Vector-valued Distribution Regression: A Simple and Consistent Approach

Zoltán Szabó

Joint work with Arthur Gretton (UCL), Barnabás Póczos (CMU), Bharath K. Sriperumbudur (PSU)

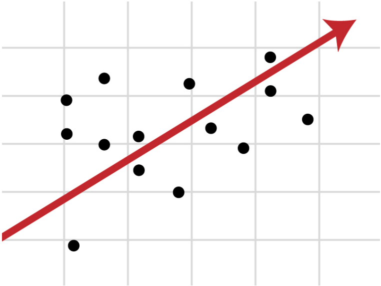
Statistical Science Seminars
October 9, 2014

Outline

- Motivation.
- Previous work.
- High-level goal.
- Definitions, algorithm, error guarantee, consistency.
- Numerical illustration.

Problem: regression on distributions

- Given: $\{(x_i, y_i)\}_{i=1}^{l}$ samples $\mathcal{H} \ni f = ?$ such that $f(x_i) \approx y_i$.



- Our interest:
  - x_i-s are distributions, but (challenge!),
  - only samples are given from x_i-s: $\{x_{i,n}\}_{n=1}^{N_i}$.

Two-stage sampled setting = bag-of-features

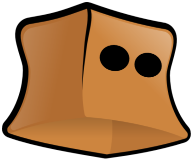
Examples:

- image = set of patches/visual descriptors,
- document = bag of words/sentences/paragraphs,
- molecule = different configurations/shapes,
- group of people on a social network: bag of friendship graphs,
- customer = his/her shopping records,
- user = set of trial time-series.

Distribution regression: wider context

Several problems are covered in machine learning and statistics:

- multi-instance learning,
- point estimation tasks without analytical formula.



Existing methods

Idea:

1. estimate distribution similarities,
2. plug them into a learning algorithm.

Approaches:

1. parametric approaches: Gaussian, MOG, exponential 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].

Existing methods+

1. (Positive definite) kernels: [Cuturi et al., 2005, Martins et al., 2009, Hein and Bousquet, 2005].
2. Divergence measures (KL, . . .): [Póczos et al., 2011].
3. Set metric based algorithms:
   1. Hausdorff metric [Edgar, 1995], and
   2. its variants [Wang and Zucker, 2000, Wu et al., 2010, Zhang and Zhou, 2009, Chen and Wu, 2012].

Existing methods: summary

- MIL dates back to [Haussler, 1999, Gärtner et al., 2002].
- There are several multi-instance methods, applications.

Existing methods: summary

- MIL dates back to [Haussler, 1999, Gärtner et al., 2002].
- There are several multi-instance methods, applications.

- One 'small' open question:

Does any of these techniques make sense?

Existing methods: "exceptions"

- APR (axis-parallel rectangles) and its variants, classification [Auer, 1998, Long and Tan, 1998, Blum and Kalai, 1998, Babenko et al., 2011, Zhang et al., 2013, Sabato and Tishby, 2012]:

$$y_i = \max(\mathbb{I}_R(x_{i,1}), \ldots, \mathbb{I}_R(x_{i,N})) \in \{0, 1\},$$

where R = unknown rectangle.

Existing methods: “exceptions”

- APR (axis-parallel rectangles) and its variants, classification [Auer, 1998, Long and Tan, 1998, Blum and Kalai, 1998, Babenko et al., 2011, Zhang et al., 2013, Sabato and Tishby, 2012]:

$$y_i = \max(\mathbb{I}_R(x_{i,1}), \ldots, \mathbb{I}_R(x_{i,N})) \in \{0, 1\},$$

  where R = unknown rectangle.
- Density based approaches, regression: KDE + kernel smoothing [Póczos et al., 2013, Oliva et al., 2014],
  - densities live on compact Euclidean domain,
  - density estimation: nuisance step.

High-level goal: set kernel

- Given (2 bags):

$$B_i := \{x_{i,n}\}_{n=1}^{N_i} \sim x_i,$$
$$B_j := \{x_{j,m}\}_{m=1}^{N_j} \sim x_j.$$

- Similarity of the bags (set/multi-instance/ensemble-, convolution kernel [Haussler, 1999, Gärtner et al., 2002]):

$$K(B_i, B_j) = \frac{1}{N_i N_j} \sum_{n=1}^{N_i} \sum_{m=1}^{N_j} k(x_{i,n}, x_{j,m}).$$

High-level goal: consistency of set kernels

- Are set kernels *consistent*, when plugged into some regression scheme?
- Our focus:

  ridge regression.

- Motivation (ridge scheme):
  1. simple algorithm.
  2. recently proved parallelizations [Zhang et al., 2014].

Story

- $\mathcal{H}$: assumed function class to capture the (x, y) relation.
- f_ρ: true regression function (might not be in $\mathcal{H}$).
- $f_{\mathcal{H}}$: "best" function from $\mathcal{H}$ ($l = \infty$, $N := N_i = \infty$).
- $\hat{f}$: estimated function from $\mathcal{H}$ based on $\{(\{x_{i,n}\}_{n=1}^N, y_i)\}_{i=1}^l$.
- Aim:
  - High probability error guarantees (λ: reg., $\mathcal{E}$: risk):

$$\mathcal{E}[\hat{f}] - \mathcal{E}[f_{\mathcal{H}}] \leq r_1(l, N, \lambda), \tag{1}$$

$$\|\hat{f} - f_\rho\|_{L^2} \leq r_2(l, N, \lambda) + r_3(\text{richness of } \mathcal{H}). \tag{2}$$

  - Consistency: $(l, N, \lambda) =?$ such that $r_i(l, N, \lambda) \to 0$ $(i = 1, 2)$.

Distribution regression: definition, solution idea

- $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{l}$: $x_i \in \mathcal{M}_1^+(\mathcal{D})$, $y_i \in Y$.
- $\hat{\mathbf{z}} = \left\{\left(\{x_{i,n}\}_{n=1}^{N}, y_i\right)\right\}_{i=1}^{l}$: $x_{i,1}, \ldots, x_{i,N} \overset{i.i.d.}{\sim} x_i$.
- Goal: learn the relation between x and y based on $\hat{\mathbf{z}}$.
- Idea:
  1. embed the distributions (using μ defined by k),
  2. apply ridge regression (determined by K).

$$\mathcal{M}_1^+(\mathcal{D}) \xrightarrow{\mu} X \subseteq H(k) \xrightarrow{f \in \mathcal{H}(K)} Y.$$

Kernel part (k, K): 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,
  - $k(a, b) = e^{-\theta\|a-b\|_2}$: Laplacian.
- In the $H = H(k)$ RKHS $(\exists!)$: $\varphi(u) = k(\cdot, u)$.

Kernel part: example domains ($\mathcal{D}$)

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

- Distributions.

Embedding step: $\mathcal{M}_1^+(\mathcal{D}) \xrightarrow{\mu} X \subseteq H(k)$

- Given: kernel $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$.
- Mean embedding of a distribution $x \in \mathcal{M}_1^+(\mathcal{D})$:

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

- Mean embedding of the empirical distribution $\hat{x}_i = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_{i,n}} \in \mathcal{M}_1^+(\mathcal{D})$:

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

Objective function: $X \xrightarrow{f \in \mathcal{H}=\mathcal{H}(K)} Y$

- Optimal ($\mathcal{H}$/measurable) in expected risk ($\mathcal{E}$) sense:

$$\mathcal{E}\left[f_{\mathcal{H}}\right] = \inf_{f \in \mathcal{H}} \mathcal{E}[f] = \inf_{f \in \mathcal{H}} \int_{X \times Y} \|f(\mu_a) - y\|_Y^2 \, \mathrm{d}\rho(\mu_a, y),$$

$$f_\rho(\mu_a) = \mathbb{E}[y|\mu_a] = \int_Y y \mathrm{d}\rho(y|\mu_a) \quad (\mu_a \in X).$$

- One-stage ($\int \to \mathbf{z}$), two-stage difficulty ($\mathbf{z} \to \hat{\mathbf{z}}$):

$$f_{\mathbf{z}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \|f(\mu_{x_i}) - y_i\|_Y^2 + \lambda \|f\|_{\mathcal{H}}^2, \tag{3}$$

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \|f(\mu_{\hat{x}_i}) - y_i\|_Y^2 + \lambda \|f\|_{\mathcal{H}}^2. \tag{4}$$

Algorithmically: ridge regression ⇒ analytical solution

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

$$(f_{\hat{\mathbf{z}}}^{\lambda} \circ \mu)(t) = [y_1, \dots, y_l](\mathbf{K} + l\lambda \mathbf{I}_l)^{-1}\mathbf{k}, \tag{5}$$

$$\mathbf{K} = [K_{ij}] = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathcal{L}(Y)^{l \times l}, \tag{6}$$

$$\mathbf{k} = \begin{bmatrix} K(\mu_{\hat{x}_1}, \mu_t) \\ \vdots \\ K(\mu_{\hat{x}_l}, \mu_t) \end{bmatrix} \in \mathcal{L}(Y)^l. \tag{7}$$

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

Assumption-1

- $\mathcal{D}$: separable, topological.
- Y: separable Hilbert.
- k:
  - bounded: $\sup_{u\in\mathcal{D}} k(u,u) \le B_k \in (0,\infty)$,
  - continuous.
- $X = \mu\left(\mathcal{M}_1^+(\mathcal{D})\right) \in \mathcal{B}(H)$.

Assumption-1 – continued

- K $[K_{\mu_a} := K(\cdot, \mu_a)]$:
  1. bounded:

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

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

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

- y is bounded: $\exists C < \infty$ such that $\|y\|_Y \leq C$ almost surely.

Assumption-1: remarks (before the ρ assumptions)

- k: bounded, continuous $\Rightarrow$
  - $\mu : (\mathcal{M}_1^+(\mathcal{D}), \mathcal{B}(\tau_w)) \to (H, \mathcal{B}(H))$ 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 $X \in \mathcal{B}(H)$.
- If $Y = \mathbb{R}$, we get the traditional boundedness of K:

$$K(\mu_a, \mu_a) \leq B_K, \quad (\forall \mu_a \in X).$$

Assumption-1: linear $K \leftrightarrow$ set kernel

- Let $Y = \mathbb{R}$ and $K(\mu_a, \mu_b) = \langle \mu_a, \mu_b \rangle_H$. Recall

$$\mu_{\hat{x}_i} = \frac{1}{N} \sum_{n=1}^{N} k(\cdot, x_{i,n}).$$

- In this case: $B_K = B_k$, $L = 1$, $h = 1$, we get the set kernel

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

Assumption-1: nonlinear K examples for $Y = \mathbb{R}$

In case of (*) and $Y = \mathbb{R}$: Hölder K-s ($\mathcal{D}$: compact, metric; μ: continuous)

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

K_t	K_i
$\left(1+\|\mu_a-\mu_b\|_H^\theta\right)^{-1}$	$\left(\|\mu_a-\mu_b\|_H^2+\theta^2\right)^{-\frac{1}{2}}$
$h = \frac{\theta}{2}$ $(\theta \leq 2)$	$h = 1$

Assumption-1 – continued: $\rho \in \mathcal{P}(b, c)$

- 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 $f_{\mathcal{H}} = T^{\frac{c-1}{2}} g$ with $\|g\|_{\mathcal{H}}^2 \leq R$ $(R > 0)$,

  where $b \in (1, \infty)$, $c \in [1, 2]$.
- Intuition: b – effective input dimension, c – smoothness of $f_{\mathcal{H}}$.

Assumption-2: Assumption-1, but with alternative ρ

Let $\tilde{T}$ be the extension of T from $\mathcal{H}$ to $L^2_{\rho_X}$:

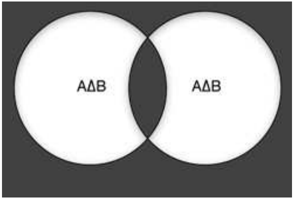
$$S_K^* : \mathcal{H} \hookrightarrow L^2_{\rho_X},$$
$$S_K : L^2_{\rho_X} \to \mathcal{H}, \quad (S_K g)(\mu_u) = \int_X K(\mu_u, \mu_t) g(\mu_t) \mathrm{d}\rho_X(\mu_t),$$
$$\tilde{T} = S_K^* S_K : L^2_{\rho_X} \to L^2_{\rho_X}.$$

Our assumptions on ρ:

- Range space assumption: $f_\rho \in Im\left(\tilde{T}^s\right)$ for some $s \geq 0$.
- $L^2_{\rho_X}$: separable.

Assumption-2: remarks

- Range space assumption:
  - $f_\rho \in Im\left(\tilde{T}^s\right)$: s captures the smoothness of f_ρ.
- $L^2_{\rho_X}$: separable $\Leftrightarrow$ measure space with $d(A,B) = \rho_X(A \triangle B)$ is so [Thomson et al., 2008].



Error guarantees, consistency (in human-readable format)

In case of

- Assumption-1: if $l \geq \lambda^{-\frac{1}{b}-1}$

$$\mathcal{E}[f_{\hat{\mathbf{z}}}^{\lambda}] - \mathcal{E}[f_{\mathcal{H}}] \leq \frac{\log^{h}(l)}{N^{h}\lambda^{3}} + \lambda^{c} + \frac{1}{l^{2}\lambda} + \frac{1}{l\lambda^{\frac{1}{b}}} \to 0$$

- Assumption-2: if $\frac{1}{\lambda^2} \leq l$

$$\left\| S_K^{*} f_{\hat{\mathbf{z}}}^{\lambda} - f_{\rho} \right\|_{L^2_{\rho_X}} \leq \frac{\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}}\lambda^{\frac{3}{2}}} + \frac{1}{\lambda\sqrt{l}} + D_{\mathcal{H}} \to D_{\mathcal{H}},$$

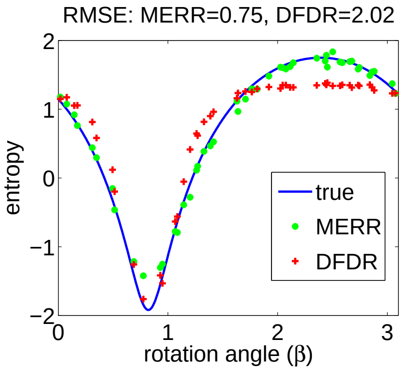
$$D_{\mathcal{H}} = \inf_{q\in\mathcal{H}} \left\| f_{\rho} - S_K^{*} q \right\|_{L^2_{\rho_X}}$$

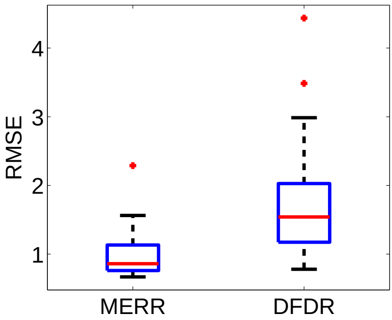
with high probability.

Demo-1 ($Y = \mathbb{R}$): Supervised entropy learning

- Problem: learn the entropy 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.

Supervised entropy learning: plots





- Bag:= multispectral satellite image over an area.
- Label of a bag:= AOD value of a highly accurate ground-based instrument.
- Performance: RMSE.
- Experience:
  - $\approx$ domain-specific, engineered methods,
  - beating state-of-the-art MI techniques.

Aerosol prediction: results

- Baseline [mixture model (EM)]: $7.5 - 8.5$ ($\pm 0.1 - 0.6$).
- Linear K:
  - single: 7.91 ($\pm$1.61).
  - ensemble: **7.86** ($\pm$**1.71**).
- Nonlinear K:
  - Single: 7.90 ($\pm$1.63),
  - Ensemble: **7.81** ($\pm$**1.64**).

Summary

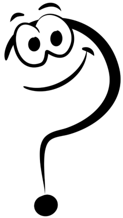
- Problem: two-stage sampled distribution regression.
- Literature: large number of heuristics.
- Contribution:
  - error guarantees, consistency for the ridge based solution.
  - specially, set kernel is consistent in regression (15-year-old open question).
- Code $\in$ ITE toolbox:

  https://bitbucket.org/szzoli/ite/

Future research directions

- Theoretical:
  - quadratic loss ($\mathcal{E}$), bounded kernels (k, K), mean embedding (μ) with i.i.d. samples ($\{x_{i,n}\}_{n=1}^{N}$): relaxation,
  - equivalent characterizations/alternative priors (ρ),
  - lower/optimal bounds,
  - error guarantees for non-point estimates.
- Practical: large-scale solvers, $dim(Y) = \infty$.

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.
- Vector-valued RKHS.
- Hausdorff metric.
- Weak topology on $\mathcal{M}_1^+(\mathcal{D})$.
- Universal kernel examples.

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.

Topology: examples

- $\tau = \{\emptyset, \mathcal{D}\}$: indiscrete topology.
- $\tau = 2^{\mathcal{D}}$: discrete topology.
- $(\mathcal{D}, d)$ metric space:
  - Open ball: $B_\epsilon(x) = \{y \in \mathcal{D} : d(x, y) < \epsilon\}$.
  - $O \subseteq \mathcal{D}$ is open if for $\forall x \in O$ $\exists \epsilon > 0$ such that $B_\epsilon(x) \subseteq O$.
  - $\tau := \{O \subseteq \mathcal{D} : O \text{ is an open subset of } \mathcal{D}\}$.

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 $A \subseteq \mathcal{D}$:

$$\bar{A} := \bigcap_{A \subseteq C \text{ closed in } \mathcal{D}} C. \tag{8}$$

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

The discrete space

- $(\mathcal{D}, 2^{\mathcal{D}})$: complete metric space.
- Discrete metric (inducing the discrete topology):

$$d(x, y) = \begin{Bmatrix} 0, \text{ if } x = y \\ 1, \text{ if } x \neq y \end{Bmatrix}. \tag{9}$$

- Discrete space: separable $\Leftrightarrow$ $|\mathcal{D}|$ is countable.

Definition:

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

$$A_{\mu_x,y} : f \mapsto \langle y, f(\mu_x)\rangle_Y \tag{10}$$

is *continuous* for $\forall \mu_x \in X, y \in Y$.

- $=$ The evaluation functional is continuous in every direction.

Riesz representation theorem $\Rightarrow$

- $\exists K_{\mu_t} \in \mathcal{L}(Y, \mathcal{H})$:

$$K(\mu_x, \mu_t)(y) = (K_{\mu_t} y)(\mu_x), \quad (\forall \mu_x, \mu_t \in X), \text{ or shortly}$$

$$K(\cdot, \mu_t)(y) = K_{\mu_t} y, \tag{11}$$

$$\mathcal{H}(K) = \overline{span}\{K_{\mu_t} y : \mu_t \in X, y \in Y\}. \tag{12}$$

Vector-valued RKHS – continued

Examples ($Y = \mathbb{R}^d$):

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

$$K(\mu_a, \mu_b) = diag(K_1(\mu_a, \mu_b), \ldots, K_d(\mu_a, \mu_b)). \tag{13}$$

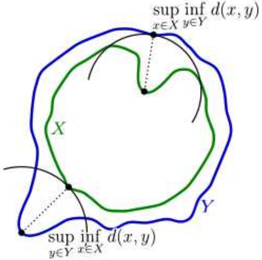
2. Combination of D_j diagonal kernels [$D_j(\mu_a, \mu_b) \in \mathbb{R}^{r \times r}$, $A_j \in \mathbb{R}^{r \times d}$]:

$$K(\mu_a, \mu_b) = \sum_{j=1}^{m} A_j^* D_j(\mu_a, \mu_b) A_j. \tag{14}$$

Existing methods: set metric based algorithms

- Hausdorff metric [Edgar, 1995]:

$$d_H(X,Y) = \max\left\{ \sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y) \right\}. \quad (15)$$



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

Weak topology on $\mathcal{M}_1^+(\mathcal{D})$

Def.: It is the weakest topology such that the

$$L_h : (\mathcal{M}_1^+(\mathcal{D}), \tau_w) \to \mathbb{R},$$
$$L_h(x) = \int_{\mathcal{D}} h(u) \mathrm{d}x(u)$$

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

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

Universal kernel examples

On every compact subset of $\mathbb{R}^d$:

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

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

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

$$k(a,b) = (1 - \langle a,b\rangle)^{\alpha}, \quad (\alpha > 0).$$

Auer, P. (1998).
Approximating hyper-rectangles: Learning and pseudorandom sets.
Journal of Computer and System Sciences, 57:376–388.

Babenko, B., Verma, N., Dollár, P., and Belongie, S. (2011).
Multiple instance learning with manifold bags.
In *International Conference on Machine Learning (ICML)*, pages 81–88.

Blum, A. and Kalai, A. (1998).
A note on learning from multiple-instance examples.
Machine Learning, 30:23–29.

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/convoluti

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.

Long, P. M. and Tan, L. (1998).
PAC learning of axis-aligned rectangles with respect to product distributions from multiple-instance examples.
Machine Learning, 30:7–21.

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. 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.

Sabato, S. and Tishby, N. (2012).
Multi-instance learning with any hypothesis class.
Journal of Machine Learning Research, 13:2999–3039.

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.

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, D., He, J., Si, L., and Lawrence, R. D. (2013).
MILEAGE: Multiple Instance LEArning with Global Embedding.
International Conference on Machine Learning (ICML; JMLR W&CP), 28:82–90.

Zhang, M.-L. and Zhou, Z.-H. (2009).
Multi-instance clustering with applications to multi-instance prediction.
Applied Intelligence, 31:47–68.

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2014).

Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates.
Technical report, University of California, Berkeley. (http://arxiv.org/abs/1305.5029).

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.