# The Minimax Rate of HSIC Estimation for Translation-Invariant Kernels*

Florian Kalinke[1] and Zoltán Szabó[2]

[1]Institute for Program and Data Structures, Karlsruhe Institute of Technology [2]Department of Statistics, London School of Economics

## Quick Summary


- Hilbert-Schmidt independence criterion (HSIC [1, 2, 3, 4]; aka. distance covariance): popular dependency measure.
- Applications: feature selection, causal discovery, independence testing, clustering, sensitivity analysis, uncertainty quantification.
- Many known estimators converge at a rate of $\mathcal{O}_P\left(n^{-1/2}\right)$.
- Contribution: For a large class of distributions and kernels on $\mathbb{R}^d$, $\mathcal{O}_P\left(n^{-1/2}\right)$ is the optimal rate.


## HSIC

- Given $X = (X_m)_{m=1}^M \sim \mathbb{P}$ on $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, $\mathcal{X}_m$ is equipped with kernel $k_m$ and feature map $\phi_{k_m} : \mathcal{X}_m \to \mathcal{H}_{k_m}$, HSIC takes the form

$$\mathrm{HSIC}_k(\mathbb{P}) = \left\| \mu_k(\mathbb{P}) - \mu_k\left(\otimes_{m=1}^M \mathbb{P}_m\right)\right\|_{\mathcal{H}_k}, \qquad k := \otimes_{m=1}^M k_m$$

with $\otimes_{m=1}^M \mathbb{P}_m$ the product of the marginal distributions $\mathbb{P}_m$, $m \in [M] := \{1, \dots, M\}$, and $\mu_k(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}[\phi_k(X)]$.

- We focus on $\mathcal{X} = \mathbb{R}^d$, $\mathcal{X}_m = \mathbb{R}^{d_m}$, $d = \sum_{m=1}^M d_m$; $\mathbb{R}^d = \times_{m=1}^M \mathbb{R}^{d_m}$.
- Gaussian kernels: $k_m(\mathbf{x}, \mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|^2_{\mathbb{R}^{d_m}}}$ $(\gamma > 0)$.
- Translation-invariant kernels: There exists $\psi_m : \mathbb{R}^{d_m} \to \mathbb{R}$ such that $k_m(\mathbf{x}, \mathbf{y}) = \psi_m(\mathbf{x} - \mathbf{y})$. Example: Gaussian kernel.

## Our Goal: 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 there exists $c > 0$ such that

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

- An estimator with a matching upper bound is called minimax-optimal.
- Wanted: $\xi_n \asymp n^{-1/2}$.

## Tool: Le Cam's Method

- Key [5]: There exist $\alpha > 0$ and a positive sequence $(s_n)_{n=1}^\infty$ such that for any fixed $n$, there exists an adversarial pair $\left(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}\right) \in \mathcal{P} \times \mathcal{P}$ s.t. (i) $\mathrm{KL}\left(\mathbb{P}^n_{\theta_1} || \mathbb{P}^n_{\theta_0}\right) \leq \alpha$, and (ii) $\left|\mathrm{HSIC}_k(\mathbb{P}_{\theta_1}) - \mathrm{HSIC}_k(\mathbb{P}_{\theta_0})\right| \geq 2s_n > 0$.
- Then, for all $n$,

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

## Our Adversarial Pair

- Let $\mathcal{G}$ be $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ Gaussians on $\mathbb{R}^d = \times_{m=1}^M \mathbb{R}^{d_m}$ 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 $i = d_1$, $j = d_1 + 1$, $\rho \in (-1, 1)$.

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

## Proof Sketch

- We use the reduction

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

- For our adversarial pair $\left(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}\right)$, one can show that

(i) $\mathrm{KL}\left(\mathbb{P}^n_{\theta_1} || \mathbb{P}^n_{\theta_0}\right) \leq \alpha := \frac{5}{4}$ for $n \geq 2$ (Gaussians $\Rightarrow$ closed-form), and

(ii) $\left|\mathrm{HSIC}_k(\mathbb{P}_{\theta_1}) - \mathrm{HSIC}_k(\mathbb{P}_{\theta_0})\right| \geq 2s_n := 2\frac{c}{\sqrt{n}} > 0$.

## Main Result

- Let $\mathcal{P}$ be 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 $n \geq 2$

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

- Gaussian $k_m$-s: $c = \frac{\gamma}{2(2\gamma+1)^{\frac{d}{4}+1}} > 0$.
- General case by Bochner's theorem ($c > 0$).

## Discussion

- Many of the existing HSIC estimators on $\mathbb{R}^d$ are minimax-optimal.
- Existing lower bounds (MMD [6], mean embedding [7], covariance operator [8]) rely on adversarial pairs that do not work for HSIC.
- Corollary: Lower bound of $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ on covariance operator estimation, recovering [8].

## References


[1] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory (ALT)*, pages 63–78, 2005.

[2] Novi Quadrianto, Le Song, and Alex Smola. Kernelized sorting. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1289–1296, 2009.

[3] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

[4] Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

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

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

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

[8] Yang Zhou, Di-Rong Chen, and Wei Huang. A class of optimal estimators for the covariance operator in reproducing kernel Hilbert spaces. *Journal of Multivariate Analysis*, 169:166–178, 2019.


*NeurIPS–2024, Vancouver, Canada. Acknowledgements. This work was supported by the German Research Foundation (DFG) Research Training Group GRK 2153: Energy Status Data — Informatics Methods for its Collection, Analysis and Exploitation, and by the pilot program Core-Informatics of the Helmholtz Association (HGF).