The Minimax Rate of HSIC Estimation

Zoltán Szabó

Joint work with:

- Florian Kalinke @ Karlsruhe Institute of Technology (KIT), Germany.

Statistics Research Showcase, LSE
June 20, 2024

Today: in a nutshell

- Hilbert-Schmidt independence criterion (HSIC; [Gretton et al., 2005]):
  - simple-to-estimate, popular dependency measure,
  - capable of handling $M \geq 2$ random variables,
  - with various successful applications,
  - a.k.a. distance covariance [Székely et al., 2007, Lyons, 2013] (when $M = 2$).

Today: in a nutshell

- Hilbert-Schmidt independence criterion (HSIC; [Gretton et al., 2005]):
  - simple-to-estimate, popular dependency measure,
  - capable of handling $M \geq 2$ random variables,
  - with various successful applications,
  - a.k.a. distance covariance [Székely et al., 2007, Lyons, 2013] (when $M = 2$).
- Existing estimators:

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

Today: in a nutshell

- Hilbert-Schmidt independence criterion (HSIC; [Gretton et al., 2005]):
  - simple-to-estimate, popular dependency measure,
  - capable of handling $M \geq 2$ random variables,
  - with various successful applications,
  - a.k.a. distance covariance [Székely et al., 2007, Lyons, 2013] (when $M = 2$).
- 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

- Hilbert-Schmidt independence criterion (HSIC; [Gretton et al., 2005]):
  - simple-to-estimate, popular dependency measure,
  - capable of handling $M \geq 2$ random variables,
  - with various successful applications,
  - a.k.a. distance covariance [Székely et al., 2007, Lyons, 2013] (when $M = 2$).
- Existing estimators:

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

Focus

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

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

[Aronszajn, 1950, Steinwart and Christmann, 2008]

- Def-1 (feature space):

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

- 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]

- Def-1 (feature space):

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

- Def-2 (reproducing kernel):

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

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

Notes

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

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, Draief et al., 2018, Bai et al., 2020, Borgwardt et al., 2020, Schulz et al., 2022, Nikolentzos and Vazirgiannis, 2023].

Mean embedding

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

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

Mean embedding, MMD

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

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

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

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

Mean embedding, MMD, HSIC

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

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

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

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

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

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

Mean embedding, MMD, HSIC

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

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

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

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

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

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

Tensor product

- Meaning of $k = \otimes_{m=1}^{M} k_m$,

$$k(x, x') = \prod_{m=1}^{M} \underbrace{k_m(x_m, x'_m)}_{\text{coordinate-wise similarity}}, \quad \left(x, x' \in \times_{m=1}^{M} \mathcal{X}_m\right).$$

Tensor product

- Meaning of $k = \otimes_{m=1}^{M} k_m$,

$$k(x, x') = \prod_{m=1}^{M} \underbrace{k_m(x_m, x'_m)}_{\text{coordinate-wise similarity}}, \quad \left(x, x' \in \times_{m=1}^{M} \mathcal{X}_m\right).$$

- Computation in $\mathcal{H}_k = \otimes_{m=1}^{M} \mathcal{H}_{k_m} = \overline{\text{Span}}(\otimes_{m=1}^{M} a_m \, : \, a_m \in \mathcal{H}_{k_m})$:

$$\left\langle \otimes_{m=1}^{M} a_m, \otimes_{m=1}^{M} b_m \right\rangle_{\mathcal{H}_k} = \prod_{m=1}^{M} \langle a_m, b_m \rangle_{\mathcal{H}_{k_m}}.$$

A few HSIC applications

- independence testing in batch [Gretton et al., 2008, Wehbe and Ramdas, 2015, Bilodeau and Nangue, 2017, Górecki et al., 2018, Pfister et al., 2018, Albert et al., 2022] and streaming settings [Podkopaev et al., 2023],
- feature selection [Camps-Valls et al., 2010, Song et al., 2012, Yamada et al., 2014, Wang et al., 2022], with apps in biomarker detection [Climente-González et al., 2019] & wind power prediction [Bouche et al., 2023],
- clustering [Song et al., 2007, Climente-González et al., 2019],
- causal discovery [Mooij et al., 2016, Pfister et al., 2018, Chakraborty and Zhang, 2019, Schölkopf et al., 2021, Kalinke and Szabó, 2023],
- sensitivity analysis [Veiga, 2015, Freitas Gustavo et al., 2023, Fellmann et al., 2023, Herrando-Pérez and Saltré, 2024],
- uncertainty quantification [Stenger et al., 2020],
- analysis of data augmentation methods for brain tumor detection [Anaya-Isaza and Mera-Jiménez, 2022],
- multimodal neural networks trained on neuroimaging data [Fedorov et al., 2024].

Validness of HSIC

- MMD_k is a metric if μ_k is injective; in this case k is called characteristic [Fukumizu et al., 2008, Sriperumbudur et al., 2010].

Validness of HSIC

- MMD_k is a metric if μ_k is injective; in this case k is called characteristic [Fukumizu et al., 2008, Sriperumbudur et al., 2010].
- Bochner theorem: for continuous bounded shift-invariant kernels

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

Theorem ([Sriperumbudur et al., 2010])

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

Validness of HSIC

- MMD_k is a metric if μ_k is injective; in this case k is called characteristic [Fukumizu et al., 2008, Sriperumbudur et al., 2010].
- Bochner theorem: for continuous bounded shift-invariant kernels

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

Theorem ([Sriperumbudur et al., 2010])

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

Theorem ([Szabó and Sriperumbudur, 2018])

$\mathrm{HSIC}_k(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{m=1}^M \mathbb{P}_m$ *iff.* $(k_m)_{m=1}^M$*-s are characteristic.*

HSIC estimation (example)

- Samples: $\hat{\mathbb{P}}_n := \{(x_1^1, \ldots, x_M^1), \ldots, (x_1^n, \ldots, x_M^n)\} \subset \mathcal{X}$. Estimator:

$$\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\,\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\right)\mathbf{1}_n + \frac{1}{n^{2M}}\prod_{m\in[M]}\mathbf{1}_n^\mathsf{T}\mathbf{K}_{k_m}\mathbf{1}_n - \frac{2}{n^{M+1}}\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\mathbf{1}_n\right),$$

$$\mathbf{K}_{k_m} = \left[k_m\left(x_m^i, x_m^j\right)\right]_{i,j\in[n]} \in \mathbb{R}^{n\times n}.$$

HSIC estimation (example)

- Samples: $\hat{\mathbb{P}}_n := \{(x_1^1, \ldots, x_M^1), \ldots, (x_1^n, \ldots, x_M^n)\} \subset \mathcal{X}$. Estimator:

$$\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\,\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\right)\mathbf{1}_n + \frac{1}{n^{2M}}\prod_{m\in[M]}\mathbf{1}_n^\mathsf{T}\mathbf{K}_{k_m}\mathbf{1}_n - \frac{2}{n^{M+1}}\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\mathbf{1}_n\right),$$

$$\mathbf{K}_{k_m} = \left[k_m\left(x_m^i, x_m^j\right)\right]_{i,j\in[n]} \in \mathbb{R}^{n\times n}.$$

- Existing estimators (upper bound):

$$|\,\mathrm{HSIC}_k(\mathbb{P}) - \widehat{\mathrm{HSIC}}_{k,n}(\mathbb{P})| = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right).$$

Our aim: lower bound

- $\hat{F}_n$: any estimator of $\mathrm{HSIC}_k(\mathbb{P})$ based on n i.i.d. samples from $\mathbb{P}$.
- A positive sequence $(\xi_n)_{n=1}^{\infty}$ is a lower bound of HSIC estimation if $\exists c > 0$:

$$\underbrace{\inf_{\hat{F}_n}}_{\text{best estimator}} \overbrace{\sup_{\mathbb{P} \in \mathcal{P}}}^{\text{worst distribution}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq c \xi_n \right\} > 0 \text{ for all } n.$$

Our aim: lower bound

- $\hat{F}_n$: any estimator of $\mathrm{HSIC}_k(\mathbb{P})$ based on n i.i.d. samples from $\mathbb{P}$.
- A positive sequence $(\xi_n)_{n=1}^{\infty}$ is a lower bound of HSIC estimation if $\exists c > 0$:

$$\underbrace{\inf_{\hat{F}_n}}_{\text{best estimator}} \overbrace{\sup_{\mathbb{P}\in\mathcal{P}}}^{\text{worst distribution}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq c\xi_n \right\} > 0 \text{ for all } n.$$

- If an estimator has matching upper bound, it is called minimax-optimal.

Our aim: lower bound

- $\hat{F}_n$: any estimator of $\mathrm{HSIC}_k(\mathbb{P})$ based on n i.i.d. samples from $\mathbb{P}$.
- A positive sequence $(\xi_n)_{n=1}^{\infty}$ is a lower bound of HSIC estimation if $\exists c > 0$:

$$\underbrace{\inf_{\hat{F}_n}}_{\text{best estimator}} \overbrace{\sup_{\mathbb{P}\in\mathcal{P}}}^{\text{worst distribution}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq c\xi_n \right\} > 0 \text{ for all } n.$$

- If an estimator has matching upper bound, it is called minimax-optimal.
- Note: minimax-optimality is meant w.r.t. a class of probability measures $\mathcal{P}$.

Tool: Le Cam's method [Cam, 1973, Tsybakov, 2009]

Key: $\exists \alpha > 0$ such that for any n fixed, there exists an adversarial pair $(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}) \in \mathcal{P} \times \mathcal{P}$ s.t.

1. $\mathrm{KL}\left(\mathbb{P}^n_{\theta_1} || \mathbb{P}^n_{\theta_0}\right) \le \alpha$, and
2. $|\mathrm{HSIC}_k(\mathbb{P}_{\theta_1}) - \mathrm{HSIC}_k(\mathbb{P}_{\theta_0})| \ge 2s_n > 0$.

Tool: Le Cam's method [Cam, 1973, Tsybakov, 2009]

Key: $\exists \alpha > 0$ such that for any n fixed, there exists an adversarial pair $(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}) \in \mathcal{P} \times \mathcal{P}$ s.t.

1. $\mathrm{KL}\left(\mathbb{P}^n_{\theta_1} || \mathbb{P}^n_{\theta_0}\right) \le \alpha$, and
2. $|\mathrm{HSIC}_k(\mathbb{P}_{\theta_1}) - \mathrm{HSIC}_k(\mathbb{P}_{\theta_0})| \ge 2 s_n > 0$.

In this case,

$$\inf_{\hat{F}_n} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \ge s_n \right\} \ge \max\left(\frac{e^{-\alpha}}{4}, \frac{1 - \sqrt{\alpha/2}}{2} \right) \text{ for all } n.$$

Tool: Le Cam's method [Cam, 1973, Tsybakov, 2009]

Key: $\exists\alpha > 0$ such that for any n fixed, there exists an adversarial pair $(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}) \in \mathcal{P} \times \mathcal{P}$ s.t.

1. $\mathrm{KL}\left(\mathbb{P}^n_{\theta_1} || \mathbb{P}^n_{\theta_0}\right) \leq \alpha$, and
2. $\left|\mathrm{HSIC}_k(\mathbb{P}_{\theta_1}) - \mathrm{HSIC}_k(\mathbb{P}_{\theta_0})\right| \geq 2s_n > 0$.

In this case,

$$\inf_{\hat{F}_n} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n \left\{ \left|\mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n\right| \geq s_n \right\} \geq \max\left(\frac{e^{-\alpha}}{4}, \frac{1-\sqrt{\alpha/2}}{2}\right) \text{ for all } n.$$

If we can achieve this with $s_n \asymp n^{-1/2}$, our goal: ✓.

Tool: Le Cam's method [Cam, 1973, Tsybakov, 2009]

Key: $\exists \alpha > 0$ such that for any n fixed, there exists an adversarial pair $(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}) \in \mathcal{P} \times \mathcal{P}$ s.t.

1. $\text{KL}\left(\mathbb{P}^n_{\theta_1} || \mathbb{P}^n_{\theta_0}\right) \le \alpha$, and
2. $|\text{HSIC}_k(\mathbb{P}_{\theta_1}) - \text{HSIC}_k(\mathbb{P}_{\theta_0})| \ge 2s_n > 0$.

In this case,

$$\inf_{\hat{F}_n} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^n \left\{ \left| \text{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \ge s_n \right\} \ge \max\left(\frac{e^{-\alpha}}{4}, \frac{1 - \sqrt{\alpha/2}}{2} \right) \text{ for all } n.$$

If we can achieve this with $s_n \asymp n^{-1/2}$, our goal: ✓.

We will assume $\mathcal{X}_m = \mathbb{R}^{d_m}$ ($m \in [M]$) in the sequel.

Towards the adversarial pair: $d = \sum_{m=1}^{M} d_m$

Let $\mathcal{G}$ be $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ Gaussians on $\mathbb{R}^d$ with covariance

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(i,j,\rho) = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & \rho & \cdots & 0 \\ 0 & \cdots & \rho & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{d \times d},$$

where $i = d_1$, $j = d_1 + 1$, $\rho \in (-1, 1)$.

Towards the adversarial pair: $d = \sum_{m=1}^{M} d_m$

Let $\mathcal{G}$ be $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ Gaussians on $\mathbb{R}^d$ with covariance

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(i,j,\rho) = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & \rho & \cdots & 0 \\ 0 & \cdots & \rho & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{d \times d},$$

where $i = d_1$, $j = d_1 + 1$, $\rho \in (-1, 1)$. If $\mathcal{G} \subseteq \mathcal{P}$, then for every $s_n > 0$:

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^n \left\{ \left| \text{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq s_n \right\} \geq \sup_{\mathbb{P} \in \mathcal{G}} \mathbb{P}^n \left\{ \left| \text{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq s_n \right\}.$$

$\Rightarrow$ We can work with the r.h.s. (to our lower bound).

KL upper bound

We choose $\mathbb{P}_{\theta_0} = \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$ and $\mathbb{P}_{\theta_1} = \mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$ with

$$\boldsymbol{\mu}_0 = \mathbf{0}_d \in \mathbb{R}^d, \qquad \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(d_1, d_1+1, 0) = \mathbf{I}_d \in \mathbb{R}^{d \times d},$$
$$\boldsymbol{\mu}_1 = \frac{1}{\sqrt{dn}}\mathbf{1}_d \in \mathbb{R}^d, \qquad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}(d_1, d_1+1, \rho_n) \in \mathbb{R}^{d \times d},$$

with $\rho_n = \frac{1}{\sqrt{n}}$.

KL upper bound

We choose $\mathbb{P}_{\theta_0} = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathbb{P}_{\theta_1} = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ with

$$\boldsymbol{\mu}_0 = \mathbf{0}_d \in \mathbb{R}^d, \qquad \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(d_1, d_1 + 1, 0) = \mathbf{I}_d \in \mathbb{R}^{d \times d},$$
$$\boldsymbol{\mu}_1 = \frac{1}{\sqrt{dn}} \mathbf{1}_d \in \mathbb{R}^d, \qquad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}(d_1, d_1 + 1, \rho_n) \in \mathbb{R}^{d \times d},$$

with $\rho_n = \frac{1}{\sqrt{n}}$. One can show that

$$\mathrm{KL}\left(\mathbb{P}^n_{\theta_1} || \mathbb{P}^n_{\theta_0}\right) \leq \alpha := \frac{5}{4} \text{ for } n \geq 2$$

KL upper bound

We choose $\mathbb{P}_{\theta_0} = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathbb{P}_{\theta_1} = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ with

$$\boldsymbol{\mu}_0 = \mathbf{0}_d \in \mathbb{R}^d, \qquad \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(d_1, d_1+1, 0) = \mathbf{I}_d \in \mathbb{R}^{d\times d},$$
$$\boldsymbol{\mu}_1 = \frac{1}{\sqrt{dn}}\mathbf{1}_d \in \mathbb{R}^d, \qquad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}(d_1, d_1+1, \rho_n) \in \mathbb{R}^{d\times d},$$

with $\rho_n = \frac{1}{\sqrt{n}}$. One can show that

$$\mathrm{KL}\left(\mathbb{P}^n_{\theta_1} || \mathbb{P}^n_{\theta_0}\right) \le \alpha := \frac{5}{4} \text{ for } n \ge 2$$

by

1. $\mathrm{KL}(\mathbb{P}||\mathbb{Q}) = \sum_{i=1}^n \mathrm{KL}(\mathbb{P}_i||\mathbb{Q}_i)$ for $\mathbb{P} = \otimes_{i=1}^n \mathbb{P}_i$, $\mathbb{Q} = \otimes_{i=1}^n \mathbb{Q}_i$,
2. the analytical formula of $\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0))$.

HSIC lower bound

- Consider the Gaussian kernel: $k(\mathbf{x},\mathbf{y}) = e^{-\frac{\gamma}{2}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbb{R}^d}}$ $(\gamma > 0)$.
- With $F(\theta) := \text{HSIC}_k(\mathbb{P}_\theta)$, and using that $\text{HSIC}_k(\mathbb{P}_{\theta_0}) = 0$:

$$\left|F(\theta_1) - \underbrace{F(\theta_0)}_{=0}\right|^2 = F^2(\theta_1) = \text{HSIC}_k^2(\mathbb{P}_{\theta_1})$$

HSIC lower bound

- Consider the Gaussian kernel: $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\gamma}{2}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbb{R}^d}}$ $(\gamma > 0)$.
- With $F(\theta) := \text{HSIC}_k(\mathbb{P}_\theta)$, and using that $\text{HSIC}_k(\mathbb{P}_{\theta_0}) = 0$:

$$\left|F(\theta_1) - \underbrace{F(\theta_0)}_{=0}\right|^2 = F^2(\theta_1) = \text{HSIC}^2_k(\mathbb{P}_{\theta_1})$$

$$= \text{MMD}^2_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d)\right)$$

HSIC lower bound

- Consider the Gaussian kernel: $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\gamma}{2}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbb{R}^d}}$ $(\gamma > 0)$.
- With $F(\theta) := \text{HSIC}_k(\mathbb{P}_\theta)$, and using that $\text{HSIC}_k(\mathbb{P}_{\theta_0}) = 0$:

$$\left|F(\theta_1) - \underbrace{F(\theta_0)}_{=0}\right|^2 = F^2(\theta_1) = \text{HSIC}_k^2(\mathbb{P}_{\theta_1})$$

$$= \text{MMD}_k^2(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d))$$

$$= \|\mu_k(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) - \mu_k(\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d))\|^2_{\mathcal{H}_k}$$

HSIC lower bound

- Consider the Gaussian kernel: $k(\mathbf{x},\mathbf{y}) = e^{-\frac{\gamma}{2}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbb{R}^d}}$ $(\gamma > 0)$.
- With $F(\theta) := \text{HSIC}_k(\mathbb{P}_\theta)$, and using that $\text{HSIC}_k(\mathbb{P}_{\theta_0}) = 0$:

$$\left|F(\theta_1) - \underbrace{F(\theta_0)}_{=0}\right|^2 = F^2(\theta_1) = \text{HSIC}_k^2(\mathbb{P}_{\theta_1})$$

$$= \text{MMD}_k^2\left(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d)\right)$$

$$= \left\|\mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\right) - \mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d)\right)\right\|^2_{\mathcal{H}_k}$$

$$= \langle (i), (i)\rangle_{\mathcal{H}_k} + \langle (ii), (ii)\rangle_{\mathcal{H}_k} - 2\langle (i), (ii)\rangle_{\mathcal{H}_k}$$

HSIC lower bound

- Consider the Gaussian kernel: $k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\gamma}{2}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbb{R}^d}}$ $(\gamma > 0)$.
- With $F(\theta) := \mathrm{HSIC}_k(\mathbb{P}_\theta)$, and using that $\mathrm{HSIC}_k(\mathbb{P}_{\theta_0}) = 0$:

$$
\begin{aligned}
\left|F(\theta_1) - \underbrace{F(\theta_0)}_{=0}\right|^2 &= F^2(\theta_1) = \mathrm{HSIC}_k^2(\mathbb{P}_{\theta_1}) \\
&= \mathrm{MMD}_k^2\left(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d)\right) \\
&= \left\|\mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\right) - \mu_k\left(\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d)\right)\right\|^2_{\mathcal{H}_k} \\
&= \langle (i), (i) \rangle_{\mathcal{H}_k} + \langle (ii), (ii) \rangle_{\mathcal{H}_k} - 2\langle (i), (ii) \rangle_{\mathcal{H}_k} \\
&\overset{(\dagger)}{=} \left[(2\gamma+1)^{d-2}\left((2\gamma+1)^2 - (2\gamma\rho_n)^2\right)\right]^{-1/2} + \left[(2\gamma+1)^d\right]^{-1/2} \\
&\quad - 2\left[(2\gamma+1)^{d-2}\left((2\gamma+1)^2 - (\gamma\rho_n)^2\right)\right]^{-1/2} =: g(\gamma, \rho_n, d)
\end{aligned}
$$

$(\dagger) \Leftarrow$ analytical formula for $\langle \mu_k(\mathbb{P}), \mu_k(\mathbb{Q}) \rangle_{\mathcal{H}_k}$ for Gaussian $\mathbb{P}$, $\mathbb{Q}$, k.

HSIC lower bound

- Consider the Gaussian kernel: $k(\mathbf{x},\mathbf{y}) = e^{-\frac{\gamma}{2}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbb{R}^d}}$ $(\gamma > 0)$.
- With $F(\theta) := \text{HSIC}_k(\mathbb{P}_\theta)$, and using that $\text{HSIC}_k(\mathbb{P}_{\theta_0}) = 0$:

$$\big|F(\theta_1) - \underbrace{F(\theta_0)}_{=0}\big|^2 = F^2(\theta_1) = \text{HSIC}_k^2(\mathbb{P}_{\theta_1})$$

$$= \text{MMD}_k^2(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d))$$

$$= \|\mu_k(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) - \mu_k(\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}_d))\|^2_{\mathcal{H}_k}$$

$$= \langle (i), (i)\rangle_{\mathcal{H}_k} + \langle (ii), (ii)\rangle_{\mathcal{H}_k} - 2\langle (i), (ii)\rangle_{\mathcal{H}_k}$$

$$\overset{(\dagger)}{=} \left[(2\gamma+1)^{d-2}\left((2\gamma+1)^2 - (2\gamma\rho_n)^2\right)\right]^{-1/2} + \left[(2\gamma+1)^d\right]^{-1/2}$$

$$- 2\left[(2\gamma+1)^{d-2}\left((2\gamma+1)^2 - (\gamma\rho_n)^2\right)\right]^{-1/2} =: g(\gamma, \rho_n, d) \overset{(\ddagger)}{\geq} \frac{c}{n} =: (2s_n)^2$$

$(\dagger) \Leftarrow$ analytical formula for $\langle \mu_k(\mathbb{P}), \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k}$ for Gaussian $\mathbb{P}$, $\mathbb{Q}$, k,

$(\ddagger) \Leftarrow$ function analysis, $\rho_n = \frac{1}{\sqrt{n}}$, $c = \frac{\gamma^2}{(2\gamma+1)^2\sqrt{(2\gamma+1)^d}} > 0$.

Result

Theorem (Lower bound for HSIC estimation on $\mathbb{R}^d$)

$\mathcal{P}$:= *any class of Borel probability measures containing the d-dimensional Gaussians,* $k = \otimes_{m=1}^{M} k_m$ *with* $k_m : \mathbb{R}^{d_m} \times \mathbb{R}^{d_m} \to \mathbb{R}$ *continuous bounded shift-invariant characteristic kernels. Then, there exists a constant* $C > 0$*, such that for any* $n \geq 2$

$$\inf_{\hat{F}_n} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq \frac{C}{\sqrt{n}} \right\} \geq \frac{1 - \sqrt{\frac{5}{8}}}{2}.$$

Result

Theorem (Lower bound for HSIC estimation on $\mathbb{R}^d$)

$\mathcal{P}$:= any class of Borel probability measures containing the d-dimensional Gaussians, $k = \otimes_{m=1}^M k_m$ with $k_m : \mathbb{R}^{d_m} \times \mathbb{R}^{d_m} \to \mathbb{R}$ continuous bounded shift-invariant characteristic kernels. Then, there exists a constant $C > 0$, such that for any $n \geq 2$

$$\inf_{\hat{F}_n} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^n \left\{ \left| \mathrm{HSIC}_k(\mathbb{P}) - \hat{F}_n \right| \geq \frac{C}{\sqrt{n}} \right\} \geq \frac{1 - \sqrt{\frac{5}{8}}}{2}.$$

Notes:

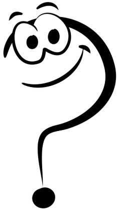
- Gaussian case: $C = \frac{\gamma}{2(2\gamma+1)^{\frac{d}{4}+1}} > 0$.
- Proof of the general case $\Leftarrow$ Bochner theorem.
- Frequently-used HSIC estimators are minimax-optimal on $\mathbb{R}^d$.

Summary

- HSIC can not be estimated faster on $\mathbb{R}^d$ than $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.
- Open: $\mathcal{X}_m \neq \mathbb{R}^d$. Note: universal $(k_m)_{m=1}^M$-s for valid HSIC_k.
- Paper on arXiv.
- ITE toolbox (`https://bitbucket.org/szzoli/ite/`).

Summary

- HSIC can not be estimated faster on $\mathbb{R}^d$ than $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.
- Open: $\mathcal{X}_m \neq \mathbb{R}^d$. Note: universal $(k_m)_{m=1}^M$-s for valid HSIC_k.
- Paper on arXiv.
- ITE toolbox (`https://bitbucket.org/szzoli/ite/`).



Supplement

- Characteristic kernels on $\mathbb{R}^d$.
- Bochner integral.

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

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

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

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

Contents

Bochner integral

[Diestel and Uhl, 1977, Dinculeanu, 2000, Steinwart and Christmann, 2008]

- Given:
  - $(\mathcal{X}, \mathcal{A}, \mu)$: σ-finite measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to \mathcal{H}$-valued function (note: Banach-valued f ✓).

Bochner integral

[Diestel and Uhl, 1977, Dinculeanu, 2000, Steinwart and Christmann, 2008]

- Given:
  - $(\mathcal{X}, \mathcal{A}, \mu)$: σ-finite measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to \mathcal{H}$-valued function (note: Banach-valued f ✓).
- For $f = \sum_{i=1}^{n} c_i \chi_{A_i}$ $(A_i \in \mathcal{A}, c_i \in \mathcal{H})$ step functions

$$\int_{\mathcal{X}} f \mathrm{d}\mu := \sum_{i=1}^{n} c_i \mu(A_i) \in \mathcal{H}.$$

Bochner integral

[Diestel and Uhl, 1977, Dinculeanu, 2000, Steinwart and Christmann, 2008]

- Given:
  - $(\mathcal{X}, \mathcal{A}, \mu)$: σ-finite measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to \mathcal{H}$-valued function (note: Banach-valued f ✓).
- For $f = \sum_{i=1}^n c_i \chi_{A_i}$ $(A_i \in \mathcal{A}, c_i \in \mathcal{H})$ step functions

$$\int_{\mathcal{X}} f \mathrm{d}\mu := \sum_{i=1}^{n} c_i \mu(A_i) \in \mathcal{H}.$$

- f measurable function is Bochner μ-integrable if
  - $\exists$ $(f_n)_{n\in\mathbb{N}}$ step functions: $\lim_{n\to\infty} \int_{\mathcal{X}} \|f - f_n\|_{\mathcal{H}} \mathrm{d}\mu = 0$.
  - In this case $\lim_{n\to\infty} \int_{\mathcal{X}} f_n \mathrm{d}\mu$ exists, $=: \int_{\mathcal{X}} f \mathrm{d}\mu$.

Bochner integral: properties

- $f : \mathcal{X} \to \mathcal{H}$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_{\mathcal{H}} \, \mathrm{d}\mu < \infty$.

Bochner integral: properties

- $f:\mathcal{X}\to\mathcal{H}$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}}\|f\|_{\mathcal{H}}\,\mathrm{d}\mu<\infty$.
- In this case $\left\|\int_{\mathcal{X}} f\mathrm{d}\mu\right\|_{\mathcal{H}}\le\int_{\mathcal{X}}\|f\|_{\mathcal{H}}\,\mathrm{d}\mu$. ('Jensen inequality')

Bochner integral: properties

- $f : \mathcal{X} \to \mathcal{H}$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_{\mathcal{H}} \,\mathrm{d}\mu < \infty$.
- In this case $\left\|\int_{\mathcal{X}} f \mathrm{d}\mu\right\|_{\mathcal{H}} \le \int_{\mathcal{X}} \|f\|_{\mathcal{H}} \,\mathrm{d}\mu$. ('Jensen inequality')
- In our context: $\mathcal{H} = \mathcal{H}_k$,

$$\mu_k(\mu) \text{ exists iff. } \int_{\mathcal{X}} \underbrace{\|k(\cdot, x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}} \mathrm{d}\mu(x) < \infty.$$

Specifically: for bounded kernel ($\sup_{x,x' \in \mathcal{X}} k(x, x') < \infty$) ✓.

Bochner integral: properties – continued

- If
  - $S : B \to B_2$: bounded linear operator,
  - $f : X \to B$: Bochner integrable, then

$S \circ f : X \to B_2$ is Bochner integrable and

$$S\left(\int_{\mathcal{X}} f \,\mathrm{d}\mu\right) = \int_{\mathcal{X}} Sf \,\mathrm{d}\mu.$$

Bochner integral: properties – continued

- If
  - $S : B \to B_2$: bounded linear operator,
  - $f : X \to B$: Bochner integrable, then

  $S \circ f : X \to B_2$ is Bochner integrable and

$$S\left(\int_{\mathcal{X}} f \mathrm{d}\mu\right) = \int_{\mathcal{X}} Sf \mathrm{d}\mu.$$

In short

$|\int f \mathrm{d}\mu| \leq \int |f| \mathrm{d}\mu$ and $c \int f \mathrm{d}\mu = \int cf \mathrm{d}\mu$ generalize nicely.

Contents

Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022).
Adaptive test of independence based on HSIC measures.
The Annals of Statistics, 50(2):858–879.

Anaya-Isaza, A. and Mera-Jiménez, L. (2022).
Data augmentation and transfer learning for brain tumor detection in magnetic resonance imaging.
IEEE Access, 10:23217–23233.

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.

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

Bilodeau, M. and Nangue, A. G. (2017).
Tests of mutual or serial independence of random vectors with applications.
Journal of Machine Learning Research, 18:1–40.

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.

Bouche, D., Flamary, R., d'Alché Buc, F., Plougonven, R., Clausel, M., Badosa, J., and Drobinski, P. (2023).
Wind power predictions from nowcasts to 4-hour forecasts: a learning approach with variable selection.
Renewable Energy.

Cam, L. L. (1973).
Convergence of estimates under dimensionality restrictions.
The Annals of Statistics, 1:38–53.

Camps-Valls, G., Mooij, J. M., and Schölkopf, B. (2010).
Remote sensing feature selection by kernel dependence measures.
IEEE Geoscience and Remote Sensing Letters, 7(3):587–591.

Chakraborty, S. and Zhang, X. (2019).
Distance metrics for measuring joint dependence with application to causal inference.

Journal of the American Statistical Association, 114(528):1638–1650.

Climente-González, H., Azencott, C.-A., Kaski, S., and Yamada, M. (2019).
Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data.
Bioinformatics, 35(14):i427–i435.

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.

Diestel, J. and Uhl, J. J. (1977).
Vector Measures.
American Mathematical Society. Providence.

Dinculeanu, N. (2000).
Vector Integration and Stochastic Integration in Banach Spaces.
Wiley.

Draief, M., Kutzkov, K., Scaman, K., and Vojnovic, M. (2018).
KONG: Kernels for ordered-neighborhood graphs.
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4055–4064.

Fedorov, A., Geenjaar, E., Wu, L., Sylvain, T., DeRamus, T. P., Luck, M., Misiura, M., Mittapalle, G., Hjelm, R. D., Plis, S. M., et al. (2024).
Self-supervised multimodal learning for group inferences from MRI data: Discovering disorder-relevant brain regions and multimodal links.
NeuroImage, 285:120485.

Fellmann, N., Blanchet-Scalliet, C., Helbert, C., Spagnol, A., and Sinoquet, D. (2023).
Kernel-based sensivity analysis for (excursion) sets.
Technical report.
(https://arxiv.org/abs/2305.09268).

Freitas Gustavo, M., Hellström, M., and Verstraelen, T. (2023).
Sensitivity analysis for ReaxFF reparametrization using the Hilbert–Schmidt independence criterion.
Journal of Chemical Theory and Computation, 19(9):2557–2573.

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.

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.

Górecki, T., Krzyśko, M., and Wolyński, W. (2018).
Independence test and canonical correlation analysis based on the alignment between kernel matrices for multivariate functional data.
Artificial Intelligence Review, pages 1–25.

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.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. (2008).
A kernel statistical test of independence.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–592.

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.

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.

Herrando-Pérez, S. and Saltré, F. (2024).

Estimating extinction time using radiocarbon dates.
Quaternary Geochronology, 79:101489.

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. (2023).
Nyström M-Hilbert-Schmidt independence criterion.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1005–1015.

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.

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.

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.

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.

Mooij, J., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016).
Distinguishing cause from effect using observational data: Methods and benchmarks.
Journal of Machine Learning Research, 17:1–102.

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.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2018).
Kernel-based tests for joint independence.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(1):5–31.

Podkopaev, A., Blöbaum, P., Kasiviswanathan, S., and Ramdas, A. (2023).
Sequential kernelized independence testing.
In *International Conference on Machine Learning (ICML)*, pages 27957–27993.

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.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021).
Toward causal representation learning.
Proceedings of the IEEE, 109(5):612–634.

Schulz, T. H., Welke, P., and Wrobel, S. (2022).
Graph filtration kernels.
In *AAAI Conference on Artifical 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. (2013).
A kernel test for three-variable interactions.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1124–1132.

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.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012).
Feature selection via dependence maximization.
Journal of Machine Learning Research, 13(1):1393–1434.

Song, L., Smola, A. J., Gretton, A., and Borgwardt, K. M. (2007).
A dependence maximization view of clustering.
In *International Conference on Machine Learning (ICML)*, pages 815–822.

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.

Stenger, J., Gamboa, F., Keller, M., and Iooss, B. (2020).
Optimal uncertainty quantification of a risk measurement from a thermal-hydraulic code using canonical moments.
International Journal for Uncertainty Quantification, 10(1).

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

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.

Tsuda, K., Kin, T., and Asai, K. (2002).
Marginalized kernels for biological sequences.
Bioinformatics, 18:268–275.

Tsybakov, A. B. (2009).
Introduction to Nonparametric Estimation.
Springer.

Veiga, S. D. (2015).
Global sensitivity analysis with dependence measures.
Journal of Statistical Computation and Simulation, 85(7):1283–1305.

Vishwanathan, S. N., Schraudolph, N., Kondor, R., and Borgwardt, K. (2010).
Graph kernels.
Journal of Machine Learning Research, 11:1201–1242.

Wang, A., Du, J., Zhang, X., and Shi, J. (2022).
Ranking features to promote diversity: An approach based on sparse distance correlation.
Technometrics, 64(3):384–395.

Watkins, C. (1999).
Dynamic alignment kernels.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 39–50.

Wehbe, L. and Ramdas, A. (2015).
Nonparametric independence testing for small sample sizes.
In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3777–3783.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M. (2014).
High-dimensional feature selection by feature-wise kernelized lasso.
Neural Computation, 26(1):185–207.