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

Xing Liu
Department of Mathematics
Imperial College London

Andrew Duncan
Department of Mathematics
Imperial College London
& Alan Turing Institute

Axel Gandy
Department of Mathematics
Imperial College London

Abstract

Kernelized Stein discrepancy (KSD) is a score-based discrepancy widely employed in goodness-of-fit tests. It is applicable even when the target distribution has an unknown normalising factor, such as in Bayesian analysis. We show theoretically and empirically that the power of the KSD test can be low when the target distribution has well-separated modes, which is due to insufficient data in regions where the score functions of the alternative and the target distributions differ the most. To improve its test power, we propose to perturb the target and alternative distributions before applying the KSD test. The perturbation uses a Markov transition kernel that leaves the target invariant but perturbs alternatives. We provide numerical evidence that the proposed approach can lead to a substantially higher power than the KSD test when the target and the alternative are mixture distributions that differ only in mixing weights.

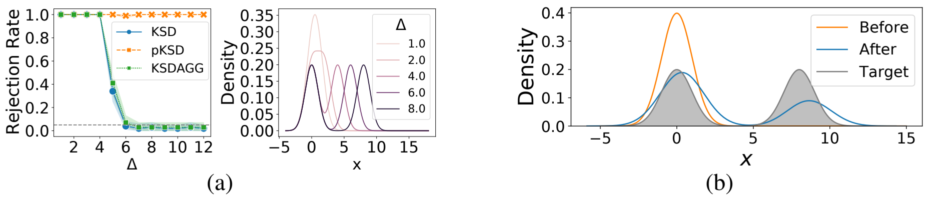
1 Introduction

Goodness-of-fit (GOF) testing concerns the question: given a *target distribution* P and a finite sample drawn from a *alternative distribution* Q, is there evidence against the null hypothesis $H_0 : Q = P$? A popular test statistic for GOF tests is kernelized Stein discrepancy (KSD), which is a score-based statistical divergence and is applicable even if the model is *unnormalised* (Chwialkowski et al., 2016; Liu et al., 2016).

However, KSD can fail to detect discrepancies when the target distribution has *well-separated* modes. For example, when Q and P are mixtures with the same components but with different mixing proportions, it has been observed that the KSD test would suffer from a low test power. This is a consequence of the *blindness* of KSD to isolated components, which is a well-known problem for KSD and other score-based applications, such as density estimation (Wenliang et al., 2019; Zhang et al., 2022) and quality measures for MCMC (Gorham et al., 2019). Neverthless, how the blindness of KSD manifests itself in GOF tests is yet to be formalised.

The contribution of this paper is twofold. First, we show theoretically and numerically on a bimodal Gaussian example that the power of KSD tests can converge to the test level as the mode separation increases (Fig. 1a). This is different from the analyses in Wenliang and Kanagawa (2020) and Gorham and Mackey (2017), which focus on the convergence of the sample KSD statistic but not the test power. Second, we address this issue by introducing a *perturbation operator*. The operator is designed to be a Markov transition kernel that leaves the target invariant but perturbs alternatives to create discrepancies that can be more easily detected by KSD. We discuss how to perform GOF tests with the proposed discrepancy, and show numerically that it can increase the test power against multi-modal targets, sometimes substantially from the nominal level to almost 1.

Score-based Method Workshop at Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) (b)

Figure 1: Power for a 1D bimodal Gaussian target distribution with mixing weight $\pi_p = 0.5$ and mode separation Δ. The alternative distribution is only the left component, from which samples are drawn. pKSD is our proposed method; the others are existing methods. (a) Rejection rates and target densities for varying Δ. (b) Density of the target distribution, as well as the density of the alternative distribution before and after 10 steps of the perturbation described in Sec. 4.

2 Background

Kernelized Stein discrepancy Denote by Q and P two Borel probability measures supported on $\mathcal{X} = \mathbb{R}^d$. We assume P admits a positive, continuously differentiable density p with respect to the Lebesgue measure. A statistical divergence that measures how well samples from Q agree with P is the Stein discrepancy (SD) (Stein, 1972; Gorham and Mackey, 2015)

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

where $\mathcal{F}$ is a set of functions on $\mathcal{X}$, and $\mathcal{A}_P$ is an operator acting on $\mathcal{F}$. A natural candidate for $\mathcal{A}_p f(x) := \langle \nabla \log p(x), f(x) \rangle + \langle \nabla, f(x) \rangle$, where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a continuously differentiable, vector-valued function. When restricting $\mathcal{F}$ to the unit ball of a *reproducing kernel Hilbert space* (RKHS) (Berlinet and Thomas-Agnan, 2011), $\mathbb{S}(Q, P; \mathcal{F})$ is a statistical discrepancy, i.e., $\mathbb{D}(Q, P; \mathcal{F}) = 0 \iff Q = P$, and (1) can be solved efficiently (Liu et al., 2016; Chwialkowski et al., 2016). Formally, let $\mathcal{H}$ be an RKHS associated with positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and $\mathcal{F}^d$ the unit ball of the d-times Cartesian product $\mathcal{H}^d := \mathcal{H} \times \cdots \mathcal{H}$. Choosing $\mathcal{F} = \mathcal{F}^d$ and the operator $\mathcal{A}_P$ yields the *(Langevin) kernelized Stein discrepancy* (KSD): $\mathbb{D}(Q, P) := \mathbb{S}(Q, P; \mathcal{F}^d)$.

Assuming the kernel k has continuous first-order derivatives with respect to both arguments, KSD attains a closed form: $\mathbb{D}(Q, P) = \mathbb{E}_{x, x' \sim Q}[u_P(x, x')]$, where $x, x' \sim Q$ are independent random variables, and u_P is the *Stein kernel*: $u_P(x, x') := s_p(x)^\top k(x, x') s_p(x') + s_p(x)^\top \nabla_{x'} k(x, x') + \nabla_x k(x, x')^\top s_p(x') + \sum_{i=1}^d \frac{\partial^2}{\partial x_i \partial x'_i} k(x, x')$ (Chwialkowski et al., 2016, Thm. 2.1). 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 (possibly intractable) normalising constant of p. Given a finite sample $\{x_i\}_{i=1}^n$ from Q, a consistent and unbiased estimator for $\mathbb{D}(Q, P)$ is the U-statistic $\hat{\mathbb{D}}_P := \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} u_P(x_i, x_j)$.

GOF testing based on KSD Under mild regularity conditions (see e.g., Chwialkowski et al. (2016, Thm. 2.1), Liu et al. (2016, Prop. 3.3)), KSD guarantees $\mathbb{D}(Q, P) = 0 \iff Q = P$, and hence testing $H_0 : Q = P$ against $H_1 : Q \neq P$ is equivalent to $H_0 : \mathbb{D}(Q, P) = 0$ against $H_1 : \mathbb{D}(Q, P) \neq 0$. The KSD test uses $\hat{\mathbb{D}}_P$ as a test statistic, and uses a bootstrap procedure to approximate the (intractable) asymptotic distribution of $\hat{\mathbb{D}}_P$ under H_0; see, e.g., Liu et al. (2016).

3 Limitations of KSD in GOF tests

As all other hypothesis tests, the KSD test can be insensitive to certain families of alternatives. One example is when the target and the alternative distributions are mixtures that differ only in the mixing weights. This is a consequence of the "blindness" of KSD as a score-based discrepancy (Wenliang and Kanagawa, 2020). It can be seen from the fact that KSD is upper-bounded by the *Fisher divergence* (Johnson, 2004): $\mathbb{F}(q, p) := \mathbb{E}_{x \sim Q}[\|s_p(x) - s_q(x)\|_2^2]$ (see Liu et al. (2016, Thm. 5.1)), so if the set where the score difference $\|s_p(x) - s_q(x)\|_2^2$ is large has low Q-probability, the Fisher divergence, thus also the KSD, will be small. The blindness of KSD has been highlighted in a number of works (Wenliang and Kanagawa, 2020; Matsubara et al., 2021; Zhang et al., 2022; Gorham et al., 2019); however, its implication to the test power in GOF testing has yet been formalised.

We show in Prop. 1 (proved in Appendix A) that this issue will cause the test power to converge to the prescribed test level as mode separation increases, unless the sample size grows exponentially fast.

Proposition 1. *Let $Q = \mathcal{N}(0, I_d)$ and $P = P_\Delta = \pi\mathcal{N}(0, I_d) + (1-\pi)\mathcal{N}(\Delta, I_d)$, respectively, where $\pi \in [0,1]$ is the mixing proportion and $\Delta \in \mathbb{R}^d$. Suppose the kernel k satisfies*

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

Let $x_1, x_2, \ldots$ be a sequence of i.i.d. samples from Q. Suppose $\Delta_1, \Delta_2, \cdots \in \mathbb{R}^d$ is a sequence with $\|\Delta_\nu\|_2 \to \infty$ as $\nu \to \infty$. Let $n_1, n_2, \cdots \in \mathbb{N}$ be such that $n_\nu = o\left(e^{\|\Delta_\nu\|_2^2/64}\right)$. Then

$$n_\nu \hat{\mathbb{D}}_{P_{\Delta_\nu}} \to_d \textstyle\sum_{j=1}^\infty c_j(Z_j^2 - 1) \quad (\nu \to \infty), \tag{3}$$

where $\hat{\mathbb{D}}_{P_{\Delta_\nu}}$ is computed using $x_1, \ldots, x_{n_\nu}$, $Z_j \sim \mathcal{N}(0,1)$ i.i.d. and $\{c_j\}_j$ are the eigenvalues of the Stein kernel u_P under Q, i.e. solutions of $c_j\phi_j(x) = \mathbb{E}_{x'\sim Q}[u_P(x,x')\phi_j(x')]$ for non-zero ϕ_j.

Remark 1.1. *The RHS of (3) is the limiting distribution of $\hat{\mathbb{D}}_{P_\Delta}$ under H_0 (see, e.g., Liu et al. (2016)); hence, Prop. 1 implies that, unless the sample size n grows prohibitively fast, the power of the KSD test will converge to the test level as the two modes of p becomes more and more separated.*

Remark 1.2. *Assumption (2) is mild 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* distantly dissipative *densities, which include Gaussian mixtures with common covariance like the example in Prop. 1, c.f. Gorham et al. (2019); Gorham and Mackey (2017); Hodgkinson et al. (2020). Prop. 1 does* not *contradict this result, as it considers a different regime where a* sequence *of target distributions is of interest.*

Fig. 1 provides numerical evidence for Prop. 1 by showing the rejection rate 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)) becomes indistinguishable from the prescribed level for $\Delta \geq 6$. The KSDAGG of Schrab et al. (2022), which uses an aggregated technique to avoid bandwidth selection via heuristics, behaved similarly. Notably, this issue persists even if the samples are drawn from *both* components but with an incorrect weight; see Fig. 3 in Appendix E.1. In contrast, our proposed test, called pKSD, achieved an almost perfect power.

4 Kernelized Stein discrepancy test with perturbation

Prop. 1 highlights the myopia of the KSD test: it can only detect "local" discrepancies in regions where we have observations. If the discrepancy is "global", such as a missing mode or incorrect mixing weight, it is necessary to turn this into a local discrepancy in order to improve the test power. We propose to do so by *perturbing both the alternative and the target distributions with a* Markov transition kernel $\mathcal{K}$ *(Robert and Casella, 2004, Chapter 6) and performing KSD test on the perturbed distributions*. Given a probability measure Q, the perturbed measure under $\mathcal{K}$ is $(\mathcal{K}Q)(\cdot) := \int_\mathcal{X} \mathcal{K}(x,\cdot)Q(dx)$. In our paper, $\mathcal{K}$ may also be an iterated composition of an underlying kernel, e.g. a Metropolis-Hastings kernel.

Perturbed kernel Stein discrepancy Given a Markov transition kernel $\mathcal{K}$, the *perturbed kernelized Stein discrepancy* (pKSD), $\mathbb{D}(Q, P; \mathcal{K})$, is defined as

$$\mathbb{D}(Q,P;\mathcal{K}) := \mathbb{D}(\mathcal{K}Q, \mathcal{K}P) = \sup\nolimits_{f\in\mathcal{F}^d} |\mathbb{E}_{x\sim\mathcal{K}Q}[\mathcal{A}_{\mathcal{K}P} f(x)]|, \tag{4}$$

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 (4) to exist. The next result (proved in Appendix B) shows that pKSD with an appropriately chosen $\mathcal{K}$ admits a closed-form solution.

Proposition 2 (pKSD close form)**.** *Assume $\mathcal{K}$ is a perturbation operator for which $\mathcal{K}P$ attains a (Lebesgue) density that is continuously differentiable. If $\mathbb{E}_{x\sim\mathcal{K}Q}[u_{\mathcal{K}P}(x,x)] < \infty$, then $\mathbb{D}(Q,P;\mathcal{K}) = \mathbb{E}_{x,x'\sim\mathcal{K}Q}[u_{\mathcal{K}P}(x,x')]$, where $u_{\mathcal{K}P}$ is the Stein kernel for $\mathcal{K}P$ as defined in Sec. 2.*

The choice of $\mathcal{K}$ is paramount. A desirable choice should ensure *(i)* the score function of $\mathcal{K}P$ is well defined and efficiently evaluatable, and *(ii)* the test can achieve a high power against alternatives with wrong mixing weights. To this end, we propose to use a transition kernel that is *P-invariant* and relies on a *jump proposal* to perturb alternatives. A transition kernel $\mathcal{K}$ is P-invariant if $\mathcal{K}P = P$. This means the score function $s_{\mathcal{K}p} = s_p$ is unchanged after perturbation, thus *(i)* is trivially satisfied. For *(ii)*, the "jump proposal" leverages local geometric information of the modes of p and proposes inter-modal jumps to create local discrepancy that KSD can detect. The resulting pKSD *no longer* separates probability measures; we will address this major limitation in the full version of the paper.

Choosing the transition kernel 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 randomly selecting two modes of p indexed by u_1, u_2, then mapping (x, u) to $(x', u') = h(x, u)$, where $u := (u_1, u_2)$, and h is a deterministic, invertible function that is differentiable with differentiable inverse. The proposed state $x' = x'(x, u)$ is hence deterministic given x and 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 $\alpha(x, x')$ is an accept-reject rule, $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise, $r(x) = 1 - \sum_{u \in \mathcal{U}} g(u) \alpha(x, x')$, and the sum is over pairs of distinct mode indices $\mathcal{U} := \{(i, j) : 1 \leq i \neq j \leq M\}$. Crucially, α is designed to satisfy the *detailed balance condition* to guarantee P-invariance. See Appendix C.2 for details.

Choosing the proposal h We use a proposal h that creates local discrepancy by making inter-modal jumps. Denote by $\mu_1, \ldots, \mu_M \in \mathbb{R}^d$ the mode vectors, and $A_1, \ldots, A_M \in \mathbb{R}^{d \times d}$ the *inverse* of the Hessian matrices of $-\log p$ at those points. Given $u = (u_1, u_2) \sim \text{Uniform}(\mathcal{U})$, the proposal is defined as $h(x, u) = (A_{u_2}^{1/2} A_{u_1}^{-1/2}(x - \theta\mu_{u_1}) + \theta\mu_{u_2}, u)$, where $\theta > 0$ is a fixed constant. Intuitively, when p is a mixture of elliptic distributions such as Gaussian or multivariate t-distributions, each A_m is the covariance matrix of a component, and h sends x from mode μ_{u_1} to the "corresponding" location in μ_{u_2}. The proposed method relies on estimates of the mode vectors and Hessians, and two hyperparameters (jump scale θ and number of transition steps T). Appendix C.3 to C.4 discuss how to estimate/tune them in practice, as well as limitations of the proposed method.

GOF Tests with pKSD The P-invariance implies that pKSD reduces to KSD under H_0. Therefore, the null distribution of the pKSD statistic remains identical to that of KSD, and the p-value can be estimated with the same bootstrap technique in KSD tests given sample $\{\tilde{x}_i\}_{i=1}^n$ from the perturbed distribution $\mathcal{K}Q$, which can be drawn by running 1-step transitions under $\mathcal{K}$ starting from $x_i \sim Q$. The complete algorithm of GOF testing with pKSD is given in Algorithm 1 in Appendix.

5 Experiments

We apply our method to hyperparameter selection for a MCMC sampler in a Bayesian inference task of sensors localisation (Tak et al., 2018). The goal is to approximate a posterior distribution on the locations of 4 sensors given noisy observations of pairwise distances. The full experimental setup is in Appendix E.2. Tak et al. (2018) approached this problem with a Metropolis algorithm called RAM, which is based on alternating repelling-attracting proposals and depends on hyperparameter $\sigma > 0$ (called *scale*). Tak et al. (2018) used $\sigma = 1.08$. We aim to test the estimation quality of RAM using different values of σ under a fixed computational budget. We see from Table 1 that, for $\sigma = 0.1$ and 1.08, KSD test and KSDAGG test only reject once or none in 10 repetitions. This is *inconsistent* with the posterior plot (Fig. 2) as some modes are clearly missing (before perturbation). On the other hand, pKSD rejects most tests by creating local discrepancy around the missing modes. In contrast, with $\sigma = 0.5$, there is no clear evidence that the posterior samples have missed any modes, and the samples before and after perturbation also look similar. The results match this observation, with at most one rejection for all three methods.

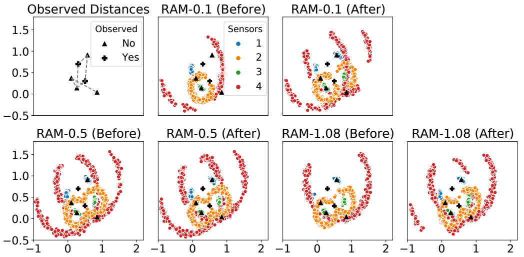


Figure 2: True and inferred locations of sensors before and after perturbation. Black triangles and crosses mark the location of the unobserved and observed sensors, respectively.

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

RAM scale	0.1	0.3	0.5	0.7	0.9	1.08	1.3
KSD	0	0	0	1	0	0	1
KSDAGG	1	0	0	2	1	1	4
pKSD (ours)	10	4	1	1	1	6	6

6 Conclusion

We formalise the failure of the KSD test when the target distribution has well-separated modes via a concrete example of bimodal Gaussian. To increase its power, we propose to perturb both the alternative and the target distributions using a Markov process, which creates local discrepancies that KSD can detect but no longer guarantee separation of probability measures. In the full version of the paper, we will extent pKSD to a statistical divergence that indeed separates measures, and include more experiments to examine the test power against different types of alternatives.

Acknowledgments and Disclosure of Funding

XL was supported by the Presidents PhD Scholarships of Imperial College London and the EPSRC StatML CDT programme EP/S023151/1. AD was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1 and EPSRC Grant EP/W006022/1, particularly the Ecosystems of Digital Twins theme within those grants & The Alan Turing Institute.

References

S. Ahn, Y. Chen, and M. Welling. Distributed and Adaptive Darting Monte Carlo through Regenerations. In C. M. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 108–116, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.

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

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

K. Chwialkowski, H. Strathmann, and A. Gretton. A Kernel Test of Goodness of Fit. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, New York, New York, USA, 20–22 Jun 2016. PMLR.

J. Gorham and L. Mackey. Measuring Sample Quality with Stein's Method. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

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

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

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

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

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

W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444.

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

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

A. Ihler, J. Fisher, R. Moses, and A. Willsky. 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.

W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton. A Linear-Time Kernel Goodness-of-Fit Test. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 261270, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

O. Johnson. *Information theory and the central limit theorem*. World Scientific, 2004.

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

Q. Liu, J. Lee, and M. Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284, New York, New York, USA, 20–22 Jun 2016. PMLR.

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

T. Matsubara, J. Knoblauch, F.-X. Briol, C. Oates, et al. Robust generalised Bayesian inference for intractable likelihoods. *arXiv preprint arXiv:2104.07359*, 2021.

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

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

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

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

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

A. Schrab, B. Guedj, and A. Gretton. KSD Aggregated Goodness-of-fit Test. *arXiv preprint arXiv:2202.00824*, 2022.

C. Stein. 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, pages 583–603. University of California Press, 1972.

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

L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of applied probability*, pages 1–9, 1998.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

L. Wenliang, D. J. Sutherland, H. Strathmann, and A. Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, pages 6737–6746. PMLR, 2019.

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

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

Appendix

A Proof of Proposition 1

Prop. 1 states that, when a sample completely misses one mode of a bimodal Gaussian target and the sample size n does not grow fast enough with the inter-modal distance $\|\Delta\|_2$, the distribution of the kernelized Stein discrepancy statistic will converge to the *null* distribution, even though the sample is clearly not representative for the target. To prove this result, we need the following lemma.

Lemma 3. *Under the same assumptions in Prop. 1, we have* $\mathbb{E}_{x\sim Q}[\|s_{p_\Delta}(x) - s_q(x)\|_2^2] = o\left(e^{-\|\Delta\|_2^2/32}\right)$.

The above lemma states that the expected square error between the score functions (i.e. the *Fisher divergence*) between Q and P_Δ decays with a rate at least exponentially fast in the inter-modal distance $\|\Delta\|_2$. We now use this lemma to prove Prop. 1; the proof of this lemma is provided later in this section.

Proof of Prop. 1. Fixing positive integer ν, we can write $n_\nu \hat{\mathbb{D}}_{P_{\Delta_\nu}} = n_\nu \hat{\mathbb{D}}_Q + n_\nu(\hat{\mathbb{D}}_{P_{\Delta_\nu}} - \hat{\mathbb{D}}_Q)$. Under the stated assumption of the kernel k and, one can check that $\mathbb{E}_{x,x'\sim Q}[u_Q(x,x')^2] < \infty$ when Q is the standard normal distribution, so we can apply Liu et al. (2016, Thm 4.1) to conclude that, as $\nu \to \infty$,

$$n_\nu \hat{\mathbb{D}}_Q \to_d \textstyle\sum_{j=1}^{\infty} c_j(z_j^2 - 1)\ , \tag{5}$$

where z_j, c_j are as defined in Prop. 1. If we could show that $n_\nu(\hat{\mathbb{D}}_{P_{\Delta_\nu}} - \hat{\mathbb{D}}_Q) \to 0$ in probability as $\nu \to \infty$, then the desired result would follow from Slutsky's Theorem (Casella and Berger, 2001) and (5).

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

$$\begin{aligned}
&\mathrm{Pr}_Q(n|\hat{\mathbb{D}}_{P_\Delta} - \hat{\mathbb{D}}_Q| \geq \epsilon) \\
&\leq \tfrac{n}{\epsilon}\mathbb{E}_{x_1,\ldots,x_n\sim Q}[|\hat{\mathbb{D}}_{P_\Delta} - \hat{\mathbb{D}}_Q|] \\
&= \tfrac{n}{\epsilon}\mathbb{E}_{x_1,\ldots,x_n\sim Q}\left|\tfrac{1}{n(n-1)}\textstyle\sum_{1\leq i\neq j\leq n} u_{P_\Delta}(x_i,x_j) - u_Q(x_i,x_j)\right| \\
&\leq \tfrac{n}{\epsilon}\tfrac{1}{n(n-1)}\textstyle\sum_{1\leq i\neq j\leq n}\mathbb{E}_{x_i,x_j\sim Q}|u_{P_\Delta}(x_i,x_j) - u_Q(x_i,x_j)| \\
&= \tfrac{n}{\epsilon}\mathbb{E}_{x,x'\sim Q}|u_{P_\Delta}(x,x') - u_Q(x,x')| \\
&\leq \tfrac{n}{\epsilon}\{\mathbb{E}_{x,x'\sim Q}|s_{p_\Delta}(x)^\intercal s_{p_\Delta}(x') - s_q(x)^\intercal s_q(x')||k(x,x')| \\
&\quad + \mathbb{E}_{x,x'\sim Q}|(s_{p_\Delta}(x) - s_q(x))^\intercal \nabla_{x'}k(x,x')| \\
&\quad + \mathbb{E}_{x,x'\sim Q}|(s_{p_\Delta}(x') - s_q(x'))^\intercal \nabla_x k(x,x')|\} \\
&\leq \tfrac{n}{\epsilon}\Big\{\left(\mathbb{E}_{x,x'\sim Q}[(s_{p_\Delta}(x)^\intercal s_{p_\Delta}(x') - s_q(x)^\intercal s_q(x'))^2]\right)^{1/2}\left(\mathbb{E}_{x,x'\sim Q}[k(x,x')^2]\right)^{1/2} \\
&\quad + \left(\mathbb{E}_{x\sim Q}[\|s_{p_\Delta}(x) - s_q(x)\|_2^2]\right)^{1/2}\left(\mathbb{E}_{x,x'\sim Q}[\|\nabla_{x'}k(x,x')\|_2^2]\right)^{1/2} \\
&\quad + \left(\mathbb{E}_{x\sim Q}[\|s_{p_\Delta}(x) - s_q(x)\|_2^2]\right)^{1/2}\left(\mathbb{E}_{x,x'\sim Q}[\|\nabla_x k(x,x')\|_2^2]\right)^{1/2}\Big\}.
\end{aligned} \tag{6}$$

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

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

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,

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

where the first line follows from the Cauchy-Schwarz inequality, and the last line holds because

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

By Lemma 3 we know that $\mathbb{E}_{x\sim Q}[\|s_{p_\Delta}(x) - s_q(x)\|_2^2] = o\left(e^{-\|\Delta\|_2^2/32}\right)$. Also, direct computation gives that $\mathbb{E}_{x\sim Q}[\|s_q(x)\|_2^2] = 1$. A similar argument shows that $T_2 = o\left(e^{-\|\Delta\|_2^2/32}\right)$ for some constant a_2. We therefore conclude that the first term of (6) is $o\left(e^{-\|\Delta\|_2^2/64}\right)$.

With a similar argument and by the boundedness assumptions on the kernel k, we conclude that the second and third terms of (6) are both $o(e^{-\|\Delta\|_2^2/64})$. Combining all the above, we conclude that there exists a universal constant a_3 such that, for sufficiently large $\|\Delta\|_2$,

$$\Pr_Q(n|\hat{\mathbb{D}}_{P_\Delta} - \hat{\mathbb{D}}_Q| \ge \epsilon) \le \tfrac{n}{\epsilon} a_3 e^{-\frac{\|\Delta\|_2^2}{64}},$$

which goes to 0 if $n = o\left(e^{\|\Delta\|_2^2/64}\right)$. This shows that $n_\nu(\hat{\mathbb{D}}_{P_{\Delta_\nu}} - \hat{\mathbb{D}}_Q) \to 0$ in probability as $\nu \to \infty$, so the desired result follows. □

Proof of Lemma 3. For any $\delta > 0$, define $B_\delta := \{x \in \mathbb{R}^d : \|x\|_2 \le \delta\}$. We have the following decomposition

$$\begin{aligned}&\mathbb{E}_{x\in Q}[\|s_{p_\Delta}(x) - s_q(x)\|_2^2]\\ &\quad = \mathbb{E}_{x\sim Q}[\delta_x(B_\delta)\|s_{p_\Delta}(x) - s_q(x)\|_2^2] + \mathbb{E}_{x\sim Q}[\delta_x(\mathbb{R}^d\backslash B_\delta)\|s_{p_\Delta}(x) - s_q(x)\|_2^2] \qquad (7)\end{aligned}$$

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

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

and

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

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

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

and the first term of (7) can be bounded as

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

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

To bound the second term of (7), we note that for $x \in \mathbb{R}^d\backslash B_\delta$,

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

Therefore,

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

by the tail probability of Gaussian distributions (e.g., Wainwright (2019, Prop. 2.5)). Combining these results we have

$$\begin{aligned}\mathbb{E}_{x\in Q}[\|s_{p_\Delta}(x) - s_q(x)\|_2^2] &\le \tfrac{(1-\pi)^2\|\Delta\|_2^2}{\pi^2\exp(-2\Delta\delta+\Delta^2)} + \|\Delta\|_2^2\exp\left(-\tfrac{\delta^2}{2}\right)\\ &= (1-\pi)^2\pi^{-2}\|\Delta\|_2^2\exp\left(-\tfrac{1}{3}\|\Delta\|_2^2\right) + \|\Delta\|_2^2\exp\left(-\tfrac{1}{18}\|\Delta\|_2^2\right),\end{aligned}$$

where the last line follows by choosing $\delta = \Delta/3$. Noting the RHS of the last inequality is $o\left(-\frac{1}{32}\|\Delta\|_2^2\right)$ completes the proof. □

Algorithm 1 Goodness-of-Fit Test with pKSD (the proposed method).

```
Require: Training set $\{x_{\text{train},j}\}_{j=1}^{n_{\text{train}}}$, test set $\{x_{\text{test},j}\}_{j=1}^{n_{\text{test}}}$, target $P$, number of transition steps $T$,
  candidate jump scales $\{\theta_1, \ldots, \theta_L\}$, sizes of training and testing sets $n_{\text{train}}$ and $n_{\text{test}}$, resp.
Output Whether the null hypothesis is rejected.
  Estimate the mode $\{\mu_1, \mu_2, \ldots, \mu_M\}$ and Hessians $\{A_1, \ldots, A_M\}$ using Algorithm 2.
  Use $\{x_{\text{train},j}\}_{j=1}^{n_{\text{train}}}$ to find $\theta^* \in \arg\max_{\theta \in \{\theta_1, \ldots, \theta_L\}} \hat{\mathbb{D}}_{P,\mathcal{K}_{\theta}^{(T)}} / \hat{\sigma}_u$.
  Perturb $\{x_{\text{test},j}\}_{j=1}^{n_{\text{test}}}$ with the selected kernel $\mathcal{K}_{\theta^*}^{(T)}$, and use (8) to compute test statistic $\hat{\mathbb{D}}_{P,\mathcal{K}_{\theta^*}^{(T)}}$.
  Compute bootstrap samples using (9) and find the $(1-\alpha)$-quantile $\hat{\gamma}_{1-\alpha}$.
  Reject the null hypothesis if $\hat{\mathbb{D}}_{P,\mathcal{K}_{\theta^*}^{(T)}} \geq \hat{\gamma}_{1-\alpha}$.
```

Algorithm 2 Finding mode vectors (Pompe et al., 2020, Algorithm 3)

```
Require: Initial points $s_1, \ldots, s_{M_0}$, small positive value $\beta$.
Output Approximates for mode vectors $\{\mu_1, \ldots, \mu_M\}$.
  Run BFGS to minimise $-\log p(x)$ with initial points $s_1, \ldots, s_{M_0}$.
  Denote the estimated 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 $\min_{j \in \{1,\ldots,M\}} \frac{1}{2}((\mu_j - m_i)^\intercal A_{\mu_j}(\mu_j - m_i) + (\mu_j - m_i)^\intercal A_i(\mu_j - m_i)) < \beta$ then
          $k := \arg\min_{j \in \{1,\ldots,M\}} \frac{1}{2}((\mu_j - m_i)^\intercal A_{\mu_j}(\mu_j - m_i) + (\mu_j - m_i)^\intercal A_i(\mu_j - m_i))$.
          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
```

B Proof of Proposition 2

Under the stated assumptions, this follows directly from Chwialkowski et al. (2016, Thm. 2.1) applied to $\mathcal{K}Q$ and $\mathcal{K}P$.

C Details of Proposed Method

C.1 Estimating pKSD

Proposition 2 is a general result and applies to any perturbation operator $\mathcal{K}$ for which the density of $\mathcal{K}P$ is well-defined. In particular, it applies to pKSD where $\mathcal{K}$ is a P-invariant Markov transition kernel (in which case $\mathcal{K}P = P$).

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. Prop.2 then suggests estimating $\mathbb{D}(Q, P; \mathcal{K})$ by the following U-statistic:

$$\hat{\mathbb{D}}_{P,\mathcal{K}} := \tfrac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} u_{\mathcal{K}P}(\tilde{x}_i, \tilde{x}_j). \tag{8}$$

Estimating the p-value In the standard KSD tests, the asymptotic distribution of the test statistic $\hat{\mathbb{D}}_P$ under H_0 has no closed form. Liu et al. (2016) proposed to approximate it using a bootstrap technique (Huskova and Janssen, 1993) via the bootstrap samples

$$\hat{\mathbb{D}}_P^b := \tfrac{1}{n^2} \sum_{1 \leq i \neq j \leq n} \left(w_i^b - 1\right)\left(w_j^b - 1\right) u_P(x_i, x_j), \tag{9}$$

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. Since pKSD reduces to KSD under H_0 due to P-invariance, the null distribution of pKSD statistic can be approximated using the same bootstrap technique, with $\{x_i\}_{i=1}^n$ replaced with the perturbed sample $\{\tilde{x}_i\}_{i=1}^n$ (see Section 4).

C.2 Choosing the transition kernel

Denoting by $\mathcal{B}(\mathcal{X})$ the Borel σ-algebra on $\mathcal{X}$, 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. Given a probability measure Q, the perturbed measure is $(\mathcal{K}Q)(A) \coloneqq \int_{\mathcal{X}} \mathcal{K}(x, A)Q(dx)$, for measurable set A.

As discussed in Section 4, we choose a transition kernel of the form

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

where $x' = x'(x, u)$ 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 $r(x) = 1 - \sum_{u \in \mathcal{U}} g(u)\alpha(x, x')$. When g is defined on a continuous space, the same argument follows by replacing the summation with integration. The accept-reject rule α is designed to satisfy the *detailed balance condition*:

$$\textstyle\int_{x \in A} \sum_{u \in \mathcal{U}} \delta_{x'}(B) p(x) g(u) \alpha(x, x') dx = \int_{x' \in B} \sum_{u' \in \mathcal{U}} \delta_x(A) p(x') g'(u') \alpha(x', x) dx', \tag{10}$$

for all $A, B \in \mathcal{B}(\mathcal{X})$. One valid choice of $\alpha(x, x')$ for which (10) holds is

$$\alpha(x, x') = \min\left(1, \tfrac{p(x')g'(u')}{p(x)g(u)} \left|\tfrac{\partial(x', u')}{\partial(x, u)}\right|\right), \tag{11}$$

where $\partial(x', u')/\partial(x, u)$ denotes the Jacobian of the transformation from (x, u) to (x', u'). See Appendix C.5 for the proof. Accept-reject rules of the form (11) are also used in Reversible-Jump MCMC (Green, 1995; Green and Hastie, 2009) and generalise the well-known MH rule, for which the determinant of the Jacobian is 1. A similar trick can be used to generalise other accept-reject rules such as the *Barker's rule* (Peskun, 1973; Tierney, 1998; Livingstone and Zanella, 2021).

C.3 Choosing the proposal

In our proposal $h(x, u) = (A_{u_2}^{1/2} A_{u_1}^{-1/2}(x - \theta\mu_{u_1}) + \theta\mu_{u_2}, u)$, two modes (indexed by $u = (u_1, u_2)$) are chosen randomly for a given current state, so a proposed state can potentially lie in a low-density region. This can hence lead to a low acceptance probability. A similar proposal considered in Pompe et al. (2020), called *deterministic jumps*, addresses this problem 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 it is guaranteed to propose a new state belonging to a different mode. However, the same trick cannot be used in our case because the augmented density no longer has a well-defined score function. In our experiment, we find that our jump proposal can achieve a significant increase in power against mixtures of elliptical distributions that disagree in the weights.

C.4 Further Details

Tuning the jump scale Fixing $T \geq 1$, we can follow the same argument in Jitkrittum et al. (2017, Prop. 4) to approximate the test power with $\hat{\mathbb{D}}_{P,\mathcal{K}_\theta^{(T)}}/\hat{\sigma}_u$, where $\hat{\sigma}_u^2 \coloneqq \frac{4}{n^3}\sum_{i=1}^n \left(\sum_{j=1}^n u_P(\tilde{x}_i, \tilde{x}_j)\right)^2 - \frac{4}{n^4}\left(\sum_{i,j=1}^n u_P(\tilde{x}_i, \tilde{x}_j)\right)^2$ with $\tilde{x}_i \sim \mathcal{K}^{(T)}Q$, which is an estimate of the asymptotic standard deviation σ_u (see also Schrab et al., 2022, Eq. 8). Since the objective is not differentiable with respect to θ, we propose to choose θ from a grid of values $\{\theta_l\}_{l=1}^L$ near 1, a heuristic we find to work well in our experiments. The objective is hence $\max_{\theta \in \{\theta_1, \ldots, \theta_L\}} \hat{\mathbb{D}}_{P,\mathcal{K}_\theta^{(T)}}/\hat{\sigma}_u$.

Estimating the mode vectors and Hessians To estimate the mode locations and Hessians at those points, we follow Pompe et al. (2020) to minimise $-\log p$ by running in parallel a sequence of optimisers initiated at different starting points. The optimiser used is BFGS (Nocedal and Wright, 2006), which returns both the local optima and approximated Hessians at those points. The optima are then merged if their Mahalanobis distance weighted by the approximated Hessian is smaller than a pre-specified threshold. In our experiments, we initialise the optimisers from a set of size n_{train}, half of which is drawn randomly from the training set and the other half sampled uniformly from a hyper cube $[L_1, U_1] \times \cdots \times [L_d, U_d]$. The full procedure is described in Appendix D.

Choice of number of transitions The number of transitions T affects how many perturbed samples are accepted across all transition steps, thus impacting the performance of the pKSD test. Intuitively, T should be set to a large value when the acceptance rate is low, which would happen if the estimates of the modes and local Hessians of the target distribution are inaccurate, or the target distribution cannot be approximated by a mixture of elliptic distributions.

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

One concern with *ii)* is that the candidate distribution might converge to the limiting distribution (which, for our choice of transition kernel, is the target distribution P) as $T \to \infty$, so a large T could drive the perturbed distribution close to P, making it harder to distinguish. This is however not a major concern, as the target distribution P is not the only limiting distribution of the transition kernel due to non-irreducibility, so the data distribution Q will not necessarily converge to P as $T \to \infty$. Although this does not offer a theoretical guarantee, some empirical evidence is provided by the experiments we conducted.

Limitations The jump proposal of the transition kernel used in pKSD is constructed specifically for targets that are mixtures of elliptic distributions and relies on estimates of the location and local geometry of modes, so it may suffer from a *low acceptance rate* for targets with more complicated geometrical structure. One future direction is to study alternative proposals that relaxes these assumptions. Another limitation is that pKSD is not a valid discrepancy, as $\mathbb{D}(Q, P; \mathcal{K}) = 0 \not\Longrightarrow Q = P$. Understanding in which distribution class pKSD does have this guarantee is another worthwhile research question.

C.5 A sufficient condition

We first give a sufficient condition for the detailed balance equation (10). For simplicity, we write $x' = x'(x, u)$ and $x = x(x', u')$, so that the dependence of x' on u and of x on x' is implicit.

Proposition 4. *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*

$$\mathcal{K}(x, A) = \textstyle\sum_{u \in \mathcal{U}} \delta_{x'}(A) g(u) \alpha(x, x') + \delta_x(A) r(x), \tag{12}$$

where $x' = x(x, u)$, $\delta_x(A) = 1$ 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 (13) if

$$p(x)g(u)\alpha(x, x') = p(x')g(u')\alpha(x', x) \left| \tfrac{\partial(x', u)}{\partial(x, u)} \right|, \tag{13}$$

Proof. The proof follows largely from Green and Hastie (2009, Sec. 1.2.1), which shows the claim when the density g is defined on a continuous space. By the invertibility of the transformation h, a change-of-variable formula can be applied to the right-hand-side of (10) to yield

$$\textstyle\int_{x \in A} \sum_{u \in \mathcal{U}} \delta_{x'}(B) p(x) g(u) \alpha(x, x') dx = \int_{x' \in B} \sum_{u' \in \mathcal{U}} \delta_x(A) p(x') g'(u') \alpha(x', x) \left| \tfrac{\partial(x', u')}{\partial(x, u)} \right| dx'.$$

Now define $\mathcal{U}_B := \{u : x' \in B\}$ and similarly for $\mathcal{U}_A$. We have that $(x, u) \in A \times \mathcal{U}_B \implies (x', u') = h(x, u) \in B \times \mathcal{U}_A$, and, similarly, $(x', u') \in B \times \mathcal{U}_A \implies (x, u) \in A \times \mathcal{U}_B$. We therefore conclude that a sufficient condition is

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

thus showing the claim. □

In particular, it follows that the detailed balance condition holds with (11) by verifying that it satisfies (13). This can be viewed as a generalisation of the Metropolis-Hastings (MH) rule $\alpha(x, x') = \min\left(1, \frac{p(x')g'(u')}{p(x)g(u)}\right)$.

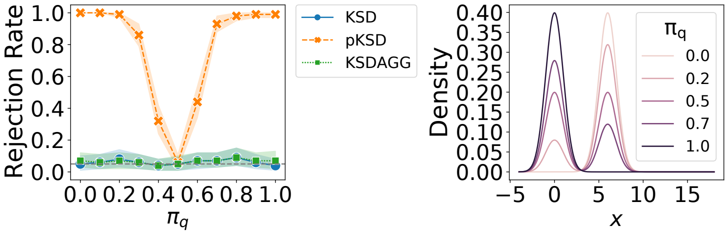


Figure 3: One-dimensional bimodal Gaussian example with $\Delta = 6$ and $\pi_p = 0.5$.

D Finding mode vectors via optimisation and merging

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

Mode merging Starting the optimisation procedure from each initial point will lead to an end point close to the true local maxima, but are numerically different from each other. Pompe et al. (2020) proposed to merge two optima points m_i and m_j if the Mahalanobis distance weighted by the averaged Hessians at those points is below a given threshold β. The full procedure is given in Algorithm 2 for completeness.

Choosing starting points The BFGS can be initiated from either a random sample uniformly drawn on a product of intervals $[L_1, U_1] \times \cdots \times [L_d, U_d]$ in $\mathcal{X}$, or simply the training set $\{x_{\text{train},i}\}_{i=1}^{n_{\text{train}}}$. 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 modes of Q and P overlaps. To combine the best of the two worlds with the same computational budget, we initialise the points by a set of size n_{train}, half of which is drawn randomly drawn from the training set and the other half initialised uniformly from $[L_1, U_1] \times \cdots \times [L_d, U_d]$.

E Experimental details

E.1 Multivariate Gaussian mixture: supplementary plots

In correct mixing ratios in 1 dimension We repeat the same experiment in Figure 1, where, instead of varying Δ, we fix $\Delta = 6$ and draw samples from the same mixture but with different mixing weights π_q. The mixing weight for the target distribution is kept at $\pi_p = 0.5$ as before. We observe that the problem of low power of KSD test and KSDAGGtest persists even if the samples are drawn from both components but with a different weight; see Figure 3 in Sec. 5.

Level and power experiments in 50 dimensions We include supplementary experiments where we study the power and level of pKSD. The target distribution is the same in the multivariate Gaussian mixture example in Sec. 5 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. 4. We can see that under the null hypothesis, all tests have the prescribed test level $\alpha = 0.05$. When samples completely miss one mode, pKSD achieves a significantly higher power for all sample sizes, whereas the power of KSD and KSDAGG remains close to the level.

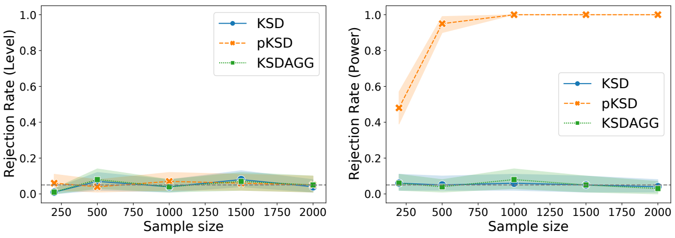


Figure 4: Level (left) and power (right) experiments with the multivariate Gaussian mixture example.

E.2 Sensor network localisation

The Bayesian inference task for sensor network localisation follows the same setup in Tak et al. (2018). This is a modified version of the example from Ihler et al. (2005), which has been used as a benchmark for MCMC samplers designed for multimodal targets (Pompe et al., 2020; Ahn et al., 2013; Lan et al., 2014). Here, six sensors $x_1, \dots, x_6$ 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 $y_{ij} \sim \mathcal{N}(\|x_i - x_j\|, 0.02^2)$. As in Tak et al. (2018), we use a diffuse bivariate Gaussian prior distribution $\mathcal{N}(0, 10^2 I_2)$ for each x_i. 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

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

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

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