The Minimax Lower Bound of Kernel Stein Discrepancy Estimation*

Jose Cribeiro-Ramallo[1] Agnideep Aich[2] Florian Kalinke[1] Ashit Baran Aich[3]
Zoltán Szabó[4]

[1]Karlsruhe Institute of Technology [2]University of Louisiana at Lafayette [3]Presidency College
[4]London School of Economics

Summary

- **Kernel Stein Discrepancy (KSD)**: popular measure to quantify the goodness-of-fit between a target and a sample distribution.
- **Applications**: model validation, learning variational models, testing, model comparison, distribution compression, and model explainability.
- **Domains**: discrete spaces, Riemannian manifolds, Hilbert spaces, point-processes, graph data.
- **Convergence rates**: $\mathcal{O}_P(n^{-1/2})$ under sub-Gaussian assumptions.
- **Contribution**: $n^{-1/2}$ rate is minimax optimal.

Example: Goodness-of-fit

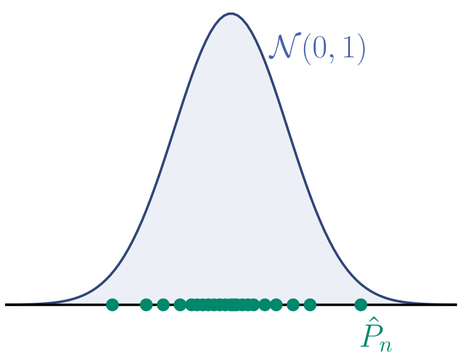

Figure 1. Gaussian—Fully known density

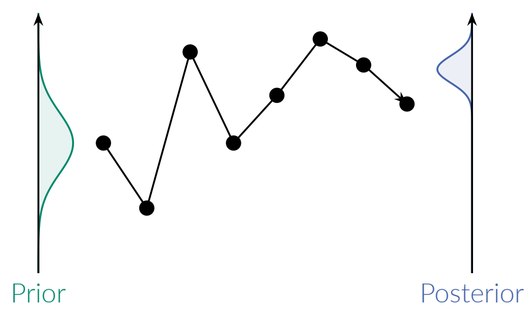

Figure 2. Posterior—Partially known density

Kernel Stein Discrepancy (KSD)

- $(\mathcal{X}, \tau)$: topological space; P_0, P: Borel probability measures on $\mathcal{X}$; $\mathcal{H}$: Hilbert space of functions on $\mathcal{X}$; $\Psi_{P_0} : \mathcal{X} \to \mathcal{H}$ mapping such that $\mathbb{E}_{P_0}[\Psi_{P_0}(X)] = 0$ holds.
- Stein kernel:

$$K_0(x, y) = \langle \Psi_{P_0}(x), \Psi_{P_0}(y) \rangle_{\mathcal{H}} \text{ for all } x, y \in \mathcal{X}.$$

- Squared KSD:

$$\mathrm{KSD}^2(P_0, P) = \mathbb{E}_{P \otimes P}\left[K_0\left(X, X'\right)\right].$$

- **Assumptions on $\mathbb{R}^d$:** $k \in \mathcal{C}^2\left(\mathbb{R}^d \times \mathbb{R}^d\right)$, $P_0 \ll \lambda$ with density p_0, $p_0 \in \mathcal{C}^1\left(\mathbb{R}^d\right)$, $p_0 > 0$, and $\lim_{\|\mathbf{x}\|_2 \to \infty} h(\mathbf{x}) p_0(\mathbf{x}) = 0$ for all $h \in \mathcal{H}_k$.
- **Assumptions on $(\mathcal{X}, \tau)$:** Ψ_{P_0} is measurable, $\mathbb{E}_{P_0}[\Psi_{P_0}(X)] = 0$, and $\mathcal{H}_{K_0}$ is separable.

Example

When $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{H} = \mathcal{H}_k$ is an RKHS, a feature map Ψ_{P_0} known to satisfy $\mathbb{E}_{P_0}[\Psi_{P_0}(X)] = 0$ is the Langevin-Stein feature map

$$\Psi_{P_0}(\mathbf{x}) \coloneqq \nabla_{\mathbf{x}}\left[\ln\left(p_0(\mathbf{x})\right)\right] k(\cdot, \mathbf{x}) + \nabla_{\mathbf{x}} k(\cdot, \mathbf{x}) \text{ for all } \mathbf{x} \in \mathbb{R}^d.$$

- **V-statistic estimator [1]:** $\widehat{\mathrm{KSD}}_V^2\left(P_0, P\right) = \frac{1}{n^2}\sum_{i,j=1}^n K_0\left(X_i, X_j\right)$; it converges with rate $n^{-1/2}$ [4].
- **Accelerated estimator [4]:** Attains the same rate of $n^{-1/2}$ under mild conditions.

Minimax Lower Bound

- **What is it?**: A lower bound of the rate attainable by any estimator $(\hat{F}_n)$ on the most challenging distribution pair (P_0, P), when $\mathrm{KSD}(P_0, P) < \infty$.
- $(a_n)_{n=1}^\infty$ is a lower bound of KSD estimation if there exists a $C > 0$ such that

$$\inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n\Big(\overbrace{\left|\mathrm{KSD}(P_0, P) - \hat{F}_n\right|}^{=:\hat{\Delta}_n} \geq C a_n\Big) > 0 \text{ for all } n \in \mathbb{N}_{>0},$$

best estimator ← → most challenging pair (P_0, P)

where

  - $\mathcal{T}$ is the space of P_0-s satisfying our assumptions, and
  - $\mathcal{S}_{P_0}$ is the space of P-s s.t. $\mathrm{KSD}(P_0, P) < \infty$.

- If there exists an estimator with a matching upper bound, we call it minimax optimal.

Main Challenge

Existing lower bounds (MMD [6], mean embedding [5], HSIC [3]) require the kernel (in this case, the Stein kernel) to be bounded. The Stein kernel is practically never bounded [2, 4].

Key Tools

Le Cam's Method [7]

- **If** for a suitable fixed $P_0 \in \mathcal{T}$, there exists an adversarial pair of distributions $(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}) \in \mathcal{S}_{P_0} \times \mathcal{S}_{P_0}$ s.t.

  (i) $\mathrm{KL}\left(\mathbb{P}_{\theta_1}^n, \mathbb{P}_{\theta_0}^n\right) \leq \alpha$,
  (ii) $\left|\mathrm{KSD}(P_0, \mathbb{P}_{\theta_1}) - \mathrm{KSD}(P_0, \mathbb{P}_{\theta_0})\right| \geq 2s_n$,

  with $\alpha > 0$, all $n \in \mathbb{N}_{>0}$, and $(s_n)_{n=1}^\infty$ positive,
- **then,** for all n,

$$\inf_{\hat{F}_n} \sup_{P \in \mathcal{S}_{P_0}} P^n\Big(\left|\mathrm{KSD}(P_0, P) - \hat{F}_n\right| \geq s_n\Big) \geq f(\alpha),$$

where $f(\alpha) = \max\left\{\exp(-\alpha)/4, (1 - \sqrt{\alpha/2})\right\} > 0$.

Bochner's Theorem [8]

The kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is continuous and translation-invariant $\iff$ $k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y}, \boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega})$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and some finite non-negative Borel measure Λ.

Local Perturbations

Given $P_0 \in \mathcal{T}$, one can define a probability measure P_n such that

$$P_n(A) = \int_A 1 + \varepsilon_n \varphi(x) \mathrm{d}P_0(x), \qquad (1)$$

with $\varepsilon_n = c/\sqrt{n}$ and some suitable $\varphi : \mathcal{X} \to \mathbb{R}$ measurable.

Adversarial Pairs

On $\mathbb{R}^d$

- Fix $P_0 = \mathcal{N}\left(\mathbf{0}_d, \mathbf{I}_d\right)$.
- The pair is

$$\left(\mathbb{P}_{\theta_0},\ \mathbb{P}_{\theta_1}\right) = \left(\mathcal{N}\left(\mathbf{0}_d, \mathbf{I}_d\right),\ \mathcal{N}\left(n^{-1/2}\mathbf{e}_j, \mathbf{I}_d\right)\right),$$

with $\mathbf{e}_j \in \mathbb{R}^d$ the j-th canonical basis vector.

On $(\mathcal{X}, \tau)$

- Fix $P_0 \in \mathcal{T}$.
- The pair is

$$\left(\mathbb{P}_{\theta_0},\ \mathbb{P}_{\theta_1}\right) = \left(P_0,\ P_n\right),$$

with P_n as in (1), and $\varphi \in \mathcal{C}_b(\mathcal{X})$ non-constant P_0-a.e. and such that $\mathbb{E}_{P_0}[\varphi(X)] = 0$.

Main Results

- $\hat{F}_n$: any estimator of $\mathrm{KSD}(P_0, P)$ using $n \in \mathbb{N}_{>0}$ samples from $\mathbb{P} \in \mathcal{S}_{\mathbb{P}_0}$.

On $\mathbb{R}^d$

Assume k is bounded, translation-invariant and characteristic. Then, there exists a universal constant $c > 0$ such that

$$\inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n\left(\hat{\Delta}_n \geq \frac{c}{\sqrt{n}}\right) > 0.$$

If k is Gaussian, that is,

$$k(\mathbf{x}, \mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|_2^2},$$

then $c = (4\gamma + 1)^{-d/4}/2$.

On $(\mathcal{X}, \tau)$

Assume there exists a $P_0 \in \mathcal{T}$ s.t. $\mathrm{KSD}(P_0, P) = 0 \iff P_0 = P$, and there exists a non-a.e. constant $\varphi_0 \in \mathcal{C}_b(\mathcal{X})$. Then, there exists an universal constant $B > 0$ such that

$$\liminf_{n \to \infty} \inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n\left(\hat{\Delta}_n \geq \frac{B}{\sqrt{n}}\right) > 0.$$

- **Implication:** The $n^{-1/2}$ rate of the V-statistic and the accelerated estimator is minimax optimal on general spaces, under mild conditions.

References

[1] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML)*, pages 2606–2615, 2016.

[2] Omar Hagrass, Bharath Sriperumbudur, and Krishnakumar Balasubramanian. Minimax optimal goodness-of-fit testing with kernel Stein discrepancy. *Bernoulli*, 2025. (accepted; preprint: `https://arxiv.org/abs/2404.08278`).

[3] Florian Kalinke and Zoltán Szabó. Nyström M-Hilbert-Schmidt independence criterion. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1005–1015, 2023.

[4] Florian Kalinke, Zoltán Szabó, and Bharath K. Sriperumbudur. Nyström kernel Stein discrepancy. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 388–396, 2025.

[5] Ilya Tolstikhin, Bharath K. Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(86):1–47, 2017.

[6] Ilya Tolstikhin, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximal mean discrepancy with radial kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1930–1938, 2016.

[7] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[8] Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, 2005.

Information Systems, IPD

*AISTATS-2026, Tangier, Morocco.

Corresponding author: jose.cribeiro@kit.edu