Nyström Kernel Stein Discrepancy*

Florian Kalinke[1], Zoltán Szabó[2], and Bharath K. Sriperumbudur[3]

[1]Karlsruhe Institute of Technology [2]London School of Economics [3]The Pennsylvania State University

Quick Summary

- Kernel Stein discrepancy (KSD; [1, 2]): powerful goodness-of-fit measure and test.
- Applications: Assessing and improving sample quality, validating MCMC methods, comparing deep generative models, ...
- Limitations: Quadratic runtime complexity.
- Main contribution: Accelerated estimator with the same convergence rate as the quadratic-time estimator.

Kernel Stein Discrepancy

- Goal: Test $H_0 : \mathbb{P} = \mathbb{Q}$ vs. $H_1 : \mathbb{P} \neq \mathbb{Q}$ for fixed known target $\mathbb{P}$ and unknown sampling distribution $\mathbb{Q}$, given samples $\hat{\mathbb{Q}}_n := \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ of $\mathbb{Q}$.
- We illustrate the method with the Langevin-Stein operator-based [3] KSD on $\mathbb{R}^d$, which is

$$S_p(\mathbb{Q}) = \left\|\mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}},$$

with kernel ($\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$)

$$\begin{aligned} h_p(\mathbf{x},\mathbf{y}) := &\left\langle \nabla_{\mathbf{x}} \log p(\mathbf{x}), \nabla_{\mathbf{y}} \log p(\mathbf{y})\right\rangle_{\mathbb{R}^d} k(\mathbf{x},\mathbf{y}) \\ &+ \left\langle \nabla_{\mathbf{y}} \log p(\mathbf{y}), \nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{y})\right\rangle_{\mathbb{R}^d} \\ &+ \left\langle \nabla_{\mathbf{x}} \log p(\mathbf{x}), \nabla_{\mathbf{y}} k(\mathbf{x},\mathbf{y})\right\rangle_{\mathbb{R}^d} + \sum_{i=1}^d \frac{\partial^2 k(\mathbf{x},\mathbf{y})}{\partial x_i \partial y_i}, \end{aligned}$$

p the (Lebesgue) density of $\mathbb{P}$, and kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.
- Note that knowledge of the derivative of the score function $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ is enough, i.e., knowledge of p up to normalization suffices.
- Existing KSD estimators take (roughly) the form

$$S_p^2\left(\hat{\mathbb{Q}}_n\right) = \frac{1}{n^2}\sum_{i,j=1}^n h_p(\mathbf{x}_i, \mathbf{x}_j),$$

and have a runtime requirement of $\mathcal{O}(n^2)$.

Nyström-based Estimator (N-KSD)

- Denote by $\tilde{\mathbb{Q}}_m := \{\{\tilde{\mathbf{x}}_i\}\}_{i=1}^m$ a subsample of $\hat{\mathbb{Q}}_n$. The Nyström estimator is

$$\tilde{S}_p^2\left(\hat{\mathbb{Q}}_n\right) = \boldsymbol{\beta}_p^\mathsf{T} \mathbf{K}_{h_p,m,m}^- \boldsymbol{\beta}_p,$$

with $\boldsymbol{\beta}_p = \frac{1}{n}\mathbf{K}_{h_p,m,n}\mathbf{1}_n \in \mathbb{R}^m$, matrices

$$\begin{aligned} \mathbf{K}_{h_p,m,m} &= \left[h_p(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)\right]_{i,j=1}^m \in \mathbb{R}^{m\times m}, \text{ and} \\ \mathbf{K}_{h_p,m,n} &= \left[h_p(\tilde{\mathbf{x}}_i, \mathbf{x}_j)\right]_{i,j=1}^{m,n} \in \mathbb{R}^{m\times n}, \end{aligned}$$

and $\mathbf{A}^-$ denoting the (Moore-Penrose) pseudo-inverse of a matrix $\mathbf{A}$.
- Runtime: $\mathcal{O}(mn + m^3)$, saving if $m = o\left(n^{2/3}\right)$.

Sub-Gaussian Assumption

- Existing Nyström analysis considers bounded kernels only. In practice, h_p is usually unbounded and existing results do not apply.
- Example: Consider $d = 1$, standard normal $p(x) \propto \exp\left(-x^2/2\right)$, and the RBF kernel $k(x,y) = \exp\left(-\gamma(x-y)^2\right)$ ($\gamma > 0$). Then

$$h_p(x,x) = x^2 + 2\gamma \overset{x\to\infty}{\to} \infty.$$

Similarly, for the IMQ kernel $k(x,y) = \left(c^2 + (x-y)^2\right)^{-\beta}$ ($\beta, c > 0$).
- For the Nyström analysis, assume that $\bar{h}_p(\cdot, X) := h_p(\cdot, X) - \mathbb{E}_{X\sim\mathbb{Q}} h_p(\cdot, X)$ with the sampling distribution $\mathbb{Q}$ is sub-Gaussian, that is,

$$\left\|\left\langle \bar{h}_p(\cdot, X), u\right\rangle_{\mathcal{H}_{h_p}}\right\|_{\psi_2} \lesssim \left\|\left\langle \bar{h}_p(\cdot, X), u\right\rangle_{\mathcal{H}_{h_p}}\right\|_{L_2(\mathbb{Q})} < \infty \qquad (1)$$

holds for all $u \in \mathcal{H}_{h_p}$, with a u-independent absolute constant in $\lesssim$, and $\|\cdot\|_{\psi_2}$ denoting the sub-Gaussian norm.

Main Results

- $\sqrt{n}$-consistency of KSD estimator: If

$$\left\|\left\|h_p(\cdot, X)\right\|_{\mathcal{H}_{h_p}}\right\|_{\psi_2} < \infty$$

holds (implied by (1)), then

$$\left|S_p(\mathbb{Q}) - S_p\left(\hat{\mathbb{Q}}_n\right)\right| = \mathcal{O}_P\left(n^{-1/2}\right).$$

- $\sqrt{n}$-consistency of N-KSD: If the sub-Gaussian property (1) holds, then

$$\left|S_p(\mathbb{Q}) - \tilde{S}_p\left(\hat{\mathbb{Q}}_n\right)\right| = O_P\left(n^{-1/2}\right),$$

given that the effective dimension $\mathcal{N}_{\mathbb{Q},\bar{h}_p}(\lambda) := \operatorname{tr}\left(C_{\mathbb{Q},\bar{h}_p,\lambda}^{-1} C_{\mathbb{Q},\bar{h}_p}\right)$ ($C_{\mathbb{Q},\bar{h}_p} := \mathbb{E}_{X\sim\mathbb{Q}}[\bar{h}_p(\cdot,X) \otimes \bar{h}_p(\cdot,X)]$; $C_{\mathbb{Q},\bar{h}_p,\lambda} := C_{\mathbb{Q},\bar{h}_p} + \lambda I$, $\lambda > 0$) either
  - decays polynomially:

$$\mathcal{N}_{\mathbb{Q},\bar{h}_p}(\lambda) \lesssim \lambda^{-\gamma}, \qquad m = \tilde{\Omega}\left(n^{1/(2-\gamma)}\right),$$

  for $\gamma \in (0,1]$ (computational savings if $\gamma < 1/2$), or
  - decays exponentially:

$$\mathcal{N}_{\mathbb{Q},\bar{h}_p}(\lambda) \lesssim \log(1 + c_1/\lambda), \qquad m = \tilde{\Omega}\left(n^{1/2}\right),$$

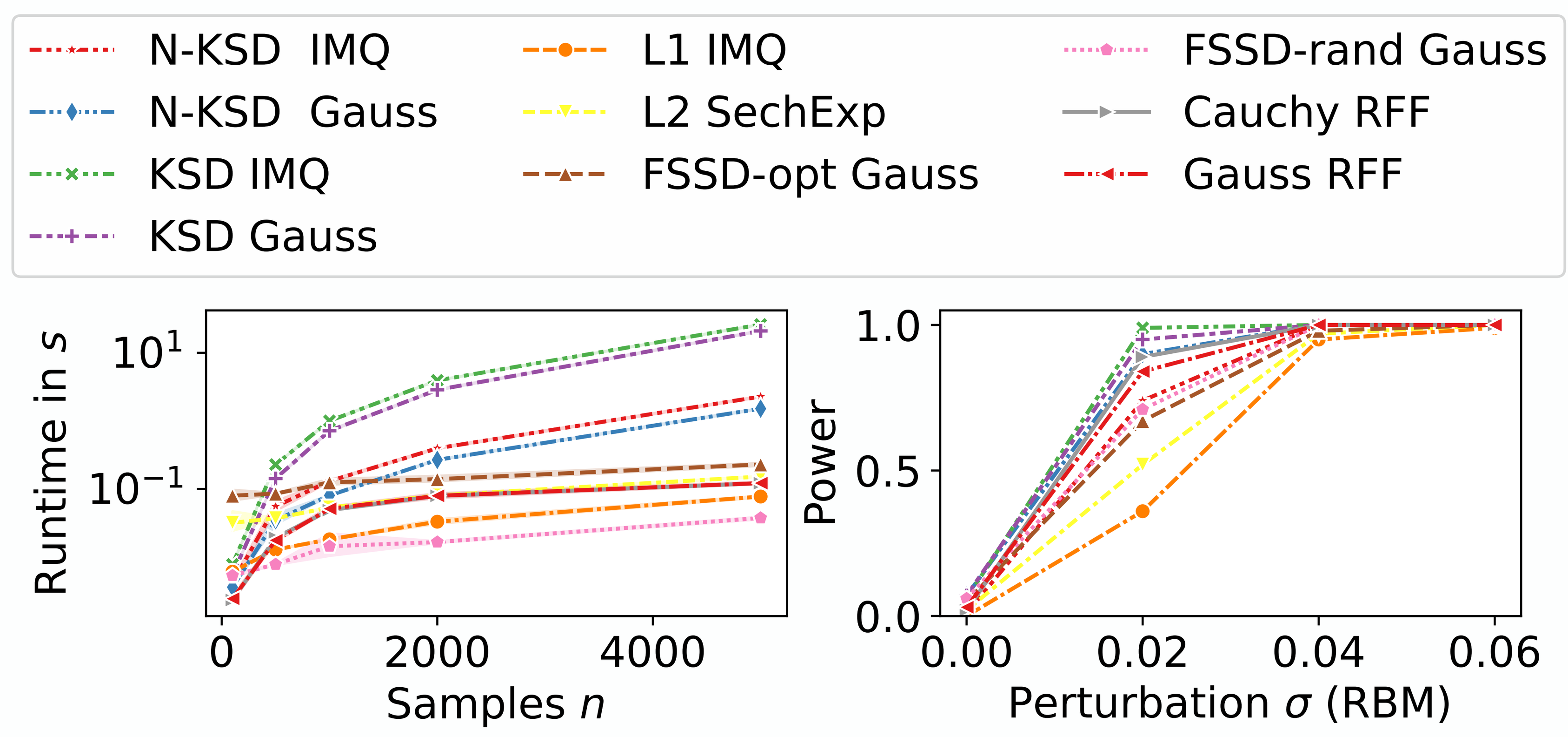
  for some $c_1 > 0$ (computational savings if n is large enough).
- The decay of the effective dimension can be linked to the decay of the eigenvalues of the covariance operator $C_{\mathbb{Q},\bar{h}_p}$ [4, Proposition 4, 5].

Discussion

- Unboundedness of the feature map handled by sub-Gaussian assumption.
- The quadratic-time and the N-KSD estimator both have $\sqrt{n}$-consistency, i.e., computational gain with no loss in statistical accuracy.
- Our results apply in the general KSD framework [5].
- Open: Weaker assumption for the Nyström case.

Goodness-of-fit Benchmark

- Runtime and power of Nyström KSD (N-KSD) and competitors on the restricted Boltzmann machine (RBM) goodness-of-fit benchmark.



- Code: `https://github.com/flopska/nystroem-ksd`

References

[1] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning (ICML)*, pages 276–284, 2016.

[2] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML)*, pages 2606–2615, 2016.

[3] Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 226–234, 2015.

[4] Andrea Della Vecchia, Jaouad Mourtada, Ernesto De Vito, and Lorenzo Rosasco. Regularized ERM on random subspaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4006–4014, 2021.

[5] Omar Hagrass, Bharath Sriperumbudur, and Krishnakumar Balasubramanian. Minimax optimal goodness-of-fit testing with kernel Stein discrepancy. *Bernoulli*, 2025. (accepted; preprint: `https://arxiv.org/abs/2404.08278`).

*AISTATS–2025, Mai Khao, Thailand. Acknowledgements. This work was supported by the pilot program Core-Informatics of the Helmholtz Association (HGF). BKS is partially supported by the National Science Foundation (NSF) CAREER award DMS-1945396.