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

François Bachoc — University of Lille
Louis Béthune — Apple
Alberto Gonzàlez-Sanz — Columbia University
Jean-Michel Loubes — Regalia INRIA

Kernel regression on $\mathbb{R}^d$

We observe i.i.d. pairs

$$(x_i, Y_i), \qquad i = 1, \dots, n.$$

- $Y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^d$.

Goal: constructing a regression function

$$\boxed{\widehat{f}_n : \mathbb{R}^d \to \mathbb{R}.}$$

RKHS $\mathcal{H}_K$: space of functions from $\mathbb{R}^d$ to $\mathbb{R}$ based on kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

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

Ridge regression = minimizing empirical square error + complexity penalty.

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

with

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

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

Kernel distribution regression

We observe i.i.d. pairs

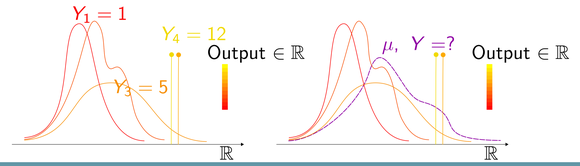
$$(\mu_i, Y_i), \qquad i = 1, \dots, n.$$

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

Goal: constructing a regression function

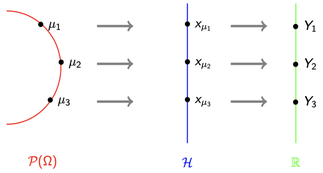
$$\widehat{f}_n : \mathcal{P}(\Omega) \to \mathbb{R},$$

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



Hilbertian embedding

$$x : \mathcal{P}(\Omega) \to \mathcal{H} \qquad \text{where } \mathcal{H} \text{ is a Hilbert space.}$$
$$\mu \mapsto x_\mu,$$



Two-stage sampling

Ridge regression on distributions:

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

with

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

- where $\mathcal{H}_K$ is the RKHS of a kernel K on $\mathcal{H}$.

Two-stage sampling: 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 $\mu_i^N = \frac{1}{N}\sum_{j=1}^{N} \delta_{X_{i,j}}$.

Approximate ridge regression:

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

with

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

Rates of convergence

Assumption: Near-unbiased condition.

$$\left\| \mathbb{E}\left[x_{\mu_i^N} \middle| \mu_i \right] - x_{\mu_i} \right\|_{\mathcal{H}} \quad \text{has order} \quad \frac{1}{N}.$$

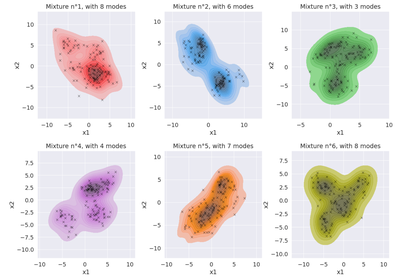
Holds for mean embedding, sliced Wasserstein and Sinkhorn.

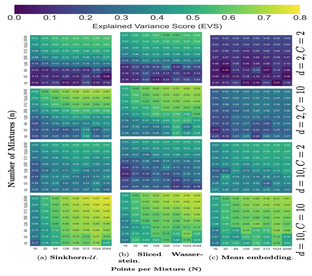
Theorem (improved rates of convergence).

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

Numerical illustration

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





Related references

- Bachoc, Béthune, Gonzàlez-Sanz, Loubes (2023). Gaussian processes on distributions based on regularized optimal transport. *AISTATS*.
- Meunier, Pontil, Ciliberto (2022). Distribution regression with sliced Wasserstein kernels. *ICML*.
- Szabó, Sriperumbudur, Póczos, Gretton (2016). Learning theory for distribution regression. *JMLR*.

Paper published in Annals of Statistics 53(4): 1753-1782 (2025) and available at `https://arxiv.org/abs/2308.14335`