Stein Points

Wilson Ye Chen [1] **Lester Mackey** [2] **Jackson Gorham** [3] **François-Xavier Briol** [4 5 6] **Chris. J. Oates** [7 6]

Abstract

An important task in computational statistics and machine learning is to approximate a posterior distribution $p(x)$ with an empirical measure supported on a set of representative points $\{x_i\}_{i=1}^n$. This paper focuses on methods where the selection of points is essentially deterministic, with an emphasis on achieving accurate approximation when n is small. To this end, we present *Stein Points*. The idea is to exploit either a greedy or a conditional gradient method to iteratively minimise a kernel Stein discrepancy between the empirical measure and $p(x)$. Our empirical results demonstrate that Stein Points enable accurate approximation of the posterior at modest computational cost. In addition, theoretical results are provided to establish convergence of the method.

1. Introduction

This paper is motivated by approximation of a Borel distribution P, defined on a topological space X, with deterministic point sets or sequences $\{x_i\}_{i=1}^n \subset X$ for $n \in \mathbb{N}$, such that

$$\tfrac{1}{n} \textstyle\sum_{i=1}^n h(x_i) \quad \to \quad \int h \, \mathrm{d}P \tag{1}$$

as $n \to \infty$ for all functions $h : X \to \mathbb{R}$ in a specified set $\mathcal{H}$. Throughout it will be assumed that P admits a density p, with respect to a reference measure, available in a form that is un-normalised (i.e., we know $q(x)$ in closed form where $p(x) = q(x)/C$ for some $C > 0$). Such problems occur in Bayesian statistics where P represents a posterior distribution, and the integral represents a posterior expectation of interest. Markov chain Monte Carlo (MCMC) methods are extensively used for this task but suffer (in terms of accuracy) from 'clustering' of the points $\{x_i\}_{i=1}^n$ when n is small. This observation motivates us to instead consider a range of goal-oriented discrete approximation methods that are designed with un-normalised densities in mind.

The problem of discrete approximation of a distribution, given its normalised density, has been considered in detail and relevant methods include quasi-Monte Carlo (QMC) (Dick & Pillichshammer, 2010), kernel herding (Chen et al., 2010; Lacoste-Julien et al., 2015), support points (Mak & Joseph, 2016; 2017), transport maps (Marzouk et al., 2016), and minimum energy methods (Johnson et al., 1990). On the other hand, the question of how to proceed with un-normalised densities has been primarily answered with increasingly sophisticated MCMC.

At the same time, recent work had led to theoretically-justified measures of sample quality in the case of an un-normalised target. In (Gorham & Mackey, 2015; Mackey & Gorham, 2016) it was shown that Stein's method can be used to construct discrepancy measures that control weak convergence of an empirical measure to a target. This was later extended in (Gorham & Mackey, 2017) to encompass a family of discrepancy measures indexed by a reproducing kernel. In the latter case, the discrepancy measure can be recognised as a maximum mean discrepancy (Smola et al., 2007). As such, one can consider discrete approximation as an optimisation problem in a Hilbert space and attempt to optimise this objective with either a greedy or a conditional gradient method. The resulting method – *Stein Points* – and its variants are proposed and studied in this work.

Our Contribution This paper makes the following contributions:

- Two algorithms are proposed for minimisation of the kernel Stein discrepancy (KSD; Chwialkowski et al., 2016; Liu et al., 2016; Gorham & Mackey, 2017); a greedy algorithm and a conditional gradient method. In each case, a convergence result of the form in Eqn. 1 is established.

- Novel kernels are proposed for the KSD, and we prove that, with these kernels, the KSD controls weak convergence of the empirical measure to the target. In other

[1] School of Mathematical and Physical Sciences, University of Technology Sydney, Australia [2] Microsoft Research New England, USA [3] Opendoor Labs, Inc., USA [4] Department of Statistics, University of Warwick, UK [5] Department of Mathematics, Imperial College London, UK [6] Alan Turing Institute, UK [7] School of Mathematics, Statistics and Physics, Newcastle University, UK. Correspondence to: Wilson Ye Chen <ye.chen@uts.edu.au>, Lester Mackey <lmackey@microsoft.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

words, the test functions h for which our results hold constitute a rich set $\mathcal{H}$.

Outline The paper proceeds as follows. In Section 2 we provide background, and in Section 3 we present the approximation methods that will be studied. Section 4 applies these methods to both simulated and real approximation problems and provides a extensive empirical comparison. All technical material is contained in Section 5, where we derive novel theoretical results for the methods we proposed. Finally we summarise our findings in Section 6.

2. Background

Throughout this section it will be assumed that X is a metric space, and we let $\mathcal{P}(X)$ denote the collection of Borel distributions on X. In this context, weak convergence of the empirical measure to P corresponds to taking the set $\mathcal{H}$ in Eqn. 1 to be the set $\mathcal{H}_{\text{CB}}$ of functions which are continuous and bounded. In this work we also consider sets $\mathcal{H}$ that correspond to stronger modes of convergence in $\mathcal{P}(X)$.

First, in 2.1, we recall how discrepancy measures are constructed. Then we recall the use of Stein's method in this context in 2.2. Formulae for KSD are presented in 2.3.

2.1. Discrepancy Measures

A *discrepancy* is a quantification of how well the points $\{x_i\}_{i=1}^n$ cover the domain X with respect to the distribution P. This framework will be developed below in reproducing kernel Hilbert spaces (RKHS; Hickernell, 1998), but the general theory of discrepancy can be found in (Dick & Pillichshammer, 2010). Note that we focus on unweighted point sets for ease of presentation, but our discussions and results generalise straightforwardly to point sets that are weighted.

Let $k : X \times X \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{K}$ of functions $\mathcal{X} \to \mathbb{R}$. That is, $\mathcal{K}$ is a Hilbert space of functions with inner product $\langle\cdot,\cdot\rangle_{\mathcal{K}}$ and induced norm $\|\cdot\|_{\mathcal{K}}$ such that, for all $x \in X$, $k(x,\cdot) \in \mathcal{K}$ and $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{K}}$ whenever $f \in \mathcal{K}$. The Cauchy-Schwarz inequality in $\mathcal{K}$ gives that

$$\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n f(x_i) - \int f \mathrm{d}P\right| \quad \leq \quad \|f\|_{\mathcal{K}}\, D_{\mathcal{K},P}\left(\{x_i\}_{i=1}^n\right)$$

where the final term

$$D_{\mathcal{K},P}\left(\{x_i\}_{i=1}^n\right) \;:=\; \left\|\tfrac{1}{n}\textstyle\sum_{i=1}^n k(x_i,\cdot) - \int k(x,\cdot)\mathrm{d}P(x)\right\|_{\mathcal{K}}$$

is the canonical discrepancy measure for the RKHS. The Bochner integral $k_P := \int k(x,\cdot)\mathrm{d}P(x) \in \mathcal{K}$ is known as the *mean embedding* of P into $\mathcal{K}$ (Smola et al., 2007). Thus, if $\mathcal{H} = B(\mathcal{K}) := \{f \in \mathcal{K} : \|f\|_{\mathcal{K}} \leq 1\}$ is the unit ball in $\mathcal{K}$, then $D_{\mathcal{K},P}\left(\{x_i\}_{i=1}^n\right) \to 0$ implies the convergence result in Eqn. 1.

The RKHS framework is now standard for QMC analysis (Dick & Pillichshammer, 2010). Its popularity derives from the fact that, when both k_P and $k_{P,P} := \int k_P \mathrm{d}P$ are explicit, the canonical discrepancy measure is also explicit:

$$D_{\mathcal{K},P}(\{x_i\}_{i=1}^n) \quad = \qquad (2)$$
$$\sqrt{k_{P,P} - \tfrac{2}{n}\textstyle\sum_{i=1}^n k_P(x_i) + \tfrac{1}{n^2}\sum_{i,j=1}^n k(x_i,x_j)}$$

Table 1 in (Briol et al., 2015) collates pairs (k,P) for which k_P and $k_{P,P}$ are explicit.

If P is a posterior distribution, so that p has unknown normalisation constant, it is unclear how the terms k_P and $k_{P,P}$ can be computed in closed form, and so similarly for the discrepancy $D_{\mathcal{K},P}$. This has so far prevented QMC and related methods such as kernel herding (Chen et al., 2010) from being used to compute posterior integrals. A solution to this problem can be found in Stein's method, presented next.

2.2. Kernel Stein Discrepancy

The method of Stein (1972) was introduced as an analytical tool for establishing convergence in distribution of random variables, but its potential for generating and analyzing computable discrepancies was developed in (Gorham & Mackey, 2015). In what follows, we recall the kernelised version of the *Stein discrepancy*, first presented for an optimally-weighted point set in 2.3.3 of (Oates et al., 2017b) and later generalised to an arbitrarily-weighted point set in (Chwialkowski et al., 2016; Liu et al., 2016; Gorham & Mackey, 2017).

Suppose that X carries the structure of a smooth manifold, and consider a linear differential operator $\mathcal{T}_P$ on X, together with a set $\mathcal{F}$ of sufficiently differentiable functions, with the following property:

$$\int \mathcal{T}_P[f]\,\mathrm{d}P \quad = \quad 0 \quad \forall f \in \mathcal{F}. \qquad (3)$$

Then $\mathcal{T}_P$ is called a *Stein operator* and $\mathcal{F}$ a *Stein set*. In the kernelised version of Stein's method, the set $\mathcal{F}$ is either an RKHS $\mathcal{K}$ with reproducing kernel $k : X \times X \to \mathbb{R}$, or the product $\mathcal{K}^d$, which contains vector-valued functions $f = (f_1,\ldots,f_d)$ with $f_j \in \mathcal{K}$ and is equipped with a norm[1] $\|f\|_{\mathcal{K}^d} = (\sum_{j=1}^d \|f_j\|_{\mathcal{K}}^2)^{1/2}$. For the case $\mathcal{F} = \mathcal{K}$, the image of $\mathcal{K}$ under a Stein operator $\mathcal{T}_P$ is denoted $\mathcal{K}_0 = \mathcal{T}_P\mathcal{K}$. The notation can be justified since, under appropriate regularity assumptions, the set $\mathcal{T}_P\mathcal{K}$ admits structure from the reproducing kernel $k_0(x,x') = \mathcal{T}_P\overline{\mathcal{T}_P}k(x,x')$ (Oates et al., 2017b). Here $\overline{\mathcal{T}_P}$ is the adjoint of the operator $\mathcal{T}_P$ and acts on the second argument x' of the kernel. If instead $\mathcal{F} = \mathcal{K}^d$, then we suppose that $\mathcal{T}_P f = \sum_{j=1}^d \mathcal{T}_{P,j} f_j$ so that the set

[1] For what follows, any vector norm can be used to combine the component norms $\|f_j\|_{\mathcal{K}}$ (Gorham & Mackey, 2017, Prop. 3).

$\mathcal{K}_0 = \mathcal{T}_P \mathcal{K}^d$ admits structure from the reproducing kernel $k_0(x,x') = \sum_{j=1}^d \mathcal{T}_{P,j} \overline{\mathcal{T}_{P,j}} k(x,x')$. In either case, we will call the reproducing kernel k_0 of $\mathcal{K}_0$ a *Stein reproducing kernel.*

Stein reproducing kernels possess the useful property that $k_{0,P} = \int k_0(x,\cdot)\mathrm{d}P = 0$ and $k_{0,P,P} = \int k_{0,P}\mathrm{d}P = 0$, so in particular both are explicit. Thus, if k_0 is a Stein reproducing kernel, then Eqn. 2 can be simplified:

$$D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n) \quad = \quad \sqrt{\tfrac{1}{n^2}\sum_{i,j=1}^n k_0(x_i,x_j)}. \qquad (4)$$

We call this quantity a *kernel Stein discrepancy* (KSD). Next, we exhibit some differential operators for which Eqn. 3 is satisfied and Eqn. 4 can be computed.

2.3. Stein Operators and Their Reproducing Kernels

The divergence theorem can be used to construct Stein operators on a manifold. For P supported on $\mathcal{X} = \mathbb{R}^d$, (Oates et al., 2017b; Gorham & Mackey, 2015; Chwialkowski et al., 2016; Liu et al., 2016; Gorham & Mackey, 2017) considered the *Langevin Stein operator*

$$\mathcal{T}_P f \quad := \quad \frac{\nabla\cdot(pf)}{p} \qquad (5)$$

where $\nabla\cdot$ is the usual divergence operator and $f \in \mathcal{K}^d$. Thus, for the Langevin Stein operator, we obtain a Stein reproducing kernel

$$\begin{aligned} k_0(x,x') \quad = \quad & \nabla_x \cdot \nabla_{x'} k(x,x') \qquad (6) \\ & + \nabla_x k(x,x') \cdot \nabla_{x'} \log p(x') \\ & + \nabla_{x'} k(x,x') \cdot \nabla_x \log p(x) \\ & + k(x,x') \nabla_x \log p(x) \cdot \nabla_{x'} \log p(x'). \end{aligned}$$

To evaluate this kernel, the normalisation constant for p is not required. Other Stein operators for the Euclidean case were developed in (Gorham et al., 2016). For P supported on a closed Riemannian manifold $\mathcal{X}$, (Oates et al., 2017a; Liu & Zhu, 2017) proposed the second order Stein operator $\mathcal{T}_P f := \frac{1}{p}\nabla\cdot(p\nabla f)$ where ∇ and $\nabla\cdot$ are, respectively, the gradient and divergence operators on the manifold and $f \in \mathcal{K}$. Other Stein operators for the general case are proposed in the supplement of (Oates et al., 2017a).

The theoretical results in (Gorham & Mackey, 2017) established that certain combinations of Stein operator $\mathcal{T}_P$ and base kernel k ensure that KSD controls weak convergence; that is, $D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n) \to 0$ implies that Eqn. 1 holds with $\mathcal{H} = \mathcal{H}_{\mathrm{CB}}$. This important result motivates our next contribution, where numerical optimisation methods are used to select points $\{x_i\}_{i=1}^n$ to approximately minimise KSD. Theoretical analysis of the proposed methods is reserved for Section 5.

3. Methods

In this paper, two algorithms to select points $\{x_i\}_{i=1}^n$ are studied in detail. The first of these is a greedy algorithm, which at each iteration attempts to minimise the KSD, whilst the second is a conditional gradient algorithm, known as *herding*, which also targets the KSD. In 3.1 and 3.2 the two algorithms are described, whilst in 3.3 some alternative approaches are briefly discussed.

3.1. Greedy Algorithm

The simplest algorithm that we consider follows a greedy strategy, whereby the first point x_1 is taken to be a global maximum of p (an operation which does not require the normalisation constant) and each subsequent point x_n is taken to be a global minimum of $D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n)$, with the KSD being viewed as a function of x_n holding $\{x_i\}_{i=1}^{n-1}$ fixed. Equivalently, at iteration $n > 1$ of the greedy algorithm, we select

$$x_n \quad \in \quad \arg\min_{x\in\mathcal{X}} \quad \frac{k_0(x,x)}{2} + \sum_{i=1}^{n-1} k_0(x_i,x). \quad (7)$$

Note that each iteration of the algorithm requires the solution of a global optimisation problem over $\mathcal{X}$; in practice we employed a numerical optimisation method, and this choice is discussed in detail in connection with the empirical results in Section 4 and the theoretical results in Section 5.

If a user has a budget of at most n points, the greedy algorithm can be run for n iterations and thereafter improved using (block) coordinate descent on the KSD objective to update an existing point x_i instead of introducing a new point. The cost of each update is equal to the cost of adding the n-th greedy Stein Point. This budget-constrained variant of the method will be called *Stein Greedy-n* in the sequel (see Section B.1.3 for more details).

3.2. Herding Algorithm

The definition of discrepancy in Section 2.1 suggests that selection of $\{x_i\}_{i=1}^n$ can be elegantly formulated as a single global optimisation problem over $\mathcal{K}_0$. Let $M(\mathcal{K}_0)$ be the *marginal polytope* of $\mathcal{K}_0$; i.e. the convex hull of the set $\{k_0(x,\cdot)\}_{x\in\mathcal{X}}$ (Wainwright & Jordan, 2008). The mean embedding $Q \mapsto k_Q$, as a map $\mathcal{P}(\mathcal{X}) \to M(\mathcal{K})$, is injective whenever the kernel k is universal and $\mathcal{X}$ is compact (Smola et al., 2007), so that in this case k_Q fully characterises Q. Results in a similar direction for Stein reproducing kernels were established in Chwialkowski et al. (2016, Theorem 2.1) and Liu et al. (2016, Proposition 3.3). Thus, as P is mapped to 0 under the embedding, we are motivated to consider non-trivial solutions to

$$\arg\min_{f\in M(\mathcal{K}_0)} J(f), \qquad J(f) := \tfrac{1}{2}\|f\|_{\mathcal{K}_0}^2. \qquad (8)$$

As might be expected, the objective function is closely related to KSD; for $f(\cdot) = \frac{1}{n}\sum_{i=1}^n k_0(x_i,\cdot)$ we have $J(f) =$

$\frac{1}{2}D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n)^2$. An iterative algorithm, called *kernel herding*, was proposed in (Chen et al., 2010) to solve problems in the form of Eqn. 8. This was later shown to be equivalent to a conditional gradient algorithm, the *Frank-Wolfe algorithm*, in (Bach et al., 2012). The canonical Frank-Wolfe algorithm, which results in an unweighted point set (as opposed to a more general weighted point set; Bach et al., 2012), is presented next.

The first point x_1 is again taken to be a global maximum of p; this corresponds to an element $f_1 = k_0(x_1, \cdot) \in M(\mathcal{K}_0)$. Then, at iteration $n > 1$, the convex combination $f_n = \frac{n-1}{n} f_{n-1} + \frac{1}{n} f_n^* \in M(\mathcal{K}_0)$ is constructed where the element f_n^* encodes a direction of steepest descent:

$$f_n \quad \in \quad \arg\min_{f \in M(\mathcal{K}_0)} \langle f, \mathrm{D}J(f_{n-1}) \rangle_{\mathcal{K}_0} ,$$

where $\mathrm{D}J(f)$ is the representer of the Fréchet derivative of J at f. Given that minimisation of a linear objective over a convex set can be restricted to the boundary of that set, it follows that $f_n^* = k(x_n, \cdot)$ for some $x_n \in X$. Thus, at iteration $n > 1$ of the proposed algorithm, we select

$$x_n \quad \in \quad \arg\min_{x \in X} \quad \textstyle\sum_{i=1}^{n-1} k_0(x_i, x) \tag{9}$$

to obtain $f_n(\cdot) = \frac{1}{n} \sum_{i=1}^n k_0(x_i, \cdot)$, the embedding of the empirical distribution of $\{x_i\}_{i=1}^n$. As in the standard kernel herding algorithm of (Chen et al., 2010), each iteration in practice requires the solution of a global optimisation problem over X.

Compared to Eqn. 7, the greedy algorithm is seen to be a regularised version of herding with regulariser $\frac{1}{2}k_0(x,x)$. The two algorithms coincide if $k_0(x,x)$ is independent of x; however, this is typically not true for a Stein reproducing kernel. The computational cost of either method is $O(n^2)$; thus we anticipate applications in which evaluation of $p(x)$ (and its gradient) constitute the principal computational bottleneck. The performance of both algorithms is studied empirically in Section 4 and theoretically in Section 5. In a similar manner to *Stein Greedy-n*, a budget-constrained variant of the above method can be considered, which we call *Stein Herding-n* in the sequel.

3.3. Other Algorithms

The output of either of our algorithms will be called *Stein Points*. These are *extensible* point sequence $S_n = (x_i)_{i=1}^n$, meaning that S_n can be incrementally extended $S_n = (S_{n-1}, x_n)$ as required. Another recently proposed extensible method is the (sequential) minimum energy design (MED) of (Joseph et al., 2015; 2017), here used as a benchmark.

For some problems the number of points n will be fixed in advance and the aim will instead be to select a single optimal point set $\{x_i\}_{i=1}^n$. This alternative problem demands different methodologies, and a promising method in this direction is Stein variational gradient descent (SVGD-n; Liu & Wang, 2016; Liu, 2017). A natural point set analogue of our approach would be to optimise KSD for n fixed. This approach was considered for other discrepancy measures in (Oettershagen, 2017), where the Newton method was used. We instead employ our budget-constrained algorithms Stein Greedy-n and Stein Herding-n for this use case.

4. Results

In this section, the proposed greedy and herding algorithms are empirically assessed and compared. In 4.2 a Gaussian mixture problem is studied in detail, whilst in 4.3 and 4.4, respectively, the methods are applied to approximate the parameter posterior in a non-parametric regression model and an IGARCH model. First, in 4.1 we provide details on the experimental protocol.

4.1. Experimental Protocol

Here we describe the parameters and settings that were varied in the experiments that are presented.

Stein Operator To limit scope, we focus on the case $X = \mathbb{R}^d$ and always take $\mathcal{T}_P$ to be the Langevin Stein operator in Eqn. 5.

Choice of Kernel For the kernel k in Eqn. 6 we considered one standard choice – the inverse multi-quadric (IMQ) kernel – together with two novel alternatives:

(k_1) (IMQ) $k_1(x, x') = \left(\alpha + \|x - x'\|_2^2\right)^\beta$

(k_2) (inverse log) $k_2(x, x') = \left(\alpha + \log(1 + \|x - x'\|_2^2)\right)^{-1}$

(k_3) (IMQ score)
$$k_3(x, x') = \left(\alpha + \|\nabla \log p(x) - \nabla \log p(x')\|_2^2\right)^\beta .$$

In all cases $\alpha > 0$ and $\beta \in (-1, 0)$. To limit scope, in what follows we considered a finite number of judiciously selected configurations for α, β, though in principle these could be optimised as in (Jitkrittum et al., 2017). The best set of parameter values was selected for each algorithm and each target distribution, where the possible values were $\alpha \in \{0.1\eta, 0.5\eta, \eta, 2\eta, 4\eta, 8\eta\}$ and $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, with $\eta > 0$ problem-dependent (see the Supplement). The IMQ kernel, together with the Langevin Stein operator, was proven in Gorham & Mackey (2017, Theorem 8) to provide a KSD that controls weak convergence. Similar results for novel kernels k_2 and k_3 are established in Section 5.

Numerical Optimisation Method Any optimisation procedure could be used to (approximately) solve the global optimisation problem embedded in each iteration of the proposed algorithms. In our experiments, we considered the following numerical methods, for which full details appear in the Section B.2.

1. Nelder-Mead (`NM`): At iteration n, parallel runs of Nelder-Mead were employed, initialised at draws from a Gaussian mixture proposal centred on the current point set $\Pi = \frac{1}{n-1}\sum_{i=1}^{n-1}\mathcal{N}(x_i, \lambda I)$ with problem-specific $\lambda > 0$.
2. Monte Carlo (`MC`): The optimisation problem at iteration n was solved over a sample of points drawn from the same Gaussian mixture proposal Π.
3. Grid search (`GS`): Through brute force, the optimisation problem at iteration n was solved over a regular grid of width $\frac{1}{\sqrt{n}}$. This required $O(n^{-\frac{d}{2}})$ points; if required, the domain was first truncated with a large bounding box.

Performance Assessment To obtain a reasonably objective assessment, we focused on the 1-Wasserstein distance between the empirical measure and P:

$$W_P(\{x_i\}_{i=1}^n) \quad = \quad \sup_{h \in \mathcal{H}_{\text{Lip}}} \left| \frac{1}{n}\sum_{i=1}^n h(x_i) - \int h \mathrm{d}P \right|,$$

where $\mathcal{H}_{\text{Lip}}$ is the set of all function $h : X \to \mathbb{R}$ with Lipschitz constant $\text{Lip}(h) \leq 1$. By replacing P with the empirical measure $P_N = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$ for $y_i \overset{\text{iid}}{\sim} P$, the expected error from using $W_{P_N}(\{x_i\}_{i=1}^n)$ in lieu of $W_P(\{x_i\}_{i=1}^n)$ converges at a $N^{-\frac{1}{2}}\log N$ rate for $d = 2$ and $N^{-\frac{1}{d}}$ rate for $d > 2$ (Fournier & Guillin, 2015). By employing L_1-spanners, the approximation $W_{P_N}(\{x_i\}_{i=1}^n)$ can be computed in $O((n+N)^2 \log^{2d-1}(n+N))$ time (Gudmundsson et al., 2007). For all reported results, the $\{y_i\}_{i=1}^N$ were obtained by brute-force Monte Carlo methods applied to P, with N sufficiently large that approximation error can be neglected.

The computational cost associated to any given method was quantified as the total number n_{eval} of times either the log-density $\log p$ or its gradient $\nabla \log p$ were evaluated. This can be justified since in most applications the 'parameter to data' map dominates the computational cost associated with the likelihood.

Benchmarks Two existing methods were used as a benchmark:

1. The MED method of (Joseph et al., 2015; 2017) relies on numerical optimisation methods to minimise an energy measure $\mathcal{E}_{\delta,P}(\{x_i\}_{i=1}^n)$, adapted to P. This measure has one tuning parameter $\delta \in [1, \infty)$. See Section B.1.1 of the Supplement for full detail.
2. The SVGD method of (Liu & Wang, 2016; Liu, 2017) performs a version of gradient descent on the Kullback-Leibler divergence, described in Section B.1.2 of the Supplement.

To avoid confounding of the empirical results by incomparable algorithm parameters, (1) the collection of numerical optimisation methods used for KSD were also used for MED, and (2) the same collection of kernels $k_1, \ldots, k_3$ was considered for SVGD as was used for KSD. Note that, apart from standard Monte Carlo, none of the methods considered in these experiments are re-parametrisation invariant.

4.2. Gaussian Mixture Test

For our first test, we considered a Gaussian mixture model

$$P \quad = \quad \tfrac{1}{2}\mathcal{N}(\mu_1, \Sigma_1) + \tfrac{1}{2}\mathcal{N}(\mu_2, \Sigma_2)$$

defined on $X = \mathbb{R}^2$. Full settings for each of the methods considered are detailed in Section C.1 in the Supplement. Typical point sets are displayed over the contours of P for $\mu_1 = (-1.5, 0)$, $\mu_2 = (1.5, 0)$, $\Sigma_1 = \Sigma_2 = I$ in Figure 1. Additionally, point sets for the n point budget-constrained algorithms Stein Greedy-n and Stein Herding-n are presented in Figure 6 in the Supplement. For each of the methods shown in Figures 1 and 6, tuning parameters were varied and the overall performance was captured in Figure 2. It was observed that for (a-c) the choice of numerical optimisation method was the most influential tuning parameter, with the simpler Monte Carlo-based method being most successful. The kernels k_1, k_2 were seen to perform well, but in (a,b,d) the kernel k_3 was sometimes seen to fail.

A subjectively-selected exemplar was extracted for each method, and these 'best' results for each method are overlaid in Figure 3. The total number of points was limited to $n = 100$. In terms of our proposed methods, two qualitative regimes were observed: (i) For low computational budget $\log n_{\text{eval}} \leq 7$, the standard Monte Carlo method performed best. (ii) For a larger computational budget $7 < \log n_{\text{eval}}$, greedy Stein points were not out-performed.

Note that KSD and SVGD are based on the log target and its gradient, whilst for MED the target $p(x)$ itself is required. As a result, numerical instabilities were sometimes encountered with MED.

Next, we turned our attention to two important posterior approximation problems that occur in the real world.

4.3. Gaussian Process Regression Model

The Gaussian process (GP) model is a popular choice for uncertainty quantification in the non-parametric regression context (Rasmussen & Williams, 2006). The data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ that we considered are from a light de-

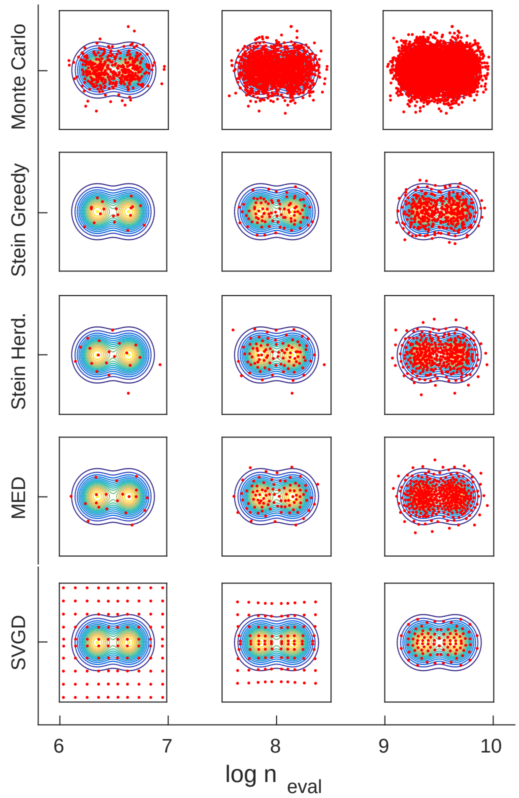
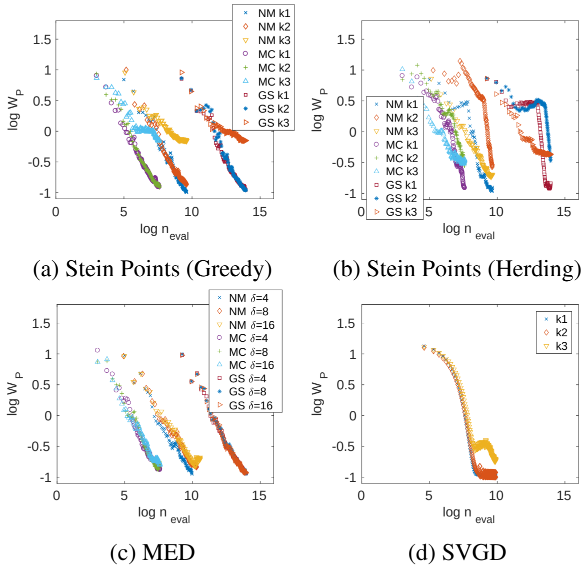


Figure 1: Typical point sets obtained in the Gaussian mixture test. [Here the left border of each sub-plot is aligned to the exact value of $\log n_{\text{eval}}$ spent to obtain each point set.]



(a) Stein Points (Greedy) (b) Stein Points (Herding)

(c) MED (d) SVGD

Figure 2: Results for the Gaussian mixture test. [Here $n = 100$. x-axis: log of the number n_{eval} of model evaluations that were used. y-axis: log of the Wasserstein distance $W_P(\{x_i\}_{i=1}^n)$ obtained. Kernel parameters α, β were optimised according to W_P in all cases, with sensitivities reported in Fig. 7 of the Supplement.]

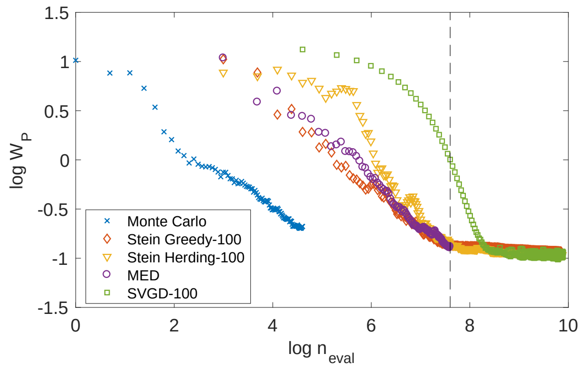


Figure 3: Combined results for the Gaussian mixture test. [Here $n = 100$. x-axis: log of the number n_{eval} of model evaluations that were used. y-axis: log of the the Wasserstein distance $W_P(\{x_i\}_{i=1}^n)$ obtained. Tuning parameters were selected to minimise W_P, as described in the main text. The dashed line indicates the point at which n Stein Points have been generated; block coordinate descent is performed thereafter to satisfy the n point budget constraint.]

tection and ranging (LIDAR) experiment (Ruppert et al., 2003). They consist of 221 realisations of an independent scalar variable x_i (distances travelled before the light is reflected back to its source) and a dependent scalar variable y_i (log-ratios of received light from two laser sources); these were modelled as $y_i = g(x_i) + \epsilon_i$, for $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ and a known value of σ. The unknown regression function g is modelled as a centred GP with covariance function $\text{cov}(x, x') = \theta_1 \exp(-\theta_2 (x - x')^2)$. The hyper-parameters $\theta_1, \theta_2 > 0$ determine the suitability of the GP model, but appropriate values will be unknown in general. In this experiment we re-parametrised $\phi_i = \log \theta_i$ and placed a standard multivariate Cauchy prior on $\phi = (\phi_1, \phi_2)$, defined on $X = \mathbb{R}^2$. The task is thus to approximate the conditional distribution $p(\phi|\mathcal{D})$. This problem is motivated by the computation of posterior predictive marginal distributions $p(y^*|x^*, \mathcal{D})$ for a new input x^*, which is defined as the integral $\int p(y^*|x^*, \phi, \mathcal{D}) p(\phi|\mathcal{D}) \mathrm{d}\phi$. Note that the density $p(\phi|\mathcal{D})$ can be differentiated, and an explicit formula is provided in Rasmussen & Williams (2006, Eqn. 5.9).

For each class of method, 'best' tuning parameters were selected and these are presented on the same plot in Figure 4a. In addition, typical point sets provided by each method are presented in Figures 8 and 9 in the Supplement. MED was not included because the method exhibited severe numerical instability on this task, as earlier discussed. Results indicated three qualitative regimes where, respectively, Monte Carlo, greedy Stein points and SVGD provided the best performance for fixed cost.

4.4. IGARCH Model

The integrated generalised autoregressive conditional heteroskedasticity (IGARCH) model is widely-used to describe financial time series (y_t) with time-varying volatility (σ_t) (Taylor, 2011). The model is as follows:

$$\begin{aligned} y_t &= \sigma_t \epsilon_t, \qquad \epsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \\ \sigma_t^2 &= \theta_1 + \theta_2 y_{t-1}^2 + (1-\theta_2)\sigma_{t-1}^2 \end{aligned}$$

with parameters $\theta = (\theta_1, \theta_2)$, $\theta_1 > 0$ and $0 < \theta_2 < 1$. The data $y = (y_t)$ that we considered were 2,000 daily percentage returns of the S&P 500 stock index (from December 6, 2005 to November 14, 2013), and an improper uniform prior was placed on θ. Thus the task was to approximate the posterior $p(\theta|y)$. Note that, whilst the domain $X = \mathbb{R}_+ \times (0,1)$ is bounded, for these data the posterior density is negligible on the boundary ∂X. This ensures that Eqn. 3 holds essentially to machine precision; see also the discussion in Oates et al. (2018, Section 3.2). For the IGARCH model, gradients $\nabla \log p(\theta|y)$ can be obtained as the solution of a recursive system of equations for $\partial \sigma_t / \partial \theta_2$.

As before, the 'best' performing of each class of method was selected and these are presented on the same plot in Figure 4b. In addition, typical point sets provided by each method are presented in Figures 12 and 13 in the Supplement. (Numerical instability again prevented results for MED from being obtained.) Results were consistent with the Gaussian mixture experiment, favouring either Monte Carlo or greedy Stein points depending on the computational budget.

5. Theoretical Results

In this section we establish two important forms of theoretical guarantees: (1) discrepancy control, i.e., $D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n) \to 0$ as $n \to \infty$ for our extensible Stein Point sequences and (2) distributional convergence control, i.e., for our kernel choices and appropriate choices of target, $D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n) \to 0$ implies that the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ converges in distribution to P.

5.1. Discrepancy Control

Earlier work has shown that, when a kernel is uniformly bounded (i.e., $\sup_{x \in X} k_0(x,x) \le R^2$), the greedy and kernel herding algorithms decrease the associated discrepancy $D_{\mathcal{K}_0,P}$ at an $O(n^{-\frac{1}{2}})$ rate (Lacoste-Julien et al., 2015; Jones, 1992). We extend these results to cover all growing, P-sub-exponential kernels.

Definition 1 (P-sub-exponential reproducing kernel). *We say a reproducing kernel k_0 is* P-sub-exponential *if*

$$\mathbb{P}_{Z \sim P}\left[k_0(Z,Z) \ge t\right] \le c_1 e^{-c_2 t}$$

for some constants $c_1, c_2 > 0$ and all $t \ge 0$.

Notably, any uniformly bounded reproducing kernel is P-sub-exponential, and, when P is a sub-Gaussian distribution, any kernel with at most quadratic growth (i.e., $k_0(x,x) = O(\|x\|_2^2)$) is also P-sub-exponential. Our first result, proved in Section A.1.1, shows that if we truncate the search domain suitably in each step, Stein Herding decreases the discrepancy at an $O(\sqrt{\log(n)/n})$ rate. This result holds even if each point x_i is selected suboptimally with error $\delta/2$. This extra degree of freedom allows a user to conduct a grid search or a search over appropriately generated random points on each step (see, e.g., Lacoste-Julien et al., 2015) and still obtain a rate of convergence.

Theorem 1 (Stein Herding Convergence). *Suppose k_0 with $k_{0,P} = 0$ is a P-sub-exponential reproducing kernel. Then there exist constants $c_1, c_2 > 0$ depending only on k_0 and P such that any point sequence $\{x_i\}_{i=1}^n$ satisfying*

$$\textstyle\sum_{i=1}^{j-1} k_0(x_i, x_j) \le \frac{\delta}{2} + \min_{x \in X: k_0(x,x) \le R_j^2} \sum_{i=1}^{j-1} k_0(x_i, x)$$

with $k_0(x_j, x_j) \le R_j^2 \in [2\log(j)/c_2, 2\log(n)/c_2]$ for each $1 \le j \le n$ also satisfies

$$D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n) \le e^{\pi/2}\sqrt{\tfrac{2\log(n)}{c_2 n} + \tfrac{c_1}{n} + \tfrac{\delta}{n}}.$$

Our next result, proved in Section A.1.2, shows that Stein Greedy decreases the discrepancy at an $O(\sqrt{\log(n)/n})$ rate whether we choose to truncate ($R_j < \infty$) or not ($R_j = \infty$). This highlights an advantage of the Stein Greedy algorithm over Stein Herding: the extra $k_0(x,x)/2$ term acts as a regularizer ensuring that no truncation is necessary. The result also accommodates points x_i selected suboptimally with error $\delta/2$.

Theorem 2 (Stein Greedy Convergence). *Suppose k_0 with $k_{0,P} = 0$ is a P-sub-exponential reproducing kernel. Then there exist constants $c_1, c_2 > 0$ depending only on k_0 and P such that any point sequence $\{x_i\}_{i=1}^n$ satisfying*

$$\begin{aligned} &\tfrac{k_0(x_j,x_j)}{2} + \textstyle\sum_{i=1}^{j-1} k_0(x_i, x_j) \\ &\qquad \le \tfrac{\delta}{2} + \min_{x \in X: k_0(x,x) \le R_j^2} \tfrac{k_0(x,x)}{2} + \textstyle\sum_{i=1}^{j-1} k_0(x_i, x) \end{aligned}$$

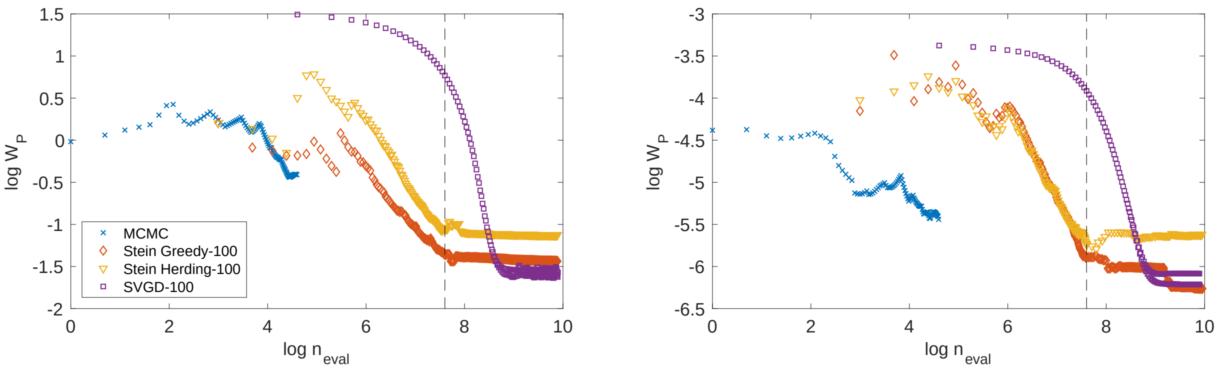
with $\sqrt{2\log(j)/c_2} \le R_j \le \infty$ for each $1 \le j \le n$ also satisfies

$$D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n) \le e^{\pi/2}\sqrt{\tfrac{2\log(n)}{c_2 n} + \tfrac{c_1}{n} + \tfrac{\delta}{n}}.$$

5.2. Distributional Convergence Control

To present our final results, we overload notation to define the KSD associated with any probability measure μ:

$$D_{\mathcal{K}_0,P}(\mu) \quad = \quad \sqrt{\mathbb{E}_{(Z,Z') \sim \mu \times \mu}\left[k_0(Z,Z')\right]}.$$



(a) Gaussian Process Test

(b) IGARCH Test

Figure 4: Combined results for the (a) Gaussian process test and (b) IGARCH test. [Here $n = 100$. x-axis: log of the number n_{eval} of model evaluations that were used. y-axis: log of the Wasserstein distance $W_P(\{x_i\}_{i=1}^n)$ obtained. Tuning parameters were selected to minimise W_P, as described in the main text. The dashed line indicates the point at which n Stein Points have been generated; block coordinate descent is performed thereafter to satisfy the n point budget constraint.]

Our original $D_{\mathcal{K}_0,P}$ definition (Eq. 4) for a point set $\{x_i\}_{i=1}^n$ is recovered when μ is the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. We also write $\mu_m \Rightarrow P$ to indicate that a sequence of probability measures $(\mu_m)_{m=1}^\infty$ converges in distribution to P.

Gorham & Mackey (2017, Thm. 8) showed that KSDs with IMQ base kernel (k_1) and Langevin Stein operator $\mathcal{T}_P$ control distributional convergence whenever P belongs to the set $\mathcal{P}$ of distantly dissipative distributions (i.e., $\langle \nabla \log p(x) - \nabla \log p(y), x - y\rangle \leq -\kappa \|x-y\|_2^2 + C$ for some $C \geq 0, \kappa > 0$) with Lipschitz $\nabla \log p$. Surprisingly, Gaussian, Matérn, and other kernels with light tails do not satisfy this property (Gorham & Mackey, 2017, Thm. 6).

Our next theorem establishes distributional convergence control for our newly introduced log inverse kernel (k_2).

Theorem 3 (Log Inverse KSD Controls Convergence)**.** *Suppose $P \in \mathcal{P}$. Consider a Stein reproducing kernel $k_0 = \mathcal{T}_P \overline{\mathcal{T}_P} k_2$ with Langevin operator $\mathcal{T}_P$ and base kernel $k_2(x,x') = (\alpha + \log(1 + \|x-x'\|_2^2))^\beta$ for $\alpha > 0$ and $\beta < 0$. If $D_{\mathcal{K}_0,P}(\mu_m) \to 0$, then $\mu_m \Rightarrow P$.*

Our final theorem, proved in Section A.3, guarantees distributional convergence control for the new IMQ score kernel (k_3) under the additional assumption that $\log p$ is strictly concave.

Theorem 4 (IMQ Score KSD Controls Convergence)**.** *Suppose $P \in \mathcal{P}$ has strictly concave log density. Consider a Stein reproducing kernel $k_0 = \mathcal{T}_P \overline{\mathcal{T}_P} k_3$ with Langevin operator $\mathcal{T}_P$ and base kernel $k_3(x,x') = (c^2 + \|\nabla \log p(x) - \nabla \log p(x')\|_2^2)^\beta$ for $c > 0$ and $\beta \in (-1, 0)$. If $D_{\mathcal{K}_0,P}(\mu_m) \to 0$, then $\mu_m \Rightarrow P$.*

6. Conclusion

This paper proposed and studied Stein Points, extensible point sequences rooted in minimisation of a KSD, building on the recent theoretical work of (Gorham & Mackey, 2017). Although we focused on KSD to limit scope, our methods could in fact be applied to any computable Stein discrepancy, even those not based on reproducing kernels (see, e.g., Gorham & Mackey, 2015; Gorham et al., 2016). Stein Points provide an interesting counterpoint to other recent work focussing on point sequences (Joseph et al., 2015; 2017) and point sets (Liu & Wang, 2016; Liu, 2017). Moreover, when X is a finite set $\{y_i\}_{i=1}^N$ (e.g., an inexpensive initial point set generated by MCMC), Stein Points provide a compact and convergent approximation to the optimally weighted probability measure $\sum_{i=1}^N w_i \delta_{y_i}$ with minimum KSD (see Section B.3 for more details).

Theoretical results were provided which guarantee the asymptotic correctness of our methods. However, we were only able to establish an $O(\sqrt{\log(n)/n})$ rate, which leaves a theoretical gap between the faster convergence that was sometimes empirically observed. Relatedly, the $O(n^2)$ computational cost could be reduced to $O(n)$ by using finite-dimensional kernels (see, e.g., Jitkrittum et al., 2017), but the associated distributional convergence control results must first be developed.

Our experiments were relatively comprehensive, but we did not consider other Stein operators, nor higher-dimensional or non-Euclidean manifolds X. Related methods not considered in this work include those based on optimal transport (Marzouk et al., 2016) and self-avoiding particle-based samplers (Robert & Mengersen, 2003). The comparison against these methods is left for future work.

Acknowledgements

WYC was supported by the ARC Centre of Excellence for Mathematical and Statistical Frontiers. FXB was supported by EPSRC [EP/L016710/1, EP/R018413/1]. CJO was supported by the Lloyd's Register Foundation programme on data-centric engineering at the Alan Turing Institute, UK. This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Bach, F., Lacoste-Julien, S., and Obozinski, G. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1355–1362, 2012.

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

Briol, F.-X., Oates, C., Girolami, M., Osborne, M., and Sejdinovic, D. Probabilistic integration: A role in statistical computation? *arXiv:1512.00933*, 2015.

Chen, Y., Welling, M., and Smola, A. Super-samples from kernel herding. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 2606–2615, 2016.

Dick, J. and Pillichshammer, F. *Digital Nets and Sequences. Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory Related Fields*, 162(3-4):707–738, 2015.

Gorham, J. and Mackey, L. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pp. 226–234, 2015.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1292–1301, 2017.

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

Gudmundsson, J., Klein, O., Knauer, C., and Smid, M. Small Manhattan networks and algorithmic applications for the earth movers distance. In *Proceedings of the 23rd European Workshop on Computational Geometry*, pp. 174–177, 2007.

Hickernell, F. A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, 67(221):299–322, 1998.

Jitkrittum, W., Xu, W., Szabo, Z., Fukumizu, K., and Gretton, A. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pp. 261–270, 2017.

Johnson, M., Moore, L., and Ylvisaker, D. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2):131–148, 1990.

Jones, L. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, 20(1):608–613, 1992.

Joseph, V., Dasgupta, T., Tuo, R., and Wu, C. Sequential exploration of complex surfaces using minimum energy designs. *Technometrics*, 57(1):64–74, 2015.

Joseph, V., Wang, D., Gu, L., Lv, S., and Tuo, R. Deterministic sampling of expensive posteriors using minimum energy designs. *arXiv:1712.08929*, 2017.

Joshi, K. *Introduction to General Topology*. New Age International, 1983.

Lacoste-Julien, S., Lindsten, F., and Bach, F. Sequential kernel herding : Frank-Wolfe optimization for particle filtering. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 544–552, 2015.

Liu, C. and Zhu, J. Riemannian Stein variational gradient descent for Bayesian inference. *arXiv:1711.11216*, 2017.

Liu, Q. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, pp. 3118–3126, 2017.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pp. 2370–2378, 2016.

Liu, Q., Lee, J., and Jordan, M. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 276–284, 2016.

Mackey, L. and Gorham, J. Multivariate Stein factors for a class of strongly log-concave distributions. *Electronic Communications in Probability*, 21(56), 2016.

Mak, S. and Joseph, V. R. Support points. *arXiv:1609.01811*, 2016.

Mak, S. and Joseph, V. R. Projected support points, with application to optimal MCMC reduction. *arXiv:1708.06897*, 2017.

Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A. *Handbook of Uncertainty Quantification*, chapter Sampling via Measure Transport: An Introduction. Springer, 2016.

Nelder, J. and Mead, R. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

Oates, C., Barp, A., and Girolami, M. Posterior integration on a Riemannian manifold. *arXiv:1712.01793*, 2017a.

Oates, C., Girolami, M., and Chopin, N. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B*, 79(3):695–718, 2017b.

Oates, C., Cockayne, J., Briol, F.-X., and Girolami, M. Convergence rates for a class of estimators based on Stein's identity. *Bernoulli*, 2018. To appear.

Oettershagen, J. *Construction of Optimal Cubature Algorithms with Applications to Econometrics and Uncertainty Quantification*. PhD thesis, University of Bonn, 2017.

Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Robert, C. and Mengersen, K. IID sampling with self-avoiding particle filters: The pinball sampler. In *Bayesian Statistics*, volume 7, chapter IID sampling with self-avoiding particle filters: The pinball sampler. Oxford University Press, 2003. Eds. Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M.

Ruppert, D., Wand, M., and Carroll, R. *Semiparametric Regression*. Number 12. Cambridge Series in Statistical and Probabilistic Mathematics, 2003.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41(5): 2263–2291, 2013.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pp. 13–31, 2007.

Spivak, M. *Calculus on Manifolds: A Modern Approach to Classical Theorems of Advanced Calculus*. Westview Press, 1965.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

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

Taylor, S. *Asset Price Dynamics, Volatility, and Prediction*. Princeton University Press, 2011.

Wainwright, M. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. 2017. URL https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Wendland, H. *Scattered Data Approximation*. Cambridge University Press, 2004.