Targeted Separation and Convergence with Kernel Discrepancies

Alessandro Barp ab2286@cam.ac.uk
University of Cambridge & The Alan Turing Institute, GB

Carl-Johann Simon-Gabriel cjsg@amazon.com
Amazon Web Services, Tübingen, Germany

Mark Girolami mag92@cam.ac.uk
University of Cambridge & The Alan Turing Institute, GB

Lester Mackey lmackey@microsoft.com
Microsoft Research, New England, US

Abstract

Kernel Stein discrepancies (KSDs) are maximum mean discrepancies (MMDs) that leverage the score information of distributions, and have grown central to a wide range of applications. In most settings, these MMDs are required to (i) separate a target P from other probability measures or even (ii) control weak convergence to P. In this article we derive new sufficient and necessary conditions that substantially broaden the known conditions for KSD separation and convergence control, and develop the first KSDs known to metrize weak convergence to P. Along the way, we highlight the implications of our results for hypothesis testing, measuring and improving sample quality, and sampling with Stein variational gradient descent.

1 Introduction

The Langevin kernel Stein discrepancy (KSD) is a maximum mean discrepancy (MMD) tailored to the target distribution P that relies on its score $\boldsymbol{s}_{\mathrm{p}} \equiv \partial \log \mathrm{p}$ to measure the integration error between P and alternative distributions. Kernel-based discrepancy measures are widely used for hypothesis testing [16, 21, 9], sampler selection and tuning [14], parameter estimation [4, 2, 11], generalized Bayesian inference [8, 24, 23, 10], discrete approximation and numerical integration [7, 6, 3], control variate design [26, 25], compression [27], and bias correction [20, 17, 27].

Each MMD uses a kernel function to measure the integration error between a pair of probability measures Q and P, and, in each setting above, their successful application relies on either P-*separation*, that is $\mathrm{MMD}(\mathrm{Q}, \mathrm{P}) > 0$ whenever $\mathrm{Q} \neq \mathrm{P}$, or P-*convergence control*, namely $\mathrm{MMD}(\mathrm{Q}_n, \mathrm{P}) \to 0$ implies $\mathrm{Q}_n \to \mathrm{P}$ weakly. Unfortunately, these properties have so far only been established under overly restrictive assumptions, e.g., for Q with continuously differentiable log densities [9, 21, 2], for P with strongly log concave tails and Lipschitz log density gradients [14], or for bounded kernels which rarely occur in the context of score-based MMD kernels [32, 31, 30, 29].

In this work, by fixing P as the target measure and allowing Q to vary, we establish new broadly applicable conditions for P-separation in Section 2, and P-convergence control in Section 3. We begin by extending the usual notions of maximum mean discrepancy and kernel Stein discrepancy to

NeurIPS 2022 Workshop on Score-Based Methods.

accommodate both arbitrary probability measures Q and unbounded kernels. Throughout we use the standard notation summarized in Appendix A.1.

Kernel Stein discrepancies Consider a (reproducing) kernel k on $\mathcal{X}$ with reproducing kernel Hilbert space $\mathcal{H}_k$ [1, 28]. We will employ the following generalized definition of MMD, for any target measure in the set of *embeddable probability measures* $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_k} \equiv \{\mathrm{Q} \in \mathcal{P} : \mathcal{H}_k \subset \mathrm{L}^1(\mathrm{Q})\}$:

$$\mathrm{MMD}_k(\mathrm{Q}, \mathrm{P}) \equiv \sup_{h \in \mathcal{B}_k : \max(h,0) \in \mathrm{L}^1(\mathrm{Q})} |\mathrm{Q}h - \mathrm{P}h| \,. \tag{1}$$

Indeed, traditionally MMD is defined as the worst-case integration error across test functions in the *whole* RKHS unit norm ball $\mathcal{B}_k$ [16]. However, the expression $\mathrm{Q}h - \mathrm{P}h$ is not well defined when either (i) both $\mathrm{Q}h$ and $\mathrm{P}h$ are infinite or (ii) h is not integrable under Q. Unfortunately, both of these cases can occur when k is unbounded. Similarly we extend the integral probability metric definition of KSD associated to the score-based Langevin Stein operator $\mathcal{S}_{\mathrm{p}}$, which corresponds to the MMD of a *Stein kernel* k_{p} under no additional assumption by Theorem A.2, see Appendix A.2 for details.

Definition 1.1 (Kernel Stein discrepancy (KSD)). *For $\mathcal{X} = \mathbb{R}^d$, a target $\mathrm{P} \in \mathcal{P}$ with density $\mathrm{p} > 0$, and matrix-valued base kernel K for which $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K)$ exists. When $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K) \subset \mathrm{L}^1(\mathrm{P})$ and $\mathrm{P}(\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K)) = \{0\}$, we define the* kernel Stein discrepancy $\mathrm{KSD}_{K,\mathrm{P}} : \mathcal{P} \to [0, \infty]$ *by*

$$\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) \equiv \sup_{v \in \mathcal{B}_K : \max(\mathcal{S}_{\mathrm{p}}(v),0) \in \mathrm{L}^1(\mathrm{Q})} |\mathrm{Q}\mathcal{S}_{\mathrm{p}}(v)| \,. \tag{2}$$

2 Conditions for separating measures

Our first goal is to identify when an MMD distinguishes P from other measures. Given a set of probability measures $\mathcal{M} \subset \mathcal{P}$, we will say that k *separates* P *from* $\mathcal{M}$ when $\mathrm{MMD}_k(\mathrm{Q}, \mathrm{P}) = 0$ implies that $\mathrm{Q} = \mathrm{P}$ for any $\mathrm{Q} \in \mathcal{M}$. When k separates P from all probability measures $\mathcal{P}$ we say simply that k is P-*separating*. We will first discuss restricted P-separation from distinguished subsets of measures $\mathcal{M} \neq \mathrm{P}$ in Sections 2.1 and 2.2, and then turn to general P-separation in Section 2.3.

2.1 Bochner P-separation with MMDs

We first characterize the kernels that separate P from the set $\mathcal{P}_{\sqrt{k}}$ of Bochner embeddable measures.

Theorem 2.1 (Bochner P-separation with MMDs). *Let k be a continuous kernel over a Radon space $\mathcal{X}$. Then k separates $\mathrm{P} \in \mathcal{P}_{\sqrt{k}}$ from $\mathcal{P}_{\sqrt{k}}$ iff, for any sequence $(\mathrm{Q}_n)_n \subset \mathcal{P}_{\sqrt{k}}$,*

$$\mathrm{Q}_n h \to \mathrm{P}h \quad \forall h \in \mathscr{C}_{\sqrt{k}} \qquad \Longleftrightarrow \qquad \begin{cases} \text{(a)} & \mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0 \\ \text{(b)} & (\mathrm{Q}_n)_n \textit{ is tight} \\ \text{(c)} & (\mathrm{Q}_n)_n \textit{ uniformly integrates } \sqrt{k}. \end{cases} \tag{3}$$

Theorem 2.1 exposes an important relationship between our two goals of separation and convergence control. In particular, when k is bounded, it implies that separating P from all probability measures is equivalent to *controlling tight* P-*convergence*, i.e., having $\mathrm{Q}_n \to \mathrm{P}$ weakly whenever $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0$ and $(\mathrm{Q}_n)_n$ is tight.

2.2 Score P-separation with KSDs

The standard practice in the KSD literature is to identify easily-verified properties of the base kernel K, target P, and alternative measure Q that ensure separation. One class of KSD separation conditions applies to measures that finitely integrate the score $\boldsymbol{s}_{\mathrm{p}}$ but additionally requires Q to have a continuously differentiable log-density [9, 2]. Our next result, removes the extraneous continuity conditions and extends P-separation to *all* measures $\mathrm{Q} \in \mathcal{P}_{\boldsymbol{s}_{\mathrm{p}}}$ under a standard separating assumption on the base kernel, $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$-*characteristicness* [30], that covers all of the translation-invariant base kernels commonly used with KSDs including Gaussian, inverse multiquadric (IMQ), log inverse, sech, Matérn, B-spline, and Wendland's compactly supported kernels.

Theorem 2.2 (Score P-separation with KSDs). *Suppose a matrix-valued kernel K with $\mathcal{H}_K \subset \mathscr{C}^1_b(\mathbb{R}^d)$ is $\mathscr{D}^1_{L^1}(\mathbb{R}^d)$-characteristic. If $\mathrm{P} \in \mathcal{P}_{K,0}$,[1] then k_{p} separates P from $\mathcal{P}_{\boldsymbol{s}_{\mathrm{p}}}$.*

[1] Here $\mathcal{P}_{K,0} \equiv \{\mathrm{Q} \in \mathcal{P} \text{ with } \mathrm{q} > 0 \colon \mathcal{S}_{\mathrm{q}}(\mathcal{H}_K) \text{ exists}, \mathcal{S}_{\mathrm{q}}(\mathcal{H}_K) \subset \mathrm{L}^1(\mathrm{Q}), \text{and } \mathrm{Q}(\mathcal{S}_{\mathrm{q}}(\mathcal{H}_K)) = \{0\}\}$.

Goodness-of-fit testing In goodness-of-fit (GOF) testing, one uses a sequence of datapoints $X_1, \ldots, X_n$ generated from a Markov chain to test whether the chain's stationary distribution Q coincides with a target distribution P. KSDs with $\mathscr{D}^1_{L^1}$-characteristic translation-invariant base kernels are commonly used as GOF test statistics, and such tests are known to consistently reject $\mathrm{Q} \neq \mathrm{P}$ whenever $\mathrm{KSD}(\mathrm{Q}, \mathrm{P}) > 0$ [9, 21, 14]. However, prior to this work, the separating condition $\mathrm{KSD}(\mathrm{Q}, \mathrm{P}) > 0$ had only been established for a restricted class of alternatives (continuous $\mathrm{Q} \in \mathcal{P}_{\sqrt{k_{\mathrm{P}}}}$ with differentiable log densities satisfying $\mathrm{Q}(\|\boldsymbol{s}_{\mathrm{p}} - \boldsymbol{s}_{\mathrm{q}}\|) < \infty$, [2, Prop. 1]) or a restricted class of targets (P with Lipschitz $\boldsymbol{s}_{\mathrm{p}}$ and strongly log concave tails, [14, Thm. 7]). The former restriction excludes discrete and discontinuous Q, as well as Q with tails heavier than P or non-differentiable densities. Meanwhile, the latter restriction excludes P with tails heavier than or lighter than a Gaussian. Theorem 2.2 in the present work ensures that $\mathrm{KSD}(\mathrm{Q}, \mathrm{P}) > 0$ for *any* $\mathrm{P} \in \mathcal{P}_{\kappa,0}$ and $\mathrm{Q} \in \mathcal{P}_{\boldsymbol{s}_{\mathrm{p}}}$. In particular, this accommodates discontinuous or non-smooth Q and all targets P for which the KSD Eq. (2) is defined. Moreover, Theorem 2.2 holds for all $\mathscr{D}^1_{L^1}$-characteristic kernels, a strict superset of the $\mathscr{C}^1_0$-universal kernels [5, Def. 4.1] assumed in prior work.

2.3 General P-separation

The results in the preceding sections only yield general P-separation when applied to bounded kernels, and indeed this has been the standard in much of the MMD literature [32, 31, 30, 29]. To accommodate the unbounded Stein kernels that often arise in KSDs, our next definition and result provide a new, convenient means to check that *unbounded* kernels separate P from $\mathcal{P}$.

Definition 2.3 (Bounded P-separating property)**.** *We say a set of functions $\mathcal{F}$ is* bounded P-separating *if $\mathrm{L}^\infty \cap \mathcal{F}$ is P-separating, i.e., if $\mathrm{Q} \in \mathcal{P}$ and $\mathrm{Q}h = \mathrm{P}h$ for all $h \in \mathrm{L}^\infty \cap \mathcal{F}$ then $\mathrm{Q} = \mathrm{P}$.*

Theorem 2.4 (Controlling tight convergence with bounded separation)**.** *If $\mathcal{H}_k$ is bounded P-separating, then k is P-separating and controls tight P-convergence.*

Our next result applies this strategy and shows that all of the translation-invariant base kernels commonly used with KSDs, including Gaussian, IMQ, log inverse, sech, Matérn, B-spline, and Wendland's compactly supported kernels, contain a sub-RKHS of bounded (in fact rapidly decaying) functions that is P-separating.

Theorem 2.5 (Controlling tight convergence with KSDs)**.** *For any translation invariant k with a spectral density bounded away from zero on compact sets, and s.t., $\mathcal{H}_k \subset \mathscr{C}^1$. Define the tilted kernel $k_a(x, y) = a(x)k(x, y)a(y)$ for each strictly positive $a \in \mathscr{C}^1$.*

1. *If $\mathrm{P} \in \mathcal{P}_{k,0}$ and $\|\boldsymbol{s}_{\mathrm{p}}\|$ has at most root exponential growth,*[2] *then the Stein kernel induced by k is bounded P-separating and controls tight P-convergence.*
2. *Moreover, if $\mathrm{P} \in \mathcal{P}_{k_a,0}$ and a, ∂a, and $a\|\boldsymbol{s}_{\mathrm{p}}\|$ have at most root exponential growth, then the Stein kernel induced by k_a is bounded P-separating and controls tight P-convergence.*

Goodness-of-fit testing, continued Theorem 2.5 extends the reach of KSD GOF testing by guaranteeing $\mathrm{KSD}(\mathrm{Q}, \mathrm{P}) > 0$ for *all* alternatives Q whenever $\|\boldsymbol{s}_{\mathrm{p}}\|$ has at most root exponential growth. Since the Stein kernels of Theorem 2.5 are also bounded P-separating, the same consistency guarantees immediately extend to the computationally efficient stochastic KSDs of [15].

3 Conditions for convergence control

Having derived sufficient conditions on the RKHS to separate measures and control tight convergence, we now present sufficient conditions to ensure that an MMD controls weak convergence to P. Hereafter, we will say that k *controls weak convergence to* P or *controls* P*-convergence* whenever $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0$ implies $\mathrm{Q}_n \to \mathrm{P}$ weakly. Moreover, we will say that k *enforces tightness* whenever $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0$ implies that $(\mathrm{Q}_n)_n$ is tight. Enforcing tightness is central to our developments as, if k controls tight weak convergence to P and enforces tightness, then it also controls weak convergence to P. We begin by introducing a new sufficient condition to ensure that MMDs enforce tightness.

[2] A function a has *at most root exponential growth* if $a(x) = O(\exp(c\sum_{i=1}^d \sqrt{|x^i|}))$ for some $c > 0$.

Theorem 3.1 (P-dominating indicators enforces tightness)**.** *Consider a set of functions $\mathcal{F} \subset \mathrm{L}^1(\mathrm{P})$. We say that $\mathcal{F}$* P-dominates indicators *if, for each $\epsilon > 0$, there exists a compact set $S \subset \mathcal{X}$ and a function $h \in \mathcal{F}$ that satisfy*

$$h - \mathrm{P}h \geq \mathbb{I}\left[S^c\right] - \epsilon. \tag{4}$$

If $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_k}$ and $\mathcal{H}_k$ P-dominates indicators then $(\mathrm{Q}_n)_n$ is tight whenever $\mathrm{MMD}_k(\mathrm{Q}_n, \mathrm{P}) \to 0$.

Corollary 3.2 (Controlling P-convergence with KSDs)**.** *Under the conditions of Theorem 2.2 or 2.5, if $\mathcal{H}_{k_\mathrm{p}}$ P-dominates indicators, then k_p controls P-convergence.*

Note prior work relied on a stronger *coercive function* condition to establish that KSDs enforce tightness [14, 6, 17]. As a first application of Corollary 3.2, we show that KSDs with IMQ base kernels enforce tightness and control convergence whenever the dissipativity rate of the target dominates the decay rate of the kernel.

Theorem 3.3 (IMQ KSDs control P-convergence)**.** *Consider a target measure $\mathrm{P} \in \mathcal{P}$ with score $\boldsymbol{s}_\mathrm{p} \in \mathscr{C}(\mathbb{R}^d) \cap \mathrm{L}^1(\mathrm{P})$. If, for some dissipativity rate $u > 1/2$ and $s, s_1, s_2 > 0$, P satisfies the* generalized dissipativity *condition*

$$-\langle \boldsymbol{s}_\mathrm{p}(x), x\rangle - s\left\|\boldsymbol{s}_\mathrm{p}(x)\right\|_1 \geq s_1\left\|x\right\|^{2u} - s_2 \quad \textit{for all} \quad x \in \mathbb{R}^d. \tag{5}$$

If $k(x,y) = (c^2 + \|x-y\|^2)^{-\gamma}$ for $c > 0$ and $\gamma \in (0, 2u-1)$, then $\mathcal{H}_{k_\mathrm{p}}$ P-dominates indicators and enforces tightness. If, in addition, $\|\boldsymbol{s}_\mathrm{p}\|$ has at most root exponential growth, then k_p controls P-convergence.

Measuring and improving sample quality Because the KSD provides a computable quality measure that requires no explicit integration under P, KSDs are now commonly used to select and tune MCMC sampling algorithms [14], generate accurate discrete approximations to P [22, 6, 7, 12], compress Markov chain output [27], and correct for biased or off-target sampling [20, 17, 27]. Each of these applications relies on KSD convergence control, but past work only established convergence control for P with Lipschitz $\boldsymbol{s}_\mathrm{p}$ and strongly log concave tails ([14, Lem. 16]; [6, Thm. 3]; [18, Thm. 3.2]). Notably, these conditions imply generalized dissipativity Eq. (5) with $u = 1$ but exclude all P with tails lighter than a Gaussian. Corollary 3.2 and Theorem 3.3 significantly relax these requirements by providing convergence control for all dissipative P with lighter-than-Laplace tails.

Much of the difficulty in analyzing KSDs stems from the fact that all known convergence-controlling KSDs are based on unbounded Stein kernels k_p. As a second illustration of the power of Corollary 3.2, Theorem 3.4 develops the first KSDs known to *metrize* P-*convergence* (i.e., $\mathrm{KSD}(\mathrm{Q}_n, \mathrm{P}) \to 0 \iff \mathrm{Q}_n \to \mathrm{P}$ weakly), by constructing **bounded** convergence-controlling Stein kernels.

Theorem 3.4 (Metrizing P-convergence with bounded Stein kernels)**.** *Consider a target measure $\mathrm{P} \in \mathcal{P}$ with score $\boldsymbol{s}_\mathrm{p}$ that, for some dissipativity rate $u > 1/2$ and $s, s_1, s_2 > 0$, satisfies the generalized dissipativity condition Eq.* (5)*. Define the Stein kernel with base kernel $K(x,y) = \mathrm{diag}\left(a(\|x\|)(x^i y^i + k(x,y))a(\|y\|)\right)$, for k characteristic to $\mathscr{D}^1_{L^1}$ with $\mathcal{H}_k \subset \mathscr{C}^1_0$ and $a(\|x\|) \equiv (c^2 + \|x\|^2)^{-\gamma}$ a tilting function with $c > 0$ and $\gamma \leq u$. The following statements hold true:*

1. *If $\mathrm{P} \in \mathcal{P}_{K,0}$, then $\mathcal{H}_{k_\mathrm{p}}$ P-dominates indicators and enforces tightness.*
2. *If $\mathrm{P} \in \mathcal{P}_{K,0}$, $\gamma \geq 0$, and $\|\boldsymbol{s}_\mathrm{p}(x)\| \leq (c^2 + \|x\|^2)^\gamma$, then k_p is bounded P-separating and controls P-convergence.*
3. *If $\|\boldsymbol{s}_\mathrm{p}(x)\| \cdot \|x\| \leq (c^2 + \|x\|^2)^\gamma$ and $\boldsymbol{s}_\mathrm{p} \in \mathscr{C}$, then $\mathcal{H}_{k_\mathrm{p}} \subset \mathscr{C}_b$ and k_p metrizes P-convergence.*

Sampling with Stein Variational Gradient Descent Stein variational gradient descent (SVGD) is a popular technique for approximating a target distribution P with a collection of n representative particles. The algorithm proceeds by iteratively updating the locations of the particles according to a simple rule determined by a user-selected KSD. [19] showed that the SVGD approximation converges weakly to P as the number of particles and iterations tend to infinity, provided that the chosen KSD controls P-convergence **and** that the Stein kernel is bounded. However, prior to this work, no bounded convergence-controlling Stein kernels were known. Theorem 3.4 therefore provides the first instance of a Stein kernel satisfying the SVGD convergence assumptions of [19].

4 Conclusion

This article derived new sufficient and necessary conditions for kernel discrepancies to enforce P-separation and control P-convergence, with important consequences in many applications, as we have highlighted in the contexts of GOF, improving sample quality and Stein variational gradient descent. We characterized all MMDs that separate P from Bochner embeddable measures, proposed novel sufficient conditions for separating all measures and enforcing tightness, strengthened all prior guarantees for KSD separation and convergence control on $\mathbb{R}^d$, and derived the first KSD known to exactly metrize (as opposed to strictly dominating) weak P-convergence on $\mathbb{R}^d$.

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2022/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to autmoatically hide this section in the anonymized submission.

References

[1] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 1950.

[2] Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum Stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pages 12964–12976, 2019.

[3] Alessandro Barp, Chris J Oates, Emilio Porcu, and Mark Girolami. A Riemann–Stein kernel method. *Bernoulli*, 2022.

[4] Francois-Xavier Briol, Alessandro Barp, Andrew B. Duncan, and Mark Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944*, 2019.

[5] Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 2010.

[6] Wilson Y. Chen, Lester Mackey, Jackson Gorham, François-Xavier Briol, and Chris J. Oates. Stein points. In *ICML*, 2018.

[7] Wilson Ye Chen, Alessandro Barp, François-Xavier Briol, Jackson Gorham, Mark Girolami, Lester Mackey, Chris Oates, et al. Stein point Markov chain Monte Carlo. *arXiv:1905.03673*, 2019.

[8] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, 2020.

[9] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *NeurIPS*, 2016.

[10] Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and François-Xavier Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *AISTATS*, 2022.

[11] Gintare K. Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.

[12] Futoshi Futami, Zhenghang Cui, Issei Sato, and Masashi Sugiyama. Bayesian posterior approximation via greedy particle optimization. In *AAAI*, 2019.

[13] Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *NeurIPS*, 2015.

[14] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *ICML*, 2017.

[15] Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic Stein discrepancies. *Advances in Neural Information Processing Systems*, 2020.

[16] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012.

[17] Liam Hodgkinson, Robert Salomone, and Fred Roosta. The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv:2001.09266*, 2020.

[18] Jonathan Huggins and Lester Mackey. Random feature Stein discrepancies. In *NeurIPS*, 2018.

[19] Qiang Liu. Stein variational gradient descent as gradient flow. In *NeurIPS*, 2017.

[20] Qiang Liu and Jason Lee. Black-box importance sampling. In *AISTATS*, 2017.

[21] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML*, 2016.

[22] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *NeurIPS*, 2016.

[23] Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, and Chris Oates. Generalised Bayesian inference for discrete intractable likelihood. *arXiv:2206.08420*, 2022.

[24] Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, Chris Oates, et al. Robust generalised Bayesian inference for intractable likelihoods. *arXiv:2104.07359*, 2021.

[25] Chris J. Oates, Jon Cockayne, François-Xavier Briol, and Mark Girolami. Convergence rates for a class of estimators based on Stein's method. *Bernoulli*, 2019.

[26] Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *arXiv:1410.2392*, 2014.

[27] Marina Riabiz, Wilson Ye Chen, Jon Cockayne, Pawel Swietach, Steven A. Niederer, Lester Mackey, and Chris. J. Oates. Optimal thinning of MCMC output. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2022.

[28] Laurent Schwartz. Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d'analyse mathématique*, 1964.

[29] Carl-Johann Simon-Gabriel, Alessandro Barp, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *arXiv:2006.09268*, 2020.

[30] Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *JMLR*, 2018.

[31] Bharath K. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 2016.

[32] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 2010.

A Appendix

A.1 Notation

Let $\mathcal{P}$ denote the set of (Borel) probability measures on a separable metric space $\mathcal{X}$. $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean norm and inner product on $\mathbb{R}^\ell$. For any function $f : \mathcal{X} \to \mathbb{R}^\ell$, we let $\mathcal{P}_f \equiv \{\mathrm{Q} \in \mathcal{P} : \|f\| \in \mathrm{L}^1(\mathrm{Q})\}$ denote the set of probability measures that finitely integrate $\|f\|$, where $\mathrm{L}^1(\mathrm{Q})$ denotes the seminormed space of functions with finite Q-integral. L^∞ will denote the normed space of bounded functions.

Let $\mathscr{C}^\ell(\mathbb{R}^d)$ denote the space of ℓ-times continuously differentiable $\mathbb{R}^d$-valued functions on $\mathcal{X}$ (i.e., $f \in \mathscr{C}^\ell(\mathbb{R}^d)$ if the partial derivatives of order ℓ of f^i exist and are continuous for $i \in [d] \equiv \{1, \ldots, d\}$). When $d = 1$ or $\ell = 0$ we will use the abbreviations $\mathscr{C}^\ell \equiv \mathscr{C}^\ell(\mathbb{R}^1)$ or $\mathscr{C}(\mathbb{R}^d) \equiv \mathscr{C}^0(\mathbb{R}^d)$. We let ∂f denote the vector of partial derivatives of a function f, and, for each multi-index p, let

$\partial^p f$ denote the p-th partial derivatives of f. Decay requirements will appear as subscripts: $\mathscr{C}_b(\mathbb{R}^d)$, and $\mathscr{C}_0(\mathbb{R}^d)$ will respectively denote the spaces of $\mathbb{R}^d$-valued continuous functions that are bounded, and vanishing at infinity. Analogously, for each function $h : \mathcal{X} \to [0, \infty)$, $\mathscr{C}_h(\mathbb{R}^d)$ and $\mathscr{C}_{0,h}(\mathbb{R}^d)$ respectively denote the spaces of $\mathbb{R}^d$-valued continuous functions f with $f/(1+h)$ bounded or vanishing at infinity. For any function of two arguments $K(y, x)$, we write $K_x \equiv K(\cdot, x)$.

For a reproducing kernel k we denote $\sqrt{k} : \mathcal{X} \to \mathbb{R}$ the function $x \mapsto \sqrt{k(x,x)}$. Matrix-valued reproducing kernels are denoted $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$, and the RKHS unit ball is denoted $\mathcal{B}_K$. The Langevin Stein operator on $\mathbb{R}^d$-valued functions is denoted $\mathcal{S}_{\mathrm{p}}(v) \equiv \frac{1}{\mathrm{p}}\nabla \cdot (\mathrm{p}v)$.

A.2 Extended definition of KSD

Consider a (reproducing) kernel k on $\mathcal{X}$ with reproducing kernel Hilbert space $\mathcal{H}_k$ [1, 28]. Traditionally, the associated kernel MMD is defined as the worst-case integration error across test functions in the RKHS unit norm ball $\mathcal{B}_k$ [16]:

$$\mathrm{MMD}_k(\mathrm{Q}, \mathrm{P}) \equiv \sup_{h \in \mathcal{B}_k} |\mathrm{Q}h - \mathrm{P}h| \,. \tag{6}$$

However, the expression $\mathrm{Q}h - \mathrm{P}h$ is not well defined when either (i) both $\mathrm{Q}h$ and $\mathrm{P}h$ are infinite or (ii) h is not integrable under Q. Unfortunately, both of these cases can occur when k is unbounded as $\mathcal{B}_k$ then necessarily contains an unbounded test function.

Since we are interested in a fixed target measure P, we address the first issue by focusing on kernels with finitely P-integrable test functions, i.e., with $\mathcal{B}_k \subset \mathrm{L}^1(\mathrm{P})$. To address the second issue, we extend the MMD definition Eq. (6) to all probability measures Q by taking the supremum only over the Q-integrable elements of $\mathcal{B}_k$, that is, h with either $h_+ \equiv \max(h, 0) \in \mathrm{L}^1(\mathrm{Q})$ or $h_- \equiv \max(-h, 0) \in \mathrm{L}^1(\mathrm{Q})$. In fact, since $\mathcal{B}_k$ is a symmetric set, considering only h with $h_+ \in \mathrm{L}^1(\mathrm{Q})$ suffices to ensure $|\mathrm{Q}h - \mathrm{P}h|$ is well defined and belongs to $[0, \infty]$.

Definition A.1 (Maximum mean discrepancy (MMD))**.** *For a given kernel k, define the set of* embeddable probability measures $\mathcal{P}_{\mathcal{H}_k} \equiv \{\mathrm{Q} \in \mathcal{P} : \mathcal{H}_k \subset \mathrm{L}^1(\mathrm{Q})\}$. *For any target measure* $\mathrm{P} \in \mathcal{P}_{\mathcal{H}_k}$, *we define the* maximum mean discrepancy $\mathrm{MMD}_k(\cdot\,, \mathrm{P}) : \mathcal{P} \to [0, \infty]$ *by*

$$\mathrm{MMD}_k(\mathrm{Q}, \mathrm{P}) \equiv \sup_{h \in \mathcal{B}_k \colon \max(h,0) \in \mathrm{L}^1(\mathrm{Q})} |\mathrm{Q}h - \mathrm{P}h| \,. \tag{7}$$

Building on the Stein discrepancy formalism of [13] and the zero-mean reproducing kernel theory of [26], [9, 21, 14] concurrently developed a special class of score-based MMDs that can be computed without any explicit integration under the target P. The Langevin KSD is defined in terms of the *Langevin Stein operator* [13],

$$\mathcal{S}_{\mathrm{p}}(v) \equiv \frac{1}{\mathrm{p}}\nabla \cdot (\mathrm{p}v),$$

as an integral probability metric over a subset of $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K)$ that ensures the supremum is well-defined.

$$\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) \equiv \sup_{v \in \mathcal{B}_K} |\mathrm{Q}\mathcal{S}_{\mathrm{p}}(v)| = \sup_{v \in \mathcal{B}_K} |\mathrm{Q}\mathcal{S}_{\mathrm{p}}(v) - \mathrm{P}\mathcal{S}_{\mathrm{p}}(v)|. \tag{8}$$

However, $\mathcal{S}_{\mathrm{p}}(v)$ is often unbounded so that as before the expression Eq. (8) need not be well defined for all $\mathrm{Q} \in \mathcal{P}$. To enable meaningful KSD evaluation for all probability measures, we follow the recipe of Definition A.1 to extend the definition of KSD to all $\mathrm{Q} \in \mathcal{P}$, see Definition 1.1.

Under additional assumptions, like Bochner embeddability of P and Q and continuous differentiability of K and p, prior work showed that the KSD Eq. (2) is equivalent to an MMD with a scalar *Stein kernel* k_{p} and that $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K)$ defines a *Stein RKHS* $\mathcal{H}_{k_{\mathrm{p}}}$ of scalar-valued functions [26, 9, 21, 14, 2]. Our next result shows that **no** additional assumptions are necessary: $\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = \mathrm{MMD}_{k_{\mathrm{p}}}(\mathrm{Q}, \mathrm{P})$ and $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K) = \mathcal{H}_{k_{\mathrm{p}}}$ whenever the left-hand side quantities are well defined.

Theorem A.2 (KSD as MMD)**.** *Consider a target* $\mathrm{P} \in \mathcal{P}$ *with density* $\mathrm{p} > 0$ *and matrix-valued base kernel K for which $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K)$ exists. Then $\mathcal{S}_{\mathrm{p}}(\mathcal{H}_K)$ is the* Stein RKHS $\mathcal{H}_{k_{\mathrm{p}}}$ *induced by the* Stein kernel $k_{\mathrm{p}}(x, y) \equiv \frac{1}{\mathrm{p}(x)\mathrm{p}(y)}\nabla_y \cdot \nabla_x \cdot (\mathrm{p}(x)K(x, y)\mathrm{p}(y))$. *Moreover, for target measures with zero-mean Stein RKHSes, i.e., for* P *in*

$$\mathcal{P}_{K,0} \equiv \{\mathrm{Q} \in \mathcal{P} \text{ with } \mathrm{q} > 0 \colon \mathcal{S}_{\mathrm{q}}(\mathcal{H}_K) \text{ exists}, \mathcal{S}_{\mathrm{q}}(\mathcal{H}_K) \subset \mathrm{L}^1(\mathrm{Q}), \text{and } \mathrm{Q}(\mathcal{S}_{\mathrm{q}}(\mathcal{H}_K)) = \{0\}\}, \tag{9}$$

the KSD matches the Stein kernel MMD: $\mathrm{KSD}_{K,\mathrm{P}}(\mathrm{Q}) = \mathrm{MMD}_{k_{\mathrm{p}}}(\mathrm{Q}, \mathrm{P})$ *for all* $\mathrm{Q} \in \mathcal{P}$.

The zero-mean condition $\mathrm{P} \in \mathcal{P}_{\kappa,0}$ ensures that all functions in the Stein RKHS integrate to zero under the target measure so that the KSD can be evaluated without any explicit integration under P. Moreover, when Q embeds into the Stein RKHS, the KSD takes on its more familiar double integral form, i.e., if $\mathrm{P} \in \mathcal{P}_{\kappa,0}$ and $\mathrm{Q} \in \mathcal{P}_{\mathcal{H}_{k_{\mathrm{P}}}}$, then $\mathrm{KSD}^2_{\kappa,\mathrm{P}}(\mathrm{Q}) = \iint k_{\mathrm{P}}(x,y)\mathrm{d}\mathrm{Q}(x)\mathrm{d}\mathrm{Q}(y)$