Using Perturbation to Improve Goodness-of-Fit Tests based on Kernelized Stein Discrepancy

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

Abstract

Kernelized Stein discrepancy (KSD) is a score-based discrepancy widely used in goodness-of-fit tests. It can be applied even when the target distribution has an unknown normalising factor, such as in Bayesian analysis. We show theoretically and empirically that the KSD test can suffer from low power when the target and the alternative distributions have the same well-separated modes but differ in mixing proportions. We propose to perturb the observed sample via Markov transition kernels, with respect to which the target distribution is invariant. This allows us to then employ the KSD test on the perturbed sample. We provide numerical evidence that with suitably chosen transition kernels the proposed approach can lead to substantially higher power than the KSD test.

1. Introduction

Stein discrepancy (SD) (Stein, 1972; Gorham & Mackey, 2015) is a statistical divergence between two probability measures based on Stein's method. More specifically, given two Borel probability measures Q and P supported on $\mathcal{X} \subset \mathbb{R}^d$, the Stein discrepancy is defined to be

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

where $\mathcal{F}$ is a set of functions on $\mathcal{X}$ and $\mathcal{A}_P$ is an operator acting on $\mathcal{F}$ such that $\mathbb{E}_{x \sim Q}[\mathcal{A}_P f(x)] = 0$ for all $f \in \mathcal{F}$ if and only if $Q \equiv P$. When $\mathcal{X} = \mathbb{R}^d$ and P admits a positive, continuously differentiable density p with respect to the Lebesgue measure, then the *Langevin-Stein operator* is the natural candidate for $\mathcal{A}_P$ which has the crucial property that it only depends on the *score function* $s_p(x) := \nabla \log p(x)$ of p, which does not require evaluation of the (possibly intractable) normalising constant of p. When a Reproducing

Kernel Hilbert Space (RKHS) (Berlinet & Thomas-Agnan, 2011) is used to construct the function class, this discrepancy is called the *kernelized Stein discrepancy* (KSD), and admits a closed-form expression. This has made KSD popular for applications involving an unnormalised density, such as Bayesian inference (Liu & Wang, 2016), goodness-of-fit testing (Liu et al., 2016; Chwialkowski et al., 2016; Jitkrittum et al., 2017), and sample quality measurement (Gorham & Mackey, 2015; 2017; Gorham et al., 2019); see Anastasiou et al. (2023) for a review.

We focus on goodness-of-fit (GOF) testing, where independent samples from a *candidate distribution Q* are observed and the goal is to test for evidence against the null hypothesis that Q matches a *target distribution P*. When the density p of P is only available in an unnormalised form (i.e. the normalisation constant is infeasible to compute) and direct sampling from P is infeasible, classical tests such as the Kolmogorov-Smirnov test (Massey, 1951) or two-sample tests (Gretton et al., 2012; Schrab et al., 2021) cannot be used, as they require either a tractable cumulative distribution function or samples from P. A GOF test based on KSD, on the other hand, does not have these limitations.

However, KSD tests may suffer from low test power when the target probability measure has well-separated modes. For example, when Q and P are mixtures of the same components but differ in the mixing proportions, the power will converge to the test level as the modes of the components become more and more separated (Fig. 1). This is because the KSD statistic can be numerically close to 0 if the region where the score functions are practically different has low Q-probability. This issue, sometimes known as the *blindness to isolated components* (Wenliang & Kanagawa, 2020), has been noted in a number of works (Gorham et al., 2019; Matsubara et al., 2022; Kanagawa et al., 2022) and addressed in some applications of KSD (Zhang et al., 2022). However, little work has been devoted to tackling this issue in the context of GOF testing.

Contributions Our contribution is twofold. First, we demonstrate theoretically and numerically with a bimodal Gaussian example that the power of KSD tests can converge to the test level when the target distribution has well-separated modes. This is different from the works

---

[1]Department of Mathematics, Imperial College London, London, UK. [2]Alan Turing Institute, London, UK. Correspondence to: Xing Liu <xing.liu16@imperial.ac.uk>.

of Gorham & Mackey (2017) and Wenliang & Kanagawa (2020), which focus on the convergence of the sample KSD but not its limiting null distribution. Second, we address this issue by introducing a *perturbation operator*, giving rise to a family of perturbation-based GOF test (Fig. 1, bottom right) which we call the *perturbed kernelized Stein discrepancy* (pKSD) test. The role of the operator is to perturb the candidate and the target distributions simultaneously to create discrepancy that can be more easily detected by KSD. We propose to use Markov transition kernels that are invariant to the target P as the perturbation operator. The P-invariance ensures the resulting GOF tests provably control the Type-I error. The transition kernel is non-irreducible and uses an inter-modal jump proposal, which can increase the test power against multi-modal alternatives, sometimes substantially from the nominal level to almost 1.

Outline Section 2 reviews kernelized Stein discrepancy. Section 3 formalises the low-power problem of the KSD test. The proposed method is presented in Section 4 and Section 5. We discuss related work in Section 6, followed by experiments in Section 7. Section 8 concludes.

Notation Throughout this article, we denote by Q, P probability measures on $\mathcal{X} = \mathbb{R}^d$ equipped with the Borel σ-algebra $\mathcal{B}(\mathcal{X})$, and assume P has a continuously differentiable, positive Lebesgue density p. We refer to Q as the *candidate* distribution and P as the *target* distribution. Our interest lies in testing $H_0 : Q = P$ against $H_1 : Q \neq P$ using a finite sample $\{x_i\}_{i=1}^n$ drawn independently from Q. We assume we can evaluate pointwise the *unnormalised* density $p^*(x) = p(x)/Z$, where Z is an unknown constant, as well as $\nabla \log p^*(x)$, which is identical to the *score function* of p, namely $s_p(x) := \nabla \log p(x) = (\nabla_{x_1} \log p(x), \ldots, \nabla_{x_d} \log p(x))^\top$.

2. Kernelized Stein Discrepancy Test

Choosing the Stein operator $\mathcal{A}_P$ in (1) to be the operator mapping continuously differentiable, vector-valued functions $f : \mathbb{R}^d \to \mathbb{R}^d$ to scalar-valued functions via $\mathcal{A}_p f(x) := \langle \nabla \log p(x), f(x) \rangle + \langle \nabla, f(x) \rangle$, one obtains a statistical divergence which only depends on the score function of P. If $f \in \mathcal{F}$ satisfies regularity conditions such as $\lim_{\|x\|_2 \to \infty} f(x)p(x) = 0$, then one can show that $\mathbb{E}_{x \sim P}[\mathcal{A}_P f(x)] = 0$, and f is said to lie in the *Stein class* of P (Liu et al., 2016, Sec. 2.2). The function class $\mathcal{F}$ is usually chosen to be *(i)* sufficiently broad so that the discrepancy separates distinct probability measures, i.e. $\mathbb{S}(Q, P; \mathcal{F}) = 0 \iff Q = P$, and *(ii)* sufficiently regular so that the right-hand-side of (1) can be efficiently solved.

To this end, Liu et al. (2016); Chwialkowski et al. (2016) proposed to let $\mathcal{F}$ be the unit ball of a *reproducing kernel Hilbert space* (RKHS) (Berlinet & Thomas-Agnan, 2011). Specifically, let $\mathcal{H}$ be an RKHS associated with positive def-
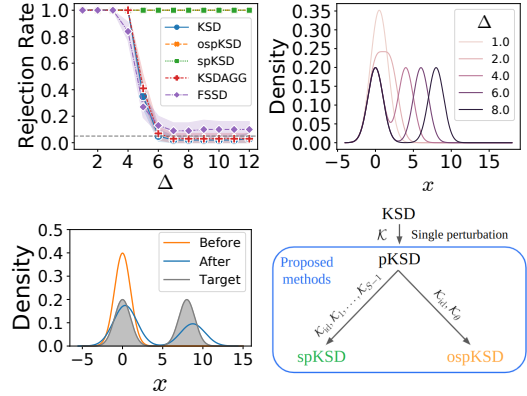

Figure 1: Power for a one-dimensional bimodal Gaussian target distribution P with mixing weight 0.5 and mode separation Δ. The candidate distribution Q is only the left component, from which 1000 samples are drawn. ospKSD and spKSD are our proposed method; the others are existing benchmarks. *Top*: Rejection rates and target densities for varying Δ; the orange and green lines overlap. *Bottom left*: Density of Q before and after 10 steps of the perturbation described in Sec. 4 and density of the target P. *Bottom right*: Connections between KSD and the proposed divergences.

inite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\mathcal{F}^d$ be the unit ball of the d-times Cartesian product $\mathcal{H}^d := \mathcal{H} \times \cdots \times \mathcal{H}$. Choosing $\mathcal{F} = \mathcal{F}^d$ and the operator $\mathcal{A}_P$ yields the *(Langevin) kernelized Stein discrepancy* (KSD): $\mathbb{D}(Q, P) := \mathbb{D}_{\mathcal{F}^d}(Q, P)$.

Assuming the kernel k has continuous first-order derivatives with respect to both arguments, Chwialkowski et al. (2016, Thm. 2.1) showed that KSD attains a closed form: $\mathbb{D}(Q, P) = \mathbb{E}_{x, x' \sim Q}[u_P(x, x')]$, where x, x' are independent random variables drawn from Q, and u_P is the *Stein kernel*: $u_P(x, x') := s_p(x)^\top k(x, x')s_p(x') + s_p(x)^\top \nabla_{x'} k(x, x') + \nabla_x k(x, x')^\top s_p(x') + \sum_{i=1}^d \frac{\partial^2}{\partial x_i \partial x_i'} k(x, x')$. Notably, u_P (hence also $\mathbb{D}(Q, P)$) depends on p only through $s_p(x) = \nabla \log p(x)$, so KSD is computable even without the knowledge of the normalising constant of p.

We will assume k lies in the *Stein class* of p (Liu et al., 2016, Def. 3.4), so that $\mathbb{D}(P, P) = 0$. When Q also admits a density q, and k is *cc-universal* (Sriperumbudur et al., 2011; 2010) or integrally strictly positive definite (Stewart, 1976, Sec. 6), KSD is *separating*, meaning that $\mathbb{D}(Q, P) = 0 \iff Q = P$, provided that $\mathbb{E}_{x \sim Q}[\|s_q(x) - s_p(x)\|^2] < \infty$ (Chwialkowski et al., 2016; Liu et al., 2016). The assumption that Q has a density can be relaxed if the target density satisfies additional tail conditions, such as distant dissipativeness (Hodgkinson et al., 2020, Proposition 4).

$\mathbb{D}(Q, P)$ can be estimated from a sample $\{x_i\}_{i=1}^n$ from Q by the following U-statistic (Serfling, 2009, Sec. 5.5):

$$\hat{\mathbb{D}}_P := \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} u_P(x_i, x_j). \tag{2}$$

The KSD test uses (2) as a test statistic. The asymptotic distribution of $\hat{\mathbb{D}}_P$ under H_0 has no closed form, but can be approximated with a bootstrap procedure (Huskova & Janssen, 1993) using the bootstrap samples

$$\hat{\mathbb{D}}_P^b := \frac{1}{n^2} \sum_{1 \le i \ne j \le n} \left(w_i^b - 1 \right) \left(w_j^b - 1 \right) u_P(x_i, x_j), \tag{3}$$

where $(w_1^b, \ldots, w_n^b) \sim \text{Mult}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right)$ follows a multinomial distribution. The test statistic $\hat{\mathbb{D}}_P$ is compared against quantiles of $\{\hat{\mathbb{D}}_P^b\}_{b=1}^B$ computed with B i.i.d. draws $(w_1^b, \ldots, w_n^b)$, and H_0 is rejected for large values of $\hat{\mathbb{D}}_P$. The resulting test achieves the desired level α asymptotically (Huskova & Janssen, 1993; Liu et al., 2016).

Many improvements over the standard KSD test have been proposed, e.g., to reduce the computational cost (Jitkrittum et al., 2017), to address the curse-of-dimensionality (Gong et al., 2021a;b), and to avoid kernel selection by adopting an aggregated testing procedure (Schrab et al., 2022).

3. Limitations of KSD Test

The KSD can be blind to certain discrepancies that are strongly visible in other metrics (e.g., in the L_2 norm). One example is mixtures of the same well-separated components, differing only by the mixing proportions (weights). In fact, the KSD will be small in settings where the score difference $\|s_p(x) - s_q(x)\|_2^2$ is large only with low Q-probability. This is because the KSD can be bounded from above by the *Fisher Divergence* (FD) $F(q, p) := \mathbb{E}_{x \sim Q}[\|s_p(x) - s_q(x)\|_2^2]$ (Liu et al., 2016, Thm. 5.1).

This is known as the "blindness" of score-based discrepancies (Wenliang & Kanagawa, 2020) such as KSD. This limitation of KSD has been highlighted in a number of works (Gorham et al., 2019; Matsubara et al., 2022; Zhang et al., 2022; Kanagawa et al., 2022); however, its implication to the test power in GOF tests has not yet been formalised.

In Prop. 3.1 (proved in Appendix A), we formally connect the blindness issue with the rates of increase of the sample size and the FD between the two distributions.

Proposition 3.1. *Let Q and P_ν, $\nu = 1, 2, \ldots$, be probability measures defined on $\mathbb{R}^d$ with positive densities q and p_ν. respectively. Assume $\mathbb{E}_{x \sim Q}[\|s_q(x)\|_2^2]$, $\mathbb{E}_{x, x' \sim Q}[u_Q(x, x')^2] < \infty$, and the kernel k satisfies*

$$\max \left\{ \mathbb{E}_{x, x' \sim Q}[|k(x, x')|], \mathbb{E}_{x, x' \sim Q}[\|\nabla_{x'} k(x, x')\|_2^2], \right.$$
$$\left. \mathbb{E}_{x, x' \sim Q}[\|\nabla_x k(x, x')\|_2^2] \right\} < \infty . \tag{4}$$

Let $x_1, x_2, \ldots$ be a sequence of i.i.d. samples from Q. Denote by $F_\nu := \mathbb{E}_{x \sim Q}[\|s_{p_\nu}(x) - s_q(x)\|_2^2]$ the Fisher Divergence between Q and P_ν. If the sequence

$n_1, n_2, \ldots \in \mathbb{N}$ *satisfies $n_\nu \to \infty$ as $\nu \to \infty$ and $n_\nu = o(1/\max(F_\nu, F_\nu^{1/2}))$, then*

$$n_\nu \hat{\mathbb{D}}_{P_\nu} \to_d \sum_{j=1}^\infty c_j(Z_j^2 - 1) \quad (\nu \to \infty), \tag{5}$$

where $\hat{\mathbb{D}}_{P_\nu}$ is the sample KSD computed using $x_1, \ldots, x_{n_\nu}$, $Z_j \sim \mathcal{N}(0, 1)$ i.i.d. and $\{c_j\}$ are the eigenvalues of the Stein kernel u_P under Q.

Remark 3.2. The RHS of (5) is the limiting distribution of $\hat{\mathbb{D}}_{P_\nu}$ under H_0 (Liu et al., 2016). Hence, this result shows that if the sample size n_ν is $o(1/\max(F_\nu, F_\nu^{1/2}))$, then the test power converges to the *nominal level* of the test.

Remark 3.3. Assumption (4) is standard and holds for Inverse Multi-Quadrics (IMQ) and Radial Basis Function (RBF) kernels when Q has a finite second moment. IMQ kernels are preferred as they have desired tail properties to ensure a *convergence determining* KSD for target densities satisfying the distantly dissipative condition (Gorham & Mackey, 2017; Hodgkinson et al., 2020). This includes Gaussian mixtures with common covariance, as well as distributions strongly log-concave outside of a compact set, such as Bayesian linear, logistic, and Huber regression posteriors with Gaussian priors, c.f., Gorham et al. (2019); Gorham & Mackey (2017). Prop. 3.1 does not contradict this result, as it considers a different regime where a *sequence* of target distributions is of interest.

Prop. 3.1 allows us to study the test power by analysing the FD. For instance, when P is a mixture of two Gaussian components and Q is one of its component, the FD decreases exponentially fast to 0 with the mode separation. Prop. 3.1 then implies that an unrealistically large sample size would be needed for the test to have a non-trivial power. This is formalised in the following result.

Theorem 3.4. *Let $Q = \mathcal{N}(0, I_d)$ and $P_\nu = \pi \mathcal{N}(0, I_d) + (1 - \pi)\mathcal{N}(\Delta_\nu, I_d)$, where $\pi \in [0, 1]$ and $\Delta_\nu \in \mathbb{R}^d$. With the same notation in Prop. 3.1 and assuming k satisfies (4), the limit (5) holds if $n_\nu = o\left(e^{\|\Delta_\nu\|_2^2/64}\right)$.*

The proof is in Appendix B. Figure 1 provides numerical evidence for Thm. 3.4 by showing the rate of rejection over 100 repetitions at level $\alpha = 0.05$. We observe that the power of the KSD test (with IMQ kernel whose bandwidth is chosen by median heuristic (Gretton et al., 2012)) approaches the prescribed level for $\Delta \ge 6$. A similarly poor performance is observed for KSDAGG (Schrab et al., 2022) and FSSD (Jitkrittum et al., 2017), two variants of KSD. In comparison, our proposed test, called ospKSD and spKSD, achieve an almost perfect power. Notably, the problem of low test power persists even if the samples are drawn from both components but with a different weight; see Figure 3 in Sec. 7.

4. KSD Test with Perturbation

We propose to increase the power of KSD test against *multi-modal alternatives* by perturbing both the candidate and the target distributions with a set of *Markov transition kernels* (Robert & Casella, 2004, Chapter 6) and performing KSD tests on the *perturbed* distributions. A Markov transition kernel is a function $\mathcal{K} : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ such that *(i)* for all $x \in \mathcal{X}$, $\mathcal{K}(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, and *(ii)* for all $A \in \mathcal{B}(\mathcal{X})$, $\mathcal{K}(\cdot, A)$ is a measurable function on $\mathcal{X}$. In our example, $\mathcal{K}$ may also be an iterated composition of an underlying kernel, e.g. a Metropolis-Hastings kernel. The perturbed measure of Q is $(\mathcal{K}Q)(\cdot) \coloneqq \int_{\mathcal{X}} \mathcal{K}(x, \cdot)Q(dx)$, and similarly for $\mathcal{K}P$.

4.1. KSD with a Single Perturbation Kernel

We first consider a *single* transition kernel $\mathcal{K}$. We define the *perturbed kernelized Stein discrepancy* (pKSD) as

$$\mathbb{D}(Q, P; \mathcal{K}) \coloneqq \mathbb{D}(\mathcal{K}Q, \mathcal{K}P)$$
$$= \sup_{f \in \mathcal{F}^d} |\mathbb{E}_{x \sim \mathcal{K}Q}[\mathcal{A}_{\mathcal{K}P}f(x)]| , \quad (6)$$

assuming $\mathcal{K}P$ admits a continuously differentiable density so that its score function is well-defined. Notably, $\mathcal{K}Q$ need not have a (Lebesgue) density for (6) to exist.

The properties of pKSD are dictated by the operator $\mathcal{K}$. A desirable choice should ensure that *(i)* pKSD is well-defined, and in particular $\mathcal{K}P$ should have a continuously differentiable density whenever P does, *(ii)* pKSD (6) can be computed efficiently, and *(iii)* the test can achieve a high power against alternatives with wrong mixing weights.

Given these *desiderata*, we propose to choose a transition kernel $\mathcal{K}$ that is *P-invariant*, i.e. $P(\cdot) = \int_{\mathcal{X}} \mathcal{K}(x, \cdot)p(x)dx$. A P-invariant kernel ensures $\mathcal{K}P = P$, so the score function $s_{\mathcal{K}p} = s_p$ is unchanged after perturbation. This means *(i)* and *(ii)* are trivially satisfied. In particular, pKSD will have a closed-form expression

$$\mathbb{D}(Q, P; \mathcal{K}) = \mathbb{E}_{x,x' \sim \mathcal{K}Q}[u_P(x, x')] ,$$

provided that $\mathbb{E}_{x \sim \mathcal{K}Q}[u_P(x, x)] < \infty$ (e.g., Chwialkowski et al. (2016, Thm. 2.1)). Moreover, the P-invariance allows a GOF test similar to the standard KSD test to be constructed, as we will elucidate in Sec. 4.2. To address *(iii)*, we employ a proposal map for $\mathcal{K}$ that "aggregates" densities across the modes of the distribution. As we will demonstrate numerically, such a proposal is sensitive to discrepancies in mixing weights.

Given i.i.d. $\{x_i\}_i^n \sim Q$, a sample $\{\tilde{x}_i\}_{i=1}^n$ from $\mathcal{K}Q$ can be drawn by running 1-step transitions under $\mathcal{K}$ starting from each x_i. pKSD can then be estimated by the U-statistic:

$$\hat{\mathbb{D}}_{P,\mathcal{K}} \coloneqq \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} u_P(\tilde{x}_i, \tilde{x}_j) . \quad (7)$$

Algorithm 1 Goodness-of-Fit Test with spKSD.

Input: Target P, observed sample $\{x_i\}_{i=1}^n$ from Q, set of transition kernels $\mathcal{S} = \{\mathcal{K}_s\}_{s=1}^S$ that includes $\mathcal{K}_{\text{id}}$ (i.e., no perturbation), number of transition steps T.
Estimate the mode $\{\mu_1, \ldots, \mu_M\}$ and Hessians $\{A_1, \ldots, A_M\}$ using Algorithm 2 in the Appendix.
For $s = 1, \ldots, S$, perturb $\{x_i\}_{i=1}^n$ with $\mathcal{K}_s$ by T steps to generate perturbed samples $\{x_i^s\}_{i=1}^n$.
Compute test statistic $\hat{\mathbb{D}}_{P,\mathcal{S}}$ using (8).
Generate bootstrap samples with (3) with u_P replaced by $\tilde{u}_P$, and find the $(1 - \alpha)$-quantile $\hat{\gamma}_{1-\alpha}$.
Reject H_0 if $\hat{\mathbb{D}}_{P,\mathcal{S}} \geq \hat{\gamma}_{1-\alpha}$.

4.2. KSD with Multiple Perturbation Kernels

A single transition kernel can be limited in improving the test power against general multi-modal alternatives. It also does *not* guarantee the *separation* property, as $\mathbb{D}(\mathcal{K}Q, \mathcal{K}P) = 0 \not\Longrightarrow Q = P$, unless $\mathcal{K}$ is injective so that $\mathcal{K}Q = \mathcal{K}P \Longrightarrow Q = P$ (such as the convolution operator). However, choosing only injective $\mathcal{K}$ would significantly restrict the class of possible options. Instead, we propose to employ a finite *collection* $\mathcal{S} = \{\mathcal{K}_s\}_{s=1}^S$ of P-invariant transition kernels, and require $\mathcal{S}$ to include the identity transition kernel $\mathcal{K}_{\text{id}}$, defined as $\mathcal{K}_{\text{id}}(x, A) = \delta_x(A)$ for all $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$, where $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. In particular, $\mathbb{D}(Q, P; \mathcal{K}_{\text{id}})$ reduces to the standard KSD. This gives rise to a separating statistical divergence which we term *sum-pKSD* (spKSD)

$$\mathbb{D}(Q, P; \mathcal{S}) \coloneqq \sum_{\mathcal{K} \in \mathcal{S}} \mathbb{D}(\mathcal{K}Q, P) ,$$

where we have overloaded $\mathbb{D}(Q, P; \mathcal{S})$ with a set $\mathcal{S}$ in place of a single transition kernel to denote spKSD. The next result (proved in Appendix C) shows that spKSD indeed separates probability measures so long as $\mathcal{K}_{\text{id}} \in \mathcal{S}$.

Proposition 4.1 (spKSD separation). *Suppose Q, P are probability measures on $\mathcal{X}$ that admit positive (Lebesgue) densities q, p, respectively. Further assume $\mathbb{E}_{x \sim \mathcal{K}Q}[u_P(x, x)] < \infty$ for all $\mathcal{K} \in \mathcal{S}$ and $\mathbb{E}_{x \sim Q}[\|s_p(x) - s_q(x)\|_2^2] < \infty$. If the kernel k is cc-universal and $\mathcal{K}_{id} \in \mathcal{S}$, then $\mathbb{D}(Q, P; \mathcal{S}) \geq 0$ with equality if and only if $Q = P$.*

The assumption that the alternative distribution Q also admits a density is common in KSD literature when proving separation (e.g., Liu et al. (2016); Chwialkowski et al. (2016); Jitkrittum et al. (2017); Gong et al. (2021b)), but it can be relaxed if P is light-tailed or distantly dissipative (Hodgkinson et al., 2020; Gorham & Mackey, 2017).

spKSD can also be written as a double expectation akin to KSD, provided $\mathbb{E}_{x \sim \mathcal{K}_s Q}[u_P(x, x)] < \infty$ for all s. This allows spKSD to be estimated given a random sample $\{x_i\}_{i=1}^n$ from Q. Formally, for each $\mathcal{K}_s \in \mathcal{S} = \{\mathcal{K}_1, \ldots, \mathcal{K}_S\}$,

a sample $\{x_i^s\}_{i=1}^n$ from $\mathcal{K}_s Q$ can be drawn by running 1-step transitions under $\mathcal{K}_s$ starting from each x_i. Denote by $x_i^{1:S} \coloneqq \mathrm{concat}(x_i^1, \ldots, x_i^S)$ the concatenation of $x_i^1, \ldots, x_i^S$ into a single vector. We propose to estimate $\mathbb{D}(Q, P; \mathcal{S})$ using the following U-statistic

$$\hat{\mathbb{D}}_{P,\mathcal{S}} \coloneqq \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \tilde{u}_P(x_i^{1:S}, x_j^{1:S}) , \quad (8)$$

where $\tilde{u}_P(x_i^{1:S}, x_j^{1:S}) \coloneqq \sum_{s=1}^S u_P(x_i^s, x_j^s)$.

4.3. GOF Testing with spKSD

Having constructed a test statistic for spKSD in the form of a U-statistic, the next result (proved in Appendix D) derives the limiting distribution of spKSD statistic under the null and alternative hypotheses. We denote by R_Q the distribution of $x_i^{1:S}$ constructed as before and use the same notations as in Prop. 4.1.

Proposition 4.2 (Asymptotic distributions of spKSD). *Suppose the assumptions in Prop. 4.1 hold, and further assume $\mathbb{E}_{w,w' \sim R_Q}[\tilde{u}_P(w, w')^2] < \infty$. Let $\{z_j\}_{j \ge 1}$ be independent draws from $\mathcal{N}(0,1)$ and denote by $\{c_j\}_{j \ge 1}$ the eigenvalues of $\tilde{u}_P$ under R_Q, i.e., the solutions of $c_j \phi_j(\cdot) = \mathbb{E}_{w \sim R_Q}[\tilde{u}_P(\cdot, w) \phi_j(w)]$ for non-zero ϕ_j. As $n \to \infty$,*

(i) Under H_0 : $Q = P$, we have $n\hat{\mathbb{D}}_{P,\mathcal{S}} \to_d \sum_{j=1}^\infty c_j(z_j^2 - 1)$.

(ii) Under H_1 : $Q \ne P$, we have $\sigma_u^2 \coloneqq 4\mathrm{Var}_{w \sim R_Q}(\mathbb{E}_{w' \sim R_Q}[\tilde{u}_P(w, w')]) > 0$, and $\sqrt{n}(\hat{\mathbb{D}}_{P,\mathcal{S}} - \mathbb{D}(Q, P; \mathcal{S})) \to_d \mathcal{N}(0, \sigma_u^2)$.

Prop. 4.2 assumes Q also admits a Lebesgue density; when it does not, the stated results still hold true if we additionally assume *i)* the conditions on Q in Prop. 4.1 for KSD to separate probability measures, and *ii)* $R_Q(A) > 0$ whenever $R_P(A) > 0$ for any measurable set $A \subset \mathcal{X}^S$.

Similarly to the case with the standard KSD, the cumulative density function of the limiting distribution under H_0 has no closed-form expression, but the same bootstrap technique can be employed to estimate the p-value using the perturbed samples. The complete algorithm of goodness-of-fit testing with pKSD is given in Algorithm 1.

5. A Transition Kernel for Multi-Modal Alternatives

We consider transition kernels of the Metropolis-Hastings (MH) type (Metropolis et al., 1953; Hastings, 1970). At a current state x, a new state x' is proposed by first generating a d_u-dimensional random vector u from some known density g, then mapping to $x' = h(x|u)$, where $h(\cdot|u)$ is some deterministic, invertible function that is differentiable with differentiable inverse. The proposed state x' is hence a deterministic function given x and u.

We choose in this paper a density g defined on some discrete space $\mathcal{U}$. The transition kernel is

$$\mathcal{K}(x, A) = \sum_{u \in \mathcal{U}} \delta_{x'}(A)g(u)\alpha(x, x') + \delta_x(A)r(x),$$

where $x' = h(x|u)$ is the proposed state, $\alpha(x, x')$ is an accept-reject rule that guarantees P-invariance, $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise, and $r(x) = 1 - \sum_{u \in \mathcal{U}} g(u)\alpha(x, x')$. The accept-reject rule $\alpha(x, x')$ is designed to satisfy the *detailed balance condition*:

$$\int_{x \in A} \sum_{u \in \mathcal{U}} \delta_{x'}(B)p(x)g(u)\alpha(x, x')dx$$
$$= \int_{x' \in B} \sum_{u' \in \mathcal{U}} \delta_x(A)p(x')g(u')\alpha(x', x)dx', \quad (9)$$

for all $A, B \in \mathcal{B}(\mathcal{X})$. One valid choice is

$$\alpha(x, x') = \min\left(1, \frac{p(x')g(u')}{p(x)g(u)}\left|\frac{\partial h(x|u)}{\partial x}\right|\right), \quad (10)$$

if $x' = h(x|u)$ and $x = h^{-1}(x'|u')$ for some $u, u' \in \mathcal{U}$, and zero otherwise. Here, $\partial h(x|u)/\partial x$ denotes the Jacobian of the transformation from x to x'. Appendix E proves that $\alpha(x, x')$ indeed satisfies (9). The accept-reject rule (10) resembles those used in Reversible-Jump MCMC (Green, 1995; Green & Hastie, 2009) and generalises the well-known MH rule, for which the determinant of the Jacobian is 1.

5.1. Choosing the Proposal Density

We propose a jump proposal $h(x|u)$ that superposes masses at each mode of p. Our choice is motivated by Markov kernels used in the optimisation-based MCMC literature, specifically the *deterministic jumps* proposal in Pompe et al. (2020). New states are proposed by randomly selecting a mapping from a set of candidates that are constructed using the location and geometry of the modes of p. The resulting kernel is *not* irreducible, so the limiting distribution is not necessarily P. Non-irreducibility is essential for the proposed test to work since, under the alternative, the transition kernel should perturb Q to some other distribution for which the KSD between P and the perturbed distribution becomes larger compared with the KSD with the un-perturbed one. This is in contrast to MCMC, which requires irreducibility so that asymptotically the chain can sample from the target distribution.

Denote by $\mu_1, \ldots, \mu_M \in \mathbb{R}^d$ the modes of the density p, and $A_1, \ldots, A_M \in \mathbb{R}^{d \times d}$ the *inverse* of the Hessian matrices at those points; how to estimate these quantities will be discussed later. When p is a mixture of elliptic distributions such as Gaussian or multivariate t-distributions, each A_m can be viewed as the covariance matrix of a component. When the Hessians do not exist (e.g., $-\log p$ is not twice differentiable), we can set $A_m = I_d$ and the remaining discussion still follows.
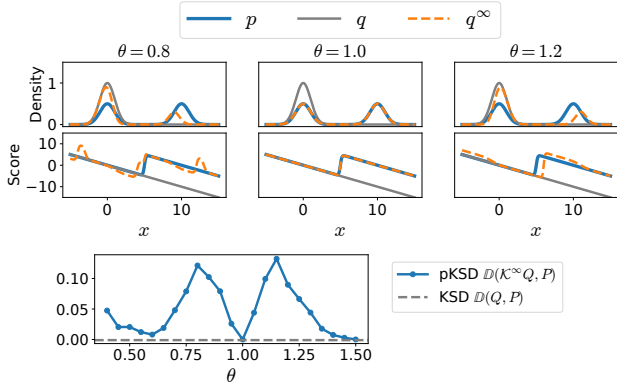
Figure 2: Top: Densities and score functions of p, q and the limiting distribution q^∞ in (11). Bottom: pKSD with different jump scales θ, compared with KSD.

For a current state x, our proposal randomly selects a pair of modes and attempts to map x from one mode to the "corresponding" point x' in the other. Formally, let $u = (u_1, u_2) \sim \text{Unif}(\{(i,j) : 1 \le i \ne j \le M\})$ be a uniform random vector over the index set of all $M(M-1)$ pairs of *distinct* (and ordered) modes, i.e., $g(u) = 1/(M(M-1))$ for all u. Given a fixed constant $\theta > 0$, the proposal map is

$$h(x|u) = h_\theta(x|u) = A_{u_2}^{1/2} A_{u_1}^{-1/2}(x - \theta\mu_{u_1}) + \theta\mu_{u_2} ,$$

with the inverse map $h^{-1}(x'|u) = A_{u_1}^{1/2} A_{u_2}^{-1/2}(x - \theta\mu_{u_2}) + \theta\mu_{u_1}$. Intuitively, h sends points from mode μ_{u_1} to μ_{u_2} allowing for scaling by local Hessians, and h^{-1} performs the opposite operation. The constant θ is a hyperparameter introduced to control the scale of the jump, which can increase the ability to detect discrepancies in the mixing weights. Herein, we call θ the *jump scale*.

Given a current state, our proposal chooses two modes randomly, so a proposed state can potentially lie in a low-density region, thus leading to a low acceptance probability. Pompe et al. (2020) address this by recording an auxiliary variable for the mode index and augmenting the state space to $\mathcal{X} \times \{1, 2, \ldots, M(M-1)\}$, so that at every step the new state is *guaranteed* to lie near a mode. However, the same trick cannot be used in our case because the augmented density no longer has a well-defined score function.

5.2. Understanding the Source of Test Power

To understand the improvement in test power against multi-modal alternatives, we characterise the limiting distribution of a general distribution Q with a positive density q when we apply the perturbation with infinitely many steps (i.e., $\mathcal{K}^T$ with $T = \infty$). For simplicity, we assume $M = 2$ and $A_1 = A_2 = I_d$ are identity matrices, so that the proposal function is $h_\theta(x|u) = x - \theta(\mu_{u_1} - \mu_{u_2})$ for $x \in \mathcal{X}$ and $u = (u_1, u_2) \in \mathcal{U} = \{(i,j) : 1 \le i \ne j \le 2\}$. Thus, given a current state x, the transition kernel proposes moves to $x + \theta(\mu_1 - \mu_2)$ and $x - \theta(\mu_1 - \mu_2)$ with equal probability.

Proposition 5.1. *Under the assumptions of Sec. 5.2, the limiting distribution under $\mathcal{K}$ with the initial distribution Q is $(\mathcal{K}^\infty Q)(A) = \int_{x \in A} q^\infty(x)dx, A \in \mathcal{B}(\mathcal{X})$, where*

$$q^\infty(x) := p(x)\frac{\sum_{s \in \mathbb{Z}} q(x+s\nu)}{\sum_{k \in \mathbb{Z}} p(x+k\nu)} , \qquad (11)$$

and $\nu := \theta(\mu_1 - \mu_2)$.

A proof is in Appendix F. Prop. 5.1 shows that the limiting density under $\mathcal{K}$ is the target density p weighted by the ratio between the total masses of q and p over a discrete grid.

To understand why this helps to increase the KSD value, we first rewrite KSD as

$$\mathbb{D}(Q, P) = \mathbb{E}_{x,x' \sim Q}[\delta_{q,p}(x)^\top k(x, x')\delta_{q,p}(x')] ,$$

where $\delta_{q,p}(x) := s_q(x) - s_p(x)$ is the score difference. This holds whenever k is an integrally strictly positive definite kernel (Liu et al., 2016, Thm. 3.6). The pKSD is then

$$\mathbb{D}(Q, P; \mathcal{K}^\infty) = \mathbb{E}_{x,x' \sim Q^\infty}[\delta_{q^\infty,p}(x)^\top k(x, x')\delta_{q^\infty,p}(x')] ,$$

where, by Prop. 5.1, the score difference becomes $\delta_{q^\infty,p}(x) = s_{q^\infty}(x) - s_p(x) = \nabla \log \phi_q(x) - \nabla \log \phi_p(x)$, where $\phi_q(x) := \sum_{s \in \mathbb{Z}} q(x + s\nu)$ and similarly for ϕ_p.

The operator ϕ superposes densities along a grid, thus allowing to create local discrepancy in the high-probability regions of Q, for example by exchanging masses between modes.

As a concrete example, we consider the setup in Thm. 3.4, where $Q = \mathcal{N}(0, I_d)$, and $P = \pi\mathcal{N}(0, I_d) + (1 - \pi)\mathcal{N}(\Delta, I_d)$ for some $\pi \in (0, 1)$ and $\Delta \in \mathbb{R}^d$. The operator has created discrepancy in high-probability regions of Q, as demonstrated in Fig. 2. This also highlights the role of ν: when $\nu = \Delta$, the two components will overlap almost exactly under the perturbation, so $\delta_{q^\infty,p}(x) \approx 0$ near $x = 0$, and the KSD will remain small (Fig. 2). It is hence crucial to tune ν (equivalently, θ). One can in principle select θ by maximising the (approximate) test power, similarly to the idea in Jitkrittum et al. (2017). However, gradient-based approaches are infeasible as pKSD is *not* differentiable with respect to θ. An alternative is to use grid-search over some finite set of θ values.

5.3. Choosing the Set of Perturbations $\mathcal{S}$

It remains to choose the set of perturbations $\mathcal{S}$ in spKSD. We propose two ways to construct $\mathcal{S}$, one based on grid-search, and the other based on optimisation.

For the grid-based approach, we choose a set of values $\{\theta_s\}_{s=1}^{S-1}$ and let $\mathcal{S} = \{\mathcal{K}_{\text{id}}, \mathcal{K}_1, \ldots, \mathcal{K}_{S-1}\}$, where $\mathcal{K}_s$ is the transition kernel described in this section with jump scale θ_s. We propose to choose each θ_s close to 1, following

the observations in Sec. 5.2. We still refer to the resulting divergence as spKSD.

For the optimisation-based approach, we use only two transition kernels $\mathcal{S} = \{\mathcal{K}_{\mathrm{id}}, \mathcal{K}_\theta\}$, where $\mathcal{K}_\theta$ has jump scale θ that is tuned by maximising a proxy for the asymptotic test power. Due to the asymptotic normality proved in Prop. 4.2, we can adopt the same approach in Jitkrittum et al. (2017, Prop. 4) to approximate the asymptotic power with the ratio

$$\hat{\mathbb{D}}_{P,\mathcal{K}_\theta^T} / \hat{\sigma}_u , \tag{12}$$

where $\hat{\sigma}_u$ is an estimate of the asymptotic standard deviation σ_u is given by the square root of

$$\hat{\sigma}_u^2 := \frac{4}{n^3} \sum_{i=1}^n \left(\sum_{j=1}^n H_{i,j} \right)^2 - \frac{4}{n^4} \left(\sum_{i,j=1}^n H_{i,j} \right)^2 ,$$

with $H_{i,j} := u_P(x_i, x_j) + u_P(x_i^\theta, x_j^\theta)$ and $x_i^\theta \sim \mathcal{K}_\theta^T Q$; see also Schrab et al. (2022, Eq. 8). Since the objective (12) (in particular, $\mathcal{K}_\theta$) is not differentiable with respect to θ, we still choose θ from a pre-specified finite set $\{\theta_s\}_{s=1}^S$. The objective is hence $\max_{\theta \in \{\theta_1,...,\theta_S\}} \hat{\mathbb{D}}_{P,\mathcal{K}_\theta^T}/\hat{\sigma}_u$. We call the resulting discrepancy the *optimised sum-pKSD* (ospKSD).

Whether the grid-based or the optimisation-based method should be preferred requires trade-offs and depends on the specific problem at hand — The spKSD requires no held-out sets, but can suffer from a low test power if $\{\theta_s\}_{s=1}^{S-1}$ is poorly chosen in that most of $\mathcal{K}_s$ fail to improve the test power. On the other hand, ospKSD uses a judiciously tuned θ, but the data-splitting can also lead to a drop in test power. In our experiments, we find that spKSD tends to work better for target distributions with a simple geometry, specifically mixtures of *elliptic* distributions (Cambanis et al., 1981) (e.g., the Gaussian mixture examples). However, for distributions whose mixing components have non-elliptic contours (e.g., the mixture of t and banana example, and the sensor network localisation example), the benefit of optimisation seems to overweigh the negative impact due to data-splitting, and ospKSD outperforms spKSD.

5.4. Estimating Mode Vectors and Hessians

We estimate μ_j and A_j by the local minima and Hessians of $-\log p$. To do so, we run in parallel a sequence of BFGS optimisers (Nocedal & Wright, 2006) initiated at different starting points, following Pompe et al. (2020). BFGS is used because it returns both the local optima and approximated Hessians at those points. The optima are then merged if their weighted Mahalanobis distance is smaller than a pre-specified threshold. In our experiments, we initialise the optimisers from a set of size n_{init}, constructed either by sampling uniformly from a hyper cube $[L_1, U_1] \times \cdots \times [L_d, U_d]$ (for spKSD), or by using both randomly sampled
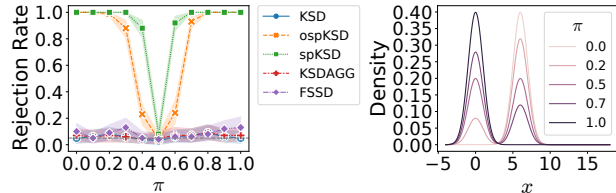

Figure 3: One-dimensional Gaussian mixture example. Samples are drawn with a different mixing weight π.

data and some training set (for ospKSD). The full procedure is described in Appendix G.1.

6. Related Work

Perturbation with convolution The idea of combining a discrepancy with perturbation has been widely studied, where the perturbation is often a convolution operator with Gaussian noise. E.g., the *spread divergence* (Zhang et al., 2020) combines Gaussian convolution with Kullback-Leibler (KL) divergence (more generally, f-divergences) to solve the issue that KL divergence is ill-defined when the distributions have undefined densities or unmatched support. In generative modelling, *denoising score matching* (Vincent, 2011) and *Noise Conditional Score Networks* (Song & Ermon, 2019) combine Gaussian convolution with score matching to improve computational efficiency or estimation quality. Notably, convolution is *not* invariant to the target distribution, thus rendering the score function intractable. This is why we chose a MH-type kernel instead.

Perturbation with convex combination In score matching, Zhang et al. (2022) addresses the blindness of Fisher Divergence by mapping the target and candidate distributions to a convex combination with a Gaussian distribution, thereby "connecting" the well-separated modes. A similar idea cannot be applied to improve the KSD test, as, similarly to convolution, the resulting target distribution no longer has a tractable score function.

Perturbation with annealing Another choice of perturbation is to anneal both distributions by raising the densities to some power, which is studied in Wenliang & Kanagawa (2020). Although the score function remains tractable under this perturbation, annealing alone cannot solve the blindness of score-based discrepancies, as noted by the authors and in Zhang et al. (2020). Moreover, sampling from the annealed candidate distribution is also non-trivial.

7. Experiments

We use 51 jump scales θ equally spaced in $[0.5, 1.5]$, a heuristic that we find works well in practice. All samples have size $n = 1000$. We compare the ospKSD
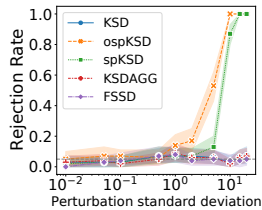
Figure 4: Mixture of 10 t and 10 banana distributions.

and spKSD tests against benchmarks including KSD test and two variants (KSDAGG and FSSD). All experiments are run with level $\alpha = 0.05$ using the IMQ kernel $k(x, y) = (1 + \|x - y\|_2^2/\lambda)^{-1/2}$, where λ is chosen to be `median`$_{i<j}\{\|x_i - x_j\|_2^2\}$. KSDAGG follows the setup in Schrab et al. (2022), and FSSD follows Jitkrittum et al. (2017). The probability of rejecting the null hypothesis is estimated by averaging the test output over 100 repetitions, except in the sensors location example, which is repeated 10 times. Translucent shades represent 95%-CIs. The number of transition steps T is selected to be 10 for the Gaussian mixture example, 100 for the mixture of t and banana distributions example, and 1000 for the sensor network localisation.

Discussions on how to choose T in practice, as well as supplementary plots and experiments, are held in Appendix H. In particular, in Appendix H.4, we include an additional experiment concerning a 50-dimensional Gaussian-Bernoulli Restricted Boltzmann Machine (RBM) (Cho et al., 2013), a latent variable model that can be viewed as a mixture of Gaussian distributions. Code for reproducing all experiments can be found at github.com/XingLLiu/pksd.

Gaussian mixture The target has density $p(x) \propto \exp\left(-\frac{1}{2}x^2\right) + 0.5 \exp\left(-\frac{1}{2}(x - 6)^2\right)$. Samples are drawn with a different mixing weight $\pi \in [0, 1]$ of the first component. The results are presented in Fig. 3. KSD, KSDAGG and FSSD all have a power close to the level 0.05 regardless of the value of π, which is not surprising due to the blindness of KSD. In comparison, ospKSD and spKSD achieve almost perfect power when π deviates from the true value 0.5. Fig. 6 in the Appendix verifies numerically that ospKSD and spKSD achieve the desired level under H_0.

Mixture of t and banana distributions We consider a mixture of 10 multivariate t-distributions and 10 banana-shaped distributions with t-tails in $d = 50$ dimensions, also studied in Pompe et al. (2020). Each component has an equal weight 0.02 and is centered randomly in $[-20, 20]^d$, giving rise to a target with sparsely located, non-elliptic modes. Samples are drawn with a different set of weights $\{w_j\}_{j=1}^{20}$ formed by sampling $\tilde{w}_j \sim \mathcal{N}(0, \sigma_s^2)$, $\sigma_s > 0$, and normalising $w_j \propto \exp(\tilde{w}_j)$. Other details are held in Appendix H.2. Fig. 4 (left) shows the results. As σ_s increases, the weights in the two distributions deviate further, so os-

Table 1: GOF tests for checking the quality of RAM samples with different scales. Reported values are the number of rejections over 10 repetitions with level 0.05.

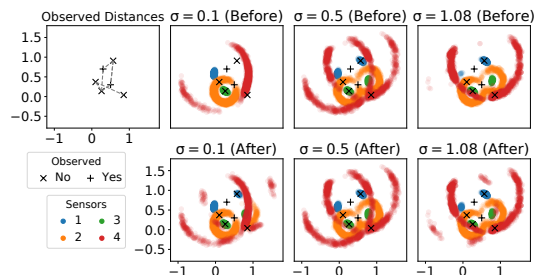| Methods		KSD	ospKSD	spKSD	KSDAGG	FSSD
	0.1	0	8	0	0	0
RAM scale σ	0.5	0	0	0	0	1
	1.08	10	8	10	10	7

Figure 5: True and inferred locations of sensors. Black plus signs mark the location of the unobserved sensors, and black crosses indicate the location of the observed ones.

pKSD and spKSD achieve a higher power. All the others perform poorly for all values of σ_s, because the components have almost no overlapping high-density regions.

Sensor network localisation Tak et al. (2018) use Bayesian methods to infer the locations of sensors from noisy distance data. This is a modification of the example in Ihler et al. (2005) that has been used as a benchmark for MCMC samplers designed for multi-modal distributions (Pompe et al., 2020; Ahn et al., 2013; Lan et al., 2014). Here, six sensors $x_1, \ldots, x_6$ are located in $[0, 1]^2$, four of which have unknown locations and the remaining two are known. We observe distance y_{ij} between two sensors x_i, x_j with probability $\exp(-\|x_i - x_j\|_2^2/(2 \times 0.3^2))$. If observed, the distance follows a Gaussian distribution $y_{ij} \sim \mathcal{N}(\|x_i - x_j\|, 0.02^2)$. Full model details are held in Appendix H.3. To draw posterior samples, Tak et al. (2018) propose the *repelling-attracting Metropolis* (RAM), which is an MCMC algorithm designed for efficient learning of multi-modal target distributions. RAM relies on a Gaussian proposal with a fixed covariance matrix $\sigma^2 I_d$ to propose new states in its uphill and downhill steps. The *scale σ* needs to be tuned to facilitate transitions between modes.

We run RAM with different scales σ, each for $420,000$ iterations. We discard the first $20,000$ particles as burn-in and thin the remaining to obtain a sample of size $n = 4000$. We then evaluate the quality of the samples by applying a GOF test. Fig. 5 shows the posterior samples generated using each σ. The samples from $\sigma = 0.5$ seem to capture all the modes of the posterior and is consistent with the results reported in Tak et al. (2018, Fig. 5), whereas samples from $\sigma = 0.1$ and 1.08 clearly miss some modes. We then compare the test results in Table 1, which reports the number

of rejections over 10 repetitions. All tests reject most runs for $\sigma = 1.08$ (the scale used in Tak et al. (2018)) and almost no run for $\sigma = 1.08$, which is consistent with the posterior plots. However, for $\sigma = 0.5$, no method except ospKSD rejected the null hypothesis, demonstrating again the ability of ospKSD to detect missing modes. Fig. 5 also shows the particles *after* perturbation by the (non-identity) transition kernel used by ospKSD, from which some particles seem to have moved to the missing high-density regions. spKSD in this case also performed poorly with no sample for $\sigma = 0.1$ being rejected, which is potentially because the benefit of not having a held-out set outweighs that of using a tuned transition kernel for this example.

8. Discussion and Conclusion

We show with a bimodal Gaussian example that GOF tests based on KSD can fail when the target has well-separated modes. To increase its power, we propose to perturb both the candidate and the target distributions using a Markov process before applying the KSD test. Empirical results suggest that our methods (ospKSD and spKSD) are more sensitive to discrepancies in the mixing weights of multimodal distributions, and can achieve remarkably high power particularly when the mixing components are elliptic distributions.

8.1. Limitations and Future Work

The ospKSD and spKSD rely heavily on accurate estimation of the mode locations and Hessians, which can be extremely challenging and computationally costly for high-dimensional problems. Additionally, the jump proposal of the transition kernel used in the proposed methods is constructed specifically for targets that are mixtures of elliptic distributions, which may be inappropriate for targets with more complicated geometrical structure. Further investigations could aim to find perturbation operators that scale better with dimensionality or that suit a wider family of target distributions.

Moreover, both spKSD and ospKSD require careful hyperparameter setting. The spKSD, as a sum-like statistic, requires a trade-off between the test power and the number of elements in the grid $\mathcal{S}$. Although the heuristic described in Section 7 is found to perform decently in our experiments, it is of practical interest to analyse the sensitivity of the test performance to the grid size both empirically and theoretically. The ospKSD, on the other hand, requires a held-out dataset to tune θ, potentially reducing test power due to datasplitting. One possible approach to mitigate this problem is to combine ospKSD with the aggregated testing framework described in Schrab et al. (2022) to avoid splitting the data.

References

Ahn, S., Chen, Y., and Welling, M. Distributed and adaptive darting Monte Carlo through regenerations. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pp. 108–116, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., Mackey, L., Oates, C. J., Reinert, G., and Swan, Y. Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1): 120 – 139, 2023. doi: 10.1214/22-STS863.

Barker, A. A. Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18:119, April 1965. doi: 10.1071/PH650119.

Bartle, R. G. *The elements of integration and Lebesgue measure*. John Wiley & Sons, 2014.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Cambanis, S., Huang, S., and Simons, G. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(81)90082-8.

Casella, G. and Berger, R. *Statistical inference*. Duxbury Resource Center, June 2001. ISBN 0534243126.

Cho, K. H., Raiko, T., and Ilin, A. Gaussian-Bernoulli deep Boltzmann machine. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2013.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2606–2615, New York, New York, USA, 20–22 Jun 2016. PMLR.

Gong, W., Li, Y., and Hernández-Lobato, J. M. Sliced kernelized Stein discrepancy. In *International Conference on Learning Representations*, 2021a.

Gong, W., Zhang, K., Li, Y., and Hernandez-Lobato, J. M. Active slices for sliced Stein discrepancy. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3766–3776. PMLR, 18–24 Jul 2021b.

Gorham, J. and Mackey, L. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pp. 1292–1301. PMLR, 2017.

Gorham, J., Duncan, A. B., Vollmer, S. J., and Mackey, L. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884 – 2928, 2019. doi: 10.1214/19-AAP1467.

Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Green, P. J. and Hastie, D. I. Reversible jump MCMC. *Genetics*, 155(3):1391–1403, 2009.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97–109, 1970. ISSN 00063444.

Hird, M., Livingstone, S., and Zanella, G. A fresh take on 'Barker dynamics' for MCMC. *arXiv preprint arXiv:2012.09731*, 2020.

Hodgkinson, L., Salomone, R., and Roosta, F. The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv preprint arXiv:2001.09266*, 2020.

Huskova, M. and Janssen, P. Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, 21(4):1811 – 1823, 1993. doi: 10.1214/aos/1176349399.

Ihler, A., Fisher, J., Moses, R., and Willsky, A. Nonparametric belief propagation for self-localization of sensor networks. *IEEE Journal on Selected Areas in Communications*, 23(4):809–819, 2005. doi: 10.1109/JSAC.2005.843548.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. A linear-time kernel goodness-of-fit test. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 261–270, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Kanagawa, H., Gretton, A., and Mackey, L. Controlling moments with kernel Stein discrepancies. *arXiv preprint arXiv:2211.05408*, 2022.

Lan, S., Streets, J., and Shahbaba, B. Wormhole Hamiltonian Monte Carlo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Liu, Q., Lee, J., and Jordan, M. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 276–284, New York, New York, USA, 20–22 Jun 2016. PMLR.

Livingstone, S. and Zanella, G. The Barker proposal: Combining robustness and efficiency in gradient-based MCMC. *Journal of the Royal Statistical Society: Series B*, 2021.

Massey, F. J. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951. doi: 10.1080/01621459.1951.10500769.

Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022, 2022.

Melchior, J., Wang, N., and Wiskott, L. Gaussian-binary restricted Boltzmann machines for modeling natural image statistics. *PloS one*, 12(2):e0171015, 2017.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Meyn, S. P. and Tweedie, R. L. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Nocedal, J. and Wright, S. *Numerical optimization*. Springer series in operations research and financial engineering. Springer, New York, NY, 2. ed. edition, 2006. ISBN 978-0-387-30303-1.

Peskun, P. H. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

Pompe, E., Holmes, C., and Łatuszyński, K. A framework for adaptive MCMC targeting multimodal distributions. *The Annals of Statistics*, 48(5):2930–2952, 2020.

Robert, C. and Casella, G. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer New York, New York, NY, 2nd ed. 2004. edition, 2004. ISBN 1-4757-4145-6.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. MMD aggregated two-Sample test. *arXiv preprint arXiv:2110.15073*, 2021.

Schrab, A., Guedj, B., and Gretton, A. KSD aggregated goodness-of-fit test. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022.

Serfling, R. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 9780470317198.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. On the relation between universality, characteristic kernels and RKHS embedding of measures. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 773–780, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (70):2389–2410, 2011.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pp. 583–603. University of California Press, 1972.

Stewart, J. Positive definite functions and generalizations, an historical survey. *Rocky Mountain Journal of Mathematics*, 6(3):409 – 434, 1976. doi: 10.1216/RMJ-1976-6-3-409.

Tak, H., Meng, X.-L., and van Dyk, D. A. A repelling–attracting Metropolis algorithm for multimodality. *Journal of Computational and Graphical Statistics*, 27(3):479–490, 2018.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Wainwright, M. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. ISBN 9781108498029.

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

Zhang, M., Hayes, P., Bird, T., Habib, R., and Barber, D. Spread divergence. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11106–11116. PMLR, 13–18 Jul 2020.

Zhang, M., Key, O., Hayes, P., Barber, D., Paige, B., and Briol, F.-X. Towards healing the blindness of score matching. *arXiv preprint arXiv:2209.07396*, 2022.

A. Proof of Proposition 3.1

Fixing positive integer ν, we can write $n_\nu \hat{\mathbb{D}}_{P_\nu} = n_\nu \hat{\mathbb{D}}_Q + n_\nu(\hat{\mathbb{D}}_{P_\nu} - \hat{\mathbb{D}}_Q)$. Under the stated assumptions on the kernel k and that $\mathbb{E}_{x,x'\sim Q}[u_Q(x,x')^2] < \infty$, we can apply Liu et al. (2016, Thm 4.1) to conclude that, as $\nu \to \infty$,

$$n_\nu \hat{\mathbb{D}}_Q \to_d \sum_{j=1}^\infty c_j(z_j^2 - 1)\,,$$

where z_j, c_j are as defined in Prop. 3.1. If we could furthermore show that $n_\nu(\hat{\mathbb{D}}_{P_\nu} - \hat{\mathbb{D}}_Q) \to 0$ in probability as $\nu \to \infty$, then the desired result would follow from Slutsky's Theorem (see, e.g., Casella & Berger (2001)).

To prove the convergence in probability, we fix $\epsilon > 0$ and denote by $\Pr_Q$ the probability under Q. We also omit the dependence of n on ν for brevity. The Markov inequality yields

$$
\begin{aligned}
&\Pr_Q(n|\hat{\mathbb{D}}_{P_\nu} - \hat{\mathbb{D}}_Q| \geq \epsilon) \\
&\leq \tfrac{n}{\epsilon} \mathbb{E}_{x_1,\ldots,x_n \sim Q}[|\hat{\mathbb{D}}_{P_\nu} - \hat{\mathbb{D}}_Q|] \\
&= \tfrac{n}{\epsilon} \mathbb{E}_{x_1,\ldots,x_n \sim Q} \left| \tfrac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} u_{P_\nu}(x_i, x_j) - u_Q(x_i, x_j) \right| \\
&\leq \tfrac{n}{\epsilon} \tfrac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \mathbb{E}_{x_i, x_j \sim Q} |u_{P_\nu}(x_i, x_j) - u_Q(x_i, x_j)| \\
&= \tfrac{n}{\epsilon} \mathbb{E}_{x,x' \sim Q} |u_{P_\nu}(x, x') - u_Q(x, x')| \\
&\leq \tfrac{n}{\epsilon} \{ \mathbb{E}_{x,x' \sim Q} |s_{p_\nu}(x)^\top s_{p_\nu}(x') - s_q(x)^\top s_q(x')||k(x,x')| \\
&\quad + \mathbb{E}_{x,x' \sim Q} |(s_{p_\nu}(x) - s_q(x))^\top \nabla_{x'} k(x,x')| \\
&\quad + \mathbb{E}_{x,x' \sim Q} |(s_{p_\nu}(x') - s_q(x'))^\top \nabla_x k(x,x')| \} \\
&\leq \tfrac{n}{\epsilon} \Big\{ \left(\mathbb{E}_{x,x' \sim Q} \left[(s_{p_\nu}(x)^\top s_{p_\nu}(x') - s_q(x)^\top s_q(x'))^2 \right] \right)^{1/2} \left(\mathbb{E}_{x,x' \sim Q} \left[k(x,x')^2 \right] \right)^{1/2} \\
&\quad + \left(\mathbb{E}_{x \sim Q} \left[\|s_{p_\nu}(x) - s_q(x)\|_2^2 \right] \right)^{1/2} \left(\mathbb{E}_{x,x' \sim Q} \left[\|\nabla_{x'} k(x,x')\|_2^2 \right] \right)^{1/2} \\
&\quad + \left(\mathbb{E}_{x \sim Q} \left[\|s_{p_\nu}(x) - s_q(x)\|_2^2 \right] \right)^{1/2} \left(\mathbb{E}_{x,x' \sim Q} \left[\|\nabla_x k(x,x')\|_2^2 \right] \right)^{1/2} \Big\}.
\end{aligned}
\tag{13}
$$

We bound each of the three terms individually. For the first term, we have

$$
\begin{aligned}
&\mathbb{E}_{x,x' \sim Q} \left[\left(s_{p_\nu}(x)^\top s_{p_\nu}(x') - s_q(x)^\top s_q(x') \right)^2 \right] \\
&= \mathbb{E}_{x,x' \sim Q} \left[\left(s_{p_\nu}(x)^\top (s_{p_\nu}(x') - s_q(x')) + (s_{p_\nu}(x) - s_q(x))^\top s_q(x') \right)^2 \right] \\
&\leq 2 \underbrace{\mathbb{E}_{x,x' \sim Q} \left[\left(s_{p_\nu}(x)^\top (s_{p_\nu}(x') - s_q(x')) \right)^2 \right]}_{=:T_1} + 2 \underbrace{\mathbb{E}_{x,x' \sim Q} \left[\left((s_{p_\nu}(x) - s_q(x))^\top s_q(x') \right)^2 \right]}_{=:T_2},
\end{aligned}
$$

where the last line follows from the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$. Now, applying the Cauchy-Schwarz inequality gives

$$
\begin{aligned}
T_1 &\leq 2 \mathbb{E}_{x,x' \sim Q} [\|s_{p_\nu}(x)\|_2^2 \|s_{p_\nu}(x') - s_q(x')\|_2^2] \\
&\leq 2 \left(2 \mathbb{E}_{x \sim Q} [\|s_{p_\nu}(x) - s_q(x)\|_2^2] + 2 \mathbb{E}_{x \sim Q} [\|s_q(x)\|_2^2] \right) \mathbb{E}_{x' \sim Q} [\|s_{p_\nu}(x') - s_q(x')\|_2^2] \\
&\leq 4 F_\nu^2 + 4 \mathbb{E}_{x \sim Q} [\|s_q(x)\|_2^2] F_\nu,
\end{aligned}
$$

where $F_\nu := \mathbb{E}_{x \sim Q} [\|s_{p_\nu}(x) - s_q(x)\|_2^2]$ is the Fisher Divergence between P_ν and Q, and where the second line holds because

$$\|s_{p_\nu}(x)\|_2^2 = \|s_{p_\nu}(x) - s_q(x) + s_q(x)\|_2^2 \leq 2\|s_{p_\nu}(x) - s_q(x)\|_2^2 + 2\|s_q(x)\|_2^2.$$

A similar argument by Cauchy-Schwarz inequality shows that

$$T_2 \leq \mathbb{E}_{x \sim Q} [\|s_q(x)\|_2^2] F_\nu\,.$$

Combining the bounds for T_1 and T_2 yields

$$\mathbb{E}_{x,x' \sim Q} [(s_{p_\nu}(x)^\top s_{p_\nu}(x') - s_q(x)^\top s_q(x'))^2] \leq 8 F_\nu^2 + 10 \mathbb{E}_{x \sim Q} [\|s_q(x)\|_2^2] F_\nu\,.$$

By the assumed boundedness of the kernel and its gradients, there exists a positive constant $M < \infty$ depending only on k and Q such that

$$\max\left\{\mathbb{E}_{x,x'\sim Q}[|k(x,x')|],\ \mathbb{E}_{x,x'\sim Q}[\|\nabla_{x'}k(x,x')\|_2^2],\ \mathbb{E}_{x,x'\sim Q}[\|\nabla_x k(x,x')\|_2^2]\right\} \leq M\ .$$

We hence conclude from (13) that

$$\Pr_Q(n|\hat{\mathbb{D}}_{P_\nu} - \hat{\mathbb{D}}_Q| \geq \epsilon) \leq \tfrac{n}{\epsilon}\left[M^{1/2}\left(8F_\nu^2 + 10\mathbb{E}_{x\sim Q}[\|s_q(x)\|_2^2]F_\nu\right)^{1/2} + 2M^{1/2}F_\nu^{1/2}\right]$$

$$\leq \tfrac{n}{\epsilon}M^{1/2}\underbrace{\left[F_\nu^{1/2}\left(8F_\nu + 10\mathbb{E}_{x\sim Q}[\|s_q(x)\|_2^2]\right)^{1/2} + 2F_\nu^{1/2}\right]}_{=:T_3}\ . \tag{14}$$

The term T_3 is $O(\max(F_\nu, F_\nu^{1/2}))$. Therefore, if $n = n_\nu = o(1/\max(F_\nu, F_\nu^{1/2}))$, then the right hand side of (14) converges to 0 as $\nu \to \infty$, thus $n_\nu|\hat{\mathbb{D}}_{P_\nu} - \hat{\mathbb{D}}_Q| \to 0$ in probability. This completes the proof.

B. Proof of Theorem 3.4

Theorem 3.4 follows directly from Proposition 3.1 and the next lemma, which states that the Fisher Divergence between Q and P_Δ decays with a rate at least exponentially fast in the inter-modal distance $\|\Delta\|_2$.

Lemma B.1. *Under the same assumptions in Prop. 3.1, we have* $\mathbb{E}_{x\sim Q}[\|s_{p_\Delta}(x) - s_q(x)\|_2^2] = o\left(e^{-\|\Delta\|_2^2/32}\right)$.

Proof of Lemma B.1. For any $\delta > 0$, define $B_\delta := \{x \in \mathbb{R}^d : \|x\|_2 \leq \delta\}$. We have the following decomposition

$$\mathbb{E}_{x\in Q}\left[\|s_{p_\Delta}(x) - s_q(x)\|_2^2\right]$$
$$= \mathbb{E}_{x\sim Q}\left[\delta_x(B_\delta)\|s_{p_\Delta}(x) - s_q(x)\|_2^2\right] + \mathbb{E}_{x\sim Q}\left[\delta_x(\mathbb{R}^d\backslash B_\delta)\|s_{p_\Delta}(x) - s_q(x)\|_2^2\right]\ . \tag{15}$$

The rest of the proof proceeds with bounding the two terms separately. We first note that standard computation gives

$$\tfrac{p_\Delta(x)}{q(x)} = \tfrac{\pi\exp\left(-\frac{1}{2}\|x\|^2\right)+(1-\pi)\exp\left(-\frac{1}{2}\|x-\Delta\|^2\right)}{\exp\left(-\frac{1}{2}\|x\|^2\right)} = \pi + (1-\pi)\exp\left(\Delta^\top x - \tfrac{1}{2}\|\Delta\|^2\right)\ ,$$

and

$$\|s_{p_\Delta}(x) - s_q(x)\|_2^2 = \left\|\tfrac{(1-\pi)\Delta\exp\left(\Delta^\top x - \frac{1}{2}\|\Delta\|_2^2\right)}{\pi+(1-\pi)\exp\left(\Delta^\top x - \frac{1}{2}\|\Delta\|_2^2\right)}\right\|_2^2 = \tfrac{(1-\pi)^2\|\Delta\|_2^2}{\left(1-\pi+\pi\exp\left(-\Delta^\top x + \frac{1}{2}\|\Delta\|_2^2\right)\right)^2}\ .$$

For $x \in B_\delta$, Cauchy-Schwarz inequality implies $\Delta^\top x \leq \|\Delta\|_2\|x\|_2 \leq \delta\|\Delta\|_2$. Hence

$$\|s_{p_\Delta}(x) - s_q(x)\|_2^2 \leq \tfrac{(1-\pi)^2\|\Delta\|_2^2}{\pi^2\exp\left(-2\Delta^\top x + \|\Delta\|_2^2\right)} \leq \tfrac{(1-\pi)^2\|\Delta\|_2^2}{\pi^2\exp\left(-2\delta\|\Delta\|_2 + \|\Delta\|_2^2\right)}\ ,$$

and the first term of (15) can be bounded as

$$\mathbb{E}_{x\sim Q}\left[\delta_x(B_\delta)\|s_{p_\Delta}(x) - s_q(x)\|_2^2\right] \leq \mathbb{E}_{x\sim Q}\left[\delta_x(B_\delta)\tfrac{(1-\pi)^2\|\Delta\|_2^2}{\pi^2\exp\left(-2\delta\|\Delta\|_2+\|\Delta\|_2^2\right)}\right] \leq \tfrac{(1-\pi)^2\|\Delta\|_2^2}{\pi^2\exp\left(-2\delta\|\Delta\|_2+\|\Delta\|_2^2\right)}\ ,$$

where the last inequality follows from the fact that $\mathbb{E}_{x\sim Q}[\delta_x(B_\delta)] \leq 1$.

To bound the second term of (15), we note that for $x \in \mathbb{R}^d\backslash B_\delta$,

$$\|s_{p_\Delta}(x) - s_q(x)\|_2^2 \leq \tfrac{(1-\pi)^2\|\Delta\|_2^2}{(1-\pi)^2} = \|\Delta\|_2^2\ .$$

Therefore,

$$\mathbb{E}_{x\sim Q}\left[\delta_x(B_\delta)\|s_{p_\Delta}(x) - s_q(x)\|_2^2\right] \leq \|\Delta\|_2^2\mathbb{E}_{x\sim Q}\left[\delta_x(B_\delta)\right] \leq 5^d\|\Delta\|_2^2\exp\left(-\tfrac{\delta^2}{8}\right)\ ,$$

13

by the tail probability of the norm of centred Gaussian random vectors (see e.g. Wainwright (2019, Prop. 2.5)). Combining these results we have

$$\mathbb{E}_{x \in Q}[\|s_{p_\Delta}(x) - s_q(x)\|_2^2] \leq \frac{(1-\pi)^2 \|\Delta\|_2^2}{\pi^2 \exp(-2\delta\|\Delta\|_2 + \|\Delta\|_2^2)} + 5^d \|\Delta\|_2^2 \exp\left(-\frac{\delta^2}{8}\right)$$
$$= (1-\pi)^2 \pi^{-2} \|\Delta\|_2^2 \exp\left(-\tfrac{1}{17}\|\Delta\|_2^2\right) + 5^d \|\Delta\|_2^2 \exp\left(-\tfrac{1}{17}\|\Delta\|_2^2\right) ,$$

where the last line follows by choosing $\delta = 8\|\Delta\|_2/17$. Noting the RHS of the last inequality is $o\left(-\tfrac{1}{32}\|\Delta\|_2^2\right)$ completes the proof. $\qquad\square$

C. Proof of Proposition 4.1

Proof. The stated assumptions ensure that, for all $\mathcal{K} \in \mathcal{S}$, $\mathbb{D}(Q, P; \mathcal{K})$ is well defined, and that $\mathbb{D}(Q, P; \mathcal{K}_{id}) = \mathbb{D}(Q, P) = 0 \iff Q = P$ (see, e.g., Chwialkowski et al. (2016, Theorem 2.2)). The desired result then follows since $\mathcal{K}$ is P-invariant for all $\mathcal{K}$ and $\mathbb{D}(Q, P; \mathcal{S}) \geq \mathbb{D}(Q, P)$. $\qquad\square$

D. Proof of Proposition 4.2

By Serfling (2009, Sections 5.5.1, 5.5.2), sufficient conditions for the stated results are

C1. $\mathbb{E}_{w,w' \sim R_Q}[\tilde{u}_P(w, w')^2] < \infty$.

C2. Under H_0: $\xi_1 := \mathrm{Var}_{w \sim R_Q}(\mathbb{E}_{w' \sim R_Q}[\tilde{u}_P(w, w')]) = 0$ and $\xi_2 := \mathrm{Var}_{w,w' \sim R_Q}(\tilde{u}_P(w, w')) > 0$.

C3. Under H_1: $\xi_1 > 0$.

Now C1. holds by assumption. To show C2., we start with the decomposition where, for fixed $w = (x^1, \ldots, x^S) \in \mathcal{X}^S$,

$$\mathbb{E}_{w' \sim R_Q}[\tilde{u}_P(w, w')] = \sum_{s=1}^S \mathbb{E}_{(y^1, \ldots, y^S) \sim R_Q}[u_P(x^s, y^s)] = \sum_{s=1}^S \mathbb{E}_{y^s \sim \mathcal{K}_s Q}[u_P(x^s, y^s)] , \tag{16}$$

where the first equality holds as for $w' \sim R_Q$ we can write $w' = (y^1, \ldots, y^S)$ for some $y^s \in \mathcal{X}$ by construction, and the second equality follows since the marginal distribution of y^s is $\mathcal{K}_s Q$. When $Q = P$, each term in (16) equals to $\mathbb{E}_{y^s \sim P}[u_P(x^s, y^s)]$ by P-invariance. Now, under the assumed conditions in Prop. 4.1, the same argument in the proof of Liu et al. (2016, Theorem 4.1) shows that $\mathbb{E}_{y^s \sim P}[u_P(x^s, y^s)] = 0$. Hence, $\xi_1 = 0$.

To prove $\xi_2 > 0$ when $Q = P$, we suppose for a contradiction that $\xi_2 = 0$. We then must have $\tilde{u}_P \equiv c'$ P-almost surely for some fixed constant c'. Since P admits a positive density on $\mathcal{X}$ by assumption, this implies that $c' = 0$. On the other hand, for any probability measure Q' on $\mathcal{X}$, taking expectation with respect to $R_{Q'}$ yields

$$0 = c' = \mathbb{E}_{w,w' \sim R_{Q'}}[\tilde{u}_P(w, w')]$$
$$= \sum_{s=1}^S \mathbb{E}_{(x^1, \ldots, x^S), (y^1, \ldots, y^m) \sim R_{Q'}}[u_P(x^s, y^s)]$$
$$= \sum_{s=1}^S \mathbb{E}_{x^s, y^s \sim \mathcal{K}_s Q'}[u_P(x^s, y^s)]$$
$$= \sum_{s=1}^S \mathbb{D}(\mathcal{K}_s Q', P)$$
$$\geq \mathbb{D}(Q', P) ,$$

where the second identity follows from a similar argument in (16), and the last line holds because $\mathcal{K}_{id} \in \mathcal{S}$. This is a contradiction, as it would imply $\mathbb{D}(Q', P) = 0$ for any $Q' \neq P$.

To show C3, we prove the contrapositive by supposing $\xi_1 = 0$ and aiming to show that $Q = P$. If $\xi_1 = 0$ then there must exist a constant c for which

$$c = \mathbb{E}_{w' \sim R_Q}[\tilde{u}_P(w, w')] ,$$

for all w R_Q-almost surely. With the stated choice of $\mathcal{K}$ and the assumption that Q admits a Lebesgue density, the above identity also holds for all w R_P-almost surely, where R_P is constructed in the same way as R_Q by replacing Q with P. Taking expectation of both sides with respect to $w \sim R_P$ then yields

$$c = \mathbb{E}_{w \sim R_P} \mathbb{E}_{w' \sim R_Q}[\tilde{u}_P(w, w')] = \mathbb{E}_{w' \sim R_Q} \mathbb{E}_{w \sim R_P}[\tilde{u}_P(w, w')] ,$$

where the last equality follows from the Fubini-Toneli Theorem. Following the same argument above for $Q = P$, we conclude that $\mathbb{E}_{w \sim R_P}[\tilde{u}_P(w, w')] = 0$, and hence

$$0 = \mathbb{E}_{w' \sim R_Q}\mathbb{E}_{w \sim R_Q}[\tilde{u}_P(w, w')] = \sum_{s=1}^{S} \mathbb{D}(\mathcal{K}_s Q, P) \geq \mathbb{D}(Q, P) .$$

It follows that $\mathbb{D}(Q, P) = 0$, thus $Q = P$.

E. Validity of the Accept-Reject Rule

E.1. A Sufficient Condition

Let $\mathcal{K}$ be the Markov transition kernel studied in Sec. 5.1. We first present a sufficient condition for the detailed balance equation (9):

$$\int_{x \in A} \sum_{u \in \mathcal{U}} \delta_{x'}(B) p(x) g(u) \alpha(x, x') dx = \int_{x' \in B} \sum_{u' \in \mathcal{U}} \delta_x(A) p(x') g(u') \alpha(x', x) dx', \tag{17}$$

for all $A, B \in \mathcal{B}(\mathcal{X})$. For simplicity, we have written $x' = h(x|u)$ and $x = h^{-1}(x'|u')$, so that the dependence of x' on u and of x on x' is implicit.

Proposition E.1. *Let p be a probability density function on $\mathcal{X} \subset \mathbb{R}^d$. Suppose that h is a deterministic, invertible function that is differentiable with differentiable inverse. Furthermore, let g be a known density defined on some discrete space $\mathcal{U}$. Consider a Markov transition kernel of the form*

$$\mathcal{K}(x, A) = \sum_{u \in \mathcal{U}} \delta_{x'}(A) g(u) \alpha(x, x') + \delta_x(A) r(x), \tag{18}$$

where $x' := h(x|u)$, $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise, and $r(x) = 1 - \sum_{u \in \mathcal{U}} g(u(x, x')) \alpha(x, x')$. Then an accept-reject rule $\alpha(x, x')$ satisfies the detailed balance condition (17) if

$$p(x) g(u) \alpha(x, x') = p(x') g(u') \alpha(x', x) \left| \frac{\partial h(x|u)}{\partial x} \right| . \tag{19}$$

Proof. The proof largely imitates Green & Hastie (2009, Sec. 2.1), which shows the claim when the density g is defined on a continuous space. Defining $\mathcal{U}_B := \{u : x' = h(x|u) \in B \text{ for some } x \in \mathcal{X}\}$ and $\mathcal{U}_A := \{u : x = h^{-1}(x'|u) \in B \text{ for some } x' \in \mathcal{X}\}$, we can rewrite (17) as

$$\int_{x \in A} \sum_{u \in \mathcal{U}_B} p(x) g(u) \alpha(x, x') dx = \int_{x' \in B} \sum_{u' \in \mathcal{U}_A} p(x') g(u') \alpha(x', x) dx' .$$

Noting that $(x, u) \in A \times \mathcal{U}_B \iff (x', u') = (h(x|u), u) \in B \times \mathcal{U}_A$, and by the invertibility of the transformation h, a change-of-variable formula can be applied to the right-hand-side of (17) to yield

$$\int_{x \in A} \sum_{u \in \mathcal{U}_A} p(x) g(u) \alpha(x, x') dx = \int_{x \in A} \sum_{u \in \mathcal{U}_A} p(x') g(u) \alpha(x', x) \left| \frac{\partial h(x|u)}{\partial x} \right| dx.$$

We therefore conclude that a sufficient condition is

$$p(x) g(u) \alpha(x, x') = p(x') g(u') \alpha(x', x) \left| \frac{\partial h(x|u)}{\partial x} \right| .$$

$\square$

In particular, it follows that the detailed balance condition holds with

$$\alpha(x, x') = \min \left(1, \frac{p(x') g'(u')}{p(x) g(u)} \left| \frac{\partial h(x|u)}{\partial x} \right| \right), \tag{20}$$

by verifying that it indeed satisfies (19). This can be viewed as a generalisation of the Metropolis-Hastings (MH) rule $\alpha(x, x') = \min \left(1, \frac{p(x') g'(u')}{p(x) g(u)} \right)$.

E.2. A Class of Valid Accept-Reject Rules

Accept-reject rules of the form (20) is not the only choice that satisfies the detailed balance condition. For the standard Metropolis-Hastings transition kernel, alternative accept-reject rules have been studied (Barker, 1965; Peskun, 1973; Hird et al., 2020). We follow Hird et al. (2020) to propose a class of accept-reject rules that are valid for proposed kernels of the form (18).

Lemma E.2. *Using the same notations in Prop. E.1, define $t(x, x') \coloneqq \frac{p(x')g(u')}{p(x)g(u)}$ when $p(x)g(u) > 0$, and $t(x, x') = 0$ otherwise, where $u, u' \in \mathcal{U}$ such that $x' = h(x|u)$ and $x = h(x'|u')$. Then the equality (19) holds for*

$$\alpha(x, x') = \rho\left(\left|\frac{\partial h(x|u)}{\partial x}\right| t(x, x')\right),$$

where ρ is any function that satisfies $\rho(s) = s\rho(1/s)$, for all $s > 0$, and $\rho(0) \coloneqq 0$.

Proof. We follow the derivation in Hird et al. (2020, Eq. 4). By the definition of t, it is obvious that $t(x, x') = 1/t(x', x)$ and $p(x)g(u)t(x, x') = p(x')g(u')$. The assumption on ρ then implies

$$\begin{aligned}
p(x)g(u)\alpha(x, x') &= p(x)g(u)\rho\left(\left|\frac{\partial h(x|u)}{\partial x}\right| t(x, x')\right) \\
&= p(x)g(u)t(x, x')\left|\frac{\partial h(x|u)}{\partial x}\right| \rho\left(\frac{1}{\left|\frac{\partial h(x|u)}{\partial x}\right| t(x, x')}\right) \\
&= p(x')g(u')\left|\frac{\partial h(x|u)}{\partial x}\right| \rho\left(\left|\frac{\partial h(x'|u')}{\partial x}\right| t(x', x)\right) \\
&= p(x')\alpha(x', x)\left|\frac{\partial h(x|u)}{\partial x}\right|,
\end{aligned} \tag{21}$$

where (21) follows from the fact that $\left|\frac{\partial h(x'|u')}{\partial x'}\right| = \left|\frac{\partial h(x|u)}{\partial x}\right|^{-1}$ by the invertibility of h. $\qquad\square$

In particular, choosing $\rho(t) = \min(1, t)$ gives the generalised MH accept-reject rule (20). Another feasible choice is $g(t) = t/(1 + t)$, which leads to a generalised version of the Barker's rule (Barker, 1965; Peskun, 1973; Livingstone & Zanella, 2021)

$$\alpha(x', x) = \frac{p(x')g(u')\left|\frac{\partial h(x|u)}{\partial x}\right|}{p(x)g(u) + p(x')g(u')\left|\frac{\partial h(x|u)}{\partial x}\right|},$$

whenever $p(x)g(u) > 0$, and 0 otherwise.

F. Proof of Proposition 5.1

We first characterise the limiting distribution when the initial distribution is a point mass δ_{x_0} for any $x_0 \in \mathbb{R}^d$, then generalise the result to an arbitrary probability measure Q.

F.1. Limiting Distribution with a Point Mass Initial Distribution

Fixing $x_0 \in \mathbb{R}^d$, we define $\mathcal{I}_{x_0} \coloneqq \{x_0 + k\nu : k \in \mathbb{Z}\}$. We first identify a stationary distribution and aim to show that it is also the limiting distribution.

Lemma F.1. *The following probability mass function defines a stationary distribution under $\mathcal{K}$:*

$$r_{x_0}(x) = \frac{p(x)}{\sum_{k \in \mathbb{Z}} p(x_0 + k\nu)},$$

if $x \in \mathcal{I}_{x_0}$, and $r_{x_0}(x) = 0$, otherwise.

Proof. A sufficient condition for the detailed-balance condition in this case is

$$r_{x_0}(x)g(u)\alpha(x, x') = r_{x_0}(x')g'(u')\alpha(x', x),$$

for all $x, x' \in \mathbb{R}^d$. Fix x. Since $\alpha(x, x') = 0$ unless $x' \in \{x - \nu, x + \nu\}$, it is sufficient to check whether the above equation holds for $x' \in \{x - \nu, x + \nu\}$. For, e.g., $x' = x + \nu$,

$$
\begin{aligned}
\text{LHS} &= \frac{p(x)}{\sum_{k \in \mathbb{Z}} p(x_0 + k\nu)} g(u) \min\left(1, \frac{g(u')p(x')}{g(u)p(x)}\right) \\
&= \frac{1}{\sum_{k \in \mathbb{Z}} p(x_0 + k\nu)} g(u') \min\left(p(x), p(x')\right), && \text{as } g(u) = g(u') \text{ by definition.} \\
&= \frac{p(x')}{\sum_{k \in \mathbb{Z}} p(x_0 + k\nu)} g(u') \min\left(\frac{g(u)p(x)}{g(u')p(x')}, 1\right), && \text{again by } g(u) = g(u'). \\
&= \text{RHS.}
\end{aligned}
$$

A similar derivation for $x' = x - \nu$ completes the proof. $\qquad\square$

The next result shows that the Markov chain defined on $\mathcal{I}_{x_0}$ is irreducible and aperiodic under mild conditions.

Lemma F.2. *Given $x_0 \in \mathcal{X}$ and consider the Markov chain with initial distribution δ_{x_0}. Then*

1. *All state in $\mathcal{I}_{x_0}$ are irreducible.*

2. *A state $x \in \mathcal{I}_{x_0}$ is aperiodic if $p(x + \nu) < p(x)$ or $p(x - \nu) < p(x)$.*

Proof. To prove *1.*, it is sufficient to show that any state $x \in \mathcal{I}_{x_0}$ can reach any other state $y \in \mathcal{I}_{x_0}$ with positive probability, i.e., for any singleton $A = \{y\}$ where $y \in \mathcal{I}_{x_0}$, there exists $T \in \mathbb{N}$ so that $\mathcal{K}^T(x, A) > 0$, where

$$
\mathcal{K}(x, A) = \sum_{u \in \mathcal{U}} \delta_{x'}(A)g(u)\alpha(x, x') + \delta_x(A)r(x),
$$

where $\mathcal{U} = \{(1, 2), (2, 1)\}$, $g(u) = 1/2$ for all $u \in \mathcal{U}$, and $x' = x'(x, u) = x + \theta(\mu_{u_1} - \mu_{u_2})$ (see Section 5.2). We fix $x \in \mathcal{I}_{d_0}$ and pick $y \in \{x - \nu, x + \nu\}$, i.e., y is the point immediately to the left or right of x in $\mathcal{I}_{x_0}$. The transition probability in this case reduces to

$$
\mathcal{K}(x, \{y\}) = \tfrac{1}{2}\alpha(x, y) = \tfrac{1}{2}\min\left(1, \frac{p(y)}{p(x)}\right),
$$

which is positive as p is positive on $\mathbb{R}^d$, i.e., x can move to its left or right state in one step with positive probability. An inductive argument directly shows that x can move to *any* $y \in \mathcal{I}_{x_0}$ with positive probability in finitely many steps. This shows *1.*

To prove *2.*, we note that if $p(x + \nu) < p(x)$ or $p(x - \nu) < p(x)$, then the 1-step transition probability of starting from x and staying is non-zero. Indeed,

$$
\mathcal{K}(x, \{x\}) = 1 - \mathcal{K}(x, \{x + \nu\}) - \mathcal{K}(x, \{x - \nu\}) = 1 - \tfrac{1}{2}\alpha(x, x + \nu) - \tfrac{1}{2}\alpha(x, x - \nu).
$$

Since $p(x + \nu) < p(x)$, we have $\alpha(x, x + \nu) = \min\left(1, p(x + \nu)/p(x)\right) < 1$. Similarly, $\alpha(x, x - \nu) < 1$. Hence, $\mathcal{K}(x, \{x\}) > 0$, thus x is aperiodic. $\qquad\square$

Combining Lemma F.1 and F.2, we can identify the limiting distribution when the initial distribution is a point mass at $x \in \mathcal{I}_{x_0}$.

Proposition F.3. *If $p(x + \nu) < p(x)$ or $p(x - \nu) < p(x)$ for all $x \in \mathcal{I}_0$, then r_{x_0} is also the unique limiting distribution, i.e., $\mathcal{K}^T(x, A) \to \sum_{x' \in A} r_{x_0}(x')$, for all $x \in \mathcal{I}_{x_0}$ and $A \subset \mathcal{I}_{x_0}$. Furthermore, for a Lebesgue-measurable set $A \subset \mathcal{X}$ and a state $x \in \mathcal{X}$,*

$$
\lim_{n \to \infty} \mathcal{K}^n(x, A) = \sum_{x' \in \mathcal{I}_x} \delta_{x'}(A) \frac{p(x')}{\sum_{k \in \mathbb{Z}} p(x + k\nu)} . \tag{22}
$$

Proof. Under the stated assumption, Lemma F.2 shows that the Markov chain is irreducible and aperiodic on $\mathcal{I}_{x_0}$. Since the stationary distribution of an irreducible and aperiodic Markov chain defined on a countable space is also the unique limiting distribution (e.g., Meyn & Tweedie (2012)), the first part follows. (22) holds because, for any x, r_x is a probability mass function taking zero values outside of $\mathcal{I}_x$. $\qquad\square$

Algorithm 2 Estimating mode vectors and Hessians (Pompe et al., 2020, Algorithm 3)

Input: Initial points $s_1, \ldots, s_{M_0}$, small positive value β.
Output: Approximates for mode vectors $\{\mu_1, \ldots, \mu_M\}$.
Initialise BFGS at points $s_1, \ldots, s_{M_0}$ and run the algorithm to minimise $-\log p(x)$.
Denote the returned estimates of the local optima by $m_1, \ldots, m_{M_0}$ and their corresponding Hessian matrices by $A_1, \ldots, A_{M_0}$.
Set $\mu_1 := m_1$, $A_{\mu_1} := A_1$, $M = 1$.
for $i = 2, \ldots, M_0$ **do**
    if $\min_{j \in \{1, \ldots, M\}} \frac{1}{2}((\mu_j - m_i)^\top A_{\mu_j}(\mu_j - m_i) + (\mu_j - m_i)^\top A_i(\mu_j - m_i)) < \beta$ **then**
        $k := \arg\min_{j \in \{1, \ldots, M\}} \frac{1}{2}((\mu_j - m_i)^\top A_{\mu_j}(\mu_j - m_i) + (\mu_j - m_i)^\top A_i(\mu_j - m_i))$.
        if $p^*(\mu_k) < p^*(m_i)$ **then**
            Set $\mu_k := m_i$ and $A_{\mu_k} := A_i$.
        end if
    else
        $\mu_{M+1} := m_i$ and $A_{\mu_{M+1}} := A_i$.
        $M := M + 1$.
    end if
end for

F.2. Limiting Distribution with a General Initial Distribution

We now prove Proposition 5.1, which characterises the limiting distribution with a general initial distribution Q.

Proof of Proposition 5.1. Let $A \subset \mathcal{X}$ be Lebesgue-measurable. For any fixed $n \in \mathbb{Z}_+$, the probability of A under the n-step perturbed distribution is

$$(\mathcal{K}^n Q)(A) = \int_{x \in \mathcal{X}} \mathcal{K}^n(x, A) Q(dx) = \int_{x \in \mathcal{X}} \mathcal{K}^n(x, A) q(x) dx.$$

Since $|\mathcal{K}^n(x, A)| \leq |\mathcal{K}^n(x, \mathcal{I}_x)| \leq 1$ and $\int_{x \in \mathcal{X}} q(x) dx = 1 < \infty$, we can apply Dominated Convergence Theorem (Bartle, 2014, Section 5.6) to conclude

$$
\begin{aligned}
\lim_{n \to \infty} (\mathcal{K}^n Q)(A) &= \int_{x \in \mathcal{X}} \mathcal{K}^\infty(x, A) q(x) dx \\
&= \int_{x \in \mathcal{X}} \sum_{x' \in \mathcal{I}_x} \delta_{x'}(A) \frac{p(x')}{\sum_{k \in \mathbb{Z}} p(x + k\nu)} q(x) dx, &&\text{by Corollary F.3.} \\
&= \int_{x \in \mathcal{X}} \sum_{s \in \mathbb{Z}} \delta_{x+s\nu}(A) \frac{p(x+s\nu)}{\sum_{k \in \mathbb{Z}} p(x + k\nu)} q(x) dx \\
&= \sum_{s \in \mathbb{Z}} \int_{x \in \mathcal{X}} \delta_{x+s\nu}(A) \frac{p(x+s\nu)}{\sum_{k \in \mathbb{Z}} p(x + k\nu)} q(x) dx &&(23)\\
&= \sum_{s \in \mathbb{Z}} \int_{u \in \mathcal{X}} \delta_u(A) \frac{p(u)}{\sum_{k \in \mathbb{Z}} p(w + (k-s)\nu)} q(u - s\nu) du &&(24)\\
&= \int_{u \in \mathcal{X}} \sum_{s \in \mathbb{Z}} \delta_u(A) \frac{p(u)}{\sum_{k \in \mathbb{Z}} p(w + (k-s)\nu)} q(u - s\nu) du &&(25)\\
&= \int_{u \in \mathcal{X}} \delta_u(A) \frac{p(u)}{\sum_{k \in \mathbb{Z}} p(w + (k-s)\nu)} \sum_{s \in \mathbb{Z}} q(u - s\nu) du \\
&= \int_{u \in A} p(u) \frac{\sum_{s \in \mathbb{Z}} q(u + s\nu)}{\sum_{k \in \mathbb{Z}} p(u + k\nu)} du,
\end{aligned}
$$

where in (23) we have applied Fubini-Toneli Theorem (Bartle, 2014, Section 10.9, 10.10), (24) follows from a change of variable $u := x + s\nu$ for a given s, (25) follows from Fubini-Toneli Theorem again, and the last line holds by a re-indexing of the sums on the numerator and on the denominator. In particular, we can apply Fubini-Toneli Theorem as $\int_{x \in \mathcal{X}} \sum_{s \in \mathbb{Z}} |\delta_{x+s\nu}(A) \frac{p(x+s\nu)}{\sum_{k \in \mathbb{Z}} p(x+k\nu)}| q(x) dx \leq \int_{x \in \mathcal{X}} \sum_{s \in \mathbb{Z}} \frac{p(x+s\nu)}{\sum_{k \in \mathbb{Z}} p(x+k\nu)} q(x) dx < \infty$. $\square$

G. Implementation Details

This section holds details about the practical implementation of the spKSD and the ospKSD methods.
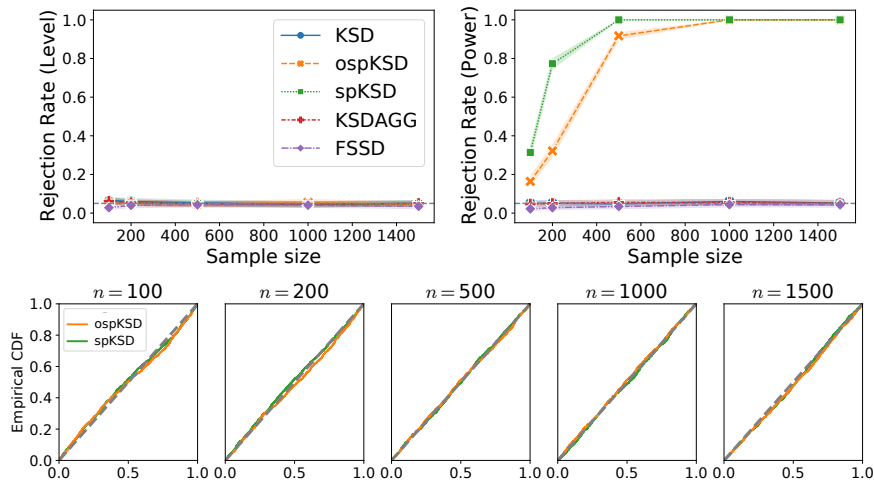
Figure 6: Level (top left) and power (top right) experiments with the multivariate Gaussian mixture example. Bottom: Empirical cumulative distribution function (CDF) of the p-values in the level experiments, where the grey dashed line is the CDF of a uniform distribution on $[0, 1]$.

G.1. Finding Mode Vectors via Optimisation and Merging

Finding local modes   In practice, the mode locations and Hessians of the density of a non-trivial target distribution are rarely available. Pompe et al. (2020) describes a general framework to estimate these quantities. It proceeds by running in parallel a sequence of optimisers initiated at different starting points. This is done by minimising the objective $-\log p$ using the BFGS algorithm (Nocedal & Wright, 2006), which returns both the local minima and the approximated Hessian at those points. In our experiments, we use the BFGS algorithm and run for at most 1000 iterations with each initial point.

Mode merging   Although the end points of the optimisation procedure starting from different initial points may lie close to the local minima, they will still be numerically different from each other. Pompe et al. (2020) proposed to merge two end points m_i and m_j if their Mahalanobis distance weighted by the averaged Hessians at those points is below a given threshold β. The full procedure is stated in Algorithm 2 for completeness.

Choosing the initial points   A set of n_{init} initial points for BFGS can be constructed either by sampling randomly from a product of intervals $[L_1, U_1] \times \cdots \times [L_d, U_d]$ in $\mathcal{X}$, or simply from a held-out training set. The first approach will allow modes not covered by the training data to be detected, and the second approach can lead to faster convergence of the optimisation algorithm when the training points lie near the modes of P. For spKSD, the first approach is used as we do not assume a held-out set is available. For ospKSD, to combine the best of the two approaches whilst maintaining the same computational budget, half of the n_{init} initial points are drawn randomly from the training set and the other half are initialised uniformly from $[L_1, U_1] \times \cdots \times [L_d, U_d]$.

G.2. Choosing the Number of Transitions T

The number of transitions T dictates the perturbed distribution, thus impacting the performance of spKSD. Intuitively, T should be set to a large value when the acceptance rate is low to ensure the limiting distribution is achieved. The spKSD could suffer from a low acceptance rate when the estimates of the modes and local Hessians of the target distribution are inaccurate, or when the target distribution cannot be approximated by a mixture of elliptic distributions.

We propose two heuristics to choose this hyper-parameter in practice: *(i)* viewing this as another hyper-parameter and tuning it using a training set by selecting from a pre-specified set of values, or *(ii)* setting it to a large value (e.g., $T = 1000$) if the computational budget allows.

In particular, we recommend a large T because, with the proposed transition kernel $\mathcal{K}$, the KSD $\mathbb{D}(\mathcal{K}^T Q, \mathcal{K}^T P)$ between the perturbed distributions does *not* necessarily decrease as T grows. This is because $\mathcal{K}^T Q$ does *not* necessarily converge to P as $T \to \infty$, since $\mathcal{K}$ is *not* irreducible (see the discussions in Section 5.1).

H. Experimental Details and Supplementary Plots

In this section, we provide detailed definition of the distributions used in each experiment, as well as supplementary figures, including a level study and a power study of the proposed methods.

H.1. Multivariate Gaussian Mixture: Supplementary Plots

We include a level study and a power study for the spKSD and ospKSD tests. The target distribution is a multivariate Gaussian mixture in 50 dimensions, with density $p(x) \propto \pi_p \exp\left(-\frac{1}{2}\|x\|_2^2\right) + (1 - \pi_p)\exp\left(-\frac{1}{2}\|x - \Delta e_1\|_2^2\right)$, where $\pi_p = 0.5$, $\Delta = 6$, and $e_1 \in \mathbb{R}^d$ is a vector with 1 in the first coordinate and 0 in others. Samples are drawn either from the same distribution (level experiment), or from only the left component (power experiment). The probability of rejection over 100 repetitions is plotted in Fig. 6. We can see from the top left plot that under the null hypothesis, all tests have the prescribed test level $\alpha = 0.05$. The plots in the bottom row further confirm the validity of the test by showing that the empirical cumulative distribution function (CDF) of the p-values under the null is indeed close to the CDF of a uniform distribution. In the top right plot, we investigate the rejection rate under the alternative hypothesis where the samples are drawn from the left mode only. We see that both ospKSD and spKSD achieve a significantly higher power than the benchmarks (KSD, KSDAGG and FSSD), whose power remains close to the level for all sample sizes.

H.2. Mixture of t and Banana Distributions

The mixture of t and banana example mostly follows the setup in Pompe et al. (2020). Each t-distribution has 7 degrees of freedom and covariance matrix $0.1\sqrt{d}I_d$. Each banana-shaped distribution has density function $p_{b,\mu} = p_\mu \circ \phi_b$, where $\phi_{b,\mu}(x) = (x_1, x_2 + bx_1^2 - 100b, x_3, \ldots, x_d)^\top$ with $b = 0.003$, and p_μ is the density of a t-distribution with 7 degrees of freedom centred at $\mu \in \mathbb{R}^d$ and with shape matrix $C = \text{diag}(100, 1, \ldots, 1) \in \mathbb{R}^{d \times d}$.

H.3. Sensors Localisation

Following Tak et al. (2018), we use a diffuse bivariate Gaussian prior distribution $\mathcal{N}(0, 10^2 I_2)$ for each $x_i \in \mathbb{R}^2$. Let w_{ij} be the binary random variable for which $w_{ij} = 1$ if the distance y_{ij} is observed and 0 otherwise. The full posterior is

$$\pi(x_1, \ldots, x_4 | y, w) \propto \exp\left(-\frac{\sum_{k=1}^4 x_k^\top x_k}{2 \times 10^2}\right) \Pi_{i<j} f_{ij}(x_i, x_j | y_{ij}, w_{ij}),$$

where $w = \{w_{ij}\}$, $y = \{y_{ij}\}$ and

$$f_{ij}(x_i, x_j | y_{ij}, w_{ij}) = \left[\exp\left(-\frac{(y_{ij} - \|x_i - x_j\|_2)^2}{2 \times 0.02^2}\right) \exp\left(\frac{-\|x_i - x_j\|_2^2}{2 \times 0.3^2}\right)\right]^{w_{ij}} \left[1 - \exp\left(\frac{-\|x_i - x_j\|_2^2}{2 \times 0.3^2}\right)\right]^{1 - w_{ij}}.$$

H.4. Gaussian-Bernoulli Restricted Boltzmann Machine

We include a supplementary experiment with Gaussian-Bernoulli Restricted Boltzmann Machines (RBMs) (Cho et al., 2013). This model is a popular benchmark for assessing GOF tests (Liu et al., 2016; Jitkrittum et al., 2017; Schrab et al., 2022). It is a latent variable model with joint density $p(x, h) \propto \exp\left(\frac{1}{2}x^\top Bh + b^\top x + c^\top h - \frac{1}{2}\|x\|_2^2\right)$, where $h \in \{-1, 1\}^{d_h}$, and B, b, c are fixed hyperparameters. The marginal density $p(x)$ can be rewritten as a mixture of Gaussian distributions:

$$p(x) = \sum_h \gamma(h) N\left(x; \tfrac{1}{2}Bh + b, I_d\right), \qquad \text{where } \gamma(h) \propto \exp\left(\frac{1}{2}\left\|\tfrac{1}{2}Bh + b\right\|_2^2 + c^\top h\right). \tag{26}$$

We consider two parameter settings: *standard* and *multi-modal*. The *standard* setting follows the setups in Liu et al. (2016), in which case the RBM is unimodal. For the target distribution P, we randomly sample the entries of b and c from a standard normal, and select the entries of B from $\{-1, 1\}$ with equal probability. We sample from a perturbed version of P where Gaussian noises with standard deviation σ are injected into the entries of B. As σ increases, the problem becomes easier, so all tests are able to reject with a high probability (Fig. 7).

For the *multi-modal* setting, we set $c = 0$ and $b = 0$, and choose B so that the modes are well-separated. Samples are drawn from the same model with a different c, which controls the mixing weights. Specifically, we choose $B = 6E$, where $E \in \mathbb{R}^{d \times d_h}$ is formed by the top d row and d_h columns of the matrix $I_{d_{\max}}$, where $d_{\max} := \max(d, d_h)$. This renders the local modes of p to be located at the corners of a hyper-cube of width 6. Due to the choice of b and B, changing c only affects the weights of the components but not their mode locations. We choose $c = 0$ for the target so that all components
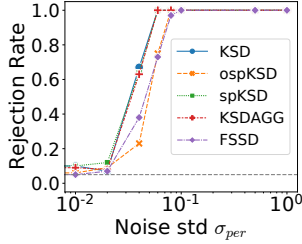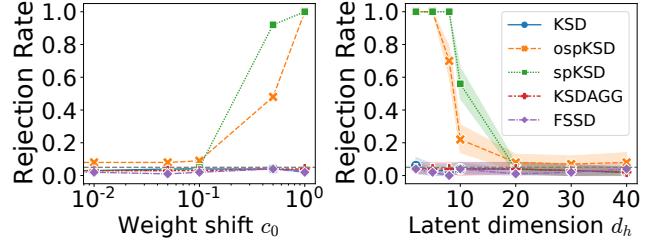
Figure 7: GB-RBM with the standard setting.

Figure 8: GB-RBM with the multi-modal setting.

have equal weights, and draw samples with from the same model with $c = (c_0, c_0, 0, \ldots, 0) \in \mathbb{R}^{d_h}$ for some c_0 using a Gibbs sampler, which we describe in the next subseciton.

We run ospKSD and spKSD with $T = 50$ steps and report the results in Fig. 8. The left plot shows the rejection probability with different values of c_0 and $d_h = 5$. As c_0 increases, the weights deviate further from those in the target, which is detected by ospKSD and spKSD as shown by the increasing power. The benchmarks again fail to detect this discrepancy due to the sparsity of the modes. We then fix $c_0 = 5$ and analyse how the performance scales with the latent dimension d_h in the right plot of Fig. 8. With a moderate d_h, the rejection probabilities of ospKSD and spKSD are significantly larger than the others. This gap vanishes for $d_h \geq 20$, since a larger d_h gives rise to more modes in p and hence more possible jump directions for each point. Therefore, the probability of proposing a "correct" move at each step declines, leading to a small acceptance rate and thus a low power.

H.4.1. SAMPLING FROM GAUSSIAN-BERNOULLI RBMs WITH WELL-SEPARATED MODES

We discuss practical considerations when sampling from the Gaussian-Bernoulli RBM with our particular choice of B (the *multi-modal* setting). Denote by GB-RBM(B, b, c) the joint density of the Gaussian-Bernoulli RBM with parameters B, b, c (see the previous section for definition).

In the *multi-modal* setting, we set $c = (6, 6, 0, \ldots, 0) \in \mathbb{R}^{d_h}$, and $B = \lambda E$, where λ is a large, positive constant, and $E \in \mathbb{R}^{d \times d_h}$ is the top $d \times d_h$ sub-matrix of $I_{d_{\max}}$, where $d_{\max} := \max(d, d_h)$. These lead to a model that is a mixture of Gaussian distributions with *distantly* located modes. More specifically, the modes are located at the corners of the hyper-cube in $\mathbb{R}^d$ with length 6.

The standard way to sample from a GB-RBM is to use a Gibbs sampler (see e.g., Melchior et al. (2017); Jitkrittum et al. (2017)). However, the Gibbs sampling can suffered from poor mixing when λ is large. This is because the modes will become more disconnected as λ increases, thus making it more challenging for the sampler to learn the mixing weights correctly. This means an impractically long burn-in period would be required for the Gibbs sampler to produce faithful samples from the ground truth.

We propose a practical method to generate faithful samples when λ is large. It is based on the observation that, with B of the above form, varying the value of λ only affects the mean locations of the Gaussian components in the GB-RBM, but *not* the mixing ratios. Indeed, for any $h \in \{-1, 1\}^{d_h}$, we have $\|Bh\|_2^2 = \lambda^2 \|Eh\|_2^2 = \lambda^2 d_h$, which is constant for all h. Substituting this into $\gamma(h)$ of (26),

$$\gamma(h) = \frac{\exp\left(\frac{1}{8}\lambda^2 d_h + c^\top h\right)}{\sum_{h'} \exp\left(\frac{1}{8}\lambda^2 d_h + c^\top h'\right)} = \frac{\exp\left(c^\top h\right)}{\sum_{h'} \exp\left(c^\top h'\right)},$$

which does not depend on λ. It follows that, given any $\lambda, \lambda' \geq 0$, if $(x, h) \sim$ GB-RBM$(\lambda E, b, c)$, then $(y, h) \sim$ GB-RBM$(\lambda' E, b, c)$, where $y := x - \frac{1}{2}(\lambda - \lambda')Eh$. This implies that we can sample from GB-RBM$(\lambda E, b, c)$ with a large λ by the following procedure:

1. Use a Gibbs sampler to sample $(x, h) \sim$ GB-RBM$(\lambda' E, b, c)$, where λ' is small.

2. Set $y = x - \frac{1}{2}(\lambda' - \lambda)Eh$.

Therefore, assuming the Gibbs sampler is capable of generating faithful samples from GB-RBM$(\lambda' E, b, c)$ for some $\lambda' \geq 0$, this will produce faithful samples from GB-RBM$(\lambda E, b, c)$ for any $\lambda \geq 0$ large. In our experiments, we used $\lambda' = 0$.