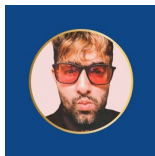
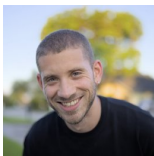


Minimax Lower Bound for KSD Estimation

Zoltán Szabó

Joint work with:

- Jose Cribeiro-Ramallo @ KIT, Germany,
- Agnideep Aich @ University of Louisiana at Lafayette, U.S.,
- Florian Kalinke @ KIT, Germany,
- Ashit Baran Aich @ formerly of Presidency College, India.



Numerical Mathematics and Data Science Seminar, School of Mathematics
University of Edinburgh
Nov. 12, 2025

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, RKHS, information theoretical measures

Kernel motivation

- 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 $\angle(\mathbf{a}, \mathbf{b})$.

Kernel motivation

- 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 $\angle(\mathbf{a}, \mathbf{b})$.
- Nonlinear features:

$$\varphi(\mathbf{a}) = (a_1^2, \sqrt{2}a_1a_2, a_2^2),$$
$$\langle \varphi(\mathbf{a}), \varphi(\mathbf{b}) \rangle = \left\langle \begin{bmatrix} a_1^2 \\ \sqrt{2}a_1a_2 \\ a_2^2 \end{bmatrix}, \begin{bmatrix} b_1^2 \\ \sqrt{2}b_1b_2 \\ b_2^2 \end{bmatrix} \right\rangle$$

Kernel motivation

- 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 $\angle(\mathbf{a}, \mathbf{b})$.
- Nonlinear features:

$$\begin{aligned}\varphi(\mathbf{a}) &= (a_1^2, \sqrt{2}a_1a_2, a_2^2), \\ \langle \varphi(\mathbf{a}), \varphi(\mathbf{b}) \rangle &= \left\langle \begin{bmatrix} a_1^2 \\ \sqrt{2}a_1a_2 \\ a_2^2 \end{bmatrix}, \begin{bmatrix} b_1^2 \\ \sqrt{2}b_1b_2 \\ b_2^2 \end{bmatrix} \right\rangle \\ &= a_1^2b_1^2 + \underbrace{\sqrt{2}\sqrt{2}}_2 a_1a_2b_1b_2 + a_2^2b_2^2\end{aligned}$$

Kernel motivation

- 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 $\angle(\mathbf{a}, \mathbf{b})$.
- Nonlinear features:

$$\begin{aligned}\varphi(\mathbf{a}) &= (a_1^2, \sqrt{2}a_1a_2, a_2^2), \\ \langle \varphi(\mathbf{a}), \varphi(\mathbf{b}) \rangle &= \left\langle \begin{bmatrix} a_1^2 \\ \sqrt{2}a_1a_2 \\ a_2^2 \end{bmatrix}, \begin{bmatrix} b_1^2 \\ \sqrt{2}b_1b_2 \\ b_2^2 \end{bmatrix} \right\rangle \\ &= a_1^2 b_1^2 + \underbrace{\sqrt{2}\sqrt{2}}_2 a_1 a_2 b_1 b_2 + a_2^2 b_2^2 \\ &= (a_1 b_1 + a_2 b_2)^2\end{aligned}$$

Kernel motivation

- 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 $\angle(\mathbf{a}, \mathbf{b})$.
- Nonlinear features:

$$\begin{aligned}\varphi(\mathbf{a}) &= (a_1^2, \sqrt{2}a_1a_2, a_2^2), \\ \langle \varphi(\mathbf{a}), \varphi(\mathbf{b}) \rangle &= \left\langle \begin{bmatrix} a_1^2 \\ \sqrt{2}a_1a_2 \\ a_2^2 \end{bmatrix}, \begin{bmatrix} b_1^2 \\ \sqrt{2}b_1b_2 \\ b_2^2 \end{bmatrix} \right\rangle \\ &= a_1^2 b_1^2 + \underbrace{\sqrt{2}\sqrt{2}}_2 a_1 a_2 b_1 b_2 + a_2^2 b_2^2 \\ &= (a_1 b_1 + a_2 b_2)^2 \\ &= \left\langle \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \right\rangle^2 = \langle \mathbf{a}, \mathbf{b} \rangle^2 =: k(\mathbf{a}, \mathbf{b}).\end{aligned}$$

Kernel motivation

- 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 $\angle(\mathbf{a}, \mathbf{b})$.
- Nonlinear features:

$$\begin{aligned}\varphi(\mathbf{a}) &= (a_1^2, \sqrt{2}a_1a_2, a_2^2), \\ \langle \varphi(\mathbf{a}), \varphi(\mathbf{b}) \rangle &= \left\langle \begin{bmatrix} a_1^2 \\ \sqrt{2}a_1a_2 \\ a_2^2 \end{bmatrix}, \begin{bmatrix} b_1^2 \\ \sqrt{2}b_1b_2 \\ b_2^2 \end{bmatrix} \right\rangle \\ &= a_1^2 b_1^2 + \underbrace{\sqrt{2}\sqrt{2}}_2 a_1 a_2 b_1 b_2 + a_2^2 b_2^2 \\ &= (a_1 b_1 + a_2 b_2)^2 \\ &= \left\langle \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \right\rangle^2 = \langle \mathbf{a}, \mathbf{b} \rangle^2 =: k(\mathbf{a}, \mathbf{b}).\end{aligned}$$

$\langle \mathbf{x}, \mathbf{x}' \rangle^p = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle$: Explicit computation would be heavy!

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

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

- 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]

- 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):

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

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

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

- 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):

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

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

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

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

- 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):

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

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

- 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):

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

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

Note:

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

Kernel examples on $\mathbb{R}^d$ ($c \geq 0, p \in \mathbb{Z}^+$)

$$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + c)^p$$

Kernel examples on $\mathbb{R}^d$ ($\gamma > 0$, $c \geq 0$, $p \in \mathbb{Z}^+$)

$$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + c)^p,$$

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

$$k_C(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + \gamma \|\mathbf{x} - \mathbf{y}\|_2^2},$$

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

$$k_L(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_1},$$

$$k_{\tilde{e}}(\mathbf{x}, \mathbf{y}) = e^{\gamma \langle \mathbf{x}, \mathbf{y} \rangle}.$$

Kernel examples on $\mathbb{R}^d$ ($\gamma, \sigma, \nu > 0$, $c \geq 0$, $p \in \mathbb{Z}^+$)

$$\begin{aligned}k_p(\mathbf{x}, \mathbf{y}) &= (\langle \mathbf{x}, \mathbf{y} \rangle + c)^p, & k_G(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}, \\k_e(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2}, & k_L(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_1}, \\k_C(\mathbf{x}, \mathbf{y}) &= \frac{1}{1 + \gamma \|\mathbf{x} - \mathbf{y}\|_2^2}, & k_{\tilde{e}}(\mathbf{x}, \mathbf{y}) &= e^{\gamma \langle \mathbf{x}, \mathbf{y} \rangle}.\end{aligned}$$

Or the flexible Matérn family:

$$k_M(\mathbf{x}, \mathbf{y}) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} \|\mathbf{x} - \mathbf{y}\|_2}{\sigma} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu} \|\mathbf{x} - \mathbf{y}\|_2}{\sigma} \right),$$

where

- K_ν : modified Bessel function of the second kind of order ν ,
- Specific cases: For $\nu = \frac{1}{2}$ one gets $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x} - \mathbf{y}\|_2}{\sigma}}$.
Gaussian kernel: $\nu \rightarrow \infty$.

Some kernel-enriched domains: $(\mathcal{X}, k)$

- **Strings**
[Watkins, 1999, Lodhi et al., 2002, Leslie et al., 2002, Kuang et al., 2004, Leslie and Kuang, 2004, Saigo et al., 2004, Cuturi and Vert, 2005],
- **time series**
[Rüping, 2001, Cuturi et al., 2007, Cuturi, 2011, Király and Oberhauser, 2019],
- **trees** [Collins and Duffy, 2001, Kashima and Koyanagi, 2002],
- **groups** and specifically **rankings** [Cuturi et al., 2005, Jiao and Vert, 2016],
- **sets** [Haussler, 1999, Gärtner et al., 2002, Balanca and Herbin, 2012, Fellmann et al., 2023], **probability distributions**
[Berlinet and Thomas-Agnan, 2004, Hein and Bousquet, 2005, Smola et al., 2007, Sriperumbudur et al., 2010],
- various **generative models** [Jaakkola and Haussler, 1999, Tsuda et al., 2002, Seeger, 2002, Jebara et al., 2004],
- **fuzzy domains** [Guevara et al., 2017], or
- **graphs** [Kondor and Lafferty, 2002, Gärtner et al., 2003, Kashima et al., 2003, Borgwardt and Kriegel, 2005, Shervashidze et al., 2009, Vishwanathan et al., 2010, Kondor and Pan, 2016, Bai et al., 2020, Borgwardt et al., 2020, Schulz et al., 2022, Nikolentzos and Vazirgiannis, 2023].

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k : **bounded** [$\sup_{x,y \in \mathcal{X}} k(x,y) \leq C$] $\Rightarrow \forall f \in \mathcal{H}_k$ is **bounded**

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k : **bounded** [$\sup_{x,y \in \mathcal{X}} k(x,y) \leq C$] $\Rightarrow \forall f \in \mathcal{H}_k$ is **bounded**:

$$|f(x)| \stackrel{\text{repr. prop.}}{=} \left| \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} \right| \stackrel{\text{CBS}}{\leq} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot, x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k : **bounded** [$\sup_{x,y \in \mathcal{X}} k(x,y) \leq C$] $\Rightarrow \forall f \in \mathcal{H}_k$ is **bounded**:

$$|f(x)| \stackrel{\text{repr. prop.}}{=} \left| \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} \right| \stackrel{\text{CBS}}{\leq} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot, x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k : **continuous** $\Rightarrow \mathcal{H}_k$: **separable** [$\ell^2(\mathbb{N}_0)$].

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k : **bounded** [$\sup_{x,y \in \mathcal{X}} k(x,y) \leq C$] $\Rightarrow \forall f \in \mathcal{H}_k$ is **bounded**:

$$|f(x)| \stackrel{\text{repr. prop.}}{=} \left| \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} \right| \stackrel{\text{CBS}}{\leq} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot, x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k : **continuous** $\Rightarrow \mathcal{H}_k$: **separable** [$\ell^2(\mathbb{N}_0)$].
- k : **bounded and continuous** $\Rightarrow \forall f \in \mathcal{H}_k$ is **bounded & continuous**.

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k : **bounded** [$\sup_{x,y \in \mathcal{X}} k(x,y) \leq C$] $\Rightarrow \forall f \in \mathcal{H}_k$ is **bounded**:

$$|f(x)| \stackrel{\text{repr. prop.}}{=} \left| \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} \right| \stackrel{\text{CBS}}{\leq} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot, x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k : **continuous** $\Rightarrow \mathcal{H}_k$: **separable** [$\ell^2(\mathbb{N}_0)$].
- k : **bounded and continuous** $\Rightarrow \forall f \in \mathcal{H}_k$ is **bounded & continuous**.
- $k \in \mathcal{C}^m \Rightarrow \forall f \in \mathcal{H}_k$ is **m -times continuously differentiable**.

Properties of k control that of $\mathcal{H}_k$

[Steinwart and Christmann, 2008, Chapter 4]:

- k : **bounded** [$\sup_{x,y \in \mathcal{X}} k(x,y) \leq C$] $\Rightarrow \forall f \in \mathcal{H}_k$ is **bounded**:

$$|f(x)| \stackrel{\text{repr. prop.}}{=} \left| \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} \right| \stackrel{\text{CBS}}{\leq} \|f\|_{\mathcal{H}_k} \underbrace{\|k(\cdot, x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}}.$$

- k : **continuous** $\Rightarrow \mathcal{H}_k$: **separable** [$\ell^2(\mathbb{N}_0)$].
- k : **bounded and continuous** $\Rightarrow \forall f \in \mathcal{H}_k$ is **bounded & continuous**.
- $k \in \mathcal{C}^m \Rightarrow \forall f \in \mathcal{H}_k$ is **m -times continuously differentiable**.
- k : **analytic** $\Rightarrow \forall f \in \mathcal{H}_k$ is **analytic**.

Towards mean embeddings: distribution representation

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

Towards mean embeddings: distribution representation

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

- Cumulative density function:

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

Towards mean embeddings: distribution representation

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

- Cumulative density function:

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

- Characteristic function:

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

Towards mean embeddings: distribution representation

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

- Cumulative density function:

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

- Characteristic function:

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

- Moment generating function:

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

Towards mean embeddings: distribution representation

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

- Cumulative density function:

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

- Characteristic function:

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

- Moment generating function:

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

Trick

φ : on any kernel-endowed domain!

Mean embedding

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

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

Mean embedding, MMD

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

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

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

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

Mean embedding, MMD, HSIC

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

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

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

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

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

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

Mean embedding, MMD, HSIC

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

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

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

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

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

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

- M MD:

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

- M MD:

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

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

- MMD:

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

- $\in$ 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].

$\dagger$ [Sejdinovic et al., 2013b].

MMD, HSIC: information theoretical & statistical relations

- MMD:

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

- $\in$ 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].

$\dagger$ [Sejdinovic et al., 2013b].

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

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

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

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

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

(*) : Bochner's theorem.

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

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

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

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

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

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

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

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

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

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

Theorem ([Sriperumbudur et al., 2010])

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

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

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

kernel name	$k_0(x)$	$\hat{k}_0(\omega)$	$\text{supp}(\hat{k}_0)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2 \omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma x }$	$\sqrt{\frac{2}{\pi}} \frac{\sigma}{\sigma^2 + \omega^2}$	$\mathbb{R}$
B_{2n+1} -spline	$*^{2n+2} \chi_{[-\frac{1}{2}, \frac{1}{2}]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}} \frac{\sin^{2n+2}(\frac{\omega}{2})}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}} \chi_{[-\sigma, \sigma]}(\omega)$	$[-\sigma, \sigma]$
Poisson	$\frac{1 - \sigma^2}{\sigma^2 - 2\sigma \cos(x) + 1}$	$\sqrt{2\pi} \sum_{j=-\infty}^{\infty} \sigma^{ j } \delta(\omega - j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin(\frac{(2n+1)x}{2})}{\sin(\frac{x}{2})}$	$\sqrt{2\pi} \sum_{j=-\infty}^{\infty} \delta(\omega - j)$	$\{0, \pm 1, \pm 2, \dots, \pm n\}$
Fejér	$\frac{1}{n+1} \frac{\sin^2(\frac{(n+1)x}{2})}{\sin^2(\frac{x}{2})}$	$\sqrt{2\pi} \sum_{j=-n}^n \left(1 - \frac{ j }{n+1}\right) \delta(\omega - j)$	$\{0, \pm 1, \pm 2, \dots, \pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}} [\delta(\omega - \sigma) + \delta(\omega + \sigma)]$	$\{-\sigma, \sigma\}$

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

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

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

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

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

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

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

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

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:

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

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 $\lim_{\|\mathbf{x}\|_2 \rightarrow \infty} h(\mathbf{x})p_0(\mathbf{x}) = 0$ for all $h \in \mathcal{H}_k$ (to the mean zero property).
[$\Leftarrow p_0$: bounded, $\lim_{\|\mathbf{x}\|_2 \rightarrow \infty} h(\mathbf{x}) = 0$ for all $h \in \mathcal{H}_k$]

Langevin-Stein operator

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

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

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 the 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 $C > 0$.

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

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

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

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

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

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

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

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

$$\begin{aligned}\mathbb{E}_{P_0}[(\mathcal{A}_{p_0} h)(X)] &= \int_{\mathbb{R}} \frac{[p_0(x) h(x)]'}{p_0(x)} p_0(x) dx \\&= \int_{\mathbb{R}} [p_0(x) h(x)]' dx\end{aligned}$$

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

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

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

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

$$\begin{aligned}\mathbb{E}_{P_0}[(\mathcal{A}_{p_0} h)(X)] &= \int_{\mathbb{R}} \frac{[p_0(x) h(x)]'}{p_0(x)} p_0(x) dx \\&= \int_{\mathbb{R}} [p_0(x) h(x)]' dx = [p_0(x) h(x)]_{x=-\infty}^{x=\infty} = 0.\end{aligned}$$

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

Langevin-Stein KSD

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

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

with the so-called **Stein kernel**

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**

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

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 $\lim_{\|\mathbf{x}\|_2 \rightarrow \infty} h(\mathbf{x})p_0(\mathbf{x}) = 0$ 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** [Sarvmali 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

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

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

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: $\{\{\tilde{X}_1, \dots, \tilde{X}_m\}\} \Rightarrow$ 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:

- 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:

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

Minimax estimation

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

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

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:

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

Aim (after Markov inequality)

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

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

Minimax estimation

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

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

Aim (after Markov inequality)

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

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

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

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

Theorem

Assume:

- $A1$

Theorem

Assume:

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

Theorem

Assume:

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

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

Theorem

Assume:

- $A1$,
- k : bounded, 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, 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 $\mathcal{X}$
- $\mathcal{C}_b(\mathcal{X}) \setminus \overbrace{\mathcal{F}_c(\mathcal{X})} \neq \emptyset$,
- *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

Requirements to the Le Cam's method:

- ① $KSD(P_0, P_n) \geq Cn^{-1/2}$ with $C > 0$,
- ② $KL(P_0^n \| P_n^n) \leq \alpha$ 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 $\text{KL}(P_0^n \| P_n^n) \leq cM$, $M := \mathbb{E}_{P_0}[\varphi^2(X)]$ [**ln 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.
- 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 on arXiv.
- 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.



Bai, L., Cui, L., Rossi, L., Xu, L., Bai, X., and Hancock, E.
(2020).

Local-global nested graph kernels using nested complexity
traces.

Pattern Recognition Letters, 134:87–95.



Balanca, P. and Herbin, E. (2012).

A set-indexed Ornstein-Uhlenbeck process.

Electronic Communications in Probability, 17:1–14.



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.



Borgwardt, K., Ghisu, E., Llinares-López, F., O’Bray, L., and
Riec, B. (2020).

Graph kernels: State-of-the-art and future challenges.
Foundations and Trends in Machine Learning,
13(5-6):531–712.



Borgwardt, K. M. and Kriegel, H.-P. (2005).

Shortest-path kernels on graphs.

In *International Conference on Data Mining (ICDM)*, pages 74–81.



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.



Collins, M. and Duffy, N. (2001).

Convolution kernels for natural language.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 625–632.



Cuturi, M. (2011).

Fast global alignment kernels.

In *International Conference on Machine Learning (ICML)*, pages 929–936.



Cuturi, M., Fukumizu, K., and Vert, J.-P. (2005).

Semigroup kernels on measures.

Journal of Machine Learning Research, 6:1169–1198.



Cuturi, M. and Vert, J.-P. (2005).

The context-tree kernel for strings.

Neural Networks, 18(8):1111–1123.



Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. (2007).

A kernel for time series based on global alignments.

In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 413–416.

-  Fellmann, N., Blanchet-Scalliet, C., Helbert, C., Spagnol, A., and Sinoquet, D. (2023).
Kernel-based sensitivity analysis for (excursion) sets.
Technical report.
(<https://arxiv.org/abs/2305.09268>).
-  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.
-  Gärtner, T., Flach, P., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.

In *International Conference on Machine Learning (ICML)*, pages 179–186.



Gärtner, T., Flach, P., and Wrobel, S. (2003).

On graph kernels: Hardness results and efficient alternatives.
Learning Theory and Kernel Machines, pages 129–143.



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.

-  Guevara, J., Hirata, R., and Canu, S. (2017).
Cross product kernels for fuzzy set similarity.
In *International Conference on Fuzzy Systems (FUZZ-IEEE)*,
pages 1–6.
-  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>).
-  Haussler, D. (1999).
Convolution kernels on discrete structures.
Technical report, University of California at Santa Cruz.
(<http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf>).
-  Hein, M. and Bousquet, O. (2005).

Hilbertian metrics and positive definite kernels on probability measures.

In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136–143.



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>).



Jaakkola, T. S. and Haussler, D. (1999).

Exploiting generative models in discriminative classifiers.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 487–493.



Jebara, T., Kondor, R., and Howard, A. (2004).

Probability product kernels.

Journal of Machine Learning Research, 5:819–844.



Jiao, Y. and Vert, J.-P. (2016).

The Kendall and Mallows kernels for permutations.
In *International Conference on Machine Learning (ICML)*,
volume 37, pages 2982–2990.



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>).

-  Kashima, H. and Koyanagi, T. (2002).
Kernels for semi-structured data.
In *International Conference on Machine Learning (ICML)*,
pages 291–298.
-  Kashima, H., Tsuda, K., and Inokuchi, A. (2003).
Marginalized kernels between labeled graphs.
In *International Conference on Machine Learning (ICML)*,
pages 321–328.
-  Király, F. J. and Oberhauser, H. (2019).
Kernels for sequentially ordered data.
Journal of Machine Learning Research, 20:1–45.
-  Klebanov, L. (2005).
N-Distances and Their Applications.
Charles University, Prague.
-  Kondor, R. and Pan, H. (2016).
The multiscale Laplacian graph kernel.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 2982–2990.



Kondor, R. I. and Lafferty, J. (2002).

Diffusion kernels on graphs and other discrete input.

In *International Conference on Machine Learning (ICML)*, pages 315–322.



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.



Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. (2004).

Profile-based string kernels for remote homology detection and motif extraction.

Journal of Bioinformatics and Computational Biology, 13(4):527–550.



Leslie, C., Eskin, E., and Noble, W. S. (2002).

The spectrum kernel: A string kernel for SVM protein classification.

Biocomputing, pages 564–575.



Leslie, C. and Kuang, R. (2004).

Fast string kernels using inexact matching for protein sequences.

Journal of Machine Learning Research, 5:1435–1455.



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.



Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002).

Text classification using string kernels.

Journal of Machine Learning Research, 2:419–444.



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.



Nikolentzos, G. and Vazirgiannis, M. (2023).

Graph alignment kernels using Weisfeiler and Leman hierarchies.

In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2019–2034.



Quadrianto, N., Song, L., and Smola, A. (2009).

Kernelized sorting.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 1289–1296.



Rüping, S. (2001).

SVM kernels for time series analysis.

Technical report, University of Dortmund.

(<http://www.stefan-rueping.de/publications/rueping-2001-a.pdf>).



Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. (2004).

Protein homology detection using string alignment kernels.

Bioinformatics, 20(11):1682–1689.



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.
-  Schulz, T. H., Welke, P., and Wrobel, S. (2022).
Graph filtration kernels.
In AAAI Conference on Artificial Intelligence (AAAI), pages 8196–8203.
-  Seeger, M. (2002).
Covariance kernels from Bayesian generative models.
In Advances in Neural Information Processing Systems (NIPS), pages 905–912.

-  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.
-  Shervashidze, N., Vishwanathan, S. V. N., Petri, T., Mehlhorn, K., and Borgwardt, K. M. (2009).
Efficient graphlet kernels for large graph comparison.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 488–495.
-  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.



Tsuda, K., Kin, T., and Asai, K. (2002).

Marginalized kernels for biological sequences.

Bioinformatics, 18:268–275.



Vishwanathan, S. N., Schraudolph, N., Kondor, R., and Borgwardt, K. (2010).

Graph kernels.

Journal of Machine Learning Research, 11:1201–1242.



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.



Watkins, C. (1999).

Dynamic alignment kernels.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 39–50.



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.