Measuring Sample Quality with Kernels

Jackson Gorham [1] **Lester Mackey** [2]

Abstract

Approximate Markov chain Monte Carlo (MCMC) offers the promise of more rapid sampling at the cost of more biased inference. Since standard MCMC diagnostics fail to detect these biases, researchers have developed computable *Stein discrepancy* measures that provably determine the convergence of a sample to its target distribution. This approach was recently combined with the theory of reproducing kernels to define a closed-form *kernel Stein discrepancy* (KSD) computable by summing kernel evaluations across pairs of sample points. We develop a theory of weak convergence for KSDs based on Stein's method, demonstrate that commonly used KSDs fail to detect non-convergence even for Gaussian targets, and show that kernels with slowly decaying tails provably determine convergence for a large class of target distributions. The resulting convergence-determining KSDs are suitable for comparing biased, exact, and deterministic sample sequences and simpler to compute and parallelize than alternative Stein discrepancies. We use our tools to compare biased samplers, select sampler hyperparameters, and improve upon existing KSD approaches to one-sample hypothesis testing and sample quality improvement.

1. Introduction

When Bayesian inference and maximum likelihood estimation (Geyer, 1991) demand the evaluation of intractable expectations $\mathbb{E}_P[h(Z)] = \int p(x)h(x)dx$ under a target distribution P, Markov chain Monte Carlo (MCMC) methods (Brooks et al., 2011) are often employed to approximate these integrals with asymptotically correct sample averages $\mathbb{E}_{Q_n}[h(X)] = \frac{1}{n}\sum_{i=1}^n h(x_i)$. However, many exact MCMC methods are computationally expensive, and recent years have seen the introduction of biased MCMC procedures (see, e.g., Welling & Teh, 2011; Ahn et al., 2012; Korattikara et al., 2014) that exchange asymptotic correctness for increased sampling speed.

Since standard MCMC diagnostics, like mean and trace plots, pooled and within-chain variance measures, effective sample size, and asymptotic variance (Brooks et al., 2011), do not account for asymptotic bias, Gorham & Mackey (2015) defined a new family of sample quality measures – the *Stein discrepancies* – that measure how well $\mathbb{E}_{Q_n}$ approximates $\mathbb{E}_P$ while avoiding explicit integration under P. Gorham & Mackey (2015); Mackey & Gorham (2016); Gorham et al. (2016) further showed that specific members of this family – the *graph Stein discrepancies* – were (a) efficiently computable by solving a linear program and (b) convergence-determining for large classes of targets P. Building on the zero mean reproducing kernel theory of Oates et al. (2016b), Chwialkowski et al. (2016) and Liu et al. (2016) later showed that other members of the Stein discrepancy family had a closed-form solution involving the sum of kernel evaluations over pairs of sample points.

This closed form represents a significant practical advantage, as no linear program solvers are necessary, and the computation of the discrepancy can be easily parallelized. However, as we will see in Section 3.2, not all *kernel Stein discrepancies* are suitable for our setting. In particular, in dimension $d \geq 3$, the kernel Stein discrepancies previously recommended in the literature fail to detect when a sample is not converging to the target. To address this shortcoming, we develop a theory of weak convergence for the kernel Stein discrepancies analogous to that of (Gorham & Mackey, 2015; Mackey & Gorham, 2016; Gorham et al., 2016) and design a class of kernel Stein discrepancies that provably control weak convergence for a large class of target distributions.

After formally describing our goals for measuring sample quality in Section 2, we outline our strategy, based on Stein's method, for constructing and analyzing practical quality measures at the start of Section 3. In Section 3.1, we define our family of closed-form quality measures – the kernel Stein discrepancies (KSDs) – and establish several

[1]Stanford University, Palo Alto, CA USA [2]Microsoft Research New England, Cambridge, MA USA. Correspondence to: Jackson Gorham <jgorham@stanford.edu>, Lester Mackey <lmackey@microsoft.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

appealing practical properties of these measures. We analyze the convergence properties of KSDs in Sections 3.2 and 3.3, showing that previously proposed KSDs fail to detect non-convergence and proposing practical convergence-determining alternatives. Section 4 illustrates the value of convergence-determining kernel Stein discrepancies in a variety of applications, including hyperparameter selection, sampler selection, one-sample hypothesis testing, and sample quality improvement. Finally, in Section 5, we conclude with a discussion of related and future work.

Notation We will use μ to denote a generic probability measure and $\Rightarrow$ to denote the weak convergence of a sequence of probability measures. We will use $\|\cdot\|_r$ for $r \in [1, \infty]$ to represent the ℓ^r norm on $\mathbb{R}^d$ and occasionally refer to a generic norm $\|\cdot\|$ with associated dual norm $\|a\|^* \triangleq \sup_{b \in \mathbb{R}^d, \|b\|=1} \langle a, b \rangle$ for vectors $a \in \mathbb{R}^d$. We let e_j be the j-th standard basis vector. For any function $g : \mathbb{R}^d \to \mathbb{R}^{d'}$, we define $M_0(g) \triangleq \sup_{x \in \mathbb{R}^d} \|g(x)\|_2$, $M_1(g) \triangleq \sup_{x \neq y} \|g(x) - g(y)\|_2 / \|x - y\|_2$, and ∇g as the gradient with components $(\nabla g(x))_{jk} \triangleq \nabla_{x_k} g_j(x)$. We further let $g \in C^m$ indicate that g is m times continuously differentiable and $g \in C_0^m$ indicate that $g \in C^m$ and $\nabla^l g$ is vanishing at infinity for all $l \in \{0, \dots, m\}$. We define $C^{(m,m)}$ (respectively, $C_b^{(m,m)}$ and $C_0^{(m,m)}$) to be the set of functions $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $(x, y) \mapsto \nabla_x^l \nabla_y^l k(x, y)$ continuous (respectively, continuous and uniformly bounded, continuous and vanishing at infinity) for all $l \in \{0, \dots, m\}$.

2. Quality measures for samples

Consider a target distribution P with continuously differentiable (Lebesgue) density p supported on all of $\mathbb{R}^d$. We assume that the *score function* $b \triangleq \nabla \log p$ can be evaluated[1] but that, for most functions of interest, direct integration under P is infeasible. We will therefore approximate integration under P using a *weighted sample* $Q_n = \sum_{i=1}^n q_n(x_i) \delta_{x_i}$ with sample points $x_1, \dots, x_n \in \mathbb{R}^d$ and q_n a probability mass function. We will make no assumptions about the origins of the sample points; they may be the output of a Markov chain or even deterministically generated.

Each Q_n offers an approximation $\mathbb{E}_{Q_n}[h(X)] = \sum_{i=1}^n q_n(x_i) h(x_i)$ for each intractable expectation $\mathbb{E}_P[h(Z)]$, and our aim is to effectively compare the quality of the approximation offered by any two samples targeting P. In particular, we wish to produce a quality measure that *(i)* identifies when a sequence of samples is converging to the target, *(ii)* determines when a sequence of samples is not converging to the target, and *(iii)* is efficiently computable. Since our interest is in approximating expectations, we will consider discrepancies quantifying the maximum expectation error over a class of test functions $\mathcal{H}$:

$$d_{\mathcal{H}}(Q_n, P) \triangleq \sup_{h \in \mathcal{H}} |\mathbb{E}_P[h(Z)] - \mathbb{E}_{Q_n}[h(X)]|. \quad (1)$$

When $\mathcal{H}$ is large enough, for any sequence of probability measures $(\mu_m)_{m \geq 1}$, $d_{\mathcal{H}}(\mu_m, P) \to 0$ only if $\mu_m \Rightarrow P$. In this case, we call (1) an *integral probability metric* (IPM) (Müller, 1997). For example, when $\mathcal{H} = BL_{\|\cdot\|_2} \triangleq \{h : \mathbb{R}^d \to \mathbb{R} \,|\, M_0(h) + M_1(h) \leq 1\}$, the IPM $d_{BL_{\|\cdot\|_2}}$ is called the *bounded Lipschitz* or *Dudley metric* and exactly metrizes convergence in distribution. Alternatively, when $\mathcal{H} = \mathcal{W}_{\|\cdot\|_2} \triangleq \{h : \mathbb{R}^d \to \mathbb{R} \,|\, M_1(h) \leq 1\}$ is the set of 1-Lipschitz functions, the IPM $d_{\mathcal{W}_{\|\cdot\|}}$ in (1) is known as the Wasserstein metric.

An apparent practical problem with using the IPM $d_{\mathcal{H}}$ as a sample quality measure is that $\mathbb{E}_P[h(Z)]$ may not be computable for $h \in \mathcal{H}$. However, if $\mathcal{H}$ were chosen such that $\mathbb{E}_P[h(Z)] = 0$ for all $h \in \mathcal{H}$, then no explicit integration under P would be necessary. To generate such a class of test functions and to show that the resulting IPM still satisfies our desiderata, we follow the lead of Gorham & Mackey (2015) and consider Charles Stein's method for characterizing distributional convergence.

3. Stein's method with kernels

Stein's method (Stein, 1972) provides a three-step recipe for assessing convergence in distribution:

1. Identify a *Stein operator* $\mathcal{T}$ that maps functions $g : \mathbb{R}^d \to \mathbb{R}^d$ from a domain $\mathcal{G}$ to real-valued functions $\mathcal{T}g$ such that

   $$\mathbb{E}_P[(\mathcal{T}g)(Z)] = 0 \text{ for all } g \in \mathcal{G}.$$

   For any such Stein operator and *Stein set* $\mathcal{G}$, Gorham & Mackey (2015) defined the *Stein discrepancy* as

   $$\mathcal{S}(\mu, \mathcal{T}, \mathcal{G}) \triangleq \sup_{g \in \mathcal{G}} |\mathbb{E}_\mu[(\mathcal{T}g)(X)]| = d_{\mathcal{T}\mathcal{G}}(\mu, P) \quad (2)$$

   which, crucially, avoids explicit integration under P.

2. Lower bound the Stein discrepancy by an IPM $d_{\mathcal{H}}$ known to dominate weak convergence. This can be done once for a broad class of target distributions to ensure that $\mu_m \Rightarrow P$ whenever $\mathcal{S}(\mu_m, \mathcal{T}, \mathcal{G}) \to 0$ for a sequence of probability measures $(\mu_m)_{m \geq 1}$ (Desideratum *(ii)*).

3. Provide an upper bound on the Stein discrepancy ensuring that $\mathcal{S}(\mu_m, \mathcal{T}, \mathcal{G}) \to 0$ under suitable convergence of μ_m to P (Desideratum *(i)*).

[1] No knowledge of the normalizing constant is needed.

While Stein's method is principally used as a mathematical tool to prove convergence in distribution, we seek, in the spirit of (Gorham & Mackey, 2015; Gorham et al., 2016), to harness the Stein discrepancy as a practical tool for measuring sample quality. The subsections to follow develop a specific, practical instantiation of the abstract Stein's method recipe based on reproducing kernel Hilbert spaces. An empirical analysis of the Stein discrepancies recommended by our theory follows in Section 4.

3.1. Selecting a Stein operator and a Stein set

A standard, widely applicable univariate Stein operator is the *density method operator* (see Stein et al., 2004; Chatterjee & Shao, 2011; Chen et al., 2011; Ley et al., 2017),

$$(\mathcal{T}g)(x) \triangleq \frac{1}{p(x)}\frac{d}{dx}(p(x)g(x)) = g(x)b(x) + g'(x).$$

Inspired by the generator method of Barbour (1988; 1990) and Götze (1991), Gorham & Mackey (2015) generalized this operator to multiple dimensions. The resulting *Langevin Stein operator*

$$(\mathcal{T}_P g)(x) \triangleq \frac{1}{p(x)}\langle \nabla, p(x)g(x)\rangle = \langle g(x), b(x)\rangle + \langle \nabla, g(x)\rangle$$

for functions $g : \mathbb{R}^d \to \mathbb{R}^d$ was independently developed, without connection to Stein's method, by Oates et al. (2016b) for the design of Monte Carlo control functionals. Notably, the Langevin Stein operator depends on P only through its score function $b = \nabla \log p$ and hence is computable even when the normalizing constant of p is not. While our work is compatible with other practical Stein operators, like the family of diffusion Stein operators defined in (Gorham et al., 2016), we will focus on the Langevin operator for the sake of brevity.

Hereafter, we will let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the reproducing kernel of a reproducing kernel Hilbert space (RKHS) $\mathcal{K}_k$ of functions from $\mathbb{R}^d \to \mathbb{R}$. That is, $\mathcal{K}_k$ is a Hilbert space of functions such that, for all $x \in \mathbb{R}^d$, $k(x, \cdot) \in \mathcal{K}_k$ and $f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{K}_k}$ whenever $f \in \mathcal{K}_k$. We let $\|\cdot\|_{\mathcal{K}_k}$ be the norm induced from the inner product on $\mathcal{K}_k$.

With this definition, we define our *kernel Stein set* $\mathcal{G}_{k,\|\cdot\|}$ as the set of vector-valued functions $g = (g_1, \dots, g_d)$ such that each component function g_j belongs to $\mathcal{K}_k$ and the vector of their norms $\|g_j\|_{\mathcal{K}_k}$ belongs to the $\|\cdot\|^*$ unit ball:[2]

$$\mathcal{G}_{k,\|\cdot\|} \triangleq \{g = (g_1, \dots, g_d) \mid \|v\|^* \leq 1 \text{ for } v_j \triangleq \|g_j\|_{\mathcal{K}_k}\}.$$

The following result, proved in Section B, establishes that this is an acceptable domain for $\mathcal{T}_P$.

Proposition 1 (Zero mean test functions)**.** *If $k \in C_b^{(1,1)}$ and $\mathbb{E}_P[\|\nabla \log p(Z)\|_2] < \infty$, then $\mathbb{E}_P[(\mathcal{T}_P g)(Z)] = 0$ for all $g \in \mathcal{G}_{k,\|\cdot\|}$.*

[2] Our analyses and algorithms support each g_j belonging to a different RKHS $\mathcal{K}_{k_j}$, but we will not need that flexibility here.

The Langevin Stein operator and kernel Stein set together define our quality measure of interest, the *kernel Stein discrepancy* (KSD) $\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_{k,\|\cdot\|})$. When $\|\cdot\| = \|\cdot\|_2$, this definition recovers the KSD proposed by Chwialkowski et al. (2016) and Liu et al. (2016). Our next result shows that, for any $\|\cdot\|$, the KSD admits a closed-form solution.

Proposition 2 (KSD closed form)**.** *Suppose $k \in C^{(1,1)}$, and, for each $j \in \{1, \dots d\}$, define the* Stein kernel

$$\begin{aligned} k_0^j(x, y) &\triangleq \frac{1}{p(x)p(y)}\nabla_{x_j}\nabla_{y_j}(p(x)k(x, y)p(y)) \\ &= b_j(x)b_j(y)k(x, y) + b_j(x)\nabla_{y_j}k(x, y) \\ &\quad + b_j(y)\nabla_{x_j}k(x, y) + \nabla_{x_j}\nabla_{y_j}k(x, y). \end{aligned} \tag{3}$$

If $\sum_{j=1}^d \mathbb{E}_\mu\left[k_0^j(X, X)^{1/2}\right] < \infty$, then $\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_{k,\|\cdot\|}) = \|w\|$ where $w_j \triangleq \sqrt{\mathbb{E}_{\mu\times\mu}[k_0^j(X, \tilde{X})]}$ with $X, \tilde{X} \overset{\text{iid}}{\sim} \mu$.

The proof is found in Section C. Notably, when μ is the discrete measure $Q_n = \sum_{i=1}^n q_n(x_i)\delta_{x_i}$, the KSD reduces to evaluating each k_0^j at pairs of support points as $w_j = \sqrt{\sum_{i,i'=1}^n q_n(x_i)k_0^j(x_i, x_{i'})q_n(x_{i'})}$, a computation which is easily parallelized over sample pairs and coordinates j.

Our Stein set choice was motivated by the work of Oates et al. (2016b) who used the sum of Stein kernels $k_0 = \sum_{j=1}^d k_0^j$ to develop nonparametric control variates. Each term w_j in Proposition 2 can also be viewed as an instance of the maximum mean discrepancy (MMD) (Gretton et al., 2012) between μ and P measured with respect to the Stein kernel k_0^j. In standard uses of MMD, an arbitrary kernel function is selected, and one must be able to compute expectations of the kernel function under P. Here, this requirement is satisfied automatically, since our induced kernels are chosen to have mean zero under P.

For clarity we will focus on the specific kernel Stein set choice $\mathcal{G}_k \triangleq \mathcal{G}_{k,\|\cdot\|_2}$ for the remainder of the paper, but our results extend directly to KSDs based on any $\|\cdot\|$, since all KSDs are equivalent in a strong sense:

Proposition 3 (Kernel Stein set equivalence)**.** *Under the assumptions of Proposition 2, there are constants $c_d, c_d' > 0$ depending only on d and $\|\cdot\|$ such that $c_d\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_{k,\|\cdot\|}) \leq \mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_{k,\|\cdot\|_2}) \leq c_d'\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_{k,\|\cdot\|})$.*

The short proof is found in Section D.

3.2. Lower bounding the kernel Stein discrepancy

We next aim to establish conditions under which the KSD $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_k) \to 0$ only if $\mu_m \Rightarrow P$ (Desideratum *(ii)*). Recently, Gorham et al. (2016) showed that the Langevin graph Stein discrepancy dominates convergence in distribution whenever P belongs to the class $\mathcal{P}$ of *distantly dissipative* distributions with Lipschitz score function b:

Definition 4 (Distant dissipativity (Eberle, 2015; Gorham et al., 2016)). *A distribution P is* distantly dissipative *if $\kappa_0 \triangleq \liminf_{r\to\infty} \kappa(r) > 0$ for*

$$\kappa(r) = \inf\{-2\tfrac{\langle b(x)-b(y), x-y\rangle}{\|x-y\|_2^2} : \|x-y\|_2 = r\}. \quad (4)$$

Examples of distributions in $\mathcal{P}$ include finite Gaussian mixtures with common covariance and all distributions strongly log-concave outside of a compact set, including Bayesian linear, logistic, and Huber regression posteriors with Gaussian priors (see Gorham et al., 2016, Section 4). Moreover, when $d = 1$, membership in $\mathcal{P}$ is sufficient to provide a lower bound on the KSD for most common kernels including the Gaussian, Matérn, and inverse multiquadric kernels.

Theorem 5 (Univariate KSD detects non-convergence). *Suppose that $P \in \mathcal{P}$ and $k(x, y) = \Phi(x - y)$ for $\Phi \in C^2$ with a non-vanishing generalized Fourier transform. If $d = 1$, then $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_k) \to 0$ only if $\mu_m \Rightarrow P$.*

The proof in Section E provides a lower bound on the KSD in terms of an IPM known to dominate weak convergence. However, our next theorem shows that in higher dimensions $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k)$ can converge to 0 without the sequence $(Q_n)_{n\geq 1}$ converging to any probability measure. This deficiency occurs even when the target is Gaussian.

Theorem 6 (KSD fails with light kernel tails). *Suppose $k \in C_b^{(1,1)}$ and define the kernel decay rate*

$$\gamma(r) \triangleq \sup\{\max(|k(x,y)|, \|\nabla_x k(x,y)\|_2, |\langle \nabla_x, \nabla_y k(x,y)\rangle|) : \|x-y\|_2 \geq r\}.$$

If $d \geq 3$, $P = \mathcal{N}(0, I_d)$, and $\gamma(r) = o(r^{-\alpha})$ for $\alpha \triangleq (\frac{1}{2} - \frac{1}{d})^{-1}$, then $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$ does not imply $Q_n \Rightarrow P$.

Theorem 6 implies that KSDs based on the commonly used Gaussian kernel, Matérn kernel, and compactly supported kernels of Wendland (2004, Theorem 9.13) all fail to detect non-convergence when $d \geq 3$. In addition, KSDs based on the inverse multiquadric kernel ($k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$) for $\beta < -1$ fail to detect non-convergence for any $d > 2\beta/(\beta + 1)$. The proof in Section F shows that the violating sample sequences $(Q_n)_{n\geq 1}$ are simple to construct, and we provide an empirical demonstration of this failure to detect non-convergence in Section 4.

The failure of the KSDs in Theorem 6 can be traced to their inability to enforce *uniform tightness*. A sequence of probability measures $(\mu_m)_{m\geq 1}$ is uniformly tight if for every $\epsilon > 0$, there is a finite number $R(\epsilon)$ such that $\limsup_m \mu_m(\|X\|_2 > R(\epsilon)) \leq \epsilon$. Uniform tightness implies that no mass in the sequence of probability measures escapes to infinity. When the kernel k decays more rapidly than the score function grows, the KSD ignores excess mass in the tails and hence can be driven to zero by a non-tight sequence of increasingly diffuse probability measures. The following theorem demonstrates uniform tightness is the missing piece to ensure weak convergence.

Theorem 7 (KSD detects tight non-convergence). *Suppose that $P \in \mathcal{P}$ and $k(x, y) = \Phi(x-y)$ for $\Phi \in C^2$ with a non-vanishing generalized Fourier transform. If $(\mu_m)_{m\geq 1}$ is uniformly tight, then $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_k) \to 0$ only if $\mu_m \Rightarrow P$.*

Our proof in Section G explicitly lower bounds the KSD $\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k)$ in terms of the bounded Lipschitz metric $d_{BL_{\|\cdot\|}}(\mu, P)$, which exactly metrizes weak convergence.

Ideally, when a sequence of probability measures is not uniformly tight, the KSD would reflect this divergence in its reported value. To achieve this, we consider the inverse multiquadric (IMQ) kernel $k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$ for some $\beta < 0$ and $c > 0$. While KSDs based on IMQ kernels fail to determine convergence when $\beta < -1$ (by Theorem 6), our next theorem shows that they automatically enforce tightness and detect non-convergence whenever $\beta \in (-1, 0)$.

Theorem 8 (IMQ KSD detects non-convergence). *Suppose $P \in \mathcal{P}$ and $k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$ for $c > 0$ and $\beta \in (-1, 0)$. If $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_k) \to 0$, then $\mu_m \Rightarrow P$.*

The proof in Section H provides a lower bound on the KSD in terms of the bounded Lipschitz metric $d_{BL_{\|\cdot\|}}(\mu, P)$. The success of the IMQ kernel over other common characteristic kernels can be attributed to its slow decay rate. When $P \in \mathcal{P}$ and the IMQ exponent $\beta > -1$, the function class $\mathcal{T}_P\mathcal{G}_k$ contains unbounded (coercive) functions. These functions ensure that the IMQ KSD $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_k)$ goes to 0 only if $(\mu_m)_{m\geq 1}$ is uniformly tight.

3.3. Upper bounding the kernel Stein discrepancy

The usual goal in upper bounding the Stein discrepancy is to provide a rate of convergence to P for particular approximating sequences $(\mu_m)_{m=1}^\infty$. Because we aim to directly compute the KSD for arbitrary samples Q_n, our chief purpose in this section is to ensure that the KSD $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_k)$ will converge to zero when μ_m is converging to P (Desideratum *(i)*).

Proposition 9 (KSD detects convergence). *If $k \in C_b^{(2,2)}$ and $\nabla \log p$ is Lipschitz with $\mathbb{E}_P[\|\nabla \log p(Z)\|_2^2] < \infty$, then $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_k) \to 0$ whenever the Wasserstein distance $d_{\mathcal{W}_{\|\cdot\|_2}}(\mu_m, P) \to 0$.*

Proposition 9 applies to common kernels like the Gaussian, Matérn, and IMQ kernels, and its proof in Section I provides an explicit upper bound on the KSD in terms of the Wasserstein distance $d_{\mathcal{W}_{\|\cdot\|_2}}$. When $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ for $x_i \overset{\text{iid}}{\sim} \mu$, (Liu et al., 2016, Thm. 4.1) further implies that $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \Rightarrow \mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k)$ at an $O(n^{-1/2})$ rate under continuity and integrability assumptions on μ.

4. Experiments

We next conduct an empirical evaluation of the KSD quality measures recommended by our theory, recording all timings on an Intel Xeon CPU E5-2650 v2 @ 2.60GHz. Throughout, we will refer to the KSD with IMQ base kernel $k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$, exponent $\beta = -\frac{1}{2}$, and $c = 1$ as the IMQ KSD. Code reproducing all experiments can be found on the Julia (Bezanson et al., 2014) package site `https://jgorham.github.io/SteinDiscrepancy.jl/`.

4.1. Comparing discrepancies

Our first, simple experiment is designed to illustrate several properties of the IMQ KSD and to compare its behavior with that of two preexisting discrepancy measures, the Wasserstein distance $d_{\mathcal{W}_{\|\cdot\|_2}}$, which can be computed for simple univariate targets (Vallender, 1974), and the spanner graph Stein discrepancy of Gorham & Mackey (2015). We adopt a bimodal Gaussian mixture with $p(x) \propto e^{-\frac{1}{2}\|x+\Delta e_1\|_2^2} + e^{-\frac{1}{2}\|x-\Delta e_1\|_2^2}$ and $\Delta = 1.5$ as our target P and generate a first sample point sequence i.i.d. from the target and a second sequence i.i.d. from one component of the mixture, $\mathcal{N}(-\Delta e_1, I_d)$. As seen in the left panel of Figure 1 where $d = 1$, the IMQ KSD decays at an $n^{-0.51}$ rate when applied to the first n points in the target sample and remains bounded away from zero when applied to the to the single component sample. This desirable behavior is closely mirrored by the Wasserstein distance and the graph Stein discrepancy.

The middle panel of Figure 1 records the time consumed by the graph and kernel Stein discrepancies applied to the i.i.d. sample points from P. Each method is given access to d cores when working in d dimensions, and we use the released code of Gorham & Mackey (2015) with the default Gurobi 6.0.4 linear program solver for the graph Stein discrepancy. We find that the two methods have nearly identical runtimes when $d = 1$ but that the KSD is 10 to 1000 times faster when $d = 4$. In addition, the KSD is straightforwardly parallelized and does not require access to a linear program solver, making it an appealing practical choice for a quality measure.

Finally, the right panel displays the optimal Stein functions, $g_j(y) = \frac{\mathbb{E}_{Q_n}\left[b_j(X)k(X,y)+\nabla_{x_j}k(X,y)\right]}{\mathcal{S}(Q_n,\mathcal{T}_P,\mathcal{G}_k)}$, recovered by the IMQ KSD when $d = 1$ and $n = 10^3$. The associated test functions $h(y) = (\mathcal{T}_P g)(y) = \frac{\sum_{j=1}^d \mathbb{E}_{Q_n}\left[k_0^j(X,y)\right]}{\mathcal{S}(Q_n,\mathcal{T}_P,\mathcal{G}_k)}$ are the mean-zero functions under P that best discriminate the target P and the sample Q_n. As might be expected, the optimal test function for the single component sample features large magnitude values in the oversampled region far from the missing mode.

4.2. The importance of kernel choice

Theorem 6 established that kernels with rapidly decaying tails yield KSDs that can be driven to zero by off-target sample sequences. Our next experiment provides an empirical demonstration of this issue for a multivariate Gaussian target $P = \mathcal{N}(0, I_d)$ and KSDs based on the popular Gaussian ($k(x, y) = e^{-\|x-y\|_2^2/2}$) and Matérn ($k(x, y) = (1 + \sqrt{3}\|x - y\|_2)e^{-\sqrt{3}\|x-y\|_2}$) radial kernels.

Following the proof Theorem 6 in Section F, we construct an off-target sequence $(Q_n)_{n\geq 1}$ that sends $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k)$ to 0 for these kernel choices whenever $d \geq 3$. Specifically, for each n, we let $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ where, for all i and j, $\|x_i\|_2 \leq 2n^{1/d}\log n$ and $\|x_i - x_j\|_2 \geq 2\log n$. To select these sample points, we independently sample candidate points uniformly from the ball $\{x : \|x\|_2 \leq 2n^{1/d}\log n\}$, accept any points not within $2\log n$ Euclidean distance of any previously accepted point, and terminate when n points have been accepted.

For various dimensions, Figure 2 displays the result of applying each KSD to the off-target sequence $(Q_n)_{n\geq 1}$ and an "on-target" sequence of points sampled i.i.d. from P. For comparison, we also display the behavior of the IMQ KSD which provably controls tightness and dominates weak convergence for this target by Theorem 8. As predicted, the Gaussian and Matérn KSDs decay to 0 under the off-target sequence and decay more rapidly as the dimension d increases; the IMQ KSD remains bounded away from 0.

4.3. Selecting sampler hyperparameters

The approximate slice sampler of DuBois et al. (2014) is a biased MCMC procedure designed to accelerate inference when the target density takes the form $p(x) \propto \pi(x)\prod_{l=1}^L \pi(y_l|x)$ for $\pi(\cdot)$ a prior distribution on $\mathbb{R}^d$ and $\pi(y_l|x)$ the likelihood of a datapoint y_l. A standard slice sampler must evaluate the likelihood of all L datapoints to draw each new sample point x_i. To reduce this cost, the approximate slice sampler introduces a tuning parameter ϵ which determines the number of datapoints that contribute to an approximation of the slice sampling step; an appropriate setting of this parameter is imperative for accurate inference. When ϵ is too small, relatively few sample points will be generated in a given amount of sampling time, yielding sample expectations with high Monte Carlo variance. When ϵ is too large, the large approximation error will produce biased samples that no longer resemble the target.

To assess the suitability of the KSD for tolerance parameter selection, we take as our target P the bimodal Gaussian mixture model posterior of (Welling & Teh, 2011). For an array of ϵ values, we generated 50 independent approximate slice sampling chains with batch size 5, each with a

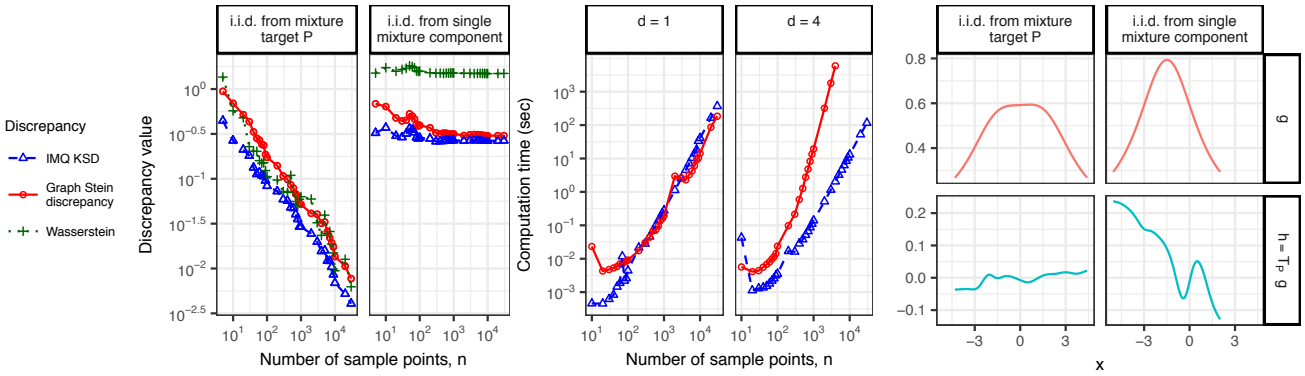

Figure 1. **Left:** For $d = 1$, comparison of discrepancy measures for samples drawn i.i.d. from either the bimodal Gaussian mixture target P or a single mixture component (see Section 4.1). **Middle:** On-target discrepancy computation time using d cores in d dimensions. **Right:** For $n = 10^3$ and $d = 1$, the Stein functions g and discriminating test functions $h = \mathcal{T}_P g$ which maximize the KSD.

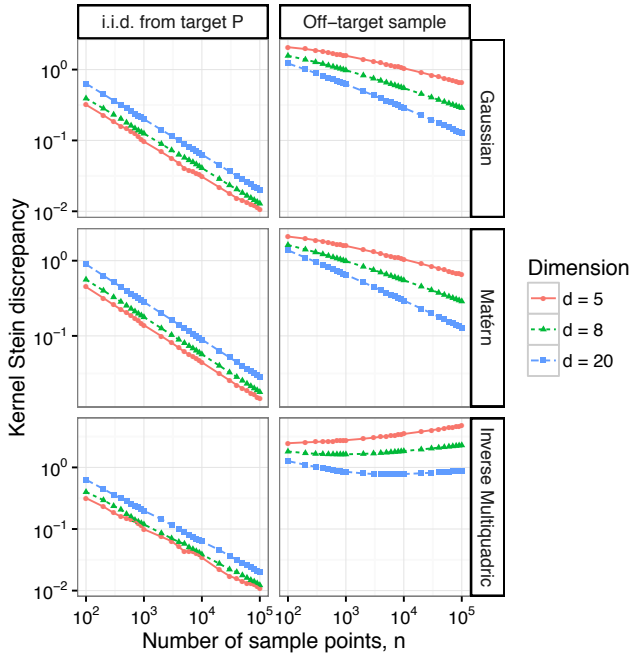

Figure 2. Gaussian and Matérn KSDs are driven to 0 by an off-target sequence that does not converge to the target $P = \mathcal{N}(0, I_d)$ (see Section 4.2). The IMQ KSD does not share this deficiency.

budget of 148000 likelihood evaluations, and plotted the median IMQ KSD and effective sample size (ESS, a standard sample quality measure based on asymptotic variance (Brooks et al., 2011)) in Figure 3. ESS, which does not detect Markov chain bias, is maximized at the largest hyperparameter evaluated ($\epsilon = 10^{-1}$), while the KSD is minimized at an intermediate value ($\epsilon = 10^{-2}$). The right panel of Figure 3 shows representative samples produced by several settings of ϵ. The sample produced by the ESS-selected chain is significantly overdispersed, while the sample from $\epsilon = 0$ has minimal coverage of the second mode due to its small sample size. The sample produced by the KSD-selected chain best resembles the posterior target. Using 4 cores, the longest KSD computation with $n = 10^3$ sample points took $0.16s$.

4.4. Selecting samplers

Ahn et al. (2012) developed two biased MCMC samplers for accelerated posterior inference, both called Stochastic Gradient Fisher Scoring (SGFS). In the full version of SGFS (termed SGFS-f), a $d \times d$ matrix must be inverted to draw each new sample point. Since this can be costly for large d, the authors developed a second sampler (termed SGFS-d) in which only a diagonal matrix must be inverted to draw each new sample point. Both samplers can be viewed as discrete-time approximations to a continuous-time Markov process that has the target P as its stationary distribution; however, because no Metropolis-Hastings correction is employed, neither sampler has the target as its stationary distribution. Hence we will use the KSD – a quality measure that accounts for asymptotic bias – to evaluate and choose between these samplers.

Specifically, we evaluate the SGFS-f and SGFS-d samples produced in (Ahn et al., 2012, Sec. 5.1). The target P is a Bayesian logistic regression with a flat prior, conditioned on a dataset of 10^4 MNIST handwritten digit images. From each image, the authors extracted 50 random projections of the raw pixel values as covariates and a label indicating whether the image was a 7 or a 9. After discarding the first half of sample points as burn-in, we obtained regression coefficient samples with 5×10^4 points and $d = 51$ dimensions (including the intercept term). Figure 4 displays the IMQ KSD applied to the first n points in each sample. As external validation, we follow the protocol of Ahn et al. (2012) to find the bivariate marginal means and 95% confidence ellipses of each sample that align best and worst with those of a surrogate ground truth sample obtained from a

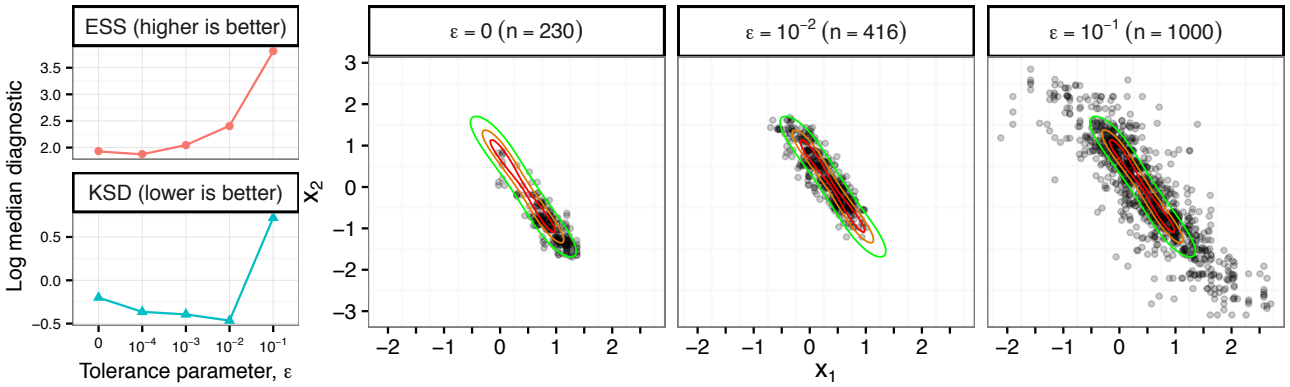

Figure 3. **Left:** Median hyperparameter selection criteria across 50 independent approximate slice sampler sample sequences (see Section 4.3); IMQ KSD selects $\epsilon = 10^{-2}$; effective sample size selects $\epsilon = 10^{-1}$. **Right:** Representative approximate slice sampler samples requiring 148000 likelihood evaluations with posterior equidensity contours overlaid; n is the associated sample size.

Hamiltonian Monte Carlo chain with 10^5 iterates. Both the KSD and the surrogate ground truth suggest that the moderate speed-up provided by SGFS-d ($0.0017s$ per sample vs. $0.0019s$ for SGFS-f) is outweighed by the significant loss in inferential accuracy. However, the KSD assessment does not require access to an external trustworthy ground truth sample. The longest KSD computation took $400s$ using 16 cores.

4.5. Beyond sample quality comparison

While our investigation of the KSD was motivated by the desire to develop practical, trustworthy tools for sample quality comparison, the kernels recommended by our theory can serve as drop-in replacements in other inferential tasks that make use of kernel Stein discrepancies.

4.5.1. One-sample hypothesis testing

Chwialkowski et al. (2016) recently used the KSD $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k)$ to develop a hypothesis test of whether a given sample from a Markov chain was drawn from a target distribution P (see also Liu et al., 2016). However, the authors noted that the KSD test with their default Gaussian base kernel k experienced a considerable loss of power as the dimension d increased. We recreate their experiment and show that this loss of power can be avoided by using our default IMQ kernel with $\beta = -\frac{1}{2}$ and $c = 1$. Following (Chwialkowski et al., 2016, Section 4) we draw $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ and $u_i \overset{\text{iid}}{\sim} \text{Unif}[0, 1]$ to generate a sample $(x_i)_{i=1}^n$ with $x_i = z_i + u_i\, e_1$ for $n = 500$ and various dimensions d. Using the authors' code (modified to include an IMQ kernel), we compare the power of the Gaussian KSD test, the IMQ KSD test, and the standard normality test of Baringhaus & Henze (1988) (B&H) to discern whether the sample $(x_i)_{i=1}^{500}$ came from the null distribution $P = \mathcal{N}(0, I_d)$. The results, averaged over 400 simulations, are shown in Table 1. Notably, the IMQ KSD experiences no power degradation over this range of dimensions, thus improving on both the Gaussian KSD and the standard B&H normality tests.

Table 1. Power of one sample tests for multivariate normality, averaged over 400 simulations (see Section 4.5.1)

	d=2	d=5	d=10	d=15	d=20	d=25
B&H	1.0	1.0	1.0	0.91	0.57	0.26
Gaussian	1.0	1.0	0.88	0.29	0.12	0.02
IMQ	1.0	1.0	1.0	1.0	1.0	1.0

4.5.2. Improving sample quality

Liu & Lee (2016) recently used the KSD $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k)$ as a means of improving the quality of a sample. Specifically, given an initial sample Q_n supported on $x_1, \dots, x_n$, they minimize $\mathcal{S}(\tilde{Q}_n, \mathcal{T}_P, \mathcal{G}_k)$ over all measures $\tilde{Q}_n$ supported on the same sample points to obtain a new sample that better approximates P over the class of test functions $\mathcal{H} = \mathcal{T}_P \mathcal{G}_k$. In all experiments, Liu & Lee (2016) employ a Gaussian kernel $k(x, y) = e^{-\frac{1}{h}\|x-y\|_2^2}$ with bandwidth h selected to be the median of the squared Euclidean distance between pairs of sample points. Using the authors' code, we recreate the experiment from (Liu & Lee, 2016, Fig. 2b) and introduce a KSD objective with an IMQ kernel $k(x, y) = (1 + \frac{1}{h}\|x - y\|_2^2)^{-1/2}$ with bandwidth selected in the same fashion. The starting sample is given by $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ for $n = 100$, various dimensions d, and each sample point drawn i.i.d. from $P = \mathcal{N}(0, I_d)$. For the initial sample and the optimized samples produced by each KSD, Figure 5 displays the mean squared error (MSE) $\frac{1}{d}\|\mathbb{E}_P[Z] - \mathbb{E}_{\tilde{Q}_n}[X]\|_2^2$ averaged across 500 independently generated initial samples. Out of the box, the IMQ kernel produces better mean estimates than the standard Gaussian.

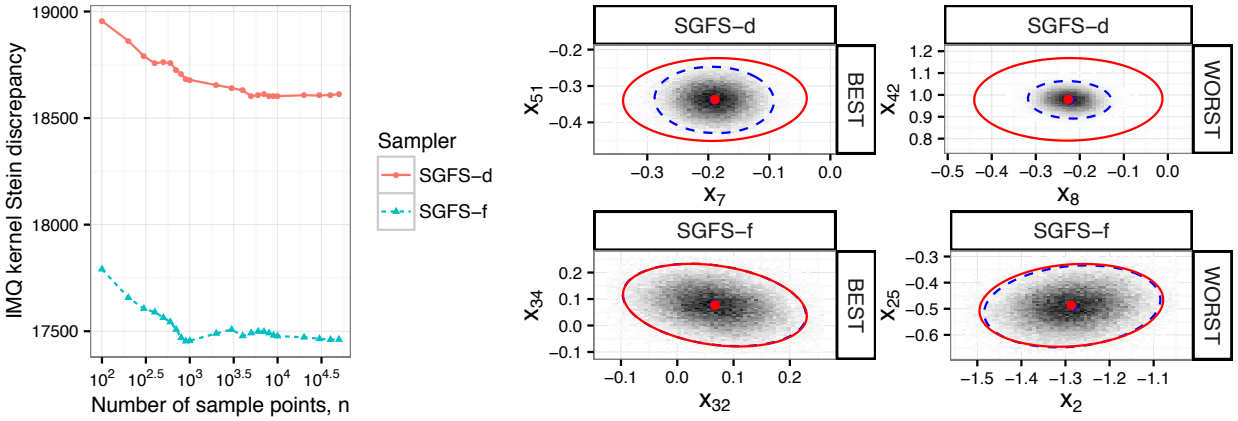

Figure 4. **Left:** Quality comparison for Bayesian logistic regression with two SGFS samplers (see Section 4.4). **Right:** Scatter plots of $n = 5 \times 10^4$ SGFS sample points with overlaid bivariate marginal means and 95% confidence ellipses (dashed blue) that align best and worst with surrogate ground truth sample (solid red).

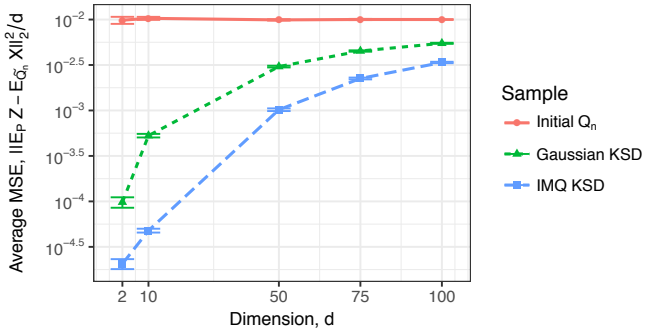

Figure 5. Average quality of mean estimates (± 2 standard errors) under optimized samples $\tilde{Q}_n$ for target $P = \mathcal{N}(0, I_d)$; MSE averaged over 500 independent initial samples (see Section 4.5.2).

5. Related and future work

The score statistic of Fan et al. (2006) and the Gibbs sampler convergence criteria of Zellner & Min (1995) detect certain forms of non-convergence but fail to detect others due to the finite number of test functions tested. For example, when $P = \mathcal{N}(0, 1)$, the score statistic (Fan et al., 2006) only monitors sample means and variances.

For an approximation μ with continuously differentiable density r, Chwialkowski et al. (2016, Thm. 2.1) and Liu et al. (2016, Prop. 3.3) established that if k is C_0-universal (Carmeli et al., 2010, Defn. 4.1) or integrally strictly positive definite (ISPD, Stewart, 1976, Sec. 6) and $\mathbb{E}_\mu[k_0(X, X) + \|\nabla \log \frac{p(X)}{r(X)}\|_2^2] < \infty$ for $k_0 \triangleq \sum_{j=1}^d k_0^j$, then $\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k) = 0$ only if $\mu = P$. However, this property is insufficient to conclude that probability measures with small KSD are close to P in any traditional sense. Indeed, Gaussian and Matérn kernels are C_0 universal and ISPD, but, by Theorem 6, their KSDs can be driven to zero by sequences not converging to P. On compact domains, where tightness is no longer an issue, the combined results of (Oates et al., 2016a, Lem. 4), (Fukumizu et al., 2007, Lem. 1), and (Simon-Gabriel & Schölkopf, 2016, Thm. 55) give conditions for a KSD to dominate weak convergence.

While assessing sample quality was our chief objective, our results may hold benefits for other applications that make use of Stein discrepancies or Stein operators. In particular, our kernel recommendations could be incorporated into the Monte Carlo control functionals framework of Oates et al. (2016b); Oates & Girolami (2015), the variational inference approaches of Liu & Wang (2016); Liu & Feng (2016); Ranganath et al. (2016), and the Stein generative adversarial network approach of Wang & Liu (2016).

In the future, we aim to leverage stochastic, low-rank, and sparse approximations of the kernel matrix and score function to produce KSDs that scale better with the number of sample and data points while still guaranteeing control over weak convergence. A reader may also wonder for which distributions outside of $\mathcal{P}$ the KSD dominates weak convergence. The following theorem, proved in Section J, shows that no KSD with a C_0 kernel dominates weak convergence when the target has a bounded score function.

Theorem 10 (KSD fails for bounded scores)**.** *If $\nabla \log p$ is bounded and $k \in C_0^{(1,1)}$, then $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$ does not imply $Q_n \Rightarrow P$.*

However, Gorham et al. (2016) developed convergence-determining graph Stein discrepancies for heavy-tailed targets by replacing the Langevin Stein operator $\mathcal{T}_P$ with *diffusion Stein operators* of the form $(\mathcal{T}g)(x) = \frac{1}{p(x)}\langle \nabla, p(x)(a(x) + c(x))g(x)\rangle$. An analogous construction should yield convergence-determining *diffusion KSDs* for P outside of $\mathcal{P}$. Our results also extend to targets P supported on a convex subset $\mathcal{X}$ of $\mathbb{R}^d$ by choosing k to satisfy $p(x)k(x, \cdot) \equiv 0$ for all x on the boundary of $\mathcal{X}$.

References

Ahn, S., Korattikara, A., and Welling, M. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proc. 29th ICML*, ICML'12, 2012.

Bachman, G. and Narici, L. *Functional Analysis*. Academic Press textbooks in mathematics. Dover Publications, 1966. ISBN 9780486402512.

Baker, J. Integration of radial functions. *Mathematics Magazine*, 72(5):392–395, 1999.

Barbour, A. D. Stein's method and Poisson process convergence. *J. Appl. Probab.*, (Special Vol. 25A):175–184, 1988. ISSN 0021-9002. A celebration of applied probability.

Barbour, A. D. Stein's method for diffusion approximations. *Probab. Theory Related Fields*, 84(3):297–322, 1990. ISSN 0178-8051. doi: 10.1007/BF01197887.

Baringhaus, L. and Henze, N. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35(1):339–348, 1988.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V.B. Julia: A fresh approach to numerical computing. *arXiv preprint arXiv:1411.1607*, 2014.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

Chatterjee, S. and Shao, Q. Nonnormal approximation by Stein's method of exchangeable pairs with application to the Curie-Weiss model. *Ann. Appl. Probab.*, 21(2):464–483, 2011. ISSN 1050-5164. doi: 10.1214/10-AAP712.

Chen, L., Goldstein, L., and Shao, Q. *Normal approximation by Stein's method*. Probability and its Applications. Springer, Heidelberg, 2011. ISBN 978-3-642-15006-7. doi: 10.1007/978-3-642-15007-4.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *Proc. 33rd ICML*, ICML, 2016.

DuBois, C., Korattikara, A., Welling, M., and Smyth, P. Approximate slice sampling for Bayesian posterior inference. In *Proc. 17th AISTATS*, pp. 185–193, 2014.

Eberle, A. Reflection couplings and contraction rates for diffusions. *Probab. Theory Related Fields*, pp. 1–36, 2015. doi: 10.1007/s00440-015-0673-1.

Fan, Y., Brooks, S. P., and Gelman, A. Output assessment for Monte Carlo simulations via the score statistic. *J. Comp. Graph. Stat.*, 15(1), 2006.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *NIPS*, volume 20, pp. 489–496, 2007.

Geyer, C. J. Markov chain Monte Carlo maximum likelihood. *Computer Science and Statistics: Proc. 23rd Symp. Interface*, pp. 156–163, 1991.

Gorham, J. and Mackey, L. Measuring sample quality with Stein's method. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Adv. NIPS 28*, pp. 226–234. Curran Associates, Inc., 2015.

Gorham, J., Duncan, A., Vollmer, S., and Mackey, L. Measuring sample quality with diffusions. *arXiv:1611.06972*, Nov. 2016.

Götze, F. On the rate of convergence in the multivariate CLT. *Ann. Probab.*, 19(2):724–739, 1991.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(1):723–773, 2012.

Herb, R. and Sally Jr., P.J. The Plancherel formula, the Plancherel theorem, and the Fourier transform of orbital integrals. In *Representation Theory and Mathematical Physics: Conference in Honor of Gregg Zuckerman's 60th Birthday, October 24–27, 2009, Yale University*, volume 557, pp. 1. American Mathematical Soc., 2011.

Korattikara, A., Chen, Y., and Welling, M. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proc. of 31st ICML*, ICML'14, 2014.

Ley, C., Reinert, G., and Swan, Y. Stein's method for comparison of univariate distributions. *Probab. Surveys*, 14: 1–52, 2017. doi: 10.1214/16-PS278.

Liu, Q. and Feng, Y. Two methods for wild variational inference. *arXiv preprint arXiv:1612.00081*, 2016.

Liu, Q. and Lee, J. Black-box importance sampling. *arXiv:1610.05247*, October 2016. To appear in AISTATS 2017.

Liu, Q. and Wang, D. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *arXiv:1608.04471*, August 2016.

Liu, Q., Lee, J., and Jordan, M. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proc. of 33rd ICML*, volume 48 of *ICML*, pp. 276–284, 2016.

Mackey, L. and Gorham, J. Multivariate Stein factors for a class of strongly log-concave distributions. *Electron. Commun. Probab.*, 21:14 pp., 2016. doi: 10.1214/16-ECP15.

Müller, A. Integral probability metrics and their generating classes of functions. *Ann. Appl. Probab.*, 29(2):pp. 429–443, 1997.

Oates, C. and Girolami, M. Control functionals for Quasi-Monte Carlo integration. *arXiv:1501.03379*, 2015.

Oates, C., Cockayne, J., Briol, F., and Girolami, M. Convergence rates for a class of estimators based on steins method. *arXiv preprint arXiv:1603.03220*, 2016a.

Oates, C. J., Girolami, M., and Chopin, N. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. n/a–n/a, 2016b. ISSN 1467-9868. doi: 10.1111/rssb.12185.

Ranganath, R., Tran, D., Altosaar, J., and Blei, D. Operator variational inference. In *Advances in Neural Information Processing Systems*, pp. 496–504, 2016.

Simon-Gabriel, C. and Schölkopf, B. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *arXiv preprint arXiv:1604.05251*, 2016.

Sriperumbudur, B. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22 (3):1839–1893, 2016.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11 (Apr):1517–1561, 2010.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pp. 583–602. Univ. California Press, Berkeley, Calif., 1972.

Stein, C., Diaconis, P., Holmes, S., and Reinert, G. Use of exchangeable pairs in the analysis of simulations. In *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser.*, pp. 1–26. Inst. Math. Statist., Beachwood, OH, 2004.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer Science & Business Media, 2008.

Stewart, J. Positive definite functions and generalizations, an historical survey. *Rocky Mountain J. Math.*, 6(3):409–434, 09 1976. doi: 10.1216/RMJ-1976-6-3-409.

Vallender, S. Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Appl.*, 18(4):784–786, 1974.

Wainwright, M. *High-dimensional statistics: A non-asymptotic viewpoint*. 2017. URL http://www.stat.berkeley.edu/~wainwrig/nachdiplom/Chap5_Sep10_2015.pdf.

Wang, D. and Liu, Q. Learning to Draw Samples: With Application to Amortized MLE for Generative Adversarial Learning. *arXiv:1611.01722*, November 2016.

Welling, M. and Teh, Y. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, 2011.

Wendland, H. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Zellner, A. and Min, C. Gibbs sampler convergence criteria. *JASA*, 90(431):921–927, 1995.