Vector-valued Distribution Regression – Keep It Simple and Consistent
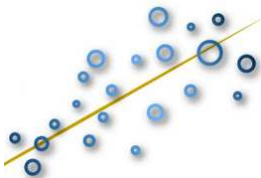
Zoltán Szabó

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

CSML Reading Group
Department of Statistics
University of Oxford
May 1, 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*.

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

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

- 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].

- 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 = $\mathbb{R}$.

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,
  - $k(a, b) = e^{-\theta\|a-b\|_2}$: 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.

Universal kernel

- Def.: $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ kernel is universal if
  - it is continuous,
  - $H(k)$ is dense in $(C(\mathcal{D}), \|\cdot\|_\infty)$.
- Examples: on compact subsets of $\mathbb{R}^d$

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

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

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

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

- Given:
    - labelled bags $\hat{\mathbf{z}} = \{(\hat{x}_i, y_i)\}_{i=1}^{l}$,
    - i^{th} bag: $\hat{x}_i = \{x_{i,1}, \ldots, x_{i,N}\} \overset{i.i.d.}{\sim} x_i \in \mathcal{M}_1^+(\mathcal{D})$, $y_i \in \mathbb{R}$.
- Task: find a $\mathcal{M}_1^+(\mathcal{D}) \to \mathbb{R}$ mapping based on $\hat{\mathbf{z}}$.

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

- Given:
    - labelled bags $\hat{\mathbf{z}} = \{(\hat{x}_i, y_i)\}_{i=1}^{l}$,
    - i^{th} bag: $\hat{x}_i = \{x_{i,1}, \ldots, x_{i,N}\} \overset{i.i.d.}{\sim} x_i \in \mathcal{M}_1^+ (\mathcal{D})$, $y_i \in \mathbb{R}$.
- Task: find a $\mathcal{M}_1^+ (\mathcal{D}) \to \mathbb{R}$ mapping based on $\hat{\mathbf{z}}$.
- Construction: distribution embedding (μ_x) + ridge regression

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

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

- Given:
  - labelled bags $\hat{\mathbf{z}} = \{(\hat{x}_i, y_i)\}_{i=1}^{l}$,
  - i^{th} bag: $\hat{x}_i = \{x_{i,1}, \ldots, x_{i,N}\} \overset{i.i.d.}{\sim} x_i \in \mathcal{M}_1^+(\mathcal{D})$, $y_i \in \mathbb{R}$.
- Task: find a $\mathcal{M}_1^+(\mathcal{D}) \to \mathbb{R}$ mapping based on $\hat{\mathbf{z}}$.
- Construction: distribution embedding (μ_x) + ridge regression

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

- Our goal: risk bound compared to the regression function

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

Contribution: analysis of the excess risk

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

$$\mathcal{R}[f] = \mathbb{E}_{(x,y)} |f(\mu_x) - y|^2 \text{ (expected risk),}$$

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

We consider two settings:

1. well-specified case: $f_{\rho} \in \mathcal{H}$,
2. misspecified case: $f_{\rho} \in L_{\rho_X}^2 \setminus \mathcal{H}$.

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 $x, \hat{x}_i \in \mathcal{M}_1^+(\mathcal{D})$:

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

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

- Linear $K \Rightarrow$ set kernel:

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

Step-2: ridge regression (analytical solution)

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

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

$$\mathbf{K} = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathbb{R}^{l \times l}, \qquad (2)$$

$$\mathbf{k} = [K(\mu_{\hat{x}_1}, \mu_t), \ldots, K(\mu_{\hat{x}_l}, \mu_t)] \in \mathbb{R}^{1 \times l}. \qquad (3)$$

Blanket assumptions

- $\mathcal{D}$: 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.
- $X = \mu\left(\mathcal{M}_1^+(\mathcal{D})\right) \in \mathcal{B}(H)$.

If in addition

1. well-specified case: f_ρ is 'c-smooth' with 'b-decaying covariance operator' and $l \geq \lambda^{-\frac{1}{b}-1}$, then

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^\lambda, f_\rho) \leq \frac{\log^h(l)}{N^h \lambda^3} + \lambda^c + \frac{1}{l^2 \lambda} + \frac{1}{l \lambda^{\frac{1}{b}}}. \tag{4}$$

2. misspecified case: f_ρ is 's-smooth', $L_{\rho_X}^2$ is separable, and $\frac{1}{\lambda^2} \leq l$, then

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^\lambda, f_\rho) \leq \frac{\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}} \lambda^{\frac{3}{2}}} + \frac{1}{\sqrt{l\lambda}} + \frac{\sqrt{\lambda^{\min(1,s)}}}{\lambda \sqrt{l}} + \lambda^{\min(1,s)}. \tag{5}$$

Performance guarantee: example

Misspecified case: assume

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

Then

1. $s = 1$ ('most difficult' task): $\mathcal{E}(f_{\hat{\mathbf{z}}}^\lambda, f_\rho) \approx t^{-0.25}$,
2. $s \to \infty$ ('simplest' problem): $\mathcal{E}(f_{\hat{\mathbf{z}}}^\lambda, f_\rho) \approx t^{-0.5}$.

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

- 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
  - $X \in \mathcal{B}(H)$.

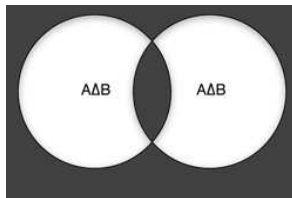In case of (*):

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

They are functions of $\|\mu_a - \mu_b\|_H \Rightarrow$ computation: similar to set kernel.

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

- Objective function:

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

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

Vector-valued output: analytical solution

Analytical solution: prediction on a new test distribution (t)

$$(f_{\hat{z}}^{\lambda} \circ \mu)(t) = \mathbf{k}(\mathbf{K} + l\lambda\mathbf{I}_l)^{-1}[y_1; \ldots; y_l], \tag{6}$$

$$\mathbf{K} = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathcal{L}(Y)^{l \times l}, \tag{7}$$

$$\mathbf{k} = [K(\mu_{\hat{x}_1}, \mu_t), \ldots, K(\mu_{\hat{x}_l}, \mu_t)] \in \mathcal{L}(Y)^{1 \times l}. \tag{8}$$

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

Boundedness and Hölder continuity of K:

1. Boundedness:

$$\|K_{\mu_a}\|_{\mathsf{HS}}^2 = Tr\left(K_{\mu_a}^* K_{\mu_a}\right) \leq B_K \in (0, \infty), \quad (\forall \mu_a \in X).$$
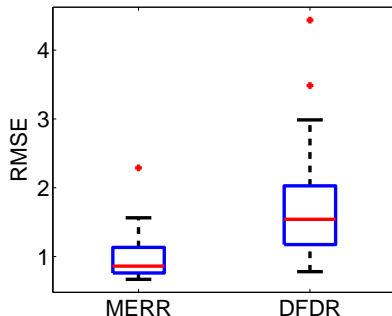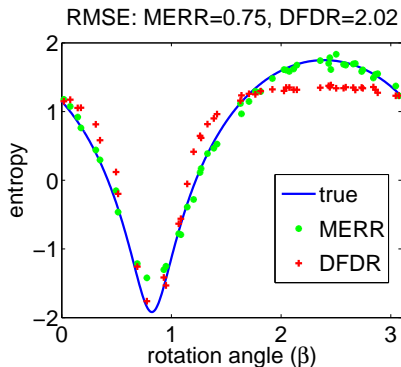
2. Hölder continuity: $\exists L > 0$, $h \in (0, 1]$ such that

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

- Problem: learn the entropy of the 1st 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.

- Performance: $100 \times$ RMSE.
- Baseline [mixture model (EM)]: $7.5 - 8.5$ ($\pm 0.1 - 0.6$).
- Linear K:
    - single: 7.91 (± 1.61).
    - ensemble: **7.86** ($\pm$**1.71**).
- Nonlinear K:
    - Single: 7.90 (± 1.63),
    - Ensemble: **7.81** ($\pm$**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 [$Y = \mathbb{R}$, $f_\rho \in \mathcal{H}$]:
    - accepted at AISTATS-2015 (oral).

- MERR code (ITE toolbox), complete 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: submitted to UAI-2015.

Thank you for the attention!

- Topological definitions, separability.
- Exact prior definitions.
- Vector-valued RKHS.
- Hausdorff metric.
- Weak topology on $\mathcal{M}_1^+(\mathcal{D})$.

- 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.

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 $A \subseteq \mathcal{D}$:

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

- $A \subseteq \mathcal{D}$ 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 $X \times Y$
  - $\alpha \leq n^b t_n \leq \beta$ $(\forall n \geq 1; \alpha > 0, \beta > 0)$,
  - $\exists g \in \mathcal{H}$ 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.

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

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

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

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

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

Definition:

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

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

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

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

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

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

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

- 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:

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

- 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 $k(u, v) \in \mathbb{R}$.

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

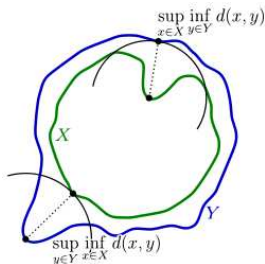$$K(\mu_a, \mu_b) = diag(K_1(\mu_a, \mu_b), \ldots, K_d(\mu_a, \mu_b)). \qquad (14)$$

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

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

- Hausdorff metric [Edgar, 1995]:

$$d_H(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\}. \quad (16)$$

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

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

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