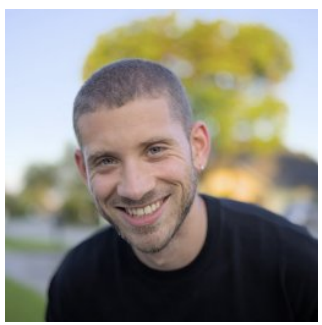
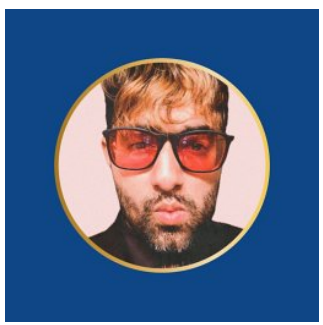
Accelerated Computation[1] and Minimax Lower Bound[2] of Kernel Stein Discrepancy

Zoltán Szabó

Joint work with:

- [2]Jose Cribeiro-Ramallo @ KIT, Germany,
- [2]Agnideep Aich @ University of Louisiana at Lafayette, U.S.,
- [1,2]Florian Kalinke @ KIT, Germany,
- [1]Bharath Sriperumbudur @ @ Pennsylvania State University, U.S.,
- [2]Ashit Baran Aich @ formerly of Presidency College, India.


Faculty of Engineering, Free University of Bozen-Bolzano

Feb. 25, 2026

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 [$n :=$ sample size]:

  convergence rate: $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$, computational complexity $= \mathcal{O}(n^2)$.

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 [$n :=$ sample size]:

  convergence rate: $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$, computational complexity $= \mathcal{O}(n^2)$.

Focus

- Question-1: Can we achieve faster convergence rate?

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 [$n :=$ sample size]:

  convergence rate: $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$, computational complexity $= \mathcal{O}(n^2)$.

Focus

- Question-1: Can we achieve faster convergence rate?
- Answer-1: No [Cribeiro-Ramallo et al., 2026].

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 [$n :=$ sample size]:

  convergence rate: $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$, computational complexity $= \mathcal{O}(n^2)$.

Focus

- Question-1: Can we achieve faster convergence rate?
- Answer-1: No [Cribeiro-Ramallo et al., 2026].
- Question-2: Can we accelerate the computation while not loosing statistically?

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 [$n :=$ sample size]:

  convergence rate: $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$, computational complexity $= \mathcal{O}(n^2)$.

Focus

- Question-1: Can we achieve faster convergence rate?
- Answer-1: No [Cribeiro-Ramallo et al., 2026].
- Question-2: Can we accelerate the computation while not loosing statistically?
- Answer-2: Yes, $O\left(n^{\frac{3}{2}}\right)$ works [Kalinke et al., 2025].

Kernel, RKHS, information theoretical measures

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

[Aronszajn, 1950, Steinwart and Christmann, 2008, Saitoh and Sawano, 2016]

In $\mathbb{R}^d$: $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i \in [d]} a_i b_i$

$\Rightarrow \|\mathbf{a} - \mathbf{b}\|_2$ and $\sphericalangle(\mathbf{a}, \mathbf{b})$, $[d] := \{1, \ldots, d\}$.

- Def-1 (feature space):

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad a, b \in \mathcal{X}.$$

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

[Aronszajn, 1950, Steinwart and Christmann, 2008, Saitoh and Sawano, 2016]

In $\mathbb{R}^d$: $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i \in [d]} a_i b_i$

$\Rightarrow \|\mathbf{a} - \mathbf{b}\|_2$ and $\sphericalangle(\mathbf{a}, \mathbf{b})$, $[d] := \{1, \ldots, d\}$.

- Def-1 (feature space):

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad a, b \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

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

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

[Aronszajn, 1950, Steinwart and Christmann, 2008, Saitoh and Sawano, 2016]

In $\mathbb{R}^d$: $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i \in [d]} a_i b_i$

$\Rightarrow \|\mathbf{a} - \mathbf{b}\|_2$ and $\sphericalangle(\mathbf{a}, \mathbf{b})$, $[d] := \{1, \ldots, d\}$.

- Def-1 (feature space):

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad a, b \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

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

- 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, Saitoh and Sawano, 2016]

In $\mathbb{R}^d$: $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i \in [d]} a_i b_i$

$\Rightarrow \|\mathbf{a} - \mathbf{b}\|_2$ and $\sphericalangle(\mathbf{a}, \mathbf{b})$, $[d] := \{1, \ldots, d\}$.

- Def-1 (feature space):

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad a, b \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

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

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n} \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

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

[Aronszajn, 1950, Steinwart and Christmann, 2008, Saitoh and Sawano, 2016]

In $\mathbb{R}^d$: $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i \in [d]} a_i b_i$

$\Rightarrow \|\mathbf{a} - \mathbf{b}\|_2$ and $\sphericalangle(\mathbf{a}, \mathbf{b})$, $[d] := \{1, \ldots, d\}$.

- Def-1 (feature space):

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad a, b \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

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

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n} \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

Note:

- $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k = \overline{\mathrm{Span}}(k(\cdot, x) : x \in \mathcal{X})$: Fourier, polynomials, splines, ...

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

[Aronszajn, 1950, Steinwart and Christmann, 2008, Saitoh and Sawano, 2016]

In $\mathbb{R}^d$: $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i \in [d]} a_i b_i$

$\Rightarrow \|\mathbf{a} - \mathbf{b}\|_2$ and $\sphericalangle(\mathbf{a}, \mathbf{b})$, $[d] := \{1, \ldots, d\}$.

- Def-1 (feature space):

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad a, b \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

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

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n} \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

Note:

- $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k = \overline{\mathrm{Span}}(k(\cdot, x) : x \in \mathcal{X})$: Fourier, polynomials, splines, ...
- It can be defined on various domains.

Towards mean embeddings: distribution representation

$$P \mapsto \mu_k(P) = \int_{\mathcal{X}} \varphi(x) \mathrm{d}P(x) .$$

Towards mean embeddings: distribution representation

$$P \mapsto \mu_k(P) = \int_{\mathcal{X}} \varphi(x) \mathrm{d}P(x) .$$

- Cumulative density function:

$$P \mapsto F_P(\mathbf{z}) = P(X < \mathbf{z}) = \int_{\mathbb{R}^d} \chi_{(-\infty, \mathbf{z})}(\mathbf{x}) \mathrm{d}P(\mathbf{x}).$$

Towards mean embeddings: distribution representation

$$P \mapsto \mu_k(P) = \int_{\mathcal{X}} \varphi(x) \mathrm{d}P(x) .$$

- Cumulative density function:

$$P \mapsto F_P(\mathbf{z}) = P(X < \mathbf{z}) = \int_{\mathbb{R}^d} \chi_{(-\infty,\mathbf{z})}(\mathbf{x}) \mathrm{d}P(\mathbf{x}).$$

- Characteristic function:

$$P \mapsto c_P(\mathbf{z}) = \int_{\mathbb{R}^d} e^{i\langle \mathbf{z},\mathbf{x} \rangle} \mathrm{d}P(\mathbf{x}).$$

Towards mean embeddings: distribution representation

$$P \mapsto \mu_k(P) = \int_{\mathcal{X}} \varphi(x) \mathrm{d}P(x) .$$

- Cumulative density function:

$$P \mapsto F_P(\mathbf{z}) = P(X < \mathbf{z}) = \int_{\mathbb{R}^d} \chi_{(-\infty,\mathbf{z})}(\mathbf{x}) \mathrm{d}P(\mathbf{x}).$$

- Characteristic function:

$$P \mapsto c_P(\mathbf{z}) = \int_{\mathbb{R}^d} e^{i\langle \mathbf{z},\mathbf{x} \rangle} \mathrm{d}P(\mathbf{x}).$$

- Moment generating function:

$$P \mapsto M_P(\mathbf{z}) = \int_{\mathbb{R}^d} e^{\langle \mathbf{z},\mathbf{x} \rangle} \mathrm{d}P(\mathbf{x}) .$$

Towards mean embeddings: distribution representation

$$P \mapsto \mu_k(P) = \int_{\mathcal{X}} \varphi(x) \mathrm{d}P(x) .$$

- Cumulative density function:

$$P \mapsto F_P(\mathbf{z}) = P(X < \mathbf{z}) = \int_{\mathbb{R}^d} \chi_{(-\infty,\mathbf{z})}(\mathbf{x}) \mathrm{d}P(\mathbf{x}).$$

- Characteristic function:

$$P \mapsto c_P(\mathbf{z}) = \int_{\mathbb{R}^d} e^{i\langle \mathbf{z},\mathbf{x} \rangle} \mathrm{d}P(\mathbf{x}).$$

- Moment generating function:

$$P \mapsto M_P(\mathbf{z}) = \int_{\mathbb{R}^d} e^{\langle \mathbf{z},\mathbf{x} \rangle} \mathrm{d}P(\mathbf{x}) .$$

Trick

φ: on any kernel-endowed domain!

Mean embedding

- Mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]:

$$\mu_k(P) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\varphi(x) \in \mathcal{H}_k} \mathrm{d}P(x) \in \mathcal{H}_k.$$

Mean embedding, MMD

- Mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]:

$$\mu_k(P) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\varphi(x) \in \mathcal{H}_k} \mathrm{d}P(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

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

Mean embedding, MMD, HSIC

- Mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]:

$$\mu_k(P) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\varphi(x) \in \mathcal{H}_k} \mathrm{d}P(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

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

- Hilbert-Schmidt independence criterion [Gretton et al., 2005] ($M = 2$), [Quadrianto et al., 2009, Sejdinovic et al., 2013a, Szabó and Sriperumbudur, 2018] ($M \geq 3$), $k := \otimes_{m=1}^{M} k_m$:

$$\mathrm{HSIC}_k(P) := \mathrm{MMD}_k\left(P, \otimes_{m=1}^{M} P_m\right)$$

Mean embedding, MMD, HSIC

- Mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]:

$$\mu_k(P) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\varphi(x) \in \mathcal{H}_k} \mathrm{d}P(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

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

- Hilbert-Schmidt independence criterion [Gretton et al., 2005] ($M = 2$), [Quadrianto et al., 2009, Sejdinovic et al., 2013a, Szabó and Sriperumbudur, 2018] ($M \geq 3$), $k := \otimes_{m=1}^M k_m$:

$$\begin{aligned}\mathrm{HSIC}_k(P) &:= \mathrm{MMD}_k\left(P, \otimes_{m=1}^M P_m\right), \\ &= \Big\| \underbrace{\mu_{\otimes_{m=1}^M k_m}(P) - \otimes_{m=1}^M \mu_{k_m}(P_m)}_{\text{cross-covariance operator}} \Big\|_{\mathcal{H}_k}.\end{aligned}$$

MMD, HSIC: information theoretical & statistical relations

- MMD:

$$\mathrm{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in B_k} \underbrace{\langle f, \mu_k(P) - \mu_k(Q) \rangle_{\mathcal{H}_k}}_{\mathbb{E}_P f(X) - \mathbb{E}_Q f(X)}$$

MMD, HSIC: information theoretical & statistical relations

- MMD:

$$\mathrm{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in B_k} \underbrace{\langle f, \mu_k(P) - \mu_k(Q) \rangle_{\mathcal{H}_k}}_{\mathbb{E}_P f(X) - \mathbb{E}_Q f(X)}$$

  - $\in$ IPMs [Zolotarev, 1983, Müller, 1997]

MMD, HSIC: information theoretical & statistical relations

- M MD:

$$\mathrm{MMD}_k(P,Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in B_k} \underbrace{\langle f, \mu_k(P) - \mu_k(Q)\rangle_{\mathcal{H}_k}}_{\mathbb{E}_P f(X) - \mathbb{E}_Q f(X)}$$

  - $\in$ IPMs [Zolotarev, 1983, Müller, 1997],
  - $\overset{\dagger}{\Leftrightarrow}$ energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005], a.k.a. N-distance [Zinger et al., 1992, Klebanov, 2005].

† [Sejdinovic et al., 2013b].

MMD, HSIC: information theoretical & statistical relations

- MMD:

$$\mathrm{MMD}_k(P,Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in B_k} \underbrace{\langle f, \mu_k(P) - \mu_k(Q)\rangle_{\mathcal{H}_k}}_{\mathbb{E}_P f(X) - \mathbb{E}_Q f(X)}$$

  - $\in$ IPMs [Zolotarev, 1983, Müller, 1997],
  - $\overset{\dagger}{\Leftrightarrow}$ energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005], a.k.a. N-distance [Zinger et al., 1992, Klebanov, 2005].

- HSIC ($M = 2$) $\overset{\dagger}{\Leftrightarrow}$ distance covariance

  [Székely et al., 2007, Székely and Rizzo, 2009, Lyons, 2013].

† [Sejdinovic et al., 2013b].

If $\mathcal{H}_k$ is 'rich', no information is lost; example

- $\mathrm{MMD}_k(P, Q) = 0 \Leftrightarrow P = Q$: k is characteristic
  [Fukumizu et al., 2008, Sriperumbudur et al., 2010].

If $\mathcal{H}_k$ is 'rich', no information is lost; example

- $\mathrm{MMD}_k(P, Q) = 0 \Leftrightarrow P = Q$: k is characteristic
  [Fukumizu et al., 2008, Sriperumbudur et al., 2010].
- For continuous bounded shift-invariant kernels on $\mathbb{R}^d$:

$$k(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}') \overset{(*)}{=} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{x}', \boldsymbol{\omega}\rangle} \mathrm{d}\Lambda(\boldsymbol{\omega}) \Rightarrow$$

$(*)$: Bochner's theorem.

If $\mathcal{H}_k$ is 'rich', no information is lost; example

- $\mathrm{MMD}_k(P, Q) = 0 \Leftrightarrow P = Q$: k is characteristic [Fukumizu et al., 2008, Sriperumbudur et al., 2010].
- For continuous bounded shift-invariant kernels on $\mathbb{R}^d$:

$$k(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}') \overset{(*)}{=} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{x}', \boldsymbol{\omega}\rangle} \mathrm{d}\Lambda(\boldsymbol{\omega}) \Rightarrow$$

$$\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \|c_P - c_Q\|_{L^2(\Lambda)}.$$

$(*)$: Bochner's theorem, c_P: characteristic function of P.

If $\mathcal{H}_k$ is 'rich', no information is lost; example

- $\mathrm{MMD}_k(P, Q) = 0 \Leftrightarrow P = Q$: k is characteristic
  [Fukumizu et al., 2008, Sriperumbudur et al., 2010].
- For continuous bounded shift-invariant kernels on $\mathbb{R}^d$:

$$k(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}') \overset{(*)}{=} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{x}', \omega \rangle} \mathrm{d}\Lambda(\omega) \Rightarrow$$

$$\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \|c_P - c_Q\|_{L^2(\Lambda)}.$$

$(*)$: Bochner's theorem, c_P: characteristic function of P.

Theorem ([Sriperumbudur et al., 2010])

k is characteristic iff. $\mathrm{supp}(\Lambda) = \mathbb{R}^d$.

Examples on $\mathbb{R}$; similarly $\mathbb{R}^d$ [Sriperumbudur et al., 2010]

For Poisson kernel: $\sigma \in (0, 1)$.

kernel name	$k_0(x)$	$\widehat{k_0}(\omega)$	supp $\left(\widehat{k_0}\right)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\lvert x\rvert}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}-spline	$*^{2n+2}\chi_{[-\frac{1}{2},\frac{1}{2}]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma, \sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\lvert j\rvert}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\delta(\omega-j)$	$\{0, \pm1, \pm2, \ldots, \pm n\}$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\lvert j\rvert}{n+1}\right)\delta(\omega-j)$	$\{0, \pm1, \pm2, \ldots, \pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma, \sigma\}$

Examples on $\mathbb{R}$; similarly $\mathbb{R}^d$ [Sriperumbudur et al., 2010]

For Poisson kernel: $\sigma \in (0, 1)$.

kernel name	$k_0(x)$	$\widehat{k_0}(\omega)$	supp $\left(\widehat{k_0}\right)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\lvert x\rvert}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}-spline	$*^{2n+2}\chi_{[-\frac{1}{2},\frac{1}{2}]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma, \sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\lvert j\rvert}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\delta(\omega-j)$	$\{0, \pm1, \pm2, \ldots, \pm n\}$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\lvert j\rvert}{n+1}\right)\delta(\omega-j)$	$\{0, \pm1, \pm2, \ldots, \pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma, \sigma\}$

For $\mathbf{x} \in \mathbb{R}^d$: $k_0(\mathbf{x}) = \prod_{j=1}^d k_0(x_j)$, $\widehat{k_0}(\omega) = \prod_{j=1}^d \widehat{k_0}(\omega_j)$.

Kernel Stein discrepancy (KSD)

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.
  - $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ kernel.

KSD on $\mathbb{R}^d$

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

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

KSD on $\mathbb{R}^d$

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

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

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

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

KSD on $\mathbb{R}^d$

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

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

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

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

Langevin-Stein operator

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

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

hence

- one assumes:
  1. $p_0 > 0$

Langevin-Stein operator

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

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

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:

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

hence

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

Langevin-Stein operator

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

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

hence

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

Why does the mean-zero property hold? ($d = 1$)

Let $h \in \mathcal{H}_k$.

$$(\mathcal{A}_{p_0} h)(x) = [\ln(p_0(x))]' h(x) + h'(x)$$

Why does the mean-zero property hold? ($d = 1$)

Let $h \in \mathcal{H}_k$.

$$
\begin{aligned}
(\mathcal{A}_{p_0} h)(x) &= [\ln(p_0(x))]' h(x) + h'(x) \\
&= \frac{p_0'(x)}{p_0(x)} h(x) + \frac{h'(x)}{p_0(x)} p_0(x)
\end{aligned}
$$

Why does the mean-zero property hold? ($d = 1$)

Let $h \in \mathcal{H}_k$.

$$
\begin{aligned}
(\mathcal{A}_{p_0} h)(x) &= [\ln(p_0(x))]' h(x) + h'(x) \\
&= \frac{p_0'(x)}{p_0(x)} h(x) + \frac{h'(x)}{p_0(x)} p_0(x) = \frac{[p_0(x) h(x)]'}{p_0(x)}.
\end{aligned}
$$

Why does the mean-zero property hold? ($d = 1$)

Let $h \in \mathcal{H}_k$.

$$(\mathcal{A}_{p_0} h)(x) = [\ln(p_0(x))]' h(x) + h'(x)$$
$$= \frac{p_0'(x)}{p_0(x)} h(x) + \frac{h'(x)}{p_0(x)} p_0(x) = \frac{[p_0(x)h(x)]'}{p_0(x)}.$$

Therefore, for any $h \in \mathcal{H}_k$

$$\mathbb{E}_{P_0}[(\mathcal{A}_{p_0} h)(X)] = \int_{\mathbb{R}} \frac{[p_0(x)h(x)]'}{p_0(x)} p_0(x)\,\mathrm{d}x$$

Why does the mean-zero property hold? ($d = 1$)

Let $h \in \mathcal{H}_k$.

$$(\mathcal{A}_{p_0} h)(x) = [\ln(p_0(x))]' h(x) + h'(x)$$
$$= \frac{p_0'(x)}{p_0(x)} h(x) + \frac{h'(x)}{p_0(x)} p_0(x) = \frac{[p_0(x)h(x)]'}{p_0(x)}.$$

Therefore, for any $h \in \mathcal{H}_k$

$$\mathbb{E}_{P_0}[(\mathcal{A}_{p_0} h)(X)] = \int_{\mathbb{R}} \frac{[p_0(x)h(x)]'}{p_0(x)} p_0(x)\,\mathrm{d}x$$
$$= \int_{\mathbb{R}} [p_0(x)h(x)]'\,\mathrm{d}x$$

Why does the mean-zero property hold? ($d = 1$)

Let $h \in \mathcal{H}_k$.

$$(\mathcal{A}_{p_0} h)(x) = [\ln(p_0(x))]' h(x) + h'(x)$$
$$= \frac{p_0'(x)}{p_0(x)} h(x) + \frac{h'(x)}{p_0(x)} p_0(x) = \frac{[p_0(x)h(x)]'}{p_0(x)}.$$

Therefore, for any $h \in \mathcal{H}_k$

$$\mathbb{E}_{P_0}[(\mathcal{A}_{p_0} h)(X)] = \int_{\mathbb{R}} \frac{[p_0(x)h(x)]'}{p_0(x)} p_0(x)\,\mathrm{d}x$$
$$= \int_{\mathbb{R}} [p_0(x)h(x)]'\,\mathrm{d}x = [p_0(x)h(x)]_{x=-\infty}^{x=\infty} = 0.$$

Hence, the assumption: $\lim_{|x|\to\infty} p_0(x)h(x) = 0$.

Langevin-Stein KSD

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

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

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:

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

with the so-called Stein kernel

$$\begin{aligned} K_0(\mathbf{x}, \mathbf{y}) = &\left\langle \nabla_{\mathbf{x}} \ln(p_0(\mathbf{x})), \nabla_{\mathbf{y}} \ln(p_0(\mathbf{y})) \right\rangle k(\mathbf{x}, \mathbf{y}) \\ &+ \left\langle \nabla_{\mathbf{y}} \ln(p_0(\mathbf{y})), \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{y}) \right\rangle \\ &+ \left\langle \nabla_{\mathbf{x}} \ln(p_0(\mathbf{x})), \nabla_{\mathbf{y}} k(\mathbf{x}, \mathbf{y}) \right\rangle \\ &+ \sum_{j=1}^{d} \frac{\partial^2 k(\mathbf{x}, \mathbf{y})}{\partial x_j \partial y_j}. \end{aligned}$$

Langevin-Stein KSD: assumptions so far (=: A1)

1. $P_0 \ll \lambda_d$, with pdf
   1. $p_0 > 0$,
   2. $p_0 \in \mathcal{C}^1(\mathbb{R}^d)$,
   3. $\lim_{\|\mathbf{x}\|_2 \to \infty} h(\mathbf{x}) p_0(\mathbf{x}) = 0$ 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 $\mathcal{X}$,
- $\Psi_{P_0} : \mathcal{X} \to \mathcal{H}$ 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 $\mathcal{X}$,
- $\Psi_{P_0} : \mathcal{X} \to \mathcal{H}$ measurable s.t. the mean-zero property holds:

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

Then, an IPM construction gives

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

KSD on general domains

Key to KSD

mean-zero property.

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

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

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

Then, an IPM construction gives

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

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

KSD estimators

Thanks to

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

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

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

KSD estimators: Nyström acceleration

Thanks to

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

2. Nyström KSD estimator:
   1. subsample with replacement: $\{\{\tilde{X}_1, \ldots, \tilde{X}_m\}\} \Rightarrow$ subspace:

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

KSD estimators: Nyström acceleration

Thanks to

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

2. Nyström KSD estimator:
   1. subsample with replacement: $\{\{\tilde{X}_1, \ldots, \tilde{X}_m\}\} \Rightarrow$ subspace:

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

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

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

KSD estimators: Nyström acceleration – continued

2. Nyström KSD estimator:
    3. resulting squared KSD estimator:

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

KSD estimators: Nyström acceleration – continued

2. Nyström KSD estimator:
   3. resulting squared KSD estimator:

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

   4. simple form in terms of Gram matrices:

$$\widehat{\mathrm{KSD}}_N^2(P_0, P) = \beta^\top \mathbf{K}_{m,m}^{-} \beta, \qquad \beta = \frac{1}{n} \mathbf{K}_{m,n} \mathbf{1}_n,$$

$$\mathbf{K}_{m,m} = \left[K_0\left(\tilde{X}_a, \tilde{X}_b\right)\right]_{a,b=1}^{m}, \quad \mathbf{K}_{m,n} = \left[K_0\left(\tilde{X}_a, X_b\right)\right]_{a,b=1}^{m,n}.$$

KSD estimators: Nyström acceleration – continued

2. Nyström KSD estimator:
   3. resulting squared KSD estimator:

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

   4. simple form in terms of Gram matrices:

$$\widehat{\mathrm{KSD}}_N^2(P_0, P) = \beta^\top \mathbf{K}_{m,m}^{-} \beta, \qquad \beta = \frac{1}{n} \mathbf{K}_{m,n} \mathbf{1}_n,$$
$$\mathbf{K}_{m,m} = \left[K_0\left(\tilde{X}_a, \tilde{X}_b\right)\right]_{a,b=1}^{m}, \quad \mathbf{K}_{m,n} = \left[K_0\left(\tilde{X}_a, X_b\right)\right]_{a,b=1}^{m,n}.$$

Both converge at a rate $\mathcal{O}(n^{-1/2})$ [Kalinke et al., 2025].

Can we get faster?

Challenge

Existing minimax lower bounds:

1. maximum mean discrepancy [Tolstikhin et al., 2016],
2. mean embedding [Tolstikhin et al., 2017],
3. covariance operator [Zhou et al., 2019],
4. Hilbert-Schmidt independence criterion [Kalinke and Szabó, 2024].

Challenge

Existing minimax lower bounds:

1. maximum mean discrepancy [Tolstikhin et al., 2016],
2. mean embedding [Tolstikhin et al., 2017],
3. covariance operator [Zhou et al., 2019],
4. Hilbert-Schmidt independence criterion [Kalinke and Szabó, 2024].

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:

$$R_n^* = \overbrace{\inf_{\hat{F}_n}}^{\text{best estimator}} \overbrace{\underbrace{\sup_{P_0 \in \mathcal{T}}}_{\text{A1}} \underbrace{\sup_{P \in \mathcal{S}_{P_0}}}_{\substack{\text{KSD}(P,P_0)<\infty \\ \Leftrightarrow \mathbb{E}_P \sqrt{K_0(X,X)}<\infty}}}^{\text{worst case}} \underbrace{\mathbb{E}_{P^n} \left| \overbrace{\hat{F}_n\left(X_{1:n}\right)}^{\text{estimator}} - \overbrace{\text{KSD}(P_0, P)}^{\text{target}} \right|}_{\text{expected error}}.$$

Minimax estimation

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

$$R_n^* = \overbrace{\inf_{\hat{F}_n}}^{\text{best estimator}} \overbrace{\underbrace{\sup_{P_0 \in \mathcal{T}}}_{\text{A1}} \underbrace{\sup_{P \in \mathcal{S}_{P_0}}}_{\substack{\text{KSD}(P,P_0)<\infty \\ \Leftrightarrow \mathbb{E}_P \sqrt{K_0(X,X)}<\infty}}}^{\text{worst case}} \underbrace{\mathbb{E}_{P^n} \left| \overbrace{\hat{F}_n\left(X_{1:n}\right)}^{\text{estimator}} - \overbrace{\text{KSD}(P_0, P)}^{\text{target}} \right|}_{\text{expected error}}.$$

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:

$$R_n^* = \overbrace{\inf_{\hat{F}_n}}^{\text{best estimator}} \overbrace{\underbrace{\sup_{P_0 \in \mathcal{T}}}_{\text{A1}} \underbrace{\sup_{P \in \mathcal{S}_{P_0}}}_{\substack{\text{KSD}(P,P_0)<\infty \\ \Leftrightarrow \mathbb{E}_P \sqrt{K_0(X,X)}<\infty}}}^{\text{worst case}} \underbrace{\mathbb{E}_{P^n} \left| \overbrace{\hat{F}_n\left(X_{1:n}\right)}^{\text{estimator}} - \overbrace{\text{KSD}(P_0, P)}^{\text{target}} \right|}_{\text{expected error}}.$$

Aim (after Markov inequality)

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

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

Minimax estimation

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

$$R_n^* = \overbrace{\inf_{\hat{F}_n}}^{\text{best estimator}} \underbrace{\overbrace{\sup_{P_0 \in \mathcal{T}}}^{\text{A1}} \underbrace{\sup_{P \in \mathcal{S}_{P_0}}}_{\substack{\text{KSD}(P,P_0)<\infty \\ \Leftrightarrow \mathbb{E}_P \sqrt{K_0(X,X)}<\infty}}}^{\text{worst case}} \underbrace{\mathbb{E}_{P^n} \left| \overbrace{\hat{F}_n\left(X_{1:n}\right)}^{\text{estimator}} - \overbrace{\text{KSD}(P_0, P)}^{\text{target}} \right|}_{\text{expected error}}.$$

Aim (after Markov inequality)

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

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

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: shift-invariant ($\Rightarrow$ bounded; with A1 $\Rightarrow$ Bochner kernel)*

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

Theorem

Assume:

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

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

Theorem

Assume:

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

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

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

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

Theorem

Assume:

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

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

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

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 $\mathcal{X}$ P_0*-a.s.*
- $\mathcal{C}_b(\mathcal{X}) \setminus \overbrace{\mathcal{F}_c(\mathcal{X})} \neq \emptyset$

Main result on general domain

Theorem

Assume:

- *A1',*
  constant functions on $\mathcal{X}$ P_0*-a.s.*
- $\mathcal{C}_b(\mathcal{X}) \setminus \overbrace{\mathcal{F}_c(\mathcal{X})} \neq \emptyset,$
- *KSD is valid.*

Main result on general domain

Theorem

Assume:

- *A1',*
  constant functions on $\mathcal{X}$ P_0-a.s.
- $\mathcal{C}_b(\mathcal{X}) \setminus \overbrace{\mathcal{F}_c(\mathcal{X})} \neq \emptyset$,
- *KSD is valid.*

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

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

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

Requirements to the Le Cam's method:

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

$$P_n(A) = \int_A 1 + \epsilon_n \varphi(x) \mathrm{d}P_0(x), \quad \forall A \in \mathcal{B}(\mathcal{X}),$$

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

$$P_n(A) = \int_A 1 + \epsilon_n \varphi(x) \mathrm{d}P_0(x), \quad \forall A \in \mathcal{B}(\mathcal{X}),$$

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. $\mathrm{KSD}(P_0, P_n) = cn^{-1/2} C_\varphi > 0$ [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

$$P_n(A) = \int_A 1 + \epsilon_n \varphi(x) \mathrm{d}P_0(x), \quad \forall A \in \mathcal{B}(\mathcal{X}),$$

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

Summary

- Focus: quantifying goodness-of-fit with KSD.
- KSD can not be estimated faster than $n^{-1/2}$.
- Paper: acceleration, minimax lower bound of KSD.
- ITE toolbox (`https://bitbucket.org/szzoli/ite/`):
  - various information theoretical estimators.

Summary

- Focus: quantifying goodness-of-fit with KSD.
- KSD can not be estimated faster than $n^{-1/2}$.
- Paper: acceleration, minimax lower bound of KSD.
- ITE toolbox (`https://bitbucket.org/szzoli/ite/`):
  - various information theoretical estimators.


Aronszajn, N. (1950).
Theory of reproducing kernels.
Transactions of the American Mathematical Society, 68:337–404.


Baringhaus, L. and Franz, C. (2004).
On a new multivariate two-sample test.
Journal of Multivariate Analysis, 88:190–206.


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.


Berlinet, A. and Thomas-Agnan, C. (2004).
Reproducing Kernel Hilbert Spaces in Probability and Statistics.

Kluwer.


Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. (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.


Cribeiro-Ramallo, J., Aich, A., Kalinke, F., Aich, A. B., and Szabó, Z. (2026).

The minimax lower bound of kernel Stein discrepancy estimation.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, page (accepted; preprint: https://arxiv.org/abs/2510.15058).


Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008).
Kernel measures of conditional dependence.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 498–496.


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.


Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2012).
A kernel two-sample test.
Journal of Machine Learning Research, 13(25):723–773.


Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005).
Measuring statistical dependence with Hilbert-Schmidt norms.
In *Algorithmic Learning Theory (ALT)*, pages 63–78.


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).


Klebanov, L. (2005).
N-Distances and Their Applications.
Charles University, Prague.


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.


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.


Lyons, R. (2013).

Distance covariance in metric spaces.
The Annals of Probability, 41:3284–3305.


Müller, A. (1997).
Integral probability metrics and their generating classes of functions.
Advances in Applied Probability, 29:429–443.


Quadrianto, N., Song, L., and Smola, A. (2009).
Kernelized sorting.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1289–1296.


Saitoh, S. and Sawano, Y. (2016).
Theory of Reproducing Kernels and Applications.
Springer Singapore.


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.


Sejdinovic, D., Gretton, A., and Bergsma, W. (2013a).
A kernel test for three-variable interactions.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1124–1132.


Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013b).
Equivalence of distance-based and RKHS-based statistics in hypothesis testing.
Annals of Statistics, 41:2263–2291.


Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007).
A Hilbert space embedding for distributions.

In *Algorithmic Learning Theory (ALT)*, pages 13–31.


Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. (2010).
Hilbert space embeddings and metrics on probability measures.

Journal of Machine Learning Research, 11:1517–1561.


Steinwart, I. and Christmann, A. (2008).
Support Vector Machines.
Springer.


Szabó, Z. and Sriperumbudur, B. K. (2018).
Characteristic and universal tensor product kernels.
Journal of Machine Learning Research, 18(233):1–29.


Székely, G. and Rizzo, M. (2004).
Testing for equal distributions in high dimension.
InterStat, 5:1249–1272.


Székely, G. and Rizzo, M. (2005).

A new test for multivariate normality.
Journal of Multivariate Analysis, 93:58–80.


Székely, G. J. and Rizzo, M. L. (2009).
Brownian distance covariance.
The Annals of Applied Statistics, 3:1236–1265.


Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007).
Measuring and testing dependence by correlation of distances.
The Annals of Statistics, 35:2769–2794.


Tolstikhin, I., Sriperumbudur, B., and Muandet, K. (2017).
Minimax estimation of kernel mean embeddings.
Journal of Machine Learning Research, 18:1–47.


Tolstikhin, I., Sriperumbudur, B., and Schölkopf, B. (2016).
Minimax estimation of maximal mean discrepancy with radial kernels.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1930–1938.


Wang, C., Chen, W. Y., Kanagawa, H., and Oates, C. J. (2023).
Stein Π-importance sampling.
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 71948–71994.


Wynne, G., aj J. Kasprzak, M., 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.


Zinger, A., Kakosyan, A., and Klebanov, L. (1992).
A characterization of distributions by mean values of statistics and certain probabilistic metrics.
Journal of Soviet Mathematics.


Zolotarev, V. (1983).
Probability metrics.
Theory of Probability and its Applications, 28:278–302.