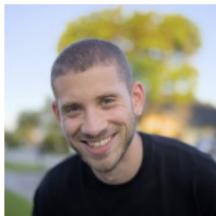


# The Minimax Lower Bound of Kernel Stein Discrepancy Estimation

Zoltán Szabó

Joint work with:

- Jose Cribreiro-Ramallo [@](#) KIT, Germany,
- Agnideep Aich [@](#) University of Louisiana at Lafayette, U.S.,
- Florian Kalinke [@](#) KIT, Germany,
- Ashit Baran Aich [@](#) formerly of Presidency College, India.



## Today: in a nutshell

- Kernel Stein discrepancy (KSD;  
[Chwialkowski et al., 2016, Liu et al., 2016, Hagrass et al., 2025]):
  - simple-to-estimate, popular goodness-of-fit measure,
  - defined on both  $\mathbb{R}^d$  and general domains,
  - with various successful applications.

## Today: in a nutshell

- Kernel Stein discrepancy (KSD;  
[Chwialkowski et al., 2016, Liu et al., 2016, Hagrass et al., 2025]):
  - simple-to-estimate, popular goodness-of-fit measure,
  - defined on both  $\mathbb{R}^d$  and general domains,
  - with various successful applications.
- Existing estimators:

convergence rate:  $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ ,  $n$  = sample size.

# Today: in a nutshell

- Kernel Stein discrepancy (KSD;  
[Chwialkowski et al., 2016, Liu et al., 2016, Hagrass et al., 2025]):
  - simple-to-estimate, popular goodness-of-fit measure,
  - defined on both  $\mathbb{R}^d$  and general domains,
  - with various successful applications.
- Existing estimators:

convergence rate:  $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ ,  $n$  = sample size.

## Focus

- Question: Can we go faster?

# Today: in a nutshell

- Kernel Stein discrepancy (KSD;  
[Chwialkowski et al., 2016, Liu et al., 2016, Hagrass et al., 2025]):
  - simple-to-estimate, popular goodness-of-fit measure,
  - defined on both  $\mathbb{R}^d$  and general domains,
  - with various successful applications.
- Existing estimators:

convergence rate:  $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ ,  $n$  = sample size.

## Focus

- Question: Can we go faster?
- Answer: No.

# Kernel (generalization of $\mathbf{a}^\top \mathbf{b}$ ), RKHS

[Aronszajn, 1950, Steinwart and Christmann, 2008]

- Def-1 (feature space):

$$k(a, b) = \langle \Phi(a), \Phi(b) \rangle_{\mathcal{H}}.$$

# Kernel (generalization of $\mathbf{a}^\top \mathbf{b}$ ), RKHS

[Aronszajn, 1950, Steinwart and Christmann, 2008]

- Def-1 (feature space):

$$k(a, b) = \langle \Phi(a), \Phi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, b) \in \mathcal{H}, \quad f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}.$$

# Kernel (generalization of $\mathbf{a}^\top \mathbf{b}$ ), RKHS

[Aronszajn, 1950, Steinwart and Christmann, 2008]

- Def-1 (feature space):

$$k(a, b) = \langle \Phi(a), \Phi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, b) \in \mathcal{H}, \quad f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix):  $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n} \succeq 0$ .

# Kernel (generalization of $\mathbf{a}^\top \mathbf{b}$ ), RKHS

[Aronszajn, 1950, Steinwart and Christmann, 2008]

- Def-1 (feature space):

$$k(a, b) = \langle \Phi(a), \Phi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, b) \in \mathcal{H}, \quad f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix):  $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n} \succeq 0$ .

## Notes

- $k \xleftrightarrow{1:1} \mathcal{H}_k = \overline{\text{Span}}(k(\cdot, x) : x \in \mathcal{X})$ : Fourier analysis, polynomials, splines, ...
- Examples:  $k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p$ ,  $k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}$ .

# Towards Langevin-Stein KSD on $\mathbb{R}^d$

- Aim: compare 2 distributions  $P_0, P \in \mathcal{M}_1^+(\mathbb{R}^d)$ .
- Assumptions:
  - $P_0$ : target distribution; fixed and known,

# Towards Langevin-Stein KSD on $\mathbb{R}^d$

- Aim: compare 2 distributions  $P_0, P \in \mathcal{M}_1^+(\mathbb{R}^d)$ .
- Assumptions:
  - $P_0$ : **target distribution**; fixed and known,
  - $P$ : **sampling distribution**; unknown, but we have samples from it.

# Towards Langevin-Stein KSD on $\mathbb{R}^d$

- Aim: compare 2 distributions  $P_0, P \in \mathcal{M}_1^+(\mathbb{R}^d)$ .
- Assumptions:
  - $P_0$ : **target distribution**; fixed and known,
  - $P$ : **sampling distribution**; unknown, but we have samples from it.
  - $P_0, P \ll \lambda_d$ : corresponding pdf-s =  $p_0, p$ .

# Towards Langevin-Stein KSD on $\mathbb{R}^d$

- Aim: compare 2 distributions  $P_0, P \in \mathcal{M}_1^+(\mathbb{R}^d)$ .
- Assumptions:
  - $P_0$ : target distribution; fixed and known,
  - $P$ : sampling distribution; unknown, but we have samples from it.
  - $P_0, P \ll \lambda_d$ : corresponding pdf-s =  $p_0, p$ .
  - $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  kernel.

# KSD on $\mathbb{R}^d$

KSD is a specific IPM:  $\mathcal{F} = \left\{ \mathcal{A}_{p_0} \mathbf{f} : \mathbf{f} \in B(\mathcal{H}_k^d) \right\}$

$$\text{KSD}(P_0, P) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{P_0}[f(X)] - \mathbb{E}_P[f(X)]|$$

KSD is a specific IPM:  $\mathcal{F} = \left\{ \mathcal{A}_{p_0} \mathbf{f} : \mathbf{f} \in B(\mathcal{H}_k^d) \right\}$

$$\begin{aligned} \text{KSD}(P_0, P) &= \sup_{\mathbf{f} \in \mathcal{F}} |\mathbb{E}_{P_0}[f(X)] - \mathbb{E}_P[f(X)]| \\ &= \sup_{\mathbf{f} \in B(\mathcal{H}_k^d)} |\mathbb{E}_{P_0}[(\mathcal{A}_{p_0} \mathbf{f})(X)] - \mathbb{E}_P[(\mathcal{A}_{p_0} \mathbf{f})(X)]| \end{aligned}$$

with  $\mathcal{A}_{p_0}$  designed to satisfy the **mean-zero property**:

$$\mathbb{E}_{P_0}[(\mathcal{A}_{p_0} \mathbf{f})(X)] = 0 \text{ for all } \mathbf{f} \in B(\mathcal{H}_k^d).$$

KSD is a specific IPM:  $\mathcal{F} = \left\{ \mathcal{A}_{p_0} \mathbf{f} : \mathbf{f} \in B(\mathcal{H}_k^d) \right\}$

$$\begin{aligned} \text{KSD}(P_0, P) &= \sup_{\mathbf{f} \in \mathcal{F}} |\mathbb{E}_{P_0}[f(X)] - \mathbb{E}_P[f(X)]| \\ &= \sup_{\mathbf{f} \in B(\mathcal{H}_k^d)} |\mathbb{E}_{P_0}[(\mathcal{A}_{p_0} \mathbf{f})(X)] - \mathbb{E}_P[(\mathcal{A}_{p_0} \mathbf{f})(X)]| \\ &= \sup_{\mathbf{f} \in B(\mathcal{H}_k^d)} \mathbb{E}_P[(\mathcal{A}_{p_0} \mathbf{f})(X)], \end{aligned}$$

with  $\mathcal{A}_{p_0}$  designed to satisfy the **mean-zero property**:

$$\mathbb{E}_{P_0}[(\mathcal{A}_{p_0} \mathbf{f})(X)] = 0 \text{ for all } \mathbf{f} \in B(\mathcal{H}_k^d).$$

# Langevin-Stein operator

For instance, the Langevin-Stein operator does the job:

$$(\mathcal{A}_{p_0} \mathbf{f})(\mathbf{x}) = \langle \nabla_{\mathbf{x}} \ln(p_0(\mathbf{x})), \mathbf{f}(\mathbf{x}) \rangle_2 + \sum_{j=1}^d \frac{\partial f_j(\mathbf{x})}{\partial x_j},$$

hence

- one assumes:

①  $p_0 > 0$

# Langevin-Stein operator

For instance, the Langevin-Stein operator does the job:

$$(\mathcal{A}_{p_0} \mathbf{f})(\mathbf{x}) = \langle \nabla_{\mathbf{x}} \ln(p_0(\mathbf{x})), \mathbf{f}(\mathbf{x}) \rangle_2 + \sum_{j=1}^d \frac{\partial f_j(\mathbf{x})}{\partial x_j},$$

hence

- one assumes:

- ①  $p_0 > 0,$
- ②  $p_0 \in \mathcal{C}^1(\mathbb{R}^d)$

# Langevin-Stein operator

For instance, the Langevin-Stein operator does the job:

$$(\mathcal{A}_{p_0} \mathbf{f})(\mathbf{x}) = \langle \nabla_{\mathbf{x}} \ln(p_0(\mathbf{x})), \mathbf{f}(\mathbf{x}) \rangle_2 + \sum_{j=1}^d \frac{\partial f_j(\mathbf{x})}{\partial x_j},$$

hence

- one assumes:
  - ①  $p_0 > 0$ ,
  - ②  $p_0 \in \mathcal{C}^1(\mathbb{R}^d)$ ,
  - ③  $\lim_{\|\mathbf{x}\|_2 \rightarrow \infty} h(\mathbf{x}) p_0(\mathbf{x}) = 0$  for all  $h \in \mathcal{H}_k$  (to mean zero property).  
[ $\Leftarrow p_0$ : bounded,  $\lim_{\|\mathbf{x}\|_2 \rightarrow \infty} h(\mathbf{x}) = 0$  for all  $h \in \mathcal{H}_k$ ]

# Langevin-Stein operator

For instance, the Langevin-Stein operator does the job:

$$(\mathcal{A}_{p_0} \mathbf{f})(\mathbf{x}) = \langle \nabla_{\mathbf{x}} \ln(p_0(\mathbf{x})), \mathbf{f}(\mathbf{x}) \rangle_2 + \sum_{j=1}^d \frac{\partial f_j(\mathbf{x})}{\partial x_j},$$

hence

- one assumes:
  - ①  $p_0 > 0$ ,
  - ②  $p_0 \in \mathcal{C}^1(\mathbb{R}^d)$ ,
  - ③  $\lim_{\|\mathbf{x}\|_2 \rightarrow \infty} h(\mathbf{x}) p_0(\mathbf{x}) = 0$  for all  $h \in \mathcal{H}_k$  (to mean zero property).  
[ $\Leftarrow p_0$ : bounded,  $\lim_{\|\mathbf{x}\|_2 \rightarrow \infty} h(\mathbf{x}) = 0$  for all  $h \in \mathcal{H}_k$ ]
- $Cp_0$ : OK with  $C > 0$ .

## Langevin-Stein KSD

Assuming further that  $k \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d)$ , KSD takes a nice form:

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

with the so-called **Stein kernel**

# Langevin-Stein KSD

Assuming further that  $k \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d)$ , KSD takes a nice form:

$$\text{KSD}^2(P_0, P) = \mathbb{E}_{P \otimes P} [\textcolor{red}{K}_0(X, X')],$$

with the so-called **Stein kernel**

$$\begin{aligned}\textcolor{red}{K}_0(\mathbf{x}, \mathbf{y}) &= \langle \nabla_{\mathbf{x}} \ln(p_0(\mathbf{x})), \nabla_{\mathbf{y}} \ln(p_0(\mathbf{y})) \rangle_2 k(\mathbf{x}, \mathbf{y}) \\ &\quad + \langle \nabla_{\mathbf{y}} \ln(p_0(\mathbf{y})), \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{y}) \rangle_2 \\ &\quad + \langle \nabla_{\mathbf{x}} \ln(p_0(\mathbf{x})), \nabla_{\mathbf{y}} k(\mathbf{x}, \mathbf{y}) \rangle_2 \\ &\quad + \sum_{j=1}^d \frac{\partial^2 k(\mathbf{x}, \mathbf{y})}{\partial x_j \partial y_j}.\end{aligned}$$

# Langevin-Stein KSD: assumptions so far ( $\equiv$ A1)

- ①  $P_0 \ll \lambda_d$ , with pdf
  - ①  $p_0 > 0$ ,
  - ②  $p_0 \in \mathcal{C}^1(\mathbb{R}^d)$ ,
  - ③  $\lim_{\|\mathbf{x}\|_2 \rightarrow \infty} h(\mathbf{x})p_0(\mathbf{x}) = 0$  for all  $h \in \mathcal{H}_k$ ,
- ②  $k \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d)$  kernel.

# KSD applications & domains

## Applications:

- **model validation** [Gorham and Mackey, 2017, Futami et al., 2019, Hodgkinson et al., 2021, Wang et al., 2023],
- **learning variational models** [Liu and Wang, 2016, Liu and Wang, 2018, Chen et al., 2018, Chen et al., 2019, Korba et al., 2020, Korba et al., 2021],
- **testing** [Liu et al., 2016, Chwialkowski et al., 2016, Schrab et al., 2022, Baum et al., 2023, Hagrass et al., 2025],
- **model comparison** [Lim et al., 2019, Kanagawa et al., 2020],
- **model explainability** [Sarvmaili et al., 2025].

# KSD applications & domains

## Applications:

- **model validation** [Gorham and Mackey, 2017, Futami et al., 2019, Hodgkinson et al., 2021, Wang et al., 2023],
- **learning variational models** [Liu and Wang, 2016, Liu and Wang, 2018, Chen et al., 2018, Chen et al., 2019, Korba et al., 2020, Korba et al., 2021],
- **testing** [Liu et al., 2016, Chwialkowski et al., 2016, Schrab et al., 2022, Baum et al., 2023, Hagrass et al., 2025],
- **model comparison** [Lim et al., 2019, Kanagawa et al., 2020],
- **model explainability** [Sarvmaili et al., 2025].

## Domains:

- **discrete spaces** [Yang et al., 2018], **Riemannian manifolds** [Xu and Matsuda, 2020, Xu and Matsuda, 2021, Barp et al., 2022], **Hilbert spaces** [Wynne et al., 2025], **point processes** [Yang et al., 2019], **graph data** [Xu and Reinert, 2021].

# KSD on general domains

Key to KSD

mean-zero property.

Let

- $(\mathcal{X}, \tau_{\mathcal{X}})$ : topological space;  $P_0, P \in \mathcal{M}_1^+(\mathcal{X})$

# KSD on general domains

Key to KSD

mean-zero property.

Let

- $(\mathcal{X}, \tau_{\mathcal{X}})$ : topological space;  $P_0, P \in \mathcal{M}_1^+(\mathcal{X})$ ,
- $\mathcal{H}$ : a Hilbert space of functions on  $\mathcal{X}$

# KSD on general domains

Key to KSD

mean-zero property.

Let

- $(\mathcal{X}, \tau_{\mathcal{X}})$ : topological space;  $P_0, P \in \mathcal{M}_1^+(\mathcal{X})$ ,
- $\mathcal{H}$ : a Hilbert space of functions on  $\mathcal{X}$ ,
- $\Psi_{P_0} : \mathcal{X} \rightarrow \mathcal{H}$  measurable s.t. the mean-zero property holds:

$$\mathbb{E}_{P_0}[\Psi_{P_0}(X)] = 0.$$

# KSD on general domains

Key to KSD

mean-zero property.

Let

- $(\mathcal{X}, \tau_{\mathcal{X}})$ : topological space;  $P_0, P \in \mathcal{M}_1^+(\mathcal{X})$ ,
- $\mathcal{H}$ : a Hilbert space of functions on  $\mathcal{X}$ ,
- $\Psi_{P_0} : \mathcal{X} \rightarrow \mathcal{H}$  measurable s.t. the mean-zero property holds:

$$\mathbb{E}_{P_0}[\Psi_{P_0}(X)] = 0.$$

Then, an IPM construction gives

$$\text{KSD}^2(P_0, P) := \mathbb{E}_{P \otimes P}[K_0(X, X')], \quad K_0(x, x') := \langle \Psi_{P_0}(x), \Psi_{P_0}(x') \rangle_{\mathcal{H}}.$$

# KSD on general domains

## Key to KSD

mean-zero property.

Let (A1': below and  $\mathcal{H}_{K_0}$  separable)

- $(\mathcal{X}, \tau_{\mathcal{X}})$ : topological space;  $P_0, P \in \mathcal{M}_1^+(\mathcal{X})$ ,
- $\mathcal{H}$ : a Hilbert space of functions on  $\mathcal{X}$ ,
- $\Psi_{P_0} : \mathcal{X} \rightarrow \mathcal{H}$  measurable s.t. the mean-zero property holds:

$$\mathbb{E}_{P_0}[\Psi_{P_0}(X)] = 0.$$

Then, an IPM construction gives

$$\text{KSD}^2(P_0, P) := \mathbb{E}_{P \otimes P}[K_0(X, X')], \quad K_0(x, x') := \langle \Psi_{P_0}(x), \Psi_{P_0}(x') \rangle_{\mathcal{H}}.$$

$$[\text{Spec.: } \mathcal{X} = \mathbb{R}^d, \mathcal{H} = \mathcal{H}_k^d, \Psi_{P_0}(\mathbf{x}) = \nabla_{\mathbf{x}} [\ln(p_0(\mathbf{x}))] k(\cdot, \mathbf{x}) + \nabla_{\mathbf{x}} k(\cdot, \mathbf{x}) \in \mathcal{H}_k^d.]$$

# KSD estimators

Thanks to

$$\text{KSD}^2(P_0, P) = \mathbb{E}_{P \otimes P}[K_0(X, X')] = \|\mathbb{E}_P[K_0(\cdot, X)]\|_{\mathcal{H}_{K_0}}^2.$$

- ① V-statistic estimator: Replacing  $P$  with  $\hat{P}_n = \frac{1}{n} \sum_{j=1}^n \delta_{X_j}$  gives

$$\widehat{\text{KSD}}_V^2(P_0, P) := \text{KSD}^2(P_0, \hat{P}_n) = \frac{1}{n^2} \sum_{a,b=1}^n K_0(X_a, X_b).$$

# KSD estimators: Nyström acceleration

Thanks to

$$\text{KSD}^2(P_0, P) = \mathbb{E}_{P \otimes P}[K_0(X, X')] = \|\mathbb{E}_P[K_0(\cdot, X)]\|_{\mathcal{H}_{K_0}}^2.$$

② Nyström KSD estimator:

① subsample with replacement:  $\{\{\tilde{X}_1, \dots, \tilde{X}_m\}\} \Rightarrow \text{subspace:}$

$$\mathcal{H}_{K_0, m} = \text{Span}(K_0(\cdot, \tilde{X}_j) : j \in [m]) \subset \mathcal{H}_{K_0}.$$

# KSD estimators: Nyström acceleration

Thanks to

$$\text{KSD}^2(P_0, P) = \mathbb{E}_{P \otimes P}[K_0(X, X')] = \|\mathbb{E}_P[K_0(\cdot, X)]\|_{\mathcal{H}_{K_0}}^2.$$

② Nyström KSD estimator:

① subsample with replacement:  $\{\{\tilde{X}_1, \dots, \tilde{X}_m\}\} \Rightarrow$  subspace:

$$\mathcal{H}_{K_0, m} = \text{Span}(K_0(\cdot, \tilde{X}_j) : j \in [m]) \subset \mathcal{H}_{K_0}.$$

② we approximate  $\mathbb{E}_{\hat{P}_n}[K_0(\cdot, X)]$  from  $\mathcal{H}_{K_0}$ , min-norm solution of

$$\min_{\alpha = (\alpha_j)_{j=1}^m \in \mathbb{R}^m} \left\| \mathbb{E}_{\hat{P}_n}[K_0(\cdot, X)] - \sum_{j=1}^m \alpha_j K_0(\cdot, \tilde{X}_j) \right\|_{\mathcal{H}_{K_0}}.$$

## KSD estimators: Nyström acceleration – continued

② Nyström KSD estimator:

③ resulting squared KSD estimator:

$$\widehat{\text{KSD}}_N^2(P_0, P) = \left\| \sum_{j=1}^m \hat{\alpha}_j K_0(\cdot, \tilde{X}_j) \right\|_{\mathcal{H}_{K_0}}^2.$$

# KSD estimators: Nyström acceleration – continued

② Nyström KSD estimator:

③ resulting squared KSD estimator:

$$\widehat{\text{KSD}}_N^2(P_0, P) = \left\| \sum_{j=1}^m \hat{\alpha}_j K_0(\cdot, \tilde{X}_j) \right\|_{\mathcal{H}_{K_0}}^2.$$

④ simple form in terms of Gram matrices:

$$\begin{aligned}\widehat{\text{KSD}}_N^2(P_0, P) &= \beta^\top \mathbf{K}_{m,m}^- \beta, & \beta &= \frac{1}{n} \mathbf{K}_{m,n} \mathbf{1}_n, \\ \mathbf{K}_{m,m} &= [K_0(\tilde{X}_a, \tilde{X}_b)]_{a,b=1}^m, & \mathbf{K}_{m,n} &= [K_0(\tilde{X}_a, X_b)]_{a,b=1}^{m,n}.\end{aligned}$$

# KSD estimators: Nyström acceleration – continued

② Nyström KSD estimator:

③ resulting squared KSD estimator:

$$\widehat{\text{KSD}}_N^2(P_0, P) = \left\| \sum_{j=1}^m \hat{\alpha}_j K_0(\cdot, \tilde{X}_j) \right\|_{\mathcal{H}_{K_0}}^2.$$

④ simple form in terms of Gram matrices:

$$\begin{aligned}\widehat{\text{KSD}}_N^2(P_0, P) &= \beta^\top \mathbf{K}_{m,m}^- \beta, & \beta &= \frac{1}{n} \mathbf{K}_{m,n} \mathbf{1}_n, \\ \mathbf{K}_{m,m} &= [K_0(\tilde{X}_a, \tilde{X}_b)]_{a,b=1}^m, & \mathbf{K}_{m,n} &= [K_0(\tilde{X}_a, X_b)]_{a,b=1}^{m,n}.\end{aligned}$$

Both (and [Li et al., 2024]) converge at a rate  $\mathcal{O}(n^{-1/2})$  [Kalinke et al., 2025].  
Can we get faster?

# Challenge

Existing minimax lower bounds:

- ② mean embedding [Tolstikhin et al., 2017]

$$\mu_k(P) := \int_{\mathcal{X}} k(\cdot, x) dP(x) \in \mathcal{H}_k.$$

# Challenge

Existing minimax lower bounds:

- ① maximum mean discrepancy [Tolstikhin et al., 2016]

$$\text{MMD}_k(P, Q) := \|\mu_k(\textcolor{red}{P}) - \mu_k(\textcolor{blue}{Q})\|_{\mathcal{H}_k}.$$

- ② mean embedding [Tolstikhin et al., 2017]

$$\mu_k(P) := \int_{\mathcal{X}} k(\cdot, x) dP(x) \in \mathcal{H}_k.$$

# Challenge

Existing minimax lower bounds:

- ① maximum mean discrepancy [Tolstikhin et al., 2016]

$$\text{MMD}_k(P, Q) := \|\mu_k(\textcolor{red}{P}) - \mu_k(\textcolor{blue}{Q})\|_{\mathcal{H}_k}.$$

- ② mean embedding [Tolstikhin et al., 2017]

$$\mu_k(P) := \int_{\mathcal{X}} k(\cdot, x) dP(x) \in \mathcal{H}_k.$$

- ③ covariance operator [Zhou et al., 2019]

# Challenge

Existing minimax lower bounds:

- ① maximum mean discrepancy [Tolstikhin et al., 2016]

$$\text{MMD}_k(P, Q) := \|\mu_k(\textcolor{red}{P}) - \mu_k(\textcolor{blue}{Q})\|_{\mathcal{H}_k}.$$

- ② mean embedding [Tolstikhin et al., 2017]

$$\mu_k(P) := \int_{\mathcal{X}} k(\cdot, x) dP(x) \in \mathcal{H}_k.$$

- ③ covariance operator [Zhou et al., 2019],
- ④ Hilbert-Schmidt independence criterion [Kalinke and Szabó, 2024]

$$\text{HSIC}_{\textcolor{red}{k}}(P) := \text{MMD}_{\textcolor{red}{k}} \left( \textcolor{red}{P}, \bigotimes_{m=1}^M P_m \right), \quad \textcolor{red}{k} := \bigotimes_{m=1}^M k_m.$$

# Challenge

Existing minimax lower bounds:

- ① maximum mean discrepancy [Tolstikhin et al., 2016]

$$\text{MMD}_k(P, Q) := \|\mu_k(\textcolor{red}{P}) - \mu_k(\textcolor{blue}{Q})\|_{\mathcal{H}_k}.$$

- ② mean embedding [Tolstikhin et al., 2017]

$$\mu_k(P) := \int_{\mathcal{X}} k(\cdot, x) dP(x) \in \mathcal{H}_k.$$

- ③ covariance operator [Zhou et al., 2019],
- ④ Hilbert-Schmidt independence criterion [Kalinke and Szabó, 2024]

$$\text{HSIC}_{\textcolor{red}{k}}(P) := \text{MMD}_{\textcolor{red}{k}} \left( \textcolor{red}{P}, \bigotimes_{m=1}^M P_m \right), \quad \textcolor{red}{k} := \bigotimes_{m=1}^M k_m.$$

However

They all assume bounded kernel; the Stein kernel  $K_0$  is (typically) *not so*.

# Minimax estimation

Minimax risk with estimator  $\hat{F}_n(X_{1:n})$  and  $X_{1:n} \sim P^n$ , with  $P_0$  given:

$$R_n^* = \inf_{\hat{F}_n} \underbrace{\sup_{P_0 \in \mathcal{T}}}_{\text{A1}} \underbrace{\sup_{P \in \mathcal{S}_{P_0}}}_{\text{KSD}(P, P_0) < \infty} \mathbb{E}_{P^n} \left[ \underbrace{\hat{F}_n(X_{1:n}) - \text{KSD}(P_0, P)}_{\text{expected error}} \right].$$

# Minimax estimation

Minimax risk with estimator  $\hat{F}_n(X_{1:n})$  and  $X_{1:n} \sim P^n$ , with  $P_0$  given:

$$R_n^* = \underbrace{\inf_{\hat{F}_n}}_{\text{best estimator}} \underbrace{\sup_{P_0 \in \mathcal{T}}}_{\substack{\text{worst case} \\ \text{A1}}} \underbrace{\sup_{P \in \mathcal{S}_{P_0}}}_{\text{KSD}(P, P_0) < \infty} \mathbb{E}_{P^n} \left[ \underbrace{\hat{F}_n(X_{1:n})}_{\text{estimator}} - \underbrace{\text{KSD}(P_0, P)}_{\text{target}} \right].$$

In short

$R_n^*$  = **smallest** possible **maximum** risk achievable by *any* estimator.

# Minimax estimation

Minimax risk with estimator  $\hat{F}_n(X_{1:n})$  and  $X_{1:n} \sim P^n$ , with  $P_0$  given:

$$R_n^* = \inf_{\hat{F}_n} \underbrace{\sup_{P_0 \in \mathcal{T}}}_{\text{best estimator}} \underbrace{\sup_{P \in \mathcal{S}_{P_0}}}_{\substack{\text{worst case} \\ \text{KSD}(P, P_0) < \infty}} \mathbb{E}_{P^n} \left[ \underbrace{\hat{F}_n(X_{1:n}) - \text{KSD}(P_0, P)}_{\substack{\text{estimator} \\ \text{target}}} \right] .$$

Aim (after Markov)

Sequence  $s_n [= \Theta(n^{-1/2})] > 0$  such that

$$s_n^{-1} R_n^* \geq \inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n \left( \left| \hat{F}_n(X_{1:n}) - \text{KSD}(P_0, P) \right| \geq s_n \right) > 0.$$

# Minimax estimation

Minimax risk with estimator  $\hat{F}_n(X_{1:n})$  and  $X_{1:n} \sim P^n$ , with  $P_0$  given:

$$R_n^* = \underbrace{\inf_{\hat{F}_n}}_{\text{best estimator}} \underbrace{\sup_{P_0 \in \mathcal{T}}}_{\substack{\text{worst case} \\ \text{A1}}} \underbrace{\sup_{P \in \mathcal{S}_{P_0}}}_{\text{KSD}(P, P_0) < \infty} \mathbb{E}_{P^n} \left[ \underbrace{\hat{F}_n(X_{1:n}) - \text{KSD}(P_0, P)}_{\substack{\text{estimator} \\ \text{target} \\ \text{expected error}}} \right].$$

Aim (after Markov)

Sequence  $s_n [= \Theta(n^{-1/2})] > 0$  such that

$$s_n^{-1} R_n^* \geq \inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n \left( \left| \hat{F}_n(X_{1:n}) - \text{KSD}(P_0, P) \right| \geq s_n \right) > 0.$$

Known KSD estimation rates:  $\mathcal{O}(n^{-1/2}) \Rightarrow$  previous estimators: rate-optimal.

# Main result on $\mathbb{R}^d$

## Theorem

Assume:

- A1

# Main result on $\mathbb{R}^d$

## Theorem

Assume:

- A1,
- $k$ : bounded and shift-invariant (with A1  $\Rightarrow$  Bochner kernel)

## Theorem

Assume:

- A1,
- $k$ : bounded and shift-invariant (with A1  $\Rightarrow$  Bochner kernel),
- characteristic.

# Main result on $\mathbb{R}^d$

## Theorem

Assume:

- A1,
- $k$ : bounded and shift-invariant (with A1  $\Rightarrow$  Bochner kernel),
- characteristic.

Then,  $\exists c > 0$  s.t. for all  $n \in \mathbb{Z}_+$ ,

$$\inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n \left( \left| \hat{F}_n(X_{1:n}) - \text{KSD}(P_0, P) \right| \geq \frac{c}{\sqrt{n}} \right) > 0.$$

# Main result on $\mathbb{R}^d$

## Theorem

Assume:

- A1,
- $k$ : bounded and shift-invariant (with A1  $\Rightarrow$  Bochner kernel),
- characteristic.

Then,  $\exists c > 0$  s.t. for all  $n \in \mathbb{Z}_+$ ,

$$\inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n \left( \left| \hat{F}_n(X_{1:n}) - \text{KSD}(P_0, P) \right| \geq \frac{c}{\sqrt{n}} \right) > 0.$$

Note: with  $k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}$ ,  $c = (4\gamma + 1)^{-d/4}/2$ .

# Main result on general domain

## Theorem

Assume:

- A1'

# Main result on general domain

## Theorem

Assume:

- A1',  
*constant functions on  $\mathcal{X}$*
- $\mathcal{C}_b(\mathcal{X}) \setminus \overbrace{\mathcal{F}_c(\mathcal{X})} \neq \emptyset$

# Main result on general domain

## Theorem

Assume:

- A1',  
*constant functions on  $\mathcal{X}$*
- $\mathcal{C}_b(\mathcal{X}) \setminus \overbrace{\mathcal{F}_c(\mathcal{X})} \neq \emptyset$ ,
- KSD is valid.

# Main result on general domain

## Theorem

Assume:

- $A1'$ ,  
*constant functions on  $\mathcal{X}$*
- $\mathcal{C}_b(\mathcal{X}) \setminus \overbrace{\mathcal{F}_c(\mathcal{X})}^{\neq \emptyset} \neq \emptyset$ ,
- $KSD$  is valid.

Then,  $\exists B > 0$  and  $n_0 \in \mathbb{Z}_+$  s.t. for all  $n \geq n_0$

$$\inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n \left( \left| \hat{F}_n(X_{1:n}) - \text{KSD}(P_0, P) \right| \geq \frac{B}{\sqrt{n}} \right) > 0.$$

# Proof idea: design of (a sequence of) adversarial pairs

Need to the Le Cam's method:

- ①  $\text{KSD}(P_0, P_n) \geq Cn^{-1/2}$  with  $C > 0$ ,
- ②  $\text{KL}(P_0^n \| P_n^n) \leq \alpha$  with  $\alpha > 0$ .



## Proof idea: general case

Like  $p_n = (1 + \epsilon_n \varphi) p_0$

## Proof idea: general case

Like  $p_n = (1 + \epsilon_n \varphi) p_0$

$P_n$  is a **perturbation** of  $P_0$

$$P_n(A) = \int_A 1 + \epsilon_n \varphi(x) dP_0(x), \quad \forall A \in \mathcal{B}(\mathcal{X}),$$

with  $\varphi \in \mathcal{C}_b(\mathcal{X}) \setminus \mathcal{F}_c(\mathcal{X})$ ,  $\mathbb{E}_{P_0}[\varphi(X)] = 0$ ,  $\varphi \not\equiv 0$ ,  $\epsilon_n = cn^{-1/2}$ ,  $c > 0$ .

## Proof idea: general case

Like  $p_n = (1 + \epsilon_n \varphi) p_0$

$P_n$  is a **perturbation** of  $P_0$

$$P_n(A) = \int_A 1 + \epsilon_n \varphi(x) dP_0(x), \quad \forall A \in \mathcal{B}(\mathcal{X}),$$

with  $\varphi \in \mathcal{C}_b(\mathcal{X}) \setminus \mathcal{F}_c(\mathcal{X})$ ,  $\mathbb{E}_{P_0}[\varphi(X)] = 0$ ,  $\varphi \not\equiv 0$ ,  $\epsilon_n = cn^{-1/2}$ ,  $c > 0$ .

Control of the terms

- ❶  $\text{KSD}(P_0, P_n) = cn^{-1/2} C_\varphi > 0$  [validness of KSD,  $\mathbb{E}_{P_0}[\varphi(X)] = 0$ ]

## Proof idea: general case

Like  $p_n = (1 + \epsilon_n \varphi) p_0$

$P_n$  is a **perturbation** of  $P_0$

$$P_n(A) = \int_A 1 + \epsilon_n \varphi(x) dP_0(x), \quad \forall A \in \mathcal{B}(\mathcal{X}),$$

with  $\varphi \in \mathcal{C}_b(\mathcal{X}) \setminus \mathcal{F}_c(\mathcal{X})$ ,  $\mathbb{E}_{P_0}[\varphi(X)] = 0$ ,  $\varphi \not\equiv 0$ ,  $\epsilon_n = cn^{-1/2}$ ,  $c > 0$ .

### Control of the terms

- ①  $\text{KSD}(P_0, P_n) = cn^{-1/2} C_\varphi > 0$  [validness of KSD,  $\mathbb{E}_{P_0}[\varphi(X)] = 0$ ],
- ②  $\text{KL}(P_0^n \| P_n^n) \leq cM$ ,  $M := \mathbb{E}_{P_0}[\varphi^2(X)]$  [In properties,  $\mathbb{E}_{P_0}[\varphi(X)] = 0$ ].

## Summary

- Focus: quantifying goodness-of-fit with KSD.
- KSD can not be estimated faster than  $n^{-1/2}$ .
- **Paper** on arXiv.

# Summary

- Focus: quantifying goodness-of-fit with KSD.
- KSD can not be estimated faster than  $n^{-1/2}$ .
- **Paper** on arXiv.



-  Aronszajn, N. (1950).  
Theory of reproducing kernels.  
*Transactions of the American Mathematical Society*,  
68:337–404.
-  Barp, A., Oates, C. J., Porcu, E., and Girolami, M. (2022).  
A Riemann–Stein kernel method.  
*Bernoulli*, 28(4):2181 – 2208.
-  Baum, J., Kanagawa, H., and Gretton, A. (2023).  
A kernel stein test of goodness of fit for sequential models.  
In *International Conference on Machine Learning (ICML)*,  
pages 1936–1953.
-  Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M.,  
Mackey, L., and Oates, C. J. (2019).  
Stein point Markov chain Monte Carlo.  
In *International Conference on Machine Learning (ICML)*,  
pages 1011–1021.

-  Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. (2018).  
Stein points.  
In *International Conference on Machine Learning (ICML)*, pages 844–853.
-  Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).  
A kernel test of goodness of fit.  
In *International Conference on Machine Learning (ICML)*, pages 2606–2615.
-  Futami, F., Cui, Z., Sato, I., and Sugiyama, M. (2019).  
Bayesian posterior approximation via greedy particle optimization.  
In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 3606–3613.
-  Gorham, J. and Mackey, L. (2017).  
Measuring sample quality with kernels.

In *International Conference on Machine Learning (ICML)*,  
pages 1292–1301.

-  Hagrass, O., Sriperumbudur, B., and Balasubramanian, K. (2025). Minimax optimal goodness-of-fit testing with kernel Stein discrepancy. *Bernoulli*. (accepted; preprint: <https://arxiv.org/abs/2404.08278>).
-  Hodgkinson, L., Salomone, R., and Roosta, F. (2021). The reproducing Stein kernel approach for post-hoc corrected sampling. Technical report. (<https://arxiv.org/abs/2001.09266>).
-  Kalinke, F. and Szabó, Z. (2024). The minimax rate of HSIC estimation for translation-invariant kernels.

In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 108468–108489.

-  Kalinke, F., Szabó, Z., and Sriperumbudur, B. K. (2025). Nyström kernel Stein discrepancy.  
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 388–396.
-  Kanagawa, H., Jitkrittum, W., Mackey, L., Fukumizu, K., and Gretton, A. (2020). A kernel Stein test for comparing latent variable models. Technical report.  
(<https://arxiv.org/abs/1907.00586>).
-  Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. (2021). Kernel Stein discrepancy descent.  
In *International Conference on Machine Learning (ICML)*, pages 5719–5730.

 Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. (2020).

A non-asymptotic analysis for Stein variational gradient descent.

In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4672–4682.

 Li, L., Dwivedi, R., and Mackey, L. (2024).

Debiased distribution compression.

In *International Conference on Machine Learning (ICML)*, pages 27675–27731.

 Lim, J. N., Yamada, M., Schölkopf, B., and Jitkrittum, W. (2019).

Kernel Stein tests for multiple model comparison.

In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2243–2253.

 Liu, Q., Lee, J., and Jordan, M. (2016).

A kernelized Stein discrepancy for goodness-of-fit tests.

In *International Conference on Machine Learning (ICML)*,  
pages 276–284.

-  Liu, Q. and Wang, D. (2016).  
Stein variational gradient descent: A general purpose Bayesian inference algorithm.  
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2378–2386.
-  Liu, Q. and Wang, D. (2018).  
Stein variational gradient descent as moment matching.  
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8854–8863.
-  Sarvmaili, M., Sajjad, H., and Wu, G. (2025).  
Data-centric prediction explanation via kernelized Stein discrepancy.  
In *International Conference on Learning Representations (ICLR)*.
-  Schrab, A., Guedj, B., and Gretton, A. (2022).

KSD aggregated goodness-of-fit test.

In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 32624–32638.

-  Steinwart, I. and Christmann, A. (2008).  
*Support Vector Machines*.  
Springer.
-  Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. (2017).  
Minimax estimation of kernel mean embeddings.  
*Journal of Machine Learning Research*, 18(86):1–47.
-  Tolstikhin, I., Sriperumbudur, B. K., and Schölkopf, B. (2016).  
Minimax estimation of maximal mean discrepancy with radial kernels.  
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1930–1938.
-  Wang, C., Chen, W. Y., Kanagawa, H., and Oates, C. J. (2023).

Stein  $\Pi$ -importance sampling.

In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 71948–71994.

-  Wynne, G., Kasprzak, M. a. J., and Duncan, A. B. (2025).  
A Fourier representation of kernel Stein discrepancy with application to goodness-of-fit tests for measures on infinite dimensional Hilbert spaces.  
*Bernoulli*, 31(2):868–893.

-  Xu, W. and Matsuda, T. (2020).  
A Stein goodness-of-fit test for directional distributions.  
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 320–330.

-  Xu, W. and Matsuda, T. (2021).  
Interpretable Stein goodness-of-fit tests on Riemannian manifold.  
In *International Conference on Machine Learning (ICML)*, pages 11502–11513.

-  Xu, W. and Reinert, G. (2021).  
A Stein goodness-of-test for exponential random graph models.  
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 415–423.
-  Yang, J., Liu, Q., Rao, V., and Neville, J. (2018).  
Goodness-of-fit testing for discrete distributions via Stein discrepancy.  
In *International Conference on Machine Learning (ICML)*, pages 5561–5570.
-  Yang, J., Rao, V. A., and Neville, J. (2019).  
A Stein-Papangelou goodness-of-fit test for point processes.  
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 226–235.
-  Zhou, Y., Chen, D.-R., and Huang, W. (2019).  
A class of optimal estimators for the covariance operator in reproducing kernel Hilbert spaces.

*Journal of Multivariate Analysis*, 169:166–178.