# Using Perturbation to Improve Goodness-of-Fit Tests based on KSD

Imperial College London

Xing Liu [1], Andrew B. Duncan [1, 2], Axel Gandy [1]

[1]Imperial College London [2]Alan Turing Institute

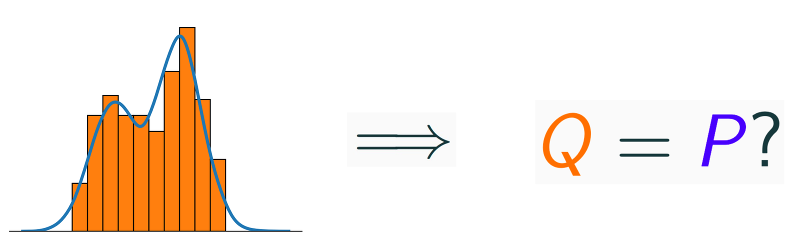
## Background

Given an alternative distribution $Q$ and a target distribution $P$ on $\mathcal{X} := \mathbb{R}^d$, we wish to test $H_0 : Q = P$ vs. $H_1 : Q \neq P$.

Assumptions:

- $P$ admits a positive, continuously differentiable Lebesgue **density** $p$ on $\mathcal{X}$, which can be evaluated **up to a normalising constant**.
- Sampling from $P$ is hard, but i.i.d. realisations $\{x_i\}_{i=1}^n \sim Q$ are available.

Example: **Bayesian analysis**, where $P$ = target posterior, and $Q$ = empirical distribution of samples drawn from a sampler targeting $P$.




## GOF tests with KSD [1, 2]

Idea: choose a **statistical divergence** $\mathbb{D}$ that satisfies $\mathbb{D}(Q, P) \geq 0$ with equality iff. $Q = P$, and test $H_0 : \mathbb{D}(Q, P) = 0$ against $H_1 : \mathbb{D}(Q, P) > 0$.

Definition (KSD) Let $\mathcal{F}$ = unit ball of a *reproducing kernel Hilbert space* (RKHS) with a p.d. kernel $k$. Define $\mathcal{A}_p f(x) := \langle \nabla_x \log p(x), f(x) \rangle + \langle \nabla, f(x) \rangle$ for continuously differentiable $f : \mathbb{R}^d \to \mathbb{R}^d$, and $s_p(x) = \nabla_x \log p(x)$. The **(Langevin) kernelized Stein discrepancy (KSD)** is

$$\mathbb{D}(Q, P) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{x \sim Q}[\mathcal{A}_P f(x)]| = \mathbb{E}_{x, x' \sim Q}[u_P(x, x')], \tag{1}$$

$$u_P(x, x') := s_p(x)^\top k(x, x') s_p(x') + s_p(x)^\top \nabla_{x'} k(x, x') + \nabla_x k(x, x')^\top s_p(x') + \sum_{i=1}^d \frac{\partial^2}{\partial x_i \partial x_i'} k(x, x') \,.$$

## Myopia of the KSD test

Setup: Consider $Q = \mathcal{N}(0, I_d)$ and a **sequence** of targets $P_\nu = \pi \mathcal{N}(0, I_d) + (1 - \pi) \mathcal{N}(\Delta_\nu, I_d)$, where $\pi \in [0, 1]$ and $\Delta_\nu \in \mathbb{R}^d$ for each $\nu = 1, 2, \ldots$. Let $\{x_i\}_{i=1}^\infty \sim Q$ i.i.d., and suppose $\|\Delta_\nu\|_2 \to \infty$ as $\nu \to \infty$.

Blindness of KSD [3] says $\mathbb{D}(Q, P_\nu) \to 0$ as $\nu \to \infty$. **What about the test power?**

Proposition 1. (Informal) Let $n_1, n_2, \cdots \in \mathbb{N}$ be such that $n_\nu = o\left(e^{\|\Delta_\nu\|_2^2/64}\right)$. Under regularity conditions,

$$n_\nu \hat{\mathbb{D}}_{P_\nu} \to_d R_{Q,k} \quad (\nu \to \infty) \tag{2}$$

where $\hat{\mathbb{D}}_{P_{\Delta_\nu}}$ is the sample KSD computed using $\{x_i\}_{i=1}^{n_\nu}$, and $R_{Q,k}$ is its limiting distribution **under** $H_0$, which depends only on $Q$ and $k$.

### What does Proposition 1 tell us?

For multi-modal target distributions, the KSD test power can converge to the **prescribed test level** unless the sample size grows **unrealistically fast** with the mode separation (**Figure 1a**).

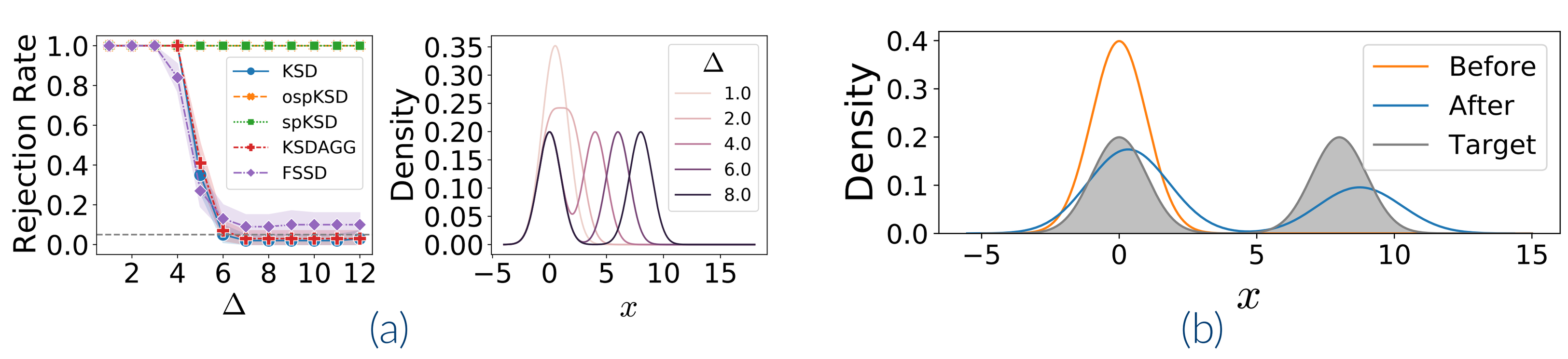



Figure 1. **Power** of the proposed (ospksd, spKSD) and benchmark tests. $P = 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(\Delta, 1)$ and samples are drawn from $Q$ = left component. **(a)** Power and target densities for varying $\Delta$. **(b)** Densities of $P$ and $Q$ before and after 10 steps of the perturbation $\mathcal{K}$.

## References

[1] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *ICML*, Proceedings of Machine Learning Research. PMLR, 2016.

[2] Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML*, Proceedings of Machine Learning Research. PMLR, 2016.

[3] L. K. Wenliang and H. Kanagawa. Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*, 2020.

Table 1. Number of rejected GOF tests over 10 repetitions with level 0.05.

| RAM scale | 0.1 | 0.5 | 1.08 |
|---|---|---|---|
| KSD | 0 | 0 | 0 |
| KSDAgg | 1 | 0 | 1 |
| FSSD | 0 | 1 | 7 |
| pKSD (ours) | 10 | 1 | 6 |

## pKSD test (proposed)

Given i.i.d. $\{x_i\}_{i=1}^n \sim Q$, partition into train set $\mathcal{D}_{\text{train}}$ and test set $\mathcal{D}_{\text{test}}$.

1. **Estimate** the location and Hessian of the modes of $p$ and **select** the optimal hyperparameters of $\mathcal{K}$ using $\mathcal{D}_{\text{train}}$.
2. **Perturb** $\mathcal{D}_{\text{test}}$ with $\mathcal{K}$, and **compute** an estimate $\hat{\mathbb{D}}(Q, P; \mathcal{K})$ for (3) as the test statistic.
3. Use a bootstrap technique to approximate the critical value $\hat{\gamma}_{1-\alpha}$.
4. Reject $H_0$ if $\hat{\mathbb{D}}(Q, P; \mathcal{K}) \geq \hat{\gamma}_{1-\alpha}$.

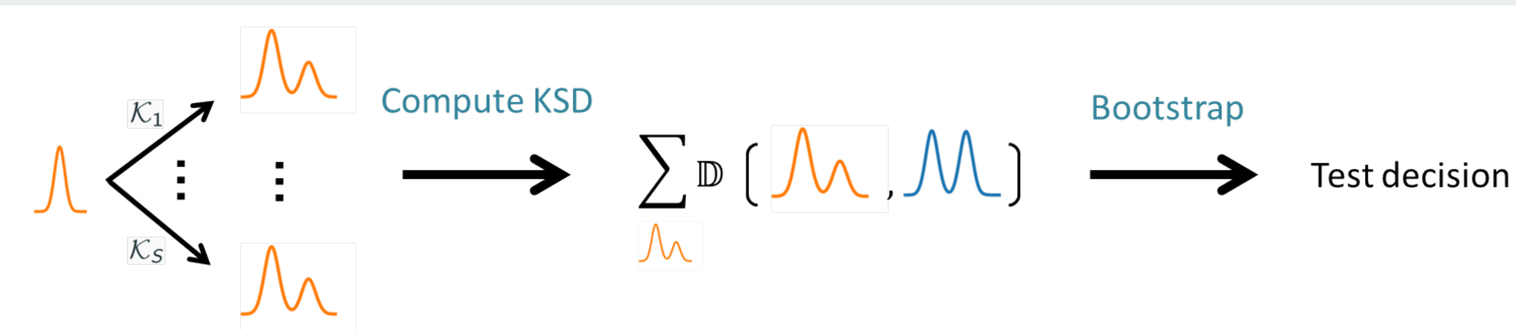



Figure 2. Illustration of GOD test with pKSD and multiple perturbations $\mathcal{K}_1, \ldots, \mathcal{K}_S$.

## Perturbed kernelized Stein discrepancy (pKSD)

Main idea: Perturb both $Q$ and $P$ with a **Markov transition kernel** $\mathcal{K}$ **that leaves** $P$ **invariant**, and perform KSD test on the perturbed distributions.

Why using $\mathcal{K}$ helps? Using estimated information about the modes, we can design $\mathcal{K}$ that **turns a "local" discrepancy such as missing modes into a "global" one** that KSD can detect (**Figure 1b**).

Definition (pKSD). Given a Markov transition kernel $\mathcal{K}$, the pKSD is

$$\mathbb{D}(Q, P; \mathcal{K}) := \mathbb{D}(\mathcal{K}Q, \mathcal{K}P) = \sup_{f \in \mathcal{F}^d} |\mathbb{E}_{x \sim \mathcal{K}Q}[\mathcal{A}_{\mathcal{K}P} f(x)]|, \tag{3}$$

where $(\mathcal{K}Q)(\cdot) := \int_{\mathcal{X}} \mathcal{K}(x, \cdot) Q(dx)$ is the perturbed measure.

### How to choose the perturbation kernel $\mathcal{K}$?

- $P$-invariance Markov transition kernel with MH correction $\to \mathcal{K}P = P$ $\to$ (3) can be computed in **closed form**.
- Jump proposal (Figure 3): $\mathcal{K}$ uses a "jump proposal" that **exchanges probability mass between pairs of modes** to create local discrepancy.
- Mode locations and local Hessians of $p$ are required to construct the jump proposal $\to$ estimated using **optimisation** (e.g., BFGS).

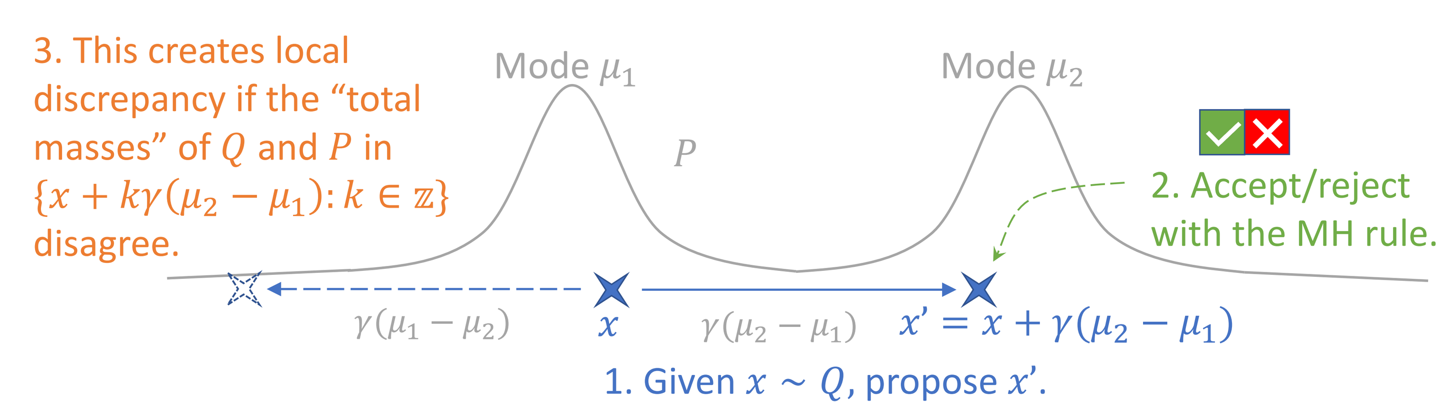



Figure 3. Schematic plot of the jump proposal used in $\mathcal{K}$.

## Experiment: assessing sample quality

Goal: Test $H_0 : Q = P$ vs. $H_1 : Q \neq P$, where $P$ = posterior distribution of a Bayesian model for inferring the **locations of sensors**, and $Q$ = sampling distribution of samples drawn from a MCMC sampler (RAM-$\sigma$, where $\sigma$ is a **tuning-parameter**).

- Figure 4: The posterior samples from some samplers (e.g., RAM-0.1 and RAM-1.08) clearly miss some modes (top row).
- Table 1: These samples are **not rejected** by KSD, KSDAgg or FSSD (benchmarks), but are **rejected** by pKSD (ours).

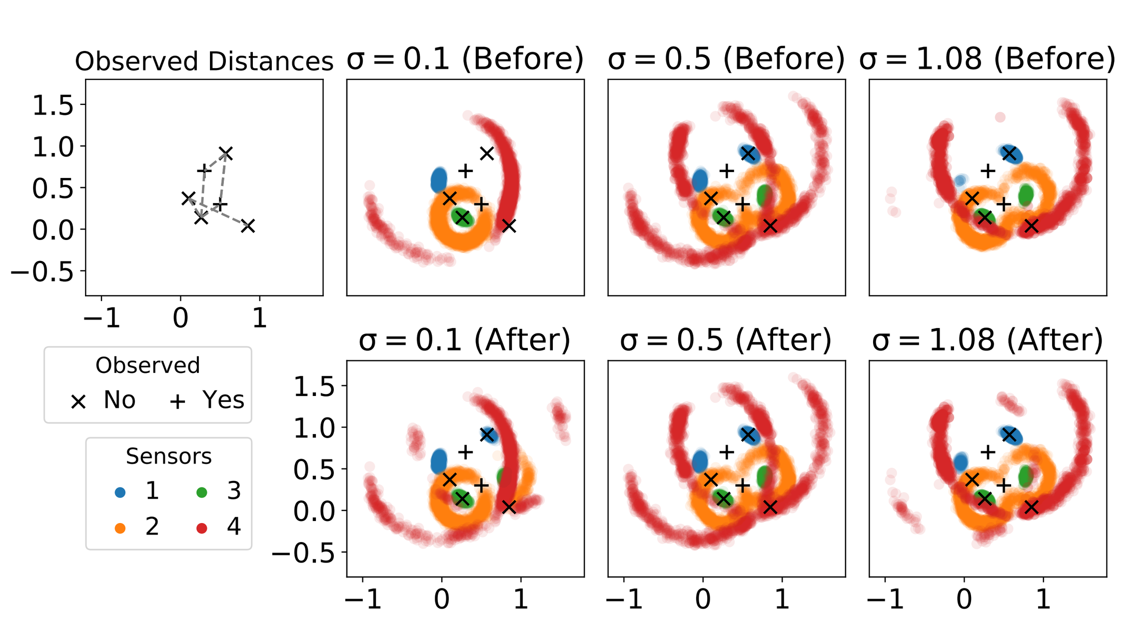



Figure 4. **Posterior plots** of inferred sensor locations before and after perturbation. **Black crosses** $\times$: unobserved sensors; **black pluses** $+$: observed sensors.



xing.liu16@imperial.ac.uk