map: $\varphi(u) = k(\cdot, u)$.
- Mean embedding of a distribution $x, \hat{x}_i \in \mathcal{P}(\mathcal{D})$:

$$\mu_x = \int_{\mathcal{D}} k(\cdot, u) \mathrm{d}x(u) \in H(k),$$

$$\mu_{\hat{x}_i} = \int_{\mathcal{D}} k(\cdot, u) \mathrm{d}\hat{x}_i(u) = \frac{1}{N} \sum_{n=1}^{N} k(\cdot, x_{i,n}).$$

- Nonlinear K example:

$$K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j}) = e^{-\frac{\|\mu_{\hat{x}_i} - \mu_{\hat{x}_j}\|_H^2}{2\sigma^2}}.$$

Step-2: ridge regression (analytical solution)

- Given:
  - training sample: $\hat{\mathbf{z}}$,
  - test distribution: t.
- Prediction on t:

$$\boxed{\begin{aligned}(f_{\hat{\mathbf{z}}}^{\lambda} \circ \mu)(t) &= \mathbf{k}(\mathbf{K} + \ell\lambda \mathbf{I}_\ell)^{-1}[y_1; \ldots; y_\ell], \\ \mathbf{K} &= [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathbb{R}^{\ell \times \ell}, \\ \mathbf{k} &= [K(\mu_{\hat{x}_1}, \mu_t), \ldots, K(\mu_{\hat{x}_\ell}, \mu_t)] \in \mathbb{R}^{1 \times \ell}.\end{aligned}} \quad \begin{aligned}(1)\\(2)\\(3)\end{aligned}$$

Blanket assumptions: both settings

- $\mathcal{D}$: separable, topological domain.
- k:
  - bounded: $\sup_{u\in\mathcal{D}} k(u,u) \le B_k \in (0,\infty)$,
  - continuous.
- K: bounded; Hölder continuous: $\exists L > 0, h \in (0,1]$ such that

$$\|K(\cdot,\mu_a) - K(\cdot,\mu_b)\|_{\mathcal{H}} \le L\,\|\mu_a - \mu_b\|_H^h.$$

- y: bounded.
- $X = \mu\left(\mathcal{P}(\mathcal{D})\right) \in \mathcal{B}(H)$.

Well-specified case: performance guarantee

- Difficulty of the task:
  - f_ρ is 'c-smooth',
  - 'b-decaying covariance operator'.
- Contribution: If $\ell \geq \lambda^{-\frac{1}{b}-1}$, then with high probability

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho) \leq \underbrace{\frac{\log^h(\ell)}{N^h\lambda^3} + \lambda^c + \frac{1}{\ell^2\lambda} + \frac{1}{\ell\lambda^{\frac{1}{b}}}}_{g(\ell,N,\lambda)}.$$

Well-specified case: performance guarantee

- Difficulty of the task:
  - f_ρ is 'c-smooth',
  - 'b-decaying covariance operator'.
- <u>Contribution</u>: If $\ell \ge \lambda^{-\frac{1}{b}-1}$, then with high probability

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho) \le \underbrace{\frac{\log^h(\ell)}{N^h \lambda^3} + \lambda^c + \frac{1}{\ell^2 \lambda} + \frac{1}{\ell \lambda^{\frac{1}{b}}}}_{g(\ell, N, \lambda)}. \qquad (4)$$

$\hat{x}_i$ $\qquad$ c-smoothness

Well-specified case: example

Assume

- b is 'large' ($1/b \approx 0$, 'small' effective input dimension),
- $h = 1$ (K: Lipschitz),
- $\boxed{1} = \boxed{2}$ in (4) $\Rightarrow \lambda$; $\ell = N^a$ $(a > 0)$,
- $t = \ell N$: total number of samples processed.

Then

1. $c = 2$ ('smooth' f_ρ): $\mathcal{E}(f_{\hat{\mathbf{z}}}^\lambda, f_\rho) \approx t^{-\frac{2}{7}}$ – faster convergence,
2. $c = 1$ ('non-smooth' f_ρ): $\mathcal{E}(f_{\hat{\mathbf{z}}}^\lambda, f_\rho) \approx t^{-\frac{1}{5}}$ – slower.

Misspecified case: performance guarantee

- Difficulty of the task:
  - f_ρ is 's-smooth' ($s > 0$).
- Contribution:
  - If $L^2_{\rho_X}$ is separable and $\frac{1}{\lambda^2} \le l$,
  - then with high probability

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^\lambda, f_\rho) \le \underbrace{\frac{\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}}\lambda^{\frac{3}{2}}} + \frac{1}{\sqrt{l\lambda}} + \frac{\sqrt{\lambda^{\min(1,s)}}}{\lambda\sqrt{l}} + \lambda^{\min(1,s)}}_{g(\ell,N,\lambda)}.$$

Misspecified case: performance guarantee

- Difficulty of the task:
  - f_ρ is 's-smooth' ($s > 0$).
- <u>Contribution</u>: If
  - $L^2_{\rho_X}$ is separable and $\frac{1}{\lambda^2} \leq l$,
  - then with high probability

$$\mathcal{E}(f^\lambda_{\hat{\mathbf{z}}}, f_\rho) \leq \underbrace{\boxed{\frac{\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}} \lambda^{\frac{3}{2}}}} + \frac{1}{\sqrt{l\lambda}} + \boxed{\frac{\sqrt{\lambda^{\min(1,s)}}}{\lambda\sqrt{l}}} + \lambda^{\min(1,s)}}_{g(\ell, N, \lambda)}. \qquad (5)$$

$\hat{x}_i$ s-smoothness

Misspecified case: example

Assume

- $s \geq 1$, $h = 1$ (K: Lipschitz),
- $\boxed{1} = \boxed{3}$ in (5) $\Rightarrow \lambda$; $\ell = N^a$ ($a > 0$)
- $t = \ell N$: total number of samples processed.

Then

1. $s = 1$ ('non-smooth' f_ρ): $\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho) \approx t^{-0.25}$ – slower,
2. $s \to \infty$ ('smooth' f_ρ): $\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho) \approx t^{-0.5}$ – faster convergence.

Notes on the assumptions: $\exists\rho$, $X \in \mathcal{B}(H)$

- k: bounded, continuous $\Rightarrow$
  - $\mu : (\mathcal{P}(\mathcal{D}), \mathcal{B}(\tau_w)) \to (H, \mathcal{B}(H))$ measurable.
  - μ measurable, $X \in \mathcal{B}(H) \Rightarrow \rho$ on $X \times Y$: well-defined.
- If (*) := $\mathcal{D}$ is compact metric, k is universal, then
  - μ is continuous, and
  - $X \in \mathcal{B}(H)$.

Notes on the assumptions: Hölder K examples

In case of (*):

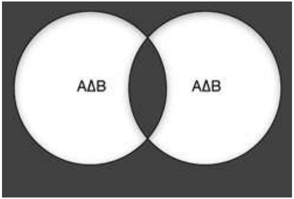
K_G	K_e	K_C
$e^{-\frac{\Vert\mu_a-\mu_b\Vert_H^2}{2\theta^2}}$ $h=1$	$e^{-\frac{\Vert\mu_a-\mu_b\Vert_H}{2\theta^2}}$ $h=\frac{1}{2}$	$\left(1+\Vert\mu_a-\mu_b\Vert_H^2/\theta^2\right)^{-1}$ $h=1$

K_t	K_i
$\left(1+\Vert\mu_a-\mu_b\Vert_H^\theta\right)^{-1}$ $h=\frac{\theta}{2}$ $(\theta\le 2)$	$\left(\Vert\mu_a-\mu_b\Vert_H^2+\theta^2\right)^{-\frac{1}{2}}$ $h=1$

Functions of $\Vert\mu_a-\mu_b\Vert_H \Rightarrow$ computation: similar to set kernel.

Notes on the assumptions: misspecified case

$L^2_{\rho_X}$: separable $\Leftrightarrow$ measure space with $d(A,B) = \rho_X(A \triangle B)$ is so [Thomson et al., 2008].



Vector-valued output: Y = separable Hilbert space

- Objective function:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \left\| f(\mu_{\hat{x}_i}) - y_i \right\|_Y^2 + \lambda \left\| f \right\|_{\mathcal{H}}^2, \quad (\lambda > 0).$$

- $K(\mu_a, \mu_b) \in \mathcal{L}(Y)$:
  - operator-valued kernel,
  - vector-valued RKHS.

Vector-valued output: analytical solution

Prediction on a new test distribution (t):

$$\boxed{\begin{aligned}(f_{\hat{\mathbf{z}}}^{\lambda} \circ \mu)(t) &= \mathbf{k}(\mathbf{K} + l\lambda \mathbf{I}_l)^{-1}[y_1; \ldots ; y_l], & (6)\\ \mathbf{K} &= [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathcal{L}(Y)^{l \times l}, & (7)\\ \mathbf{k} &= [K(\mu_{\hat{x}_1}, \mu_t), \ldots , K(\mu_{\hat{x}_l}, \mu_t)] \in \mathcal{L}(Y)^{1 \times l}. & (8)\end{aligned}}$$

Specifically: $Y = \mathbb{R} \Rightarrow \mathcal{L}(Y) = \mathbb{R}$; $Y = \mathbb{R}^d \Rightarrow \mathcal{L}(Y) = \mathbb{R}^d$.

Vector-valued output: K assumptions

Boundedness and Hölder continuity of K:

1. Boundedness:

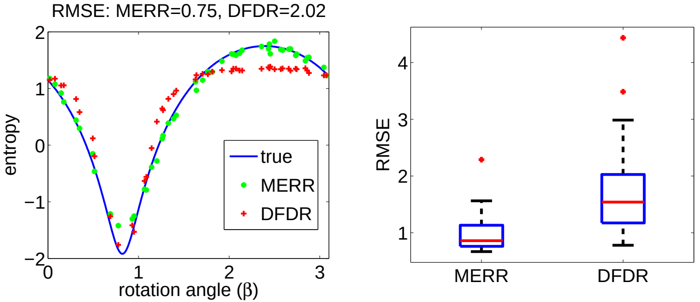
$$\|K_{\mu_a}\|_{\mathrm{HS}}^2 = Tr\left(K_{\mu_a}^* K_{\mu_a}\right) \leq B_K \in (0,\infty), \quad (\forall \mu_a \in X).$$

2. Hölder continuity: $\exists L > 0$, $h \in (0,1]$ such that

$$\|K_{\mu_a} - K_{\mu_b}\|_{\mathcal{L}(Y,\mathcal{H})} \leq L\,\|\mu_a - \mu_b\|_H^h, \quad \forall (\mu_a,\mu_b) \in X \times X.$$

Demo

- Supervised entropy learning:



- Aerosol prediction from satellite images:
  - State-of-the-art baseline: **7.5** – **8.5** ($\pm 0.1 - 0.6$).
  - MERR: **7.81** (± 1.64).

Summary

- Problem: distribution regression.
- Literature: large number of heuristics.
- Contribution:
    - a simple ridge solution is consistent,
    - specifically, the set kernel is so (15-year-old open question).
- Simplified version [$Y = \mathbb{R}$, $f_\rho \in \mathcal{H}$]:
    - AISTATS-2015 (oral).

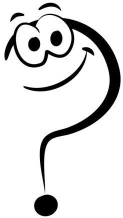
Summary – continued

- Code in ITE, extended analysis (submitted to JMLR):

```
https://bitbucket.org/szzoli/ite/
http://arxiv.org/abs/1411.2066.
```

- Closely related research directions (Bayesian world):
    - ∞-dimensional exp. family fitting,
    - just-in-time kernel EP: accepted at UAI-2015.

Thank you for the attention!



Acknowledgments: This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350. The work was carried out while Bharath K. Sriperumbudur was a research fellow in the Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge, UK.

Appendix: contents

- Topological definitions, separability.
- Prior definitions (ρ).
- Universal kernel: definition, examples.
- Vector-valued RKHS.
- Demos: further details.
- Hausdorff metric.
- Weak topology on $\mathcal{P}(\mathcal{D})$.

Topological space, open sets

- Given: $\mathcal{D} \neq \emptyset$ set.
- $\tau \subseteq 2^{\mathcal{D}}$ is called a *topology* on $\mathcal{D}$ if:
  1. $\emptyset \in \tau$, $\mathcal{D} \in \tau$.
  2. Finite intersection: $O_1 \in \tau$, $O_2 \in \tau \Rightarrow O_1 \cap O_2 \in \tau$.
  3. Arbitrary union: $O_i \in \tau$ $(i \in I) \Rightarrow \cup_{i \in I} O_i \in \tau$.

Then, $(\mathcal{D}, \tau)$ is called a *topological space*; $O \in \tau$: *open* sets.

Closed-, compact set, closure, dense subset, separability

Given: $(\mathcal{D}, \tau)$. $A \subseteq \mathcal{D}$ is

- *closed* if $\mathcal{D} \backslash A \in \tau$ (i.e., its complement is open),
- *compact* if for any family $(O_i)_{i \in I}$ of open sets with $A \subseteq \cup_{i \in I} O_i$, $\exists i_1, \ldots, i_n \in I$ with $A \subseteq \cup_{j=1}^{n} O_{i_j}$.

Closure of $A \subseteq \mathcal{D}$:

$$\bar{A} := \bigcap_{A \subseteq C \text{ closed in } \mathcal{D}} C. \tag{9}$$

- $A \subseteq \mathcal{D}$ is *dense* if $\bar{A} = \mathcal{D}$.
- $(\mathcal{D}, \tau)$ is *separable* if $\exists$ countable, dense subset of $\mathcal{D}$. Counterexample: ℓ^∞ / L^∞.

Prior (well-specified case): $\rho \in \mathcal{P}(b, c)$

- Let the $T : \mathcal{H} \to \mathcal{H}$ covariance operator be

$$T = \int_X K(\cdot, \mu_a) K^*(\cdot, \mu_a) \mathrm{d}\rho_X(\mu_a)$$

  with eigenvalues t_n $(n = 1, 2, \ldots)$.
- Assumption: $\rho \in \mathcal{P}(b, c)$ = set of distributions on $X \times Y$
  - $\alpha \leq n^b t_n \leq \beta \quad (\forall n \geq 1; \alpha > 0, \beta > 0)$,
  - $\exists g \in \mathcal{H}$ such that $f_\rho = T^{\frac{c-1}{2}} g$ with $\|g\|_{\mathcal{H}}^2 \leq R$ $(R > 0)$,

  where $b \in (1, \infty)$, $c \in [1, 2]$.
- Intuition: $1/b$ – effective input dimension, c – smoothness of f_ρ.

Prior: misspecified case

Let $\tilde{T}$ be defined as:

$$S_K^* : \mathcal{H} \hookrightarrow L^2_{\rho_X},$$

$$S_K : L^2_{\rho_X} \to \mathcal{H}, \quad (S_K g)(\mu_u) = \int_X K(\mu_u, \mu_t) g(\mu_t) \mathrm{d}\rho_X(\mu_t),$$

$$\tilde{T} = S_K^* S_K : L^2_{\rho_X} \to L^2_{\rho_X}.$$

Our range space assumption on ρ: $f_\rho \in Im\left(\tilde{T}^s\right)$ for some $s \geq 0$.

Universal kernel: definition

Assume

- $\mathcal{D}$: compact, metric,
- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ kernel is continuous.

Then

- Def-1: k is universal if $H(k)$ is dense in $(C(\mathcal{D}), \|\cdot\|_\infty)$.

Universal kernel: definition

Assume

- $\mathcal{D}$: compact, metric,
- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ kernel is continuous.

Then

- Def-1: k is universal if $H(k)$ is dense in $(C(\mathcal{D}), \|\cdot\|_\infty)$.
- Def-2: k is
  - characteristic, if $\mu : \mathcal{P}(\mathcal{D}) \to H(k)$ is injective.

Universal kernel: definition

Assume

- $\mathcal{D}$: compact, metric,
- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ kernel is continuous.

Then

- Def-1: k is universal if $H(k)$ is dense in $(C(\mathcal{D}), \|\cdot\|_\infty)$.
- Def-2: k is
  - characteristic, if $\mu : \mathcal{P}(\mathcal{D}) \to H(k)$ is injective.
  - universal, if μ is injective on the finite signed measures of $\mathcal{D}$.

Universal kernel: examples

On compact subsets of $\mathbb{R}^d$

$$k(a,b) = e^{-\frac{\|a-b\|_2^2}{2\sigma^2}}, \quad (\sigma > 0)$$

$$k(a,b) = e^{-\sigma\|a-b\|_1}, \quad (\sigma > 0)$$

$$k(a,b) = e^{\beta\langle a,b\rangle}, (\beta > 0), \text{ or more generally}$$

$$k(a,b) = f(\langle a,b\rangle), \quad f(x) = \sum_{n=0}^{\infty} a_n x^n \quad (\forall a_n > 0).$$

Vector-valued RKHS: $\mathcal{H} = \mathcal{H}(K)$

Definition:

- A $\mathcal{H} \subseteq Y^X$ Hilbert space of functions is RKHS if

$$A_{\mu_x,y} : f \in \mathcal{H} \mapsto \langle y, f(\mu_x) \rangle_Y \in \mathbb{R} \tag{10}$$

  is *continuous* for $\forall \mu_x \in X, y \in Y$.
- $=$ The evaluation functional is continuous in every direction.

Vector-valued RKHS: $\mathcal{H} = \mathcal{H}(K)$ – continued

- Riesz representation theorem $\Rightarrow \exists K(\mu_x|y) \in \mathcal{H}$:

$$\langle y, f(\mu_x)\rangle_Y = \langle K(\mu_x|y), f\rangle_{\mathcal{H}} \quad (\forall f \in \mathcal{H}). \tag{11}$$

- $K(\mu_x|y)$: linear, bounded in $y \Rightarrow K(\mu_x|y) = K_{\mu_x}(y)$ with $K_{\mu_x} \in \mathcal{L}(Y, \mathcal{H})$.

Vector-valued RKHS: $\mathcal{H} = \mathcal{H}(K)$ – continued

- Riesz representation theorem $\Rightarrow \exists K(\mu_x|y) \in \mathcal{H}$:

$$\langle y, f(\mu_x)\rangle_Y = \langle K(\mu_x|y), f\rangle_{\mathcal{H}} \quad (\forall f \in \mathcal{H}). \tag{11}$$

- $K(\mu_x|y)$: linear, bounded in $y \Rightarrow K(\mu_x|y) = K_{\mu_x}(y)$ with $K_{\mu_x} \in \mathcal{L}(Y, \mathcal{H})$.
- K construction:

$$K(\mu_x, \mu_t)(y) = (K_{\mu_t} y)(\mu_x), \quad (\forall \mu_x, \mu_t \in X), \text{ i.e.,}$$

$$K(\cdot, \mu_t)(y) = K_{\mu_t} y, \tag{12}$$

$$\mathcal{H}(K) = \overline{span}\{K_{\mu_t} y : \mu_t \in X, y \in Y\}. \tag{13}$$

Vector-valued RKHS: $\mathcal{H} = \mathcal{H}(K)$ – continued

- Riesz representation theorem $\Rightarrow \exists K(\mu_x|y) \in \mathcal{H}$:

$$\langle y, f(\mu_x)\rangle_Y = \langle K(\mu_x|y), f\rangle_{\mathcal{H}} \quad (\forall f \in \mathcal{H}). \tag{11}$$

- $K(\mu_x|y)$: linear, bounded in $y \Rightarrow K(\mu_x|y) = K_{\mu_x}(y)$ with $K_{\mu_x} \in \mathcal{L}(Y, \mathcal{H})$.
- K construction:

$$K(\mu_x, \mu_t)(y) = (K_{\mu_t} y)(\mu_x), \quad (\forall \mu_x, \mu_t \in X), \text{ i.e.,}$$
$$K(\cdot, \mu_t)(y) = K_{\mu_t} y, \tag{12}$$
$$\mathcal{H}(K) = \overline{span}\{K_{\mu_t} y : \mu_t \in X, y \in Y\}. \tag{13}$$

- Shortly: $K(\mu_x, \mu_t) \in \mathcal{L}(Y)$ generalizes $k(u, v) \in \mathbb{R}$.

Vector-valued RKHS – examples: $Y = \mathbb{R}^d$

1. $K_i : X \times X \to \mathbb{R}$ kernels ($i = 1, \ldots, d$). Diagonal kernel:

$$K(\mu_a, \mu_b) = diag(K_1(\mu_a, \mu_b), \ldots, K_d(\mu_a, \mu_b)). \tag{14}$$

2. Combination of D_j diagonal kernels [$D_j(\mu_a, \mu_b) \in \mathbb{R}^{r\times r}$, $A_j \in \mathbb{R}^{r\times d}$]:

$$K(\mu_a, \mu_b) = \sum_{j=1}^{m} A_j^* D_j(\mu_a, \mu_b) A_j. \tag{15}$$

Demo-1: supervised entropy learning

- Problem: learn the entropy of the 1^{st} coo. of (rotated) Gaussians.
- Baseline: kernel smoothing based distribution regression (applying density estimation) $=:$ DFDR.
- Performance: RMSE boxplot over 25 random experiments.
- Experience:
  - more precise than the only theoretically justified method,
  - by avoiding density estimation.

Demo-2: aerosol prediction – selected kernels

Kernel definitions ($p = 2, 3$):

$$k_G(a,b) = e^{-\frac{\|a-b\|_2^2}{2\theta^2}}, \qquad k_e(a,b) = e^{-\frac{\|a-b\|_2}{2\theta^2}}, \tag{16}$$

$$k_C(a,b) = \frac{1}{1+\frac{\|a-b\|_2^2}{\theta^2}}, \quad k_t(a,b) = \frac{1}{1+\|a-b\|_2^{\theta}}, \tag{17}$$

$$k_p(a,b) = \left(\langle a,b\rangle + \theta\right)^p, \; k_r(a,b) = 1 - \frac{\|a-b\|_2^2}{\|a-b\|_2^2+\theta}, \tag{18}$$

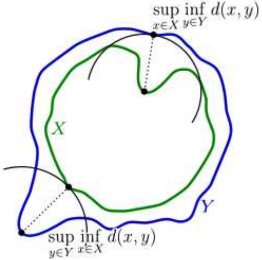
$$k_i(a,b) = \frac{1}{\sqrt{\|a-b\|_2^2+\theta^2}}, \tag{19}$$

$$k_{M,\frac{3}{2}}(a,b) = \left(1+\frac{\sqrt{3}\,\|a-b\|_2}{\theta}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\theta}}, \tag{20}$$

$$k_{M,\frac{5}{2}}(a,b) = \left(1+\frac{\sqrt{5}\,\|a-b\|_2}{\theta}+\frac{5\,\|a-b\|_2^2}{3\theta^2}\right) e^{-\frac{\sqrt{5}\|a-b\|_2}{\theta}}. \tag{21}$$

- Hausdorff metric [Edgar, 1995]:

$$d_H(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\}. \quad (22)$$



  - Metric on compact sets of metric spaces $[(M,d); X,Y\subseteq M]$.
  - 'Slight' problem: highly sensitive to outliers.

Weak topology on $\mathcal{P}(\mathcal{D})$

Def.: It is the weakest topology such that the

$$L_h : (\mathcal{P}(\mathcal{D}), \tau_w) \to \mathbb{R},$$
$$L_h(x) = \int_{\mathcal{D}} h(u) \mathrm{d}x(u)$$

mapping is continuous for all $h \in C_b(\mathcal{D})$, where

$$C_b(\mathcal{D}) = \{(\mathcal{D}, \tau) \to \mathbb{R} \text{ bounded, continuous functions}\}.$$

Chen, Y. and Wu, O. (2012).
Contextual Hausdorff dissimilarity for multi-instance clustering.
In *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 870–873.

Cuturi, M., Fukumizu, K., and Vert, J.-P. (2005).
Semigroup kernels on measures.
Journal of Machine Learning Research, 6:11691198.

Edgar, G. (1995).
Measure, Topology and Fractal Geometry.
Springer-Verlag.

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.
In *International Conference on Machine Learning (ICML)*, pages 179–186.

Haussler, D. (1999).
Convolution kernels on discrete structures.
Technical report, Department of Computer Science, University of California at Santa Cruz.
(http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf).

Hein, M. and Bousquet, O. (2005).
Hilbertian metrics and positive definite kernels on probability measures.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136–143.

Jebara, T., Kondor, R., and Howard, A. (2004).
Probability product kernels.
Journal of Machine Learning Research, 5:819–844.

Martins, A. F. T., Smith, N. A., Xing, E. P., Aguiar, P. M. Q., and Figueiredo, M. A. T. (2009).
Nonextensive information theoretical kernels on measures.

Journal of Machine Learning Research, 10:935–975.

Nielsen, F. and Nock, R. (2012).
A closed-form expression for the Sharma-Mittal entropy of exponential families.
Journal of Physics A: Mathematical and Theoretical, 45:032003.

Oliva, J., Póczos, B., and Schneider, J. (2013).
Distribution to distribution regression.
International Conference on Machine Learning (ICML; JMLR W&CP), 28:1049–1057.

Oliva, J. B., Neiswanger, W., Póczos, B., Schneider, J., and Xing, E. (2014).
Fast distribution to real regression.
International Conference on Artificial Intelligence and Statistics (AISTATS; JMLR W&CP), 33:706–714.

Póczos, B., Rinaldo, A., Singh, A., and Wasserman, L. (2013).
Distribution-free distribution regression.

International Conference on Artificial Intelligence and Statistics (AISTATS; JMLR W&CP), 31:507–515.

Póczos, B., Xiong, L., and Schneider, J. (2011).
Nonparametric divergence estimation with applications to machine learning on distributions.
In *Uncertainty in Artificial Intelligence (UAI)*, pages 599–608.

Reddi, S. J. and Póczos, B. (2014).
k-NN regression on functional data with incomplete observations.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Thomson, B. S., Bruckner, J. B., and Bruckner, A. M. (2008).
Real Analysis.
Prentice-Hall.

Wang, F., Syeda-Mahmood, T., Vemuri, B. C., Beymer, D., and Rangarajan, A. (2009).
Closed-form Jensen-Rényi divergence for mixture of Gaussians and applications to group-wise shape registration.

Medical Image Computing and Computer-Assisted Intervention, 12:648–655.

Wang, J. and Zucker, J.-D. (2000).
Solving the multiple-instance problem: A lazy learning approach.
In *International Conference on Machine Learning (ICML)*, pages 1119–1126.

Wang, Z., Lan, L., and Vucetic, S. (2012).
Mixture model for multiple instance regression and applications in remote sensing.
IEEE Transactions on Geoscience and Remote Sensing, 50:2226–2237.

Wu, O., Gao, J., Hu, W., Li, B., and Zhu, M. (2010).
Identifying multi-instance outliers.
In *SIAM International Conference on Data Mining (SDM)*, pages 430–441.

Zhang, M.-L. and Zhou, Z.-H. (2009).

Multi-instance clustering with applications to multi-instance prediction.
Applied Intelligence, 31:47–68.

Zhou, S. K. and Chellappa, R. (2006).
From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 28:917–929.