# Two-Stage Sampled Distribution Regression on Separable Topological Domains*

Zoltán Szabó[1], Arthur Gretton[1], Barnabás Póczos[2], Bharath K. Sriperumbudur[3]

[1]Gatsby Unit, University College London [2]Machine Learning Department, Carnegie Mellon University [3]Department of Statistics, Pennsylvania State University

## Problem

- Distribution regression, with two-stage sampling [1]:
  - Input = distribution, output $\in \mathbb{R}$, or more generally separable Hilbert space.
  - Challenge: we only have samples from the input distributions.
- Covered machine learning tasks include:
  - multiple instance learning (MIL),
  - point estimates of statistics (e.g., entropy or a hyperparameter).
- Existing methods: heuristics, or require density estimation (which typically scale poorly in dimension).

## Contribution

- We study an alternative, simple method: embed the distributions to a RKHS ($k$), then apply ridge regression ($K$).
- Results:
  - Consistency, convergence rate $\xrightarrow{\text{specially: } Y=\mathbb{R},\ K\text{: linear}}$
  - Set kernels [2, 3] are consistent in regression (15-year-old open problem).

## Introduction

Existing heuristics:

- parametric model fitting; kernelized Gaussian divergences; kernels on distributions; Kullback-Leibler-, Rényi-, Tsallis divergence; set (semi)metric.
- issues: parameterization may fail to hold; metric/kernel? consistent estimation? consistency in learning tasks?

Theoretically justified methods [1, 4]:

- require density estimation (often poor scaling).
- assume density, compact Euclidean domain.

## Distribution Regression

- $D(\mathfrak{X})$ distributions on domain $\mathfrak{X}$.
- $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{l} \overset{i.i.d.}{\sim} \mathcal{M}$: $(x_i, y_i) \in D(\mathfrak{X}) \times Y$.
- Given: $\hat{\mathbf{z}} = \{(\{x_{i,n}\}_{n=1}^{N}, y_i)\}_{i=1}^{l}$, where $x_{i,1}, \ldots, x_{i,N} \overset{i.i.d.}{\sim} x_i$.

---

- **Goal**: learn the relation between $(x, y)$ given $\hat{\mathbf{z}}$.
- **Idea**:

$$D(\mathfrak{X}) \xrightarrow{\mu} X (\subseteq H) \xrightarrow{f \in \mathcal{H} = \mathcal{H}(K)} Y,$$

  i.e., embed the distributions to a $H = H(k)$ RKHS on $\mathfrak{X}$, then $X \to Y$ ridge regression.
- **Notations**: $Y$ is a separable Hilbert space. $k$ is a kernel on $\mathfrak{X}$, mean embedding

$$\mu_x = \int_{\mathfrak{X}} k(\cdot, u) \mathrm{d}x(u) = \mathbb{E}_{u \sim x}[k(\cdot, u)], \qquad X = \mu\left(D(\mathfrak{X})\right).$$

  $\rho(\mu_x, y) = \rho(y|\mu_x)\rho_X(\mu_x)$; regression function of $\rho$, $\|\cdot\|_\rho$:

$$f_\rho(\mu_a) = \int_Y y \mathrm{d}\rho(y|\mu_a), \quad \|f\|_\rho^2 = \int_X \|f(\mu_a)\|_Y^2 \, \mathrm{d}\rho_X(\mu_a),$$

$\mathcal{H} = \mathcal{H}(K) = Y$-valued RKHS of $X \to Y$ functions with kernel $K: X \times X \to \mathcal{L}(Y) = \{Y \to Y \text{ bounded linear operators}\}$.

## Objective Function, Algorithm

- **Cost function** (of MERR):

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

  where $\hat{x}_i = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_{i,n}}$ is the $i^{th}$ empirical distribution.
- Analytical **solution**: prediction on a new distribution $t$

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

- **Examples**:
  - If $Y = \mathbb{R}$, then $\mathcal{L}(Y) = \mathbb{R}$.
  - If $Y = \mathbb{R}^d$, then $\mathcal{L}(Y) = \mathbb{R}^{d \times d}$.

## Intuitive Assumption

The regression function ($f_\rho$) is "sufficiently smooth" in $L^2_{\rho_X}$.

## Remarks ($Y = \mathbb{R}$)

- For linear $K(\mu_a, \mu_b) = \langle \mu_a, \mu_b \rangle_H$, we get the set kernel:

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

- On compact metric $\mathfrak{X}$ and for "rich" $H(k)$, the following $K$ functions are Hölder continuous ($h$) kernels:

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

## Error Guarantee, Consistency

If $l$ is "not too small" compared to $\lambda$ ($\frac{1}{\lambda^2} \le l$), then with high probability

$$\left\| f_{\hat{\mathbf{z}}}^{\lambda} - f_\rho \right\|_\rho \le B(l, N, \lambda) + D_{\mathcal{H}},$$

where $B(l, N, \lambda) = \frac{\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}}\lambda^{\frac{3}{2}}} + \frac{1}{\lambda\sqrt{l}}$, $D_{\mathcal{H}} = \inf_{q \in \mathcal{H}} \|f_\rho - q\|_\rho$.

**Interpretation**:

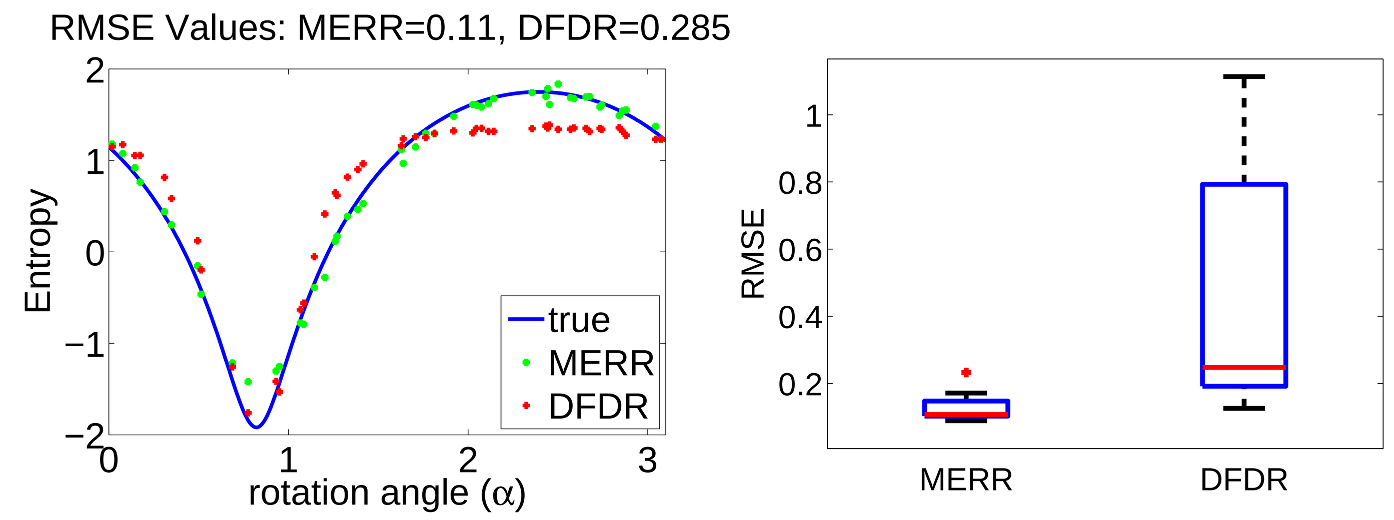
- $D_{\mathcal{H}}$: approximation error of $f_\rho$ from $\mathcal{H}$; if $\mathcal{H}$ is dense in $L^2_{\rho_X}$, then $D_{\mathcal{H}} = 0$.
- For suitable $(l, N, \lambda)$ choice $B(l, N, \lambda)$ converges to 0. Example:
  - $(l, N)$ trade-off: let $l = N^a$ with $\frac{2}{3}h \le a < h$.
  - Regularization: $\lambda = l \left[\frac{\log(l)}{N}\right]^h \to 0$.

  In this case $B(l, N, \lambda) = \frac{1}{N^{\frac{3a}{2} - h} \log^h(N)} \to 0$.

## Applications

**Supervised entropy learning**:

- Label = entropy of the distribution represented by a bag.
- MERR is more precise than the only theoretically justified method [1] (DFDR; by avoiding density estimation).




(a) Entropy of Gaussian (b) Boxplot of RMSE

**Aerosol prediction**:

- Bag = multispectral satellite image over an area.
- Label = aerosol value (highly accurate, expensive ground-based instrument).
- Performance:

| Method | 100×RMSE | ±std |
|---|---|---|
| Baseline [mixture model (EM)] | 7.5 – 8.5 | ±0.1 – 0.6 |
| MERR: linear $K$, single | 7.91 | ±1.61 |
| MERR: linear $K$, ensemble | **7.86** | ±1.71 |
| MERR: nonlinear $K$, single | 7.90 | ±1.63 |
| MERR: nonlinear $K$, ensemble | **7.81** | ±1.64 |

- MERR compares favourably to domain-specific, engineered methods (beating state-of-the art MIL techniques).

**Code**: in the ITE toolbox (`https://bitbucket.org/szzoli/ite/`).

## Acknowledgements

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.

## References

[1] Barnabás Póczos, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Distribution-free distribution regression. *AISTATS; JMLR W&CP*, 31:507–515, 2013.

[2] David Haussler. Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California at Santa Cruz, 1999.

[3] Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, and Alexander Smola. Multi-instance kernels. In *ICML*, pages 179–186, 2002.

[4] Junier B. Oliva, Willie Neiswanger, Barnabás Póczos, Jeff Schneider, and Eric Xing. Fast distribution to real regression. *AISTATS; JMLR W&CP*, 33:706–714, 2014.

*Gatsby External Review, London, UK, 29 October 2014. The ordering of the second through fourth authors is alphabetical.