Distribution Regression: Computational & Statistical Tradeoffs

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)

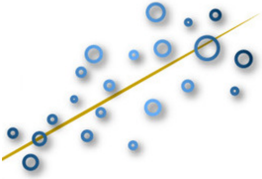
Princeton University
November 26, 2015

Outline

- Motivation: application + theory.
- Problem formulation.
- Results: computational & statistical tradeoffs.
- Numerical examples.

The task

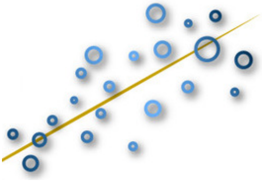
- Samples: $\{(x_i, y_i)\}_{i=1}^{\ell}$. Find $f \in \mathcal{H}$ such that $f(x_i) \approx y_i$.



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

The task

- Samples: $\{(x_i, y_i)\}_{i=1}^{\ell}$. Find $f \in \mathcal{H}$ such that $f(x_i) \approx y_i$.



- Distribution regression:
  - x_i-s are distributions,
  - available only through samples: $\{x_{i,n}\}_{n=1}^{N_i}$, labelled *bags*.
- **Goal**: computational & statistical tradeoffs implied by $N := N_i$ $(\forall i)$.

Motivation (application): aerosol prediction

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

- Relevance: climate research, sustainability.
- 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, Kandasamy et al., 2015].
5. Set metrics: Hausdorff metric [Edgar, 1995]; variants [Wang and Zucker, 2000, Wu et al., 2010, Zhang and Zhou, 2009, Chen and Wu, 2012].

Motivation: theory

- 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, Sutherland et al., 2015] + assumptions:
  1. compact Euclidean domain.
  2. output $= \mathbb{R}$ ([Oliva et al., 2013, Oliva et al., 2015]: distribution/function).

Input-output requirements

- Input: distributions on 'structured' $\mathcal{D}$ domains (kernels).

Input-output requirements

- Input: distributions on 'structured' $\mathcal{D}$ domains (kernels).
- Output:
  - simplest case: $Y = \mathbb{R}$, but

Input-output requirements

- Input: distributions on 'structured' $\mathcal{D}$ domains (kernels).
- Output:
  - simplest case: $Y = \mathbb{R}$, but
  - dependencies might matter: $Y = \mathbb{R}^d$ (or separable Hilbert).

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 ($\exists!$): $\varphi(u) = k(\cdot, u)$.

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

- 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: mean embedding (μ_x)

$$\underbrace{\mathcal{P}(\mathcal{D}) \xrightarrow{\mu=\mu(k)} X \subseteq H}_{\boxed{2}=\text{two-stage sampling}} = H(k)$$

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: mean embedding (μ_x) + ridge regression

$$\underbrace{\mathcal{P}(\mathcal{D}) \xrightarrow{\mu=\mu(k)} X \subseteq H}_{\boxed{2}=\text{two-stage sampling}} = \underbrace{H(k) \xrightarrow{f\in\mathcal{H}=\mathcal{H}(K)} \mathbb{R}}_{\boxed{1}=\text{Hilbert } \to \mathbb{R} \text{ regression}}.$$

$\boxed{1}$: Hilbert $\to \mathbb{R}$ regression, well-specified case

- Regression function, expected risk (assume for a moment: $f_\rho \in \mathcal{H}$):

$$f_\rho(\mu_x) = \int_{\mathbb{R}} y \mathrm{d}\rho(y|\mu_x), \qquad \mathcal{R}[f] = \mathbb{E}_{(x,y)} |f(\mu_x) - y|^2 .$$

1: Hilbert $\to \mathbb{R}$ regression, well-specified case

- Regression function, expected risk (assume for a moment: $f_\rho \in \mathcal{H}$):

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

- Ridge estimator:

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

1: Hilbert $\to \mathbb{R}$ regression, well-specified case

- Regression function, expected risk (assume for a moment: $f_\rho \in \mathcal{H}$):

$$f_\rho(\mu_x) = \int_{\mathbb{R}} y \mathrm{d}\rho(y|\mu_x), \qquad \mathcal{R}[f] = \mathbb{E}_{(x,y)} |f(\mu_x) - y|^2 .$$

- Ridge estimator:

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

- Excess risk:

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

- Known [Caponnetto and De Vito, 2007]: if $\rho(\mu_x, y) \in \mathcal{P}(b, c)$, then the best/achieved rate

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda}, f_{\rho}) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right) \quad (1 < b, c \in (1, 2]).$$

- Known [Caponnetto and De Vito, 2007]: if $\rho(\mu_x, y) \in \mathcal{P}(b,c)$, then the best/achieved rate

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda}, f_{\rho}) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right) \quad (1 < b, c \in (1,2]).$$

- $\rho \in \mathcal{P}(b,c)$:

$$T = \int_X K(\cdot, \mu_a) K^*(\cdot, \mu_a) \mathrm{d}\rho_X(\mu_a) : \mathcal{H} \to \mathcal{H}.$$

  - Eigenvalues of T decay as $\lambda_n = \mathcal{O}(n^{-b})$. $f_\rho \in Im\left(T^{\frac{c-1}{2}}\right)$.
  - Intuition: $1/b$ – effective input dimension, c – smoothness of f_ρ.

Our question

Can we reach this $\mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right)$ minimax rate? $N = ?$

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

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

$\boxed{2}$: mean embedding, $\mu_{x_i} \to \mu_{\hat{x}_i}$

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

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

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

$\boxed{2}$: mean embedding, $\mu_{x_i} \to \mu_{\hat{x}_i}$

- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ 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:

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

$\boxed{2}$: mean embedding, $\mu_{x_i} \to \mu_{\hat{x}_i}$

- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ 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:

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

$\boxed{2}$: ridge regression $\Rightarrow$ analytical solution

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

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

$\Rightarrow K(\mu_x, \mu_{x'}) = \left\langle K(\cdot, \mu_x), K(\cdot, \mu'_x) \right\rangle_{\mathcal{H}(K)}$ matter.

- Convergence of the mean embedding:

$$\|\mu_x - \mu_{\hat{x}}\|_{H(k)} = \mathcal{O}_p\left(\frac{1}{\sqrt{N}}\right).$$

- Hölder property of K ($0 < L$, $0 < h \leq 1$):

$$\|K(\cdot, \mu_x) - K(\cdot, \mu_{\hat{x}})\|_{\mathcal{H}(K)} \leq L \|\mu_x - \mu_{\hat{x}}\|_{H(k)}^h .$$

- $f_{\hat{z}}^{\lambda}$ depends 'nicely' on $\mu_{\hat{x}}$.

$\boxed{1}$ - $\boxed{2}$: Proof idea

By decomposing the excess risk, concentration, on $\mathcal{P}(b,c)$ we get

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}) \leq \underbrace{\frac{\log^h(\ell)}{N^h\lambda}\left(\frac{1}{\lambda^2\ell^2} + 1 + \frac{1}{\ell\lambda^{1+\frac{1}{b}}}\right)}_{\boxed{2}=\text{two-stage sampling}} + \underbrace{\lambda^c + \frac{1}{\ell^2\lambda} + \frac{1}{\ell\lambda^{\frac{1}{b}}}}_{\boxed{1}\, = H \to \mathbb{R} \text{ regression}} \to 0,$$

$$\text{s.t. } \ell\lambda^{\frac{b+1}{b}} \geq 1, \frac{\log(\ell)}{\lambda^{\frac{2}{h}}} \leq N.$$

$\boxed{1}$ - $\boxed{2}$: Proof idea

By decomposing the excess risk, concentration, on $\mathcal{P}(b, c)$ we get

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}) \leq \underbrace{\frac{\log^h(\ell)}{N^h \lambda}\left(\frac{1}{\lambda^2 \ell^2} + 1 + \frac{1}{\ell \lambda^{1+\frac{1}{b}}}\right)}_{\boxed{2}=\text{two-stage sampling}} + \underbrace{\lambda^c + \frac{1}{\ell^2 \lambda} + \frac{1}{\ell \lambda^{\frac{1}{b}}}}_{\boxed{1}=H \to \mathbb{R} \text{ regression}} \to 0,$$

$$\text{s.t. } \ell \lambda^{\frac{b+1}{b}} \geq 1, \frac{\log(\ell)}{\lambda^{\frac{2}{h}}} \leq N.$$

- Let $N = \ell^{\frac{a}{h}} \log(\ell) \Rightarrow$ 1st term $+$ constraints simplify.
- $a > 0$: needed, i.e. $N > \log(\ell)$.
- Bias-variance trick with constraint-checking $\Rightarrow$

Computational & statistical tradeoff (W)

If

- $a \leq \frac{b(c+1)}{bc+1}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{ac}{c+1}}\right)$ with $\lambda = \ell^{-\frac{a}{c+1}}$,
- $a \geq \frac{b(c+1)}{bc+1}$ then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right)$ with $\lambda = \ell^{-\frac{b}{bc+1}}$.

Computational & statistical tradeoff (W)

If

- $a \leq \frac{b(c+1)}{bc+1}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{ac}{c+1}}\right)$ with $\lambda = \ell^{-\frac{a}{c+1}}$,
- $a \geq \frac{b(c+1)}{bc+1}$ then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right)$ with $\lambda = \ell^{-\frac{b}{bc+1}}$.

Meaning (a-dependence, $N = \ell^{\frac{a}{h}} \log(\ell)$):

- 'Smaller a' = computational saving, but reduced statistical efficiency.

Computational & statistical tradeoff (W)

If

- $a \leq \frac{b(c+1)}{bc+1}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{ac}{c+1}}\right)$ with $\lambda = \ell^{-\frac{a}{c+1}}$,
- $a \geq \frac{b(c+1)}{bc+1}$ then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right)$ with $\lambda = \ell^{-\frac{b}{bc+1}}$.

Meaning (a-dependence, $N = \ell^{\frac{a}{h}} \log(\ell)$):

- 'Smaller a' = computational saving, but reduced statistical efficiency.
- Sensible choice: $a \leq \frac{b(c+1)}{bc+1} < 2$:

N sub-quadratic in ℓ achieves *one-stage sampled* minimax rate! ('=')

Computational & statistical tradeoff (W)

If

- $a \le \frac{b(c+1)}{bc+1}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right) = \mathcal{O}_p\left(\ell^{-\frac{ac}{c+1}}\right)$ with $\lambda = \ell^{-\frac{a}{c+1}}$,
- $a \ge \frac{b(c+1)}{bc+1}$ then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_\rho\right) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right)$ with $\lambda = \ell^{-\frac{b}{bc+1}}$.

Meaning (h-dependence, $N = \ell^{\frac{a}{h}} \log(\ell)$):

- smoother K kernel is rewarding = bag-size reduction; see smoothness of f_ρ.

Computational & statistical tradeoff (W)

If

- $a \leq \frac{b(c+1)}{bc+1}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{ac}{c+1}}\right)$ with $\lambda = \ell^{-\frac{a}{c+1}}$,
- $a \geq \frac{b(c+1)}{bc+1}$ then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right)$ with $\lambda = \ell^{-\frac{b}{bc+1}}$.

Meaning (c-dependence):

- $c \mapsto \frac{b(c+1)}{bc+1}$ decreasing: smaller bags are enough for easier problems.

Misspecified setting

- Relevant case: $f_\rho \in L^2_{\rho_X} \backslash \mathcal{H}$.
- f_ρ: difficulty parameter $= s \in (0,1]$, larger $s =$ easier problem.
- Proof idea:
  - ∞-D exponential family fitting [Sriperumbudur et al., 2014],
  - ridge solution.

Computational & statistical tradeoff (M)

Let $N = \ell^{\frac{2a}{h}} \log(\ell)$ $(a > 0)$. If

- $a \leq \frac{s+1}{s+2}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{2sa}{s+1}}\right)$ with $\lambda = \ell^{-\frac{a}{s+1}}$,
- $a \geq \frac{s+1}{s+2}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{2s}{s+2}}\right)$ with $\lambda = \ell^{-\frac{1}{s+2}}$.

Computational & statistical tradeoff (M)

Let $N = \ell^{\frac{2a}{h}} \log(\ell)$ $(a > 0)$. If

- $a \le \frac{s+1}{s+2}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{2sa}{s+1}}\right)$ with $\lambda = \ell^{-\frac{a}{s+1}}$,
- $a \ge \frac{s+1}{s+2}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{2s}{s+2}}\right)$ with $\lambda = \ell^{-\frac{1}{s+2}}$.

Meaning (a-dependence):

- 'Smaller a' = computational saving, but reduced statistical efficiency.

Computational & statistical tradeoff (M)

Let $N = \ell^{\frac{2a}{h}} \log(\ell)$ $(a > 0)$. If

- $a \le \frac{s+1}{s+2}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{2sa}{s+1}}\right)$ with $\lambda = \ell^{-\frac{a}{s+1}}$,
- $a \ge \frac{s+1}{s+2}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{2s}{s+2}}\right)$ with $\lambda = \ell^{-\frac{1}{s+2}}$.

Meaning (a-dependence):

- 'Smaller a' = computational saving, but reduced statistical efficiency.
- Sensible choice: $a \le \frac{s+1}{s+2} \le \frac{2}{3} \Rightarrow 2a \le \frac{4}{3} < 2$!

$\boxed{N \text{ can be sub-quadratic in } \ell \text{ again ('=')!}}$

Computational & statistical tradeoff (M)

Let $N = \ell^{\frac{2a}{h}} \log(\ell)$ $(a > 0)$. If

- $a \le \frac{s+1}{s+2}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{2sa}{s+1}}\right)$ with $\lambda = \ell^{-\frac{a}{s+1}}$,
- $a \ge \frac{s+1}{s+2}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{2s}{s+2}}\right)$ with $\lambda = \ell^{-\frac{1}{s+2}}$.

Meaning (h-dependence):

- $h \mapsto \frac{2a}{h}$ is decreasing: smoother K kernel is rewarding.

Computational & statistical tradeoff (M)

Let $N = \ell^{\frac{2a}{h}} \log(\ell)$ $(a > 0)$. If

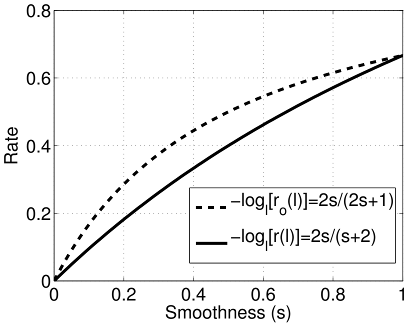
- $a \leq \frac{s+1}{s+2}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{2sa}{s+1}}\right)$ with $\lambda = \ell^{-\frac{a}{s+1}}$,
- $a \geq \frac{s+1}{s+2}$, then $\mathcal{E}\left(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}\right) = \mathcal{O}_p\left(\ell^{-\frac{2s}{s+2}}\right)$ with $\lambda = \ell^{-\frac{1}{s+2}}$.

Meaning (s-dependence): $s \mapsto \frac{2s}{s+2}$ is increasing, i.e easier task $=$ better rate,

- $s \to 0$: arbitrary slow rate.
- $s = 1$: $\ell^{-\frac{2}{3}}$ rate.

Optimality of the rate (M)

- Our rate: $r(\ell) = \ell^{-\frac{2s}{s+2}}$ – range space assumption (s).
- One-stage sampled optimal rate: $r_o(\ell) = \ell^{-\frac{2s}{2s+1}}$ [Steinwart et al., 2009],
  - range-space assumption + eigendecay constraint,
  - $\mathcal{D}$: compact metric, $Y = \mathbb{R}$.



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.

Hölder K examples (other than the linear K when h=1)

In case of compact metric $\mathcal{D}$, universal k:

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

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

Vector-valued output: similarly

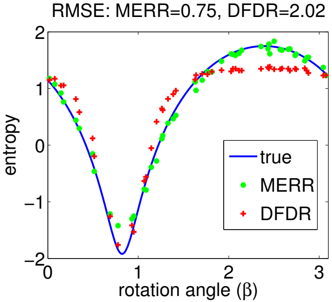
- $K(\mu_a, \mu_b) \in \mathcal{L}(Y)$.
- Prediction on a new test distribution (t):

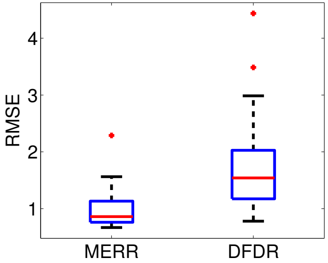
$$\boxed{\begin{aligned}(f_{\hat{\mathbf{z}}}^{\lambda} \circ \mu)(t) &= \mathbf{k}(\mathbf{K} + \ell\lambda \mathbf{I}_\ell)^{-1}[y_1; \ldots; y_\ell], \\ \mathbf{K} &= [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathcal{L}(Y)^{\ell \times \ell}, \\ \mathbf{k} &= [K(\mu_{\hat{x}_1}, \mu_t), \ldots, K(\mu_{\hat{x}_\ell}, \mu_t)] \in \mathcal{L}(Y)^{1 \times \ell}.\end{aligned}} \quad \begin{aligned}(4)\\(5)\\(6)\end{aligned}$$

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

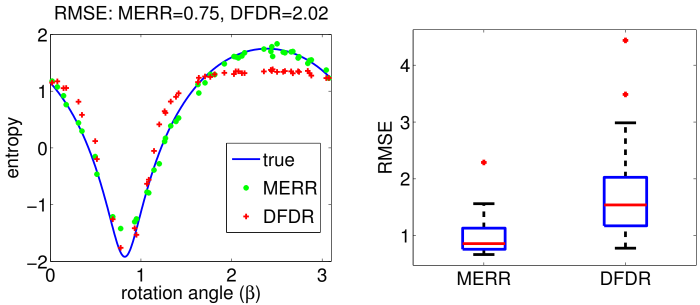
Demo

- Supervised entropy learning:





- 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

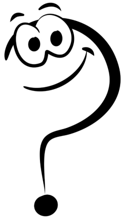
- Problem: distribution regression (k).
- Contribution:
  - computational & statistical tradeoff analysis,
  - specifically, the set kernel is consistent: 16-year-old open question,
  - minimax optimal rate is achievable: sub-quadratic bag size.

Summary

- Problem: distribution regression (k).
- Contribution:
  - computational & statistical tradeoff analysis,
  - specifically, the set kernel is consistent: 16-year-old open question,
  - minimax optimal rate is achievable: sub-quadratic bag size.
- Code in ITE, analysis submitted to JMLR:

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

Thank you for the attention!



Acknowledgments: This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350. A part of 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.

Caponnetto, A. and De Vito, E. (2007).
Optimal rates for regularized least-squares algorithm.
Foundations of Computational Mathematics, 7:331–368.

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.

Kandasamy, K., Krishnamurthy, A., Póczos, B., Wasserman, L., and Robins, J. M. (2015).
Influence functions for machine learning: Nonparametric estimators for entropies, divergences and mutual informations.
Technical report, Carnegie Mellon University.
http://arxiv.org/abs/1411.4342.

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., Neiswanger, W., Póczos, B., Xing, E., and Schneider, J. (2015).

Fast function to function regression.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 717–725.

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)*, pages 692–701.

Sriperumbudur, B. K., Fukumizu, K., Kumar, R., Gretton, A., and Hyvärinen, A. (2014).
Density estimation in infinite dimensional exponential families.
Technical report.
(http://arxiv.org/pdf/1312.3516).

Steinwart, I., Hush, D. R., and Scovel, C. (2009).
Optimal rates for regularized least squares regression.
In *Conference on Learning Theory (COLT)*.

Sutherland, D. J., Oliva, J. B., Póczos, B., and Schneider, J. (2015).
Linear-time learning on distributions with approximate kernel embeddings.
Technical report, Carnegie Mellon University.
http://arxiv.org/abs/1509.07553.

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.