Improved learning theory for kernel distribution regression with two-stage sampling

François Bachoc

Institut de Mathématiques de Toulouse
Université Paul Sabatier

Joint work with **Louis Béthune** (Apple Paris), **Alberto González-Sanz** (Columbia University) and **Jean-Michel Loubes** (IMT Toulouse)

Journée *Méthodes à noyaux à grande échelle*
GdR IASIS / RT MAIAGES
Novembre 2024

Outline

1. Distribution regression, Hilbertian embedding and two-stage sampling

2. Near-unbiased condition and improved rates

Distribution regression

We observe i.i.d. pairs

$$(\mu_i, Y_i), \qquad i = 1, \ldots, n.$$

- $Y_i \in \mathbb{R}$.
- μ_i is a probability distribution on Ω.
- Ω is compact in $\mathbb{R}^d$.

Goal: constructing a regression function

$$\widehat{f}_n : \mathcal{P}(\Omega) \to \mathbb{R},$$

- where $\mathcal{P}(\Omega)$ is the set of probability distributions on Ω.

Illustration with $d = 1$

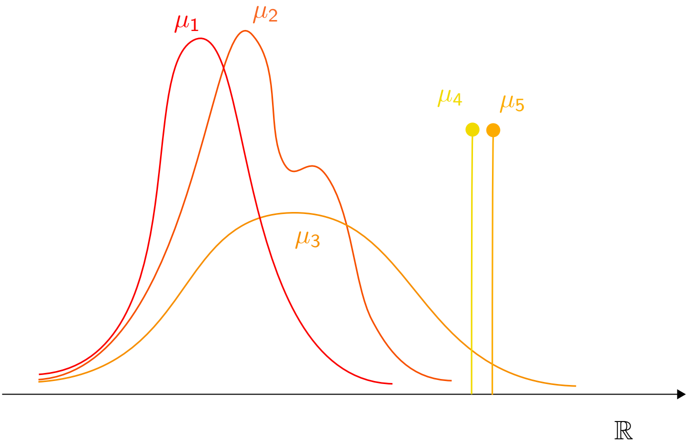


Illustration with $d = 1$

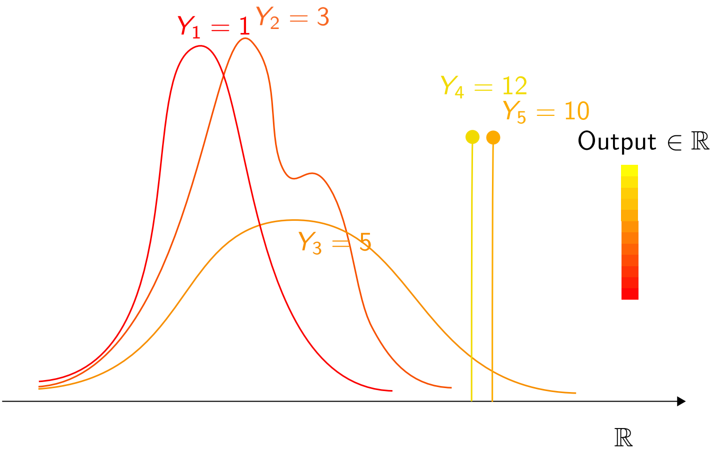
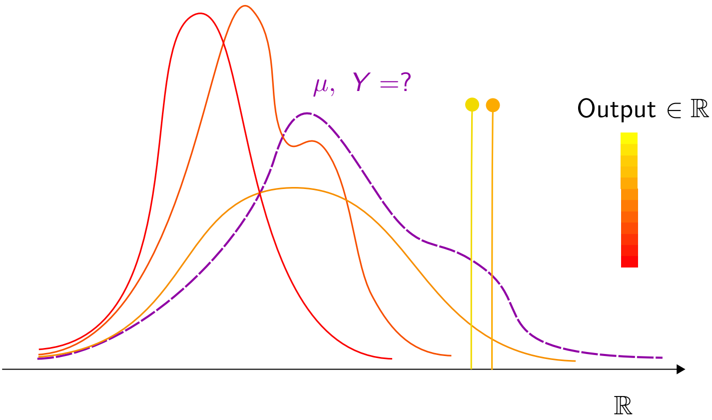


Illustration with $d = 1$



Applications

For instance.

- From $\mu \in \mathcal{P}(\Omega)$, regressing $\mathrm{entropy}(\mu) \in \mathbb{R}$.
- From a Gaussian mixture μ, regressing the number of components.
- Seeing images/textures as grids of PDF values (after renormalization) [Bachoc et al., 2023a].
- Multiple-instance learning: a label is associated to a bag of vectors [Dietterich et al., 1997, Ray and Page, 2015].
  - E.g. for drug design.
- Ecological inference: *"learning individual-level associations from aggregate data"* [Flaxman et al., 2015].
  - E.g. percentage of vote of men for Obama in 2012, county by county.

Hilbertian embedding

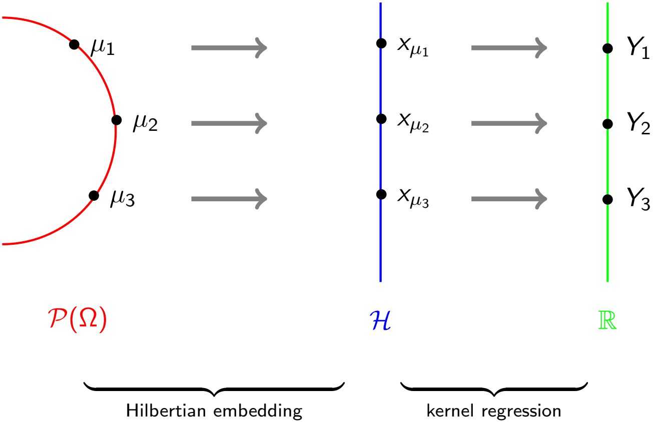
Hilbertian embedding

$$\begin{aligned} x : \mathcal{P}(\Omega) &\to \mathcal{H} \\ \mu &\mapsto x_\mu, \end{aligned}$$

where $\mathcal{H}$ is a Hilbert space.

$\Longrightarrow$ In order to use kernel regression on Hilbert spaces (see later)!

Illustration of Hilbertian embedding + kernel regression



Hilbertian embedding 1: mean embedding

Consider a kernel k on Ω.

Very quick introduction to kernels and RKHS

- $k : \Omega \times \Omega \to \mathbb{R}$.
- For any $\ell \in \mathbb{N}$, $t_1, \ldots, t_\ell \in \Omega$, the $\ell \times \ell$ matrix $[k(t_i, t_j)]$ is symmetric non-negative definite.
- There is a (unique) Hilbert space $\mathcal{H}_k$ of functions from Ω to $\mathbb{R}$,
  - with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$
  - with norm $\| \cdot \|_{\mathcal{H}_k}$

  such that
  - $\mathcal{H}_k$ contains all functions $k_t := k(t, \cdot)$ for $t \in \Omega$,
  - for all $g \in \mathcal{H}_k$, for all $t \in \Omega$, $g(t) = \langle g, k_t \rangle_{\mathcal{H}_k}$ (reproducing property).

  $\Longrightarrow$ $\mathcal{H}_k$ is the reproducing kernel Hilbert space (RKHS) of k.

Then mean embedding

$$x_\mu := \int_\Omega k_x \mathrm{d}\mu(x) = \left(t \mapsto \int_\Omega k(t, x) \mathrm{d}\mu(x) \right),$$

[Szabó et al., 2015, Szabó et al., 2016, Muandet et al., 2017].

Hilbertian embedding 2: sliced Wasserstein

The sliced Wasserstein distance [Kolouri et al., 2018, Manole et al., 2022, Meunier et al., 2022]

$$\mathcal{SW}(\mu,\nu)^2 := \int_{\mathcal{S}^{d-1}} \int_0^1 \left(F_{\mu_\theta}^{-1}(t) - F_{\nu_\theta}^{-1}(t)\right)^2 \mathrm{d}t\mathrm{d}\Lambda(\theta),$$

- with two distributions $\mu, \nu \in \mathcal{P}(\Omega)$,
- where $\mathcal{S}^{d-1}$ is the unit sphere,
- where Λ is the uniform distribution on $\mathcal{S}^{d-1}$,
- where μ_θ is the univariate distribution of $\langle \theta, X \rangle$ for $X \sim \mu$,
- where $F_{\mu_\theta}^{-1}$ is the quantile function of μ_θ.

Hilbert distance of a Hilbertian embedding

$$\mathcal{SW}(\mu,\nu)^2 = \|x_\mu - x_\nu\|_{\mathcal{H}}^2.$$

- $\mathcal{H} = \mathcal{L}^2\left(\Lambda \times \mathcal{U}([0,1])\right)$,
    - where $\mathcal{U}([0,1])$ is the uniform distribution on $[0,1]$.
- $x_\mu(\theta, t) = F_{\mu_\theta}^{-1}(t)$.

Hilbertian embedding 3: Sinkhorn distance and dual potential

Dual formulation of entropic-regularized (Sinkhorn) optimal transport [Cuturi, 2013, Genevay, 2019]

$$\sup_{h \in L^1(\mu), g \in L^1(\mathcal{U})} \int_\Omega h(x) \mathrm{d}\mu(x) + \int_\Omega g(y) \mathrm{d}\mathcal{U}(y) - \epsilon \int_{\Omega \times \Omega} e^{\frac{1}{\epsilon}\left(h(x) + g(y) - \frac{1}{2}\|x-y\|^2\right)} \mathrm{d}\mu(x) \mathrm{d}\mathcal{U}(y).$$

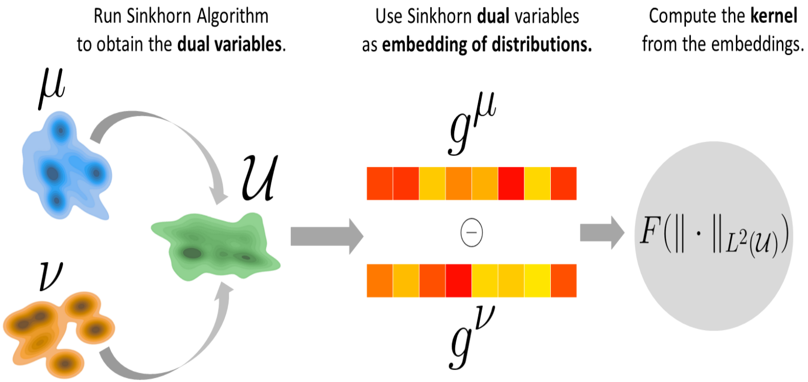
- $\epsilon > 0$ regularization parameter.
- Fixed $\mathcal{U} \in \mathcal{P}(\Omega)$ called reference measure.
- For any $\mu \in \mathcal{P}(\Omega)$.

Hilbertian embedding

There is a unique optimal $(h^\star, g^\star)$ such that $g^\star$ is centered w. r. t. $\mathcal{U}$. Also $g^\star \in L^2(\mathcal{U})$.
[Bachoc et al., 2023a]:

- $x_\mu := g^*$.
- $\mathcal{H} := L^2(\mathcal{U})$.

Illustration of Sinkhorn embedding



Kernel ridge regression on Hilbert space

- Hilbertian covariates: for $i = 1, \dots, n$, let

$$x_i := x_{\mu_i}.$$

- Kernel K on $\mathcal{H}$. E.g. squared-exponential, for $u, v \in \mathcal{H}$,

$$K(u,v) := e^{-\|u-v\|_{\mathcal{H}}^2}.$$

$\Longrightarrow$ Yields the RKHS $\mathcal{H}_K$ of functions from $\mathcal{H}$ to $\mathbb{R}$.

- Ridge regression

$$\boxed{\widehat{f}_n = \underset{f \in \mathcal{H}_K}{\mathrm{argmin}}\ R_n(f)}$$

with

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - Y_i\right)^2 + \lambda \|f\|_{\mathcal{H}_K}^2,$$

  - where $\lambda > 0$ is a regularization parameter.

Two-stage sampling

Studied in [Szabó et al., 2015, Szabó et al., 2016, Meunier et al., 2022].

- For $i = 1, \dots, n$, μ_i is unobserved.
- We observe i. i. d. $(X_{i,j})_{j=1,\dots,N}$ with $X_{i,j} \sim \mu_i$.
- We let

$$\mu_i^N = \frac{1}{N} \sum_{j=1}^{N} \delta_{X_{i,j}}$$

and

$$x_{N,i} = x_{\mu_i^N}.$$

Ridge regression with approximate covariates

$$\boxed{\widehat{f}_{n,N} = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \; R_{n,N}(f)}$$

with

$$R_{n,N}(f) := \frac{1}{n} \sum_{i=1}^{n} \left(f(x_{N,i}) - Y_i\right)^2 + \lambda \|f\|_{\mathcal{H}_K}^2.$$

1 Distribution regression, Hilbertian embedding and two-stage sampling

2 Near-unbiased condition and improved rates

Existing error bounds on $\widehat{f}_n - \widehat{f}_{n,N}$

- [Szabó et al., 2015, Szabó et al., 2016, Meunier et al., 2022] address their respective distribution regression settings.
- But their results are naturally made general.

Existing bounds

For all $s \geq 1$, conditionally to $(x_i, Y_i)_{i=1}^n$,

$$\mathbb{E}\left[\left\|\hat{f}_n - \hat{f}_{n,N}\right\|_{\mathcal{H}_K}^s\right]^{1/s} \leq \frac{\text{constant}\left(\|\hat{f}_n\|_{\mathcal{H}_K} + Y_{\max,n}\right)}{\sqrt{N}\lambda}$$

- with $Y_{\max,n} = \max_{i=1,\dots,n} |Y_i|$.

Existing error bounds are improvable?

- Proofs based on explicit expressions of $\hat{f}_n$ and $\hat{f}_{n,N}$.
- Somewhere:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_n(x_i) K_{x_i} - \hat{f}_n(x_{N,i}) K_{x_{N,i}} \right) \right\|_{\mathcal{H}_K} \leq \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{f}_n(x_i) K_{x_i} - \hat{f}_n(x_{N,i}) K_{x_{N,i}} \right\|_{\mathcal{H}_K}.$$

- But

$$\left(\hat{f}_n(x_i) K_{x_i} - \hat{f}_n(x_{N,i}) K_{x_{N,i}} \right)_{i=1}^{n}$$

  are independent conditionally on $(x_i, Y_i)_{i=1}^n$.
- Do they have approximately zero mean (in $\mathcal{H}_K$)?
- If yes, we could gain an order $\sqrt{n}$ in the upper bound.

Near-unbiased condition

In [Bachoc et al., 2023b].

Near-unbiased condition

- For $i = 1, \ldots, n$, there are random $a_{N,i}$ and $b_{N,i}$ such that

$$x_{N,i} - x_i = a_{N,i} + b_{N,i}.$$

- $\|a_{N,i}\|_{\mathcal{H}}$ has order $\frac{1}{\sqrt{N}}$.
- $\mathbb{E}[a_{N,i}|\mu_i] = 0 \in \mathcal{H}$.
- $\|b_{N,i}\|_{\mathcal{H}}$ has order $\frac{1}{N}$.

For the 3 examples of Hilbertian embedding

- Mean embedding: $b_{N,i} = 0$ (exactly unbiased).
- Sinkhorn: indeed near unbiased, relying on [González-Sanz et al., 2022].
- Sliced Wasserstein: indeed near unbiased under conditions.

Improved rates (1/2)

Main result in [Bachoc et al., 2023b].

Theorem

Up to constant

$$\sqrt{\mathbb{E}_n \left[\|\hat{f}_n - \hat{f}_{n,N}\|^2_{\mathcal{H}_K} \right]} \leq \frac{Y_{\max,n} + \|\hat{f}_n\|_{\mathcal{H}_K}}{\lambda N} + \frac{Y_{\max,n} + \|\hat{f}_n\|_{\mathcal{H}_K}}{\lambda \sqrt{n} \sqrt{N}}$$

$$+ \left(1 + \frac{\sqrt{N}}{\sqrt{n}}\right)^{-1} \left(\frac{Y_{\max,n} + \|\hat{f}_n\|_{\mathcal{H}_K}}{\lambda n} + \frac{Y_{\max,n} + \|\hat{f}_n\|_{\mathcal{H}_K}}{\lambda^2 n \sqrt{N}} \right)$$

- with $Y_{\max,n} = \max_{i=1,\dots,n} |Y_i|$,
- where $\mathbb{E}_n$ denotes the conditional expectation given $(\mu_i, Y_i)_{i=1}^n$.

The $\sqrt{n}$ we gain comes from average of independent centered variables.

Improved rates (2/2)

From previous theorem.

Corollary

- Let $n, N \to \infty$ and $\lambda \to 0$.
- Assume $1/\lambda = \mathcal{O}(\sqrt{N})$.
- Assume $n = \mathcal{O}(N)$.
- Assume $\mathbb{E}[\|\hat{f}_n\|^2_{\mathcal{H}_K}]$ and $\mathbb{E}[Y^2_{\max,n}]$ are bounded.

Then

$$\sqrt{\mathbb{E}\left[\|\hat{f}_n - \hat{f}_{n,N}\|^2_{\mathcal{H}_K}\right]} = \mathcal{O}\left(\frac{1}{\lambda\sqrt{n}\sqrt{N}}\right).$$

In comparison the methods of [Szabó et al., 2015, Szabó et al., 2016, Meunier et al., 2022] yield

$$\sqrt{\mathbb{E}\left[\|\hat{f}_n - \hat{f}_{n,N}\|^2_{\mathcal{H}_K}\right]} = \mathcal{O}\left(\frac{1}{\lambda\sqrt{N}}\right).$$

One proof ingredient

Recall

$$\hat{f}_n = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - Y_i\right)^2 + \lambda \|f\|_{\mathcal{H}_K}^2$$

and

$$\hat{f}_{n,N} = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \left(f(x_{N,i}) - Y_i\right)^2 + \lambda \|f\|_{\mathcal{H}_K}^2 .$$

Then, exploiting convexity,

$$\lambda \|\hat{f}_n - \hat{f}_{n,N}\|_{\mathcal{H}_K}^2 \leq \text{scalar bound.}$$

Another proof ingredient

We are led to bound (in $\mathbb{R}$!) terms such as

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \left\{ \left[\hat{f}_n(x_i) - \hat{f}_{n,N}(x_i) \right] - \left[\hat{f}_n(x_{N,i}) - \hat{f}_{n,N}(x_{N,i}) \right] \right\}.$$

By coupling arguments, we approximate by

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \left\{ \left[\hat{f}_n(x_i) - \tilde{f}_{n,N}(x_i) \right] - \left[\hat{f}_n(x_{N,i}) - \tilde{f}_{n,N}(x_{N,i}) \right] \right\},$$

- with $\tilde{f}_{n,N}$ constructed from new independent $(\tilde{x}_{N,i})_{i=1}^{n}$.

Hence

- conditionally to $(\tilde{x}_{N,i}, \mu_i, Y_i)_{i=1}^{n}$,
- letting $(x_{N,i})_{i=1}^{n}$ be the only remaining source of randomness,

$\Longrightarrow$ we have a sum of independent variables.

Application to sufficient N for minimax rate (1/2)

- [Caponnetto and De Vito, 2007] provide minimax rates as $n \to \infty$ with one-stage sampling (for $\hat{f}_n$).
- Target: conditional expectation function

$$f^* = \mathbb{E}\left[Y_i | x_i = \cdot\right] \qquad \text{assumed to be in } \mathcal{H}_K.$$

- We let $\mathcal{L}$ be the distribution of x_i.

Problem class on $\mathcal{H}$

Hardness of $(\mathcal{L}, K, f^*)$ measured by

- $b > 1$ effective dimension of $\mathcal{H}_K$ w. r. t. distribution $\mathcal{L}$,
- $c \in (1, 2]$ complexity of f^*.

Minimax rate

$$\sqrt{\int_{\mathcal{H}} \left(f^\star(x) - \hat{f}_n(x)\right)^2 \mathrm{d}\mathcal{L}(x)} = \mathcal{O}_{\mathbb{P}}\left(n^{-\frac{bc}{2(bc+1)}}\right).$$

- With $\lambda = n^{-\frac{b}{bc+1}}$.

Application to sufficient N for minimax rate (2/2)

In [Bachoc et al., 2023b], from our bounds:

Sufficient N for minimax rate

$$\sqrt{\int_{\mathcal{H}} \left(f^\star(x) - \hat{f}_{n,N}(x)\right)^2 \mathrm{d}\mathcal{L}(x)} = \mathcal{O}_{\mathbb{P}}\left(n^{-\frac{bc}{2(bc+1)}}\right).$$

- With $\lambda = n^{-\frac{b}{bc+1}}$.
- With $N = n^a$,

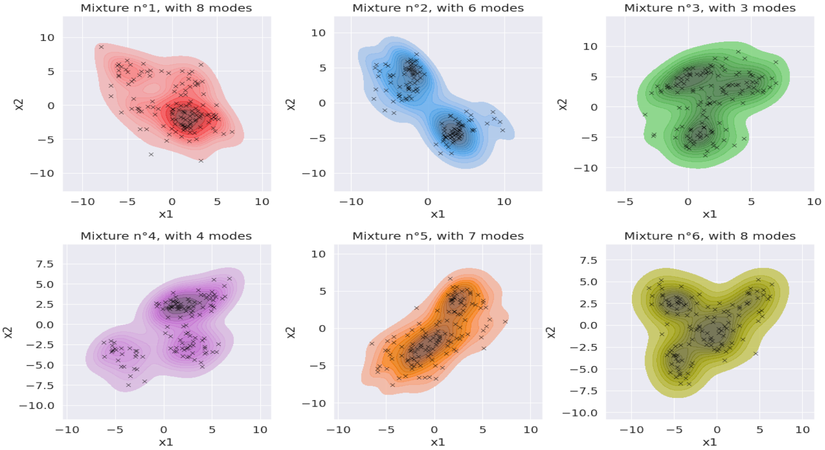
$$\begin{cases} a = \max\left(\frac{b+\frac{bc}{2}}{bc+1}, \frac{2b-1}{bc+1}, \frac{4b-bc-2}{bc+1}\right) \ (\leq 1) & \text{if } b(1-\frac{c}{2}) \leq \frac{3}{4} \\ a = \max\left(\frac{b+\frac{bc}{2}}{bc+1}, \frac{2b-\frac{1}{2}}{bc+1}\right) \ (> 1) & \text{if } b(1-\frac{c}{2}) > \frac{3}{4} \end{cases}.$$

In [Szabó et al., 2015, Szabó et al., 2016], same result for mean embedding with $N = n^{\frac{b(c+1)}{bc+1}}$,

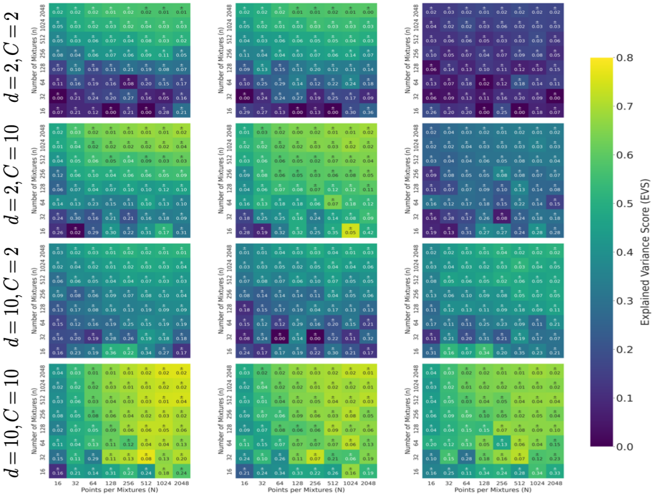
- $\frac{b(c+1)}{bc+1} > a$, for all b, c.

Numerical experiment: Gaussian mixtures (1/2)

- For $i = 1, \ldots, n$, μ_i is a Gaussian mixture with Y_i modes.
- For $i = 1, \ldots, n$ we observe N samples $X_{i,j}$ from μ_i.



Numerical experiment: Gaussian mixtures (2/2)



(a) **Sinkhorn-$\mathcal{U}$.** (b) **Sliced Wasserstein.** (c) **Mean embedding.**

$\Longrightarrow$ Score increases with n, N.

$\Longrightarrow$ There is a saturation effect of increasing N (e.g. bottom-left).

Conclusion

- Hilbertian embedding for (symmetric non-negative definite) kernels.
- Two-stage sampling as an additional source of error.
- Main contribution: tighter control of this error.
- The paper [Bachoc et al., 2023b]: `arXiv:2308.14335`.
- Paper [Bachoc et al., 2023a] on Sinkhorn kernel.
- Public Python codes (links in papers).

Thank you for your attention!

Bibliography I

Bachoc, F., Béthune, L., González-Sanz, A., and Loubes, J.-M. (2023a).
Gaussian processes on distributions based on regularized optimal transport.
In *International Conference on Artificial Intelligence and Statistics*.

Bachoc, F., Béthune, L., González-Sanz, A., and Loubes, J.-M. (2023b).
Improved learning theory for kernel distribution regression with two-stage sampling.
arXiv:2308.14335.

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

Cuturi, M. (2013).
Sinkhorn distances: Lightspeed computation of optimal transport.
Advances in Neural Information Processing Systems, 26.

Bibliography II

Dietterich, T., Lathrop, R., and Lozano-Pérez, T. (1997).
Solving the multiple instance problem with axis-parallel rectangles.
Artificial Intelligence, 89:31–71.

Flaxman, S. R., Wang, Y.-X., and Smola, A. J. (2015).
Who supported Obama in 2012? Ecological inference through distribution regression.
In *International Conference on Knowledge Discovery and Data Mining*.

Genevay, A. (2019).
Entropy-regularized optimal transport for machine learning.
PhD thesis, Paris Sciences et Lettres (ComUE).

González-Sanz, A., Loubes, J.-M., and Niles-Weed, J. (2022).
Weak limits of entropy regularized optimal transport; potentials, plans and divergences.
arXiv preprint arXiv:2207.07427.

Bibliography III

Kolouri, S., Rohde, G. K., and Hoffmann, H. (2018).
Sliced Wasserstein distance for learning Gaussian mixture models.
In *IEEE Conference on Computer Vision and Pattern Recognition*.

Manole, T., Balakrishnan, S., and Wasserman, L. (2022).
Minimax confidence intervals for the sliced Wasserstein distance.
Electronic Journal of Statistics, 16(1):2252–2345.

Meunier, D., Pontil, M., and Ciliberto, C. (2022).
Distribution regression with sliced Wasserstein kernels.
In *International Conference on Machine Learning*.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017).
Kernel mean embedding of distributions: A review and beyond.
Foundations and Trends® in Machine Learning, 10(1-2):1–141.

Ray, S. and Page, D. (2015).
Multiple instance regression.
In *International Conference on Machine Learning*.

Bibliography IV

Szabó, Z., Gretton, A., Póczos, B., and Sriperumbudur, B. (2015).
Two-stage sampled learning theory on distributions.
In *International Conference on Artificial Intelligence and Statistics*.

Szabó, Z., Sriperumbudur, B., Póczos, B., and Gretton, A. (2016).
Learning theory for distribution regression.
The Journal of Machine Learning Research, 17(1):5272–5311.