The Minimax Rate of HSIC Estimation for Translation-Invariant Kernels

Florian Kalinke[1] and Zoltán Szabó[2]

[1]Karlsruhe Institute of Technology (KIT) [2]London School of Economics (LSE)

Hilbert-Schmidt independence criterion (HSIC)

- Aka. distance covariance.
- Easy-to-estimate and popular dependency measure for $M \geqslant 2$ random variables.
- Many applications: feature selection, causal discovery, independence testing, clustering, sensitivity analysis, uncertainty quantification, independent subspace analysis, ...
- Idea: Check if the joint distribution equals the product of its marginals in RKHS.

Hilbert-Schmidt independence criterion (HSIC)

- Aka. distance covariance.
- Easy-to-estimate and popular dependency measure for $M \geqslant 2$ random variables.
- Many applications: feature selection, causal discovery, independence testing, clustering, sensitivity analysis, uncertainty quantification, independent subspace analysis, ...
- Idea: Check if the joint distribution equals the product of its marginals in RKHS.
- Formally (μ_k := mean embedding; $\mathbb{P}$:= joint measure; $\otimes_{m=1}^{M} \mathbb{P}_m$:= product of marginals):

$$\mathrm{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^{M} \mathbb{P}_m\right) = \Big\| \underbrace{\mu_k(\mathbb{P}) - \mu_k\left(\otimes_{m=1}^{M} \mathbb{P}_m\right)}_{= \text{(centered) covariance operator}} \Big\|_{\mathcal{H}_k} =: \mathrm{HSIC}_k\left(\mathbb{P}\right),$$

with $k = \otimes_{m=1}^{M} k_m$, $X = (X_m)_{m=1}^{M} \in \mathcal{X} = \times_{m=1}^{M} \mathcal{X}_m$, and $X \sim \mathbb{P}$.

Our contribution

Question:

Can HSIC be estimated faster than $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ for n samples?

Our contribution

Question:

Can HSIC be estimated faster than $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ for n samples?

Answer (our contribution):

On $\mathbb{R}^d$: **No!**

Formal statement

- We measure the independence of $X = (X_m)_{m=1}^M \in \mathbb{R}^d = \times_{m=1}^M \mathbb{R}^{d_m}$, $X \sim \mathbb{P}$.

Theorem (main result; simplified)

$\mathcal{P}$:= any class of Borel probability measures containing the d-dimensional Gaussians, $\hat{F}_n$:= any estimator of HSIC based on n samples, $k = \otimes_{m=1}^M k_m$ 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 $n \geqslant 2$

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

Formal statement

- We measure the independence of $X = (X_m)_{m=1}^M \in \mathbb{R}^d = \times_{m=1}^M \mathbb{R}^{d_m}$, $X \sim \mathbb{P}$.

Theorem (main result; simplified)

$\mathcal{P}$:= any class of Borel probability measures containing the d-dimensional Gaussians, $\hat{F}_n$:= any estimator of HSIC based on n samples, $k = \otimes_{m=1}^M k_m$ 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 $n \geqslant 2$

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

Notes:

- Proof: construct adversarial distribution pair; show that it satisfies requirements of Le Cam's method.
- Gaussian case: $c = \frac{\gamma}{2(2\gamma+1)^{\frac{d}{4}+1}} > 0$; general case: from Bochner's theorem ($c > 0$).
- Take-away: frequently-used HSIC estimators are minimax-optimal on $\mathbb{R}^d$.

Summary

- HSIC cannot be estimated faster than $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ on $\mathbb{R}^d$.
- Implies minimax-optimality of many existing estimators.
- Open: lower bounds for HSIC estimation beyond $\mathbb{R}^d$.

Questions/comments: Poster **ID 95630**.