Stein Point Markov Chain Monte Carlo

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

Abstract

An important task in machine learning and statistics is the approximation of a probability measure by an empirical measure supported on a discrete point set. Stein Points are a class of algorithms for this task, which proceed by sequentially minimising a Stein discrepancy between the empirical measure and the target and, hence, require the solution of a non-convex optimisation problem to obtain each new point. This paper removes the need to solve this optimisation problem by, instead, selecting each new point based on a Markov chain sample path. This significantly reduces the computational cost of Stein Points and leads to a suite of algorithms that are straightforward to implement. The new algorithms are illustrated on a set of challenging Bayesian inference problems, and rigorous theoretical guarantees of consistency are established.

1. Introduction

The task that we consider in this paper is to approximate a Borel probability measure P on an open and convex set $\mathcal{X} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, with an empirical measure $\hat{P}$ supported on a discrete point set $\{x_i\}_{i=1}^n \subset \mathcal{X}$. To limit scope we restrict attention to uniformly-weighted empirical measures; $\hat{P} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ where δ_x is a Dirac measure on x. The *quantisation* (Graf & Luschgy, 2007) of P by $\hat{P}$ is an important task in computational statistics and machine learning. For example, quantisation facilitates the approximation of integrals $\int_{\mathcal{X}} f \mathrm{d}P$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ using cubature rules $f \mapsto \frac{1}{n}\sum_{i=1}^n f(x_i)$. More generally, quantisation underlies a broad spectrum of algorithms for uncertainty quantification that must operate subject to a finite computational budget. Motivated by applications in Bayesian statistics, our focus is on the situation where P admits a density p with respect to the Lebesgue measure on $\mathcal{X}$ but this density can only be evaluated up to an (unknown) normalisation constant. Specifically, we assume that $p = \frac{\tilde{p}}{C}$ where $\tilde{p}$ is an un-normalised density and $C > 0$, such that both $\tilde{p}$ and $\nabla \log \tilde{p}$, where $\nabla = (\frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^d})$, can be (pointwise) evaluated at finite computational cost.

A popular approach to this task is Markov chain Monte Carlo (MCMC; Robert & Casella, 2004), where the sample path of an ergodic Markov chain with invariant distribution P constitutes a point set $\{x_i\}_{i=1}^n$. MCMC algorithms exploit a range of techniques to construct Markov transition kernels which leave P invariant, based (in general) on pointwise evaluation of $\tilde{p}$ (Metropolis et al., 1953) or (sometimes) on pointwise evaluation of $\nabla \log \tilde{p}$ and higher-order derivative information (Girolami & Calderhead, 2011). In a favourable situation, the MCMC output will be approximately independent draws from P. However, in this case the $\{x_i\}_{i=1}^n$ will typically not be a *low discrepancy* point set (Dick & Pillichshammer, 2010) and as such the quantisation of P performed by MCMC will be sub-optimal. In recent years several attempts have been made to deveop improved algorithms for quantisation in the Bayesian statistical context as an alternative to MCMC:

- **Minimum Energy Designs** (MED) In (Roshan Joseph et al., 2015; Joseph et al., 2018) it was proposed to obtain a point set $\{x_i\}_{i=1}^n$ by using a numerical optimisation method to approximately minimise an energy functional $\mathcal{E}_{\tilde{p}}(\{x_i\}_{i=1}^n)$ that depends on P only through $\tilde{p}$ rather than through p itself. Though appealing in its simplicity, MED has yet to receive a theoretical treatment that accounts for the imperfect performance of the numerical optimisation method.

- **Support Points** The method of (Mak & Joseph, 2018) first generates a large MCMC output $\{\tilde{x}_i\}_{i=1}^N$ and from this a subset $\{x_i\}_{i=1}^n$ is selected in such a way that a low-discrepancy point set is obtained. (This can be contrasted with classical *thinning* in which an arithmetic subsequence of the MCMC output is selected.)

[*] Equal contribution [1] Institute of Statistical Mathematics [2] Imperial College London [3] Alan Turing Institute [4] University of Cambridge [5] OpenDoor [6] Microsoft Research [7] Newcastle University. Correspondence to: Lester Mackey <lmackey@microsoft.com>, Chris. J. Oates <chris.oates@ncl.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

At present, a theoretical analysis that accounts for the possible poor performance of the MCMC method has not yet been announced.

- **Transport Maps and QMC** The method of (Parno, 2015) aims to learn a *transport map* $T : \mathcal{X} \to \mathcal{X}$ such that the pushforward measure $T_\# Q$ corresponds to P, where Q is a distribution for which quantisation by a point set $\{\tilde{x}_i\}_{i=1}^n$ is easily performed, for instance using quasi-Monte Carlo (QMC) (Dick & Pillichshammer, 2010). Then quantisation of P is provided by the point set $\{T(\tilde{x}_i)\}_{i=1}^n$. The flexibility in the construction of a transport map allows several algorithms to be envisaged, but an end-to-end theoretical treatment is not available at present.

- **Stein Variational Gradient Descent** (SVGD) A popular methodology due to (Liu & Wang, 2016) aims to take an arbitrary initial point set $\{x_i^0\}_{i=1}^n$ and to construct a discrete time dynamical system $x_i^t = g_{\tilde{p}}(x_1^{t-1}, \dots, x_n^{t-1})$, indexed by time t and dependent on $\tilde{p}$, such that $\lim_{t\to\infty}\{x_i^t\}_{i=1}^n$ provides a quantisation of P. This can be viewed as a discretisation of a particular gradient flow that has P as a fixed point (Liu, 2017). However, a generally applicable theoretical analysis of the SVGD method itself is not available (note that a compactness assumption on $\mathcal{X}$ was required in Liu, 2017). Note also that, unlike the other methods discussed in this section, SVGD does not readily admit an *extensible* construction; that is, the number n of points must be *a priori* fixed.

- **Stein Points** (SP) The authors of (Chen et al., 2018b) proposed to select a point set $\{x_i\}_{i=1}^n$ that approximately minimises a *kernel Stein discrepancy* (KSD; Liu et al., 2016; Chwialkowski et al., 2016; Gorham & Mackey, 2017) between the empirical measure and the target P. The KSD can be exactly computed with a finite number of pointwise evaluations of $\nabla \log \tilde{p}$ and, for the (non-convex) minimisation, a variety of numerical optimisation methods can be applied. In contrast to the other methods just discussed, SP does admit a end-to-end theoretical treatment when a grid search procedure is used as the numerical optimisation method (Thms. 1 & 2 in Chen et al., 2018b).

An empirical comparison of several of the above methods on a selection of problems arising in computational statistics was presented in (Chen et al., 2018b). The conclusion of that work was that MED and SP provided broadly similar performance-per-computational-cost at the quantisation task, where the performance was measured by the Wasserstein distance to the target and the computational cost was measured by the total number of evaluations of either $\tilde{p}$ or its gradient. In some situations, SVGD provided superior quantisation to MED and SP but this was achieved at a substantially higher computational cost. At the same time, it was observed that *all* algorithms considered provided improved quantisation compared to MCMC, but at a computational cost that was substantially higher than the corresponding cost of MCMC.

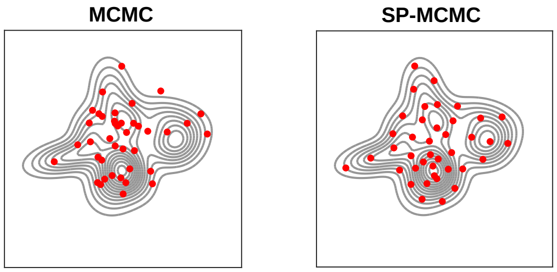


Figure 1. Illustration of Monte Carlo points (MC; left) and Stein Point Markov chain Monte Carlo (SP-MCMC; right) on a Gaussian mixture target P. SP-MCMC provides better space-filling properties than MC.

In this paper, we propose Stein Point Markov chain Monte Carlo (SP-MCMC), aiming to provide strong performance at the quantisation task (see Fig. 1) but at substantially reduced computational cost compared to the original SP method. Our contributions are summarised as follows:

- The global optimisation subroutine in SP, whose computational cost was exponential in dimension d, is replaced by a form of local search based on MCMC. This allows us to make use of efficient transition kernels for exploration of $\mathcal{X}$, which in turn improves performance in higher dimensions and reduces the overall computational cost.

- Our construction requires a new Markov chain to be initialised each time a point x_n is added, however the initial distribution of the chain does not need to coincide with P. This enables us to develop an efficient criterion for initialisation of the Markov chains, based on the introduced notion of the "most influential" point in $\{x_i\}_{i=1}^{n-1}$, as quantified by KSD. This turns our sequence of local searches into a global-like search, and also leads to automatic "mode hopping" behaviour when P is a multi-modal target.

- The consistency of SP-MCMC is established under a V-uniform ergodicity condition on the Markov kernel.

- SP-MCMC is shown, empirically, to outperform MCMC, MED, SVGD and SP when applied to posterior computation in the Bayesian statistical context.

The paper is structured as follows: In Section 2 we review the central notions of Stein's method and KSD, as well as recalling the original SP method. The novel methodology is presented in Section 3. This is assessed experimentally in Section 4 and theoretically in Section 5. Conclusions are drawn in Section 6.

2. Background

In Section 2.1 we recall the construction of KSD, then in Section 2.2 the SP method of (Chen et al., 2018b), which is based on minimisation of KSD, is discussed.

2.1. Discrepancy and Stein's Method

A *discrepancy* is a notion of how well an empirical measure, based on a point set $\{x_i\}_{i=1}^n \subset \mathcal{X}$, approximates a target P. One popular form of discrepancy is the *integral probability metric* (IPM) (Muller, 1997), which is based on a set $\mathcal{F}$ consisting of functionals on $\mathcal{X}$, and is defined as:

$$D_{\mathcal{F},P}(\{x_i\}_{i=1}^n) \;:=\; \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \int_{\mathcal{X}} f \mathrm{d}P\right| \tag{1}$$

The set $\mathcal{F}$ is required to be measure-determining in order for the IPM to be a genuine metric. Certain sets $\mathcal{F}$ lead to familiar notions, such as the Wasserstein distance, but direct computation of an IPM will generically require exact integration against P; a demand that is not met in the Bayesian context. In order to construct an IPM that *can* be computed in the Bayesian context, (Gorham & Mackey, 2015) proposed the notion of a *Stein discrepancy*, based on Stein's method (Stein, 1972). This consists of finding an operator $\mathcal{A}$, called a *Stein operator*, and a function class $\mathcal{G}$, called a *Stein class*, which satisfy the *Stein identity* $\int_{\mathcal{X}} \mathcal{A}g \mathrm{d}P = 0$ for all $g \in \mathcal{G}$. Taking $\mathcal{F} = \mathcal{A}\mathcal{G}$ to be the image of $\mathcal{G}$ under $\mathcal{A}$ in (1) leads directly to the *Stein discrepancy*:

$$D_{\mathcal{A}\mathcal{G},P}(\{x_i\}_{i=1}^n) \;=\; \sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \mathcal{A}g(x_i)\right| \tag{2}$$

A particular choice of $\mathcal{A}$ and $\mathcal{G}$ was studied in (Gorham & Mackey, 2015) with the property that exact computation can be performed based only on point-wise evaluation of $\nabla \log \tilde{p}$. The computation of this *graph Stein discrepancy* reduced to solving d independent linear programs in parallel with $O(n)$ variables and constraints.

To eliminate the the reliance on a linear program solver, (Liu et al., 2016; Chwialkowski et al., 2016; Gorham & Mackey, 2017) proposed *kernel Stein discrepancies*, alternative Stein discrepancies (2) with embarrassingly parallel, closed-form values. For the remainder we assume that $p > 0$ on $\mathcal{X}$. The canonical KSD is obtained by taking the Stein operator $\mathcal{A}$ to be the Langevin operator $\mathcal{A}g := \frac{1}{\tilde{p}}\nabla \cdot (\tilde{p}g)$ and the Stein class $\mathcal{G} = B(\mathcal{K}^d)$ to be the unit ball of a space of vector-valued functions, formed as a d-dimensional Cartesian product of scalar-valued reproducing kernel Hilbert spaces $\mathcal{K}$ (RKHS) (Berlinet & Thomas-Agnan, 2004). (Throughout we use $\nabla\cdot$ to denote divergence and $\langle\cdot,\cdot\rangle$ to denote the Euclidean inner product.) Recall that an RKHS $\mathcal{K}$ is a Hilbert space of functions with inner product $\langle\cdot,\cdot\rangle_k$ and induced norm $\|\cdot\|_k$, and there is a function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, called a *kernel*, such that $\forall x \in \mathcal{X}$, we can write the evaluation functional $f(x) = \langle f, k(\cdot,x)\rangle_k$ $\forall f \in \mathcal{K}$. It is assumed that the mixed derivatives $\partial^2 k(x,y)/\partial x^i \partial y^j$ and all lower-order derivatives are continuous and uniformly bounded. For $\mathcal{X}$ bounded, with piecewise smooth boundary denoted $\partial\mathcal{X}$, outward normal denoted n and surface element denoted $\mathrm{d}\sigma(x)$, the conditions $\oint_{\partial\mathcal{X}} k(x,x')p(x)\mathrm{n}(x)\mathrm{d}\sigma(x') = 0$, $\oint_{\partial\mathcal{X}} \nabla_x k(x,x')\cdot \mathrm{n}(x)p(x)\mathrm{d}\sigma(x') = 0$ are sufficient for the Stein identity to hold; c.f. Lemma 1 in (Oates et al., 2017). For $\mathcal{X} = \mathbb{R}^d$, a sufficient condition is $\int_{\mathcal{X}} \|\nabla \log p(x)\|_2 \mathrm{d}P(x) < \infty$; c.f. Prop. 1 of (Gorham & Mackey, 2017). The image $\mathcal{A}\mathcal{G} = B(\mathcal{K}_0)$ is the unit ball of another RKHS, denoted $\mathcal{K}_0$, whose kernel is (Oates et al., 2017):

$$\begin{aligned} k_0(x,x') \;=\; & \nabla_x \cdot \nabla_{x'} k(x,x') + \langle \nabla_x k(x,x'), \nabla_{x'} \log \tilde{p}(x')\rangle \\ & + \langle \nabla_{x'} k(x,x'), \nabla_x \log \tilde{p}(x)\rangle \\ & + k(x,x')\,\langle \nabla_x \log\tilde{p}(x), \nabla_{x'}\log\tilde{p}(x')\rangle \end{aligned} \tag{3}$$

In this case, (2) corresponds to a maximum mean discrepancy (MMD; Gretton et al., 2006) in the RKHS $\mathcal{K}_0$ and thus can be explicitly computed. The Stein identity implies that $\int_{\mathcal{X}} k_0(x,\cdot)\mathrm{d}P \equiv 0$. Thus we denote the KSD between the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and the target P (in a small abuse of notation) as

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

Under regularity assumptions (Gorham & Mackey, 2017; Chen et al., 2018b; Huggins & Mackey, 2018), the KSD controls classical weak convergence of the empirical measure to the target. This motivates selecting the $\{x_i\}_{i=1}^n$ to minimise the KSD, and to this end we now recall the SP method of (Chen et al., 2018b).

2.2. Stein Points

The *Stein Point* (SP) method due to (Chen et al., 2018b) selects points $\{x_i\}_{i=1}^n$ to approximately minimise $D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n)$. This is of course a challenging non-convex and multivariate problem in general. For this reason, two sequential strategies were proposed. The first, called *Greedy* SP, was based on greedy minimisation of KSD, whilst the second, called *Herding* SP, was based on Frank-Wolfe minimisation of KSD. In each case, at iteration $j \in \{1,\ldots,n\}$ of the algorithm, the points $\{x_i\}_{i=1}^{j-1}$ have been selected and a global search method is used to select the next point $x_j \in \mathcal{X}$. To limit scope we restrict the discussion below to Greedy SP, as this has stronger theoretical guarantees and has been shown empirically to outperform Herding SP. The convergence of Greedy SP was established in Theorem 2 of (Chen et al., 2018b) when k_0 is a

P-sub-exponential kernel (Def. 1 of Chen et al.). More precisely, assume that for some pre-specified tolerance $\delta > 0$, the resulting point sequence satisfies the following identity $\forall j \in \{1, \dots, n\}$:

$$D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^{j})^2 \leq \tfrac{\delta}{j^2} + \inf_{x\in\mathcal{X}} D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^{j-1} \cup \{x\})^2.$$

Then it was shown that $\exists\, c_1, c_2 > 0$ such that

$$D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^{n}) \leq e^{\pi/2}\sqrt{\tfrac{2\log(n)}{c_2 n} + \tfrac{c_1}{n} + \tfrac{\delta}{n}} \tag{5}$$

so that KSD is asymptotically minimised. However, a significant limitation of the SP method is that it requires a *global* (non-convex) minimisation problem over $\mathcal{X}$ to be (approximately) solved in order to select the next point. In practice, the global search at iteration j can be facilitated by a grid search over $\mathcal{X}$, but this procedure entails a computational cost that is exponential in the dimension d of $\mathcal{X}$ and even in modest dimension this becomes impractical.

The main contribution of the present paper is to re-visit the SP method and to study its behaviour when the global search is replaced with a *local* search, facilitated by a MCMC method. To proceed, two main challenges must be addressed: First, an appropriate local optimisation procedure must be developed. Second, the theoretical convergence of the modified algorithm must be established. In the next section we address the first challenge by presenting our novel methodological development.

3. Methodology

In Section 3.1 we present the novel SP-MCMC method. Then in Section 3.2 we describe how the kernel k can be pre-conditioned to improve performance in SP-MCMC.

3.1. SP-MCMC

In this paper, we propose to replace the global minimisation at iteration j of the SP method of (Chen et al., 2018b) with a local search based on a P-invariant Markov chain of length m_j, where the sequence $(m_j)_{j\in\mathbb{N}}$ is to be specified. The proposed SP-MCMC method proceeds as follows:

1. Fix an initial point $x_1 \in \mathcal{X}$.
2. For $j = 2, \dots, n$:
   i. Select an index $i^* \in \{1, \dots, j-1\}$ according to some criterion $\texttt{crit}(\{x_i\}_{i=1}^{j-1})$, to be defined.
   ii. Run a P-invariant Markov chain, initialised at x_{i^*}, for m_j iterations and denote the realised sample path as $(y_{j,l})_{l=1}^{m_j}$.
   iii. Set $x_j = y_{j,l}$ where $l \in \{1, \dots, m_j\}$ minimises $D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^{j-1} \cup \{y_{j,l}\})$.

It remains to specify the sequence $(m_j)_{j\in\mathbb{N}}$ and the criterion `crit`. Precise statements about the effect of these choices on convergence are reserved for the theoretical treatment in Section 5. For the criterion `crit`, three different approaches are considered:

- `LAST` selects the point last added: $i^* := j - 1$.
- `RAND` selects i^* uniformly at random in $\{1, \dots, j-1\}$.
- `INFL` selects i^* to be the index of a most influential point in $\{x_i\}_{i=1}^{j-1}$. Specifically, we call x_{i^*} a *most influential* point if removing it from our point set creates the greatest increase in KSD. i.e. i^* maximises $D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^{j-1} \setminus \{x_{i^*}\})$.

SP-MCMC overcomes the main limitation facing the original SP method; the global search is avoided. Indeed, the cost of simulating m_j steps of a P-invariant Markov chain will typically be just a fraction of the cost of implementing a global search method. The number of iterations m_j acts as a lever to trade-off approximation quality against computational cost, with larger m_j leading on average to an empirical measure with lower KSD. The precise relationship is elucidated in Section 5.

Remark 1 (KSD has low overhead)**.** *A large number of modern MCMC methods, such as the Metropolis-adjusted Langevin algorithm (MALA) and Hamiltonian Monte Carlo, exploit evaluations of $\nabla \log \tilde{p}$ to construct a P-invariant Markov transition kernel (Barp et al., 2018a). If such an MCMC method is used, the gradient information $\nabla \log \tilde{p}(x_{i^*})$ is computed during the course of the MCMC and can be recycled in the subsequent computation of KSD.*

Remark 2 (Automatic mode-hopping)**.** *Although the Markov chain is used only for a local search, the* initialisation *criteria* `RAND` *and* `INFL` *offer the opportunity to jump to any point in the set $\{x_i\}_{i=1}^{j-1}$ and thus can facilitate global exploration of the state space $\mathcal{X}$. The* `INFL` *criteria, in particular, favours areas of $\mathcal{X}$ that are under-represented in the point set and thus, for a multi-modal target P, one can expect "mode hopping" from near an over-represented mode to near an under-represented mode of P.*

Remark 3 (Removal of bad points)**.** *A natural extension of the SP-MCMC method allows for the possibility of removing a "bad" point from the current point set. That is, at iteration j we may decide, according to some probabilistic or deterministic schedule, to remove a point x_{i^*} that minimises $D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^{j-1} \setminus \{x_{i^*}\})$. This extension was also investigated and results are reserved for Section A.6.5.*

Remark 4 (Sequence vs set)**.** *If the number n of points is pre-specified, then after the n point is selected one can attempt to further improve the point set by applying (e.g.) co-ordinate descent to the KSD interpreted as a function $D_{\mathcal{K}_0,P} : \mathcal{X}^n \to [0,\infty)$; see (Chen et al., 2018b). To limit scope, this was not considered.*

3.2. Pre-conditioned Kernels for SP-MCMC

The original analysis of (Gorham & Mackey, 2017) focussed on the inverse multiquadric (IMQ) kernel $k(x, x') = \left(1 + \lambda^{-2}\|x - x'\|_2^2\right)^\beta$ for some length-scale parameter $\lambda > 0$ and exponent $\beta \in (-1, 0)$; alternative kernels were considered in (Chen et al., 2018b), but the IMQ kernel was observed to lead to the best empirical approximations as quantified objectively by the Wasserstein distance between the empirical measure and the target. Thus, in this paper we focus on the IMQ kernel. However, in order to improve the performance of the algorithm, we propose to allow for *pre-conditioning* of the kernel; that is, we consider

$$k(x, x') = \left(1 + \|\Lambda^{-\frac{1}{2}}(x - x')\|_2^2\right)^\beta \qquad (6)$$

for some symmetric positive definite matrix Λ. The use of pre-conditioned kernels was recently proposed in the context of SVGD in (Detommaso et al., 2018), where Λ^{-1} was taken to be an approximation to the expected Hessian $-\int \nabla_x \nabla_x^\top \log \tilde{p}(x) \mathrm{d}P(x)$ of the negative log target. Note that the matrix Λ can also form part of a MCMC transition kernel, such as the pre-conditioner matrix in MALA (Girolami & Calderhead, 2011). Sufficient conditions for when a pre-conditioned kernel ensures that KSD controls classical weak convergence of the empirical measure to the target are established in Section 5.

4. Experimental Results

In this section our attention turns to the empirical performance of SP-MCMC. The experimental protocol is explained in Section 4.1 and specific experiments are described in Sections 4.2, 4.3 and 4.4.

4.1. Experimental Protocol

To limit scope, we present a comparison of SP-MCMC to the original SP method, as well as to MCMC, MED and SVGD. All experiments involving SP-MCMC, SP or SVGD in this paper were based on the IMQ kernel in (6) with $\beta = -\frac{1}{2}$. The preconditioner matrix Λ was taken either to be a sample-based approximation to the covariance matrix of P (Secs. 4.2 and 4.3), generated by running a short MCMC, or $\Lambda \propto I$ (Sec. 4.4); however, in each experiment Λ was fixed across all methods being compared. The Markov chains used for SP-MCMC and MCMC in this work employed either a random walk Metropolis (RWM) or a MALA transition kernel, described in Appendix A.5. Our implementations of MED and SVGD are described in Appendix A.6.1.

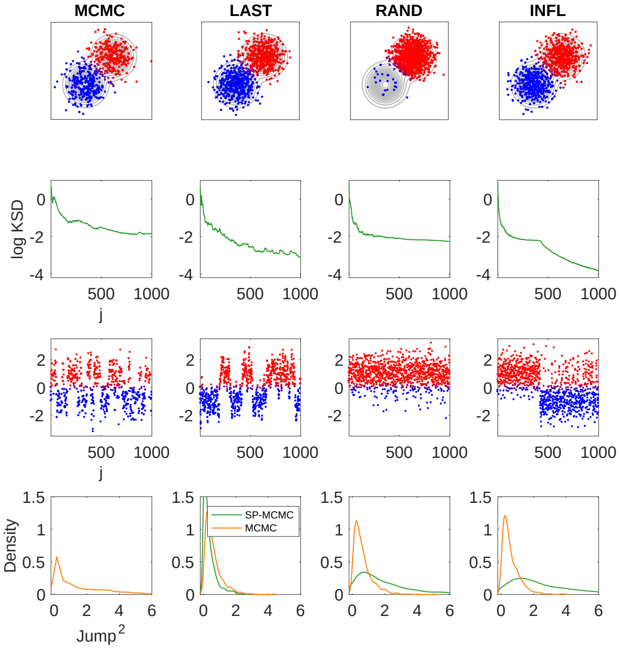


Figure 2. Gaussian mixture experiment in dimension $d = 2$. Columns (left to right): MCMC, SP-MCMC with `LAST`, SP-MCMC with `RAND`, SP-MCMC with `INFL`. Top row: Point sets of size $n = 1000$ produced by MCMC and SP-MCMC. (Point colour indicates the mode to which they are closest.) Second row: Trace plot of $\log D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^j)$ as j is varied from 1 to n. Third row: Trace plots of the sequence $(x_i)_{i=1}^n$, projected onto the first coordinate. Bottom row: Distribution of the squared jump distance $\|x_j - x_{j-1}\|_2^2$ (green) compared to the quantities $\|y_{j,m_j} - y_{j,1}\|_2^2$ associated with the Markov chains (orange) used during the course of each method.

Three experiments of increasing sophistication were considered.[1] First, in Section 4.2 we consider a simple Gaussian mixture target in order to explore SP-MCMC and investigate sensitivity to the degrees of freedom in this new method. Second, in Section 4.3 we revisit one of the experiments in (Chen et al., 2018b), in order to directly compare against SP, MCMC, MED and SVGD. Third, in Section 4.4 we consider a more challenging application to Bayesian parameter inference in an ordinary differential equation (ODE) model.

4.2. Gaussian Mixture Model

For exposition we let $\sigma^2 = 0.5$ and consider a $d = 2$ dimensional Gaussian mixture model $P = \frac{1}{2}\mathcal{N}(-1, \sigma^2 \mathrm{I}_{d\times d}) + \frac{1}{2}\mathcal{N}(1, \sigma^2 \mathrm{I}_{d\times d})$ with modes at $1 = [1, 1]$ and -1. The performance of MCMC was compared to SP-MCMC for each of the criteria `LAST`, `RAND`, `INFL`. Note that in this section we do not address computational

[1] Code to reproduce all experiments can be downloaded at `https://github.com/wilson-ye-chen/sp-mcmc`.

cost; this is examined in Secs. 4.3 and 4.4. For SP-MCMC the sequence $(m_j)_{j \in \mathbb{N}}$ was set as $m_j = 5$. Results are presented in Fig. 2 with $n = 1000$.

The point sets produced by SP-MCMC with `LAST` and `INFL` (top row) were observed to provide a better quantisation of the target P compared to MCMC, as captured by the KSD of the empirical measure to the target (second row). `RAND` did not distribute points evenly between modes and, as a result, KSD was observed to plateau in the range of n displayed. For MCMC, the proposal step-size $h > 0$ was optimised according to the recommendations in (Roberts & Rosenthal, 2001), but nevertheless the chain was observed to jump between the two components of P only infrequently (third row, colour-coded). In contrast, after an initial period where both modes are populated, SP-MCMC under the `INFL` criteria was seen to frequently jump between components of P. Finally, we note that under `INFL` the typical squared jump distance $\|x_j - x_{j-1}\|_2^2$ was greater than the analogous quantities $\|y_{j,m_j} - y_{j,1}\|_2^2$ for the underlying Markov chains that were used (bottom row), despite the latter being optimised according to the recommendations of (Roberts & Rosenthal, 2001), which supports the view that more frequent mode-hopping is a property of the `INFL` method. Based on the findings of this experiment, we focus only on `LAST` and `INFL` in the sequel. The extension where "bad" points are removed, described in Remark 3, was explored in supplemental Section A.6.5.

4.3. IGARCH Model

Next our attention turns to whether SP-MCMC improves over the original SP method and how it compares to existing methods such as MED and SVGD when computational cost is taken into account. To this end we consider an identical experiment to (Chen et al., 2018b), based on Bayesian inference for a classical integrated generalised autoregressive conditional heteroskedasticity (IGARCH) model. The IGARCH model (Taylor, 2011)

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

describes a financial time series (y_t) with time-varying volatility (σ_t). The model is parametrised by $\theta = (\theta_1, \theta_2)$, $\theta_1 > 0$ and $0 < \theta_2 < 1$ and Bayesian inference for θ is considered, based on data $y = (y_t)$ that represent 2,000 daily percentage returns of the S&P 500 stock index (from December 6, 2005 to November 14, 2013). Following (Chen et al., 2018b), an improper uniform prior was placed on θ. The domain $\mathcal{X} = \mathbb{R}_+ \times (0,1)$ is bounded and, for this example, the posterior P places negligible mass near the boundary $\partial\mathcal{X}$. This ensures that the boundary conditions described in Sec. 2.1 hold essentially to machine precision, as argued in (Chen et al., 2018b).

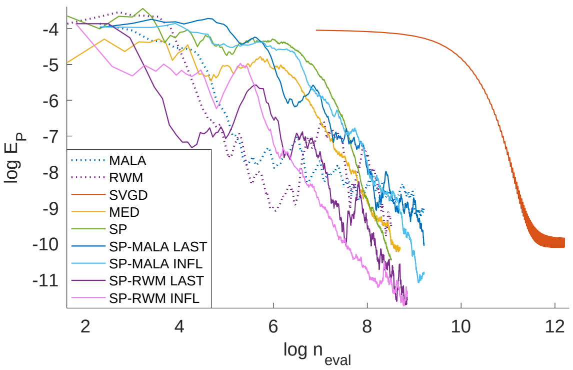


Figure 3. IGARCH experiment. The new SP-MCMC method was compared against the original SP method of (Chen et al., 2018b), as well as against MCMC, MED (Roshan Joseph et al., 2015) and SVGD (Liu & Wang, 2016). The implementation of all existing methods is described in Appendix A.6. Each method produced an empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ whose distance to the target P was quantified by the energy distance E_P. The computational cost was quantified by the number n_{eval} of times either $\tilde{p}$ or its gradient were evaluated.

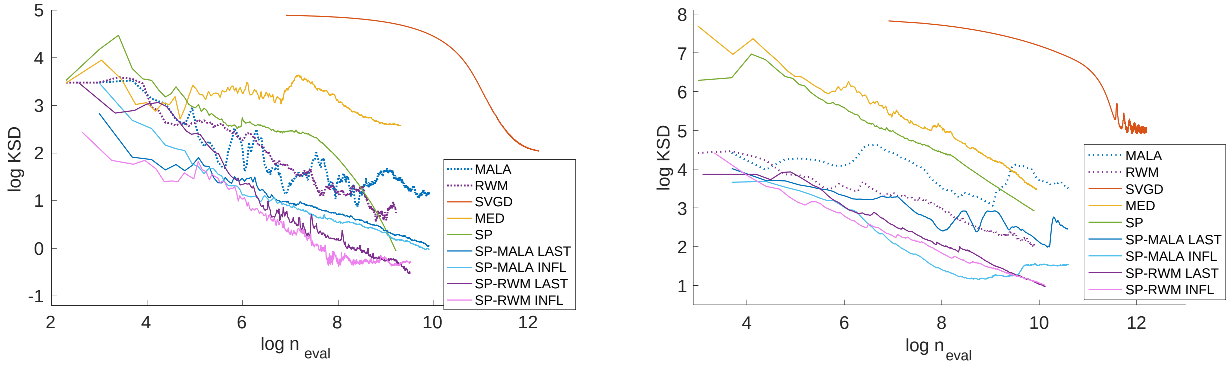
For objectivity, the *energy distance* E_P (Székely & Rizzo, 2004; Baringhaus & Franz, 2004) was used to assess closeness of all empirical measures to the target.[2] SP-MCMC was implemented with $m_j = 5 \ \forall j$. In addition to SP-MCMC, the methods SP, MED, SVGD and standard MCMC were also considered, with implementation described in Appendix A.6. All methods produced a point set of size $n = 1000$. The results, presented in Fig. 3, are indexed by the computational cost of running each method, which is a count of the total number n_{eval} of times either $\tilde{p}$ or $\nabla \log \tilde{p}$ were evaluated. It can be seen that SP-MCMC offers improved performance over the original SP method for fixed computational cost, and in turn over both MED and SVGD in this experiment. Typical point sets produced by each method are displayed in Fig. S1. The performance of the pre-conditioned kernel on this task was investigated in Appendix A.6.4.

4.4. System of Coupled ODEs

Our final example is more challenging and offers an opportunity to explore the limitations of SP-MCMC in higher dimensions. The context is an indirectly observed ODE

$$\begin{aligned} y_i &= g(u(t_i)) + \epsilon_i, & \epsilon_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \mathrm{I}) \\ \dot{u}(t) &= f_\theta(t, u), & u(0) &= u_0 \end{aligned}$$

[2] The energy distance E_P is equivalent to MMD based on the conditionally positive definite kernel $k(x,y) = -\|x-y\|_2$ (Sejdinovic et al., 2013). It was computed using a high-quality empirical approximation of P obtained from a large MCMC output.



(a) $d = 4$ (b) $d = 10$

Figure 4. ODE experiment, d-dimensional. The new SP-MCMC method was compared against the original SP method of (Chen et al., 2018b), as well as against standard MCMC, MED (Roshan Joseph et al., 2015) and SVGD (Liu & Wang, 2016). Each method produced an empirical measure $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ whose distance to the target P was quantified by the kernel Stein discrepancy (KSD). The computational cost was quantified by the number n_{eval} of times either $\tilde{p}$ or its gradient were evaluated.

and, in particular, Bayesian inference for the parameter θ in the gradient field. Here $y_i \in \mathbb{R}^p$, $u(t) \in \mathbb{R}^q$ and $\theta \in \mathbb{R}^d$ for $p, q, d \in \mathbb{N}$. For our experiment, f_θ and g comprised two instantiations of the Goodwin oscillator (Goodwin, 1965), one low-dimensional with $(q, d) = (2, 4)$ and one higher-dimensional with $(q, d) = (8, 10)$. In both cases $p = 2$, $\sigma = 0.1$ and 40 measurements were observed at uniformly-spaced time points in $[41, 80]$. The Goodwin oscillator does not permit a closed form solution, meaning that each evaluation of the likelihood function requires the numerical integration of the ODE at a non-negligible computational cost. SP-MCMC was implemented with the `INFL` criterion and $m_j = 10$ ($d = 4$), $m_j = 20$ ($d = 10$). Full details of the ODE and settings for MED and SVGD are provided in Appendix A.6.6.

In this experiment, KSD was used to assess closeness of all empirical measures to the target.[3] Naturally, SP and SP-MCMC are favoured by this choice of assessment criterion, as these methods are designed to directly minimise KSD. Therefore our main focus here is on the comparison between SP and SP-MCMC. All methods produced a point set of size $n = 1000$. Results are shown in Fig. 4a (low-dimensional) and Fig. 4b (high-dimensional). Note how the gain in performance of SP-MCMC over SP is more substantial when $d = 10$ compared to when $d = 4$, supporting our earlier intuition for the advantage of local optimisation using a Markov kernel.

5. Theoretical Results

Let Ω be a probability space on which the collection of random variables $Y_{j,l} : \Omega \to \mathcal{X}$ representing the l^{th} state of the Markov chain run at the j^{th} iteration of SP-MCMC are defined. Each of the three algorithms that we consider correspond to a different initialisation of these Markov chains and we use $\mathbb{E}$ to denote expectation over randomness in the $Y_{j,l}$. For example, the algorithm called `LAST` would set $Y_{j,1}(\omega) = x_{j-1}$. It is emphasised that the results of this section hold for *any* choice of function `crit` that takes values in $\mathcal{X}$. As a stepping-stone toward our main result, we first extend the theoretical analysis of the original SP method to the case where the global search is replaced by a Monte Carlo search based on m_i independent draws from P at iteration i of the SP method.

Theorem 1 (i.i.d. SP-MCMC Convergence)**.** *Suppose that the kernel k_0 satisfies $\int_{\mathcal{X}} k_0(x, \cdot)\mathrm{d}P(x) \equiv 0$ and $\mathbb{E}_{Z\sim P}[e^{\gamma k_0(Z,Z)}] < \infty$ for some $\gamma > 0$. Let $(m_j)_{j=1}^n \subset \mathbb{N}$ be a fixed sequence, and consider idealised Markov chains with $Y_{j,l} \overset{\text{i.i.d.}}{\sim} P$ for all $1 \leq l \leq m_j$, $j \in \mathbb{N}$. Let $\{x_i\}_{i=1}^n$ denote the output of SP-MCMC. Then, writing $a \wedge b = \min\{a, b\}$, $\exists\, C > 0$ such that*

$$\mathbb{E}\left[D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n)^2\right] \leq \frac{C}{n}\sum_{i=1}^{n}\frac{\log(n\wedge m_i)\wedge\sup_{x\in\mathcal{X}}k_0(x,x)}{n\wedge m_i}.$$

The constant C depends on k_0 and P, and the proof in Appendix A.1 makes this dependence explicit.

It follows that SP-MCMC with independent sampling from P is consistent whenever each m_j grows with n. When $m_j = m$ for all j we obtain:

$$\mathbb{E}\left[D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n)^2\right] \leq C\frac{\log(n\wedge m)\wedge\sup_{x\in\mathcal{X}}k_0(x,x)}{n\wedge m},$$

and by choosing $m = n$, we recover the rate (5) of the

[3] The more challenging nature of this experiment meant accurate computation of the energy distance was precluded, due to the fact that a sufficiently high-quality empirical approximation of P could not be obtained.

original SP algorithm which optimizes over all of $\mathcal{X}$ (Chen et al., 2018b). For bounded kernels, the result improves over the $O(1/n + 1/\sqrt{m})$ independent sampling kernel herding rate established in (Lacoste-Julien et al., 2015, App. B). Thm. 1 more generally accommodates unbounded kernels at the cost of a $\log(n \wedge m)$ factor.

The role of Thm. 1 is limited to providing a stepping stone to Thm. 2, as it is not practical to obtain exact samples from P in general. To state our result in the general case, restrict attention to $\mathcal{X} = \mathbb{R}^d$, consider a function $V : \mathcal{X} \to [1, \infty)$ and define the associated operators $\|f\|_V := \sup_{x\in\mathcal{X}} |f(x)|/V(x)$, $\|\mu\|_V := \sup_{f:\|f\|_V \leq 1} |\int f \mathrm{d}\mu|$ respectively on functions $f : \mathcal{X} \to \mathbb{R}$ and on signed measures μ on $\mathcal{X}$. A Markov chain $(Y_i)_{i\in\mathbb{N}} \subset \mathcal{X}$ with nth step transition kernel P^n is called *V-uniformly ergodic* (Meyn & Tweedie, 2012, Chap. 16) if $\exists R \in [0, \infty), \rho \in (0, 1)$ such that $\|\mathrm{P}^n(y, \cdot) - P\|_V \leq RV(y)\rho^n$ for all initial states $y \in \mathcal{X}$ and all $n \in \mathbb{N}$. The proof of the following is provided in Appendix A.2:

Theorem 2 (SP-MCMC Convergence)**.** *Suppose $\int_{\mathcal{X}} k_0(x, \cdot)\mathrm{d}P(x) \equiv 0$ with $\mathbb{E}_{Z\sim P}[e^{\gamma k_0(Z,Z)}] < \infty$ for $\gamma > 0$. For a sequence $(m_j)_{j=1}^n \subset \mathbb{N}$, let $\{x_i\}_{i=1}^n$ denote the output of SP-MCMC, based on time-homogeneous reversible Markov chains $(Y_{j,l})_{l=1}^{m_j}$, $j \in \mathbb{N}$, generated using the same V-uniformly ergodic transition kernel. Define $V_\pm(s) := \sup_{x:k_0(x,x)\leq s^2} k_0(x, x)^{1/2}V(x)^{\pm 1}$ and $S_i = \sqrt{2\log(n \wedge m_i)/\gamma}$. Then $\exists\, C > 0$ such that*

$$\mathbb{E}\left[D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n)^2\right] \leq \tfrac{C}{n}\sum_{i=1}^n \tfrac{S_i^2}{n} + \tfrac{V_+(S_i)V_-(S_i)}{m_i}.$$

We give an example of verifying the preconditions of Thm. 2 for MALA. Let $\mathcal{P}$ denote the set of distantly dissipative[4] distributions with $\nabla \log p$ Lipschitz on $\mathcal{X} = \mathbb{R}^d$. Let $C_b^{(r,r)}$ 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 and uniformly bounded for $l \in \{0, \ldots, r\}$. Let $q(x, y)$ be a density for the proposal distribution of MALA, and let $\alpha(x, y)$ denote the acceptance probability for moving from x to y, given that y has been proposed. Let $A(x) = \{y \in \mathcal{X} : \alpha(x, y) = 1\}$ denote the region where proposals are always accepted and let $R(x) = \mathcal{X} \setminus A(x)$. Let $I(x) := \{y : \|y\|_2 \leq \|x\|_2\}$. MALA is said to be *inwardly convergent* (Roberts & Tweedie, 1996, Sec. 4) if

$$\lim_{\|x\|_2\to\infty} \int_{A(x)\Delta I(x)} q(x, y)\mathrm{d}y = 0 \qquad (7)$$

where $A\Delta B$ denotes the symmetric set difference $(A\cup B)\setminus(A \cap B)$. The proof of the following is provided in Appendix A.3:

Theorem 3 (SP-MALA Convergence)**.** *Suppose k_0 has the form (3), based on a kernel $k \in C_b^{(1,1)}$ and a target $P \in \mathcal{P}$ such that $\int_{\mathcal{X}} k_0(x, \cdot)\mathrm{d}P(x) \equiv 0$. Let $(m_j)_{j=1}^n \subset \mathbb{N}$ be a fixed sequence and let $\{x_i\}_{i=1}^n$ denote the output of SP-MCMC, based on Markov chains $(Y_{j,l})_{l=1}^{m_j}$, $j \in \mathbb{N}$, generated using MALA transition kernel with step size h sufficiently small. Assume P is such that MALA is inwardly convergent. Then MALA is V-uniformly ergodic for $V(x) = 1 + \|x\|_2$ and $\exists\, C > 0$ such that*

$$\mathbb{E}\left[D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n)^2\right] \leq \tfrac{C}{n}\sum_{i=1}^n \tfrac{\log(n\wedge m_i)}{n\wedge m_i}.$$

Our final result, proved in Appendix A.4, establishes that the pre-conditioner kernel proposed in Sec. 3.2 can control weak congergence to P when the pre-conditionner Λ is symmetric positive definite (denoted $\Lambda \succ 0$). It is a generalisation of Thm. 8 of Gorham & Mackey (2017), who treated the special case of $\Lambda = I$:

Theorem 4 (Pre-conditioned IMQ KSD Controls Convergence)**.** *Suppose k_0 is a Stein kernel (3) for a target $P \in \mathcal{P}$ and a pre-conditioned IMQ base kernel (6) with $\beta \in (-1, 0)$ and $\Lambda \succ 0$. If $D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n) \to 0$ then $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ converges weakly to P.*

6. Conclusion

This paper proposed fundamental improvements to the SP method of (Chen et al., 2018b), establishing, in particular, that the global search used to select each point can be replaced with a finite-length sample path from an MCMC method. The convergence of the proposed SP-MCMC method was established, with an explicit bound provided on the KSD in terms of the V-uniform ergodicity of the Markov transition kernel.

Potential extensions to our SP-MCMC method include the use of fast approximate Markov kernels for P (such as the unadjusted Langevin algorithm; see Appendix A.3), fast approximations to KSD (Jitkrittum et al., 2017; Huggins & Mackey, 2018), exploitation of conditional independence structure in P (Wang et al., 2018; Zhuo et al., 2018) and extension to a general Riemannian manifold $\mathcal{X}$ (Liu & Zhu, 2018; Barp et al., 2018b). One could also attempt to use our MCMC optimization approach to accelerate related algorithms such as kernel herding (Chen et al., 2010; Bach et al., 2012; Lacoste-Julien et al., 2015). Other recent approaches to quantisation in the Bayesian context include (Futami et al., 2018; Hu et al., 2018; Frogner & Poggio, 2018; Zhang et al., 2018; Chen et al., 2018a; Li et al., 2019), and an assessment of the relative performance of these methods would be of interest. However, we note that these approaches are not accompanied by the same level of theoretical guarantees that we have established.

[4] The target P is said to be *distantly dissipative* (Eberle, 2016; Gorham et al., 2019) if $\kappa_0 \triangleq \liminf_{r\to\infty} \kappa(r) > 0$ for $\kappa(r) = \inf\left\{-2\frac{\langle\nabla\log[\tilde{p}(x)-\tilde{p}(y)], x-y\rangle}{\|x-y\|_2^2} : \|x - y\|_2 = r\right\}$.

Acknowledgements

The authors are grateful to the reviewers for their critical feedback on the manuscript. WYC was supported by the Australian Research Council Centre of Excellence for Mathematics and Statistical Frontiers. AB was supported by a Roth Scholarship from the Department of Mathematics at Imperial College London, UK. WYC, AB, FXB, MG and CJO were supported by the Lloyd's Register Foundation programme on data-centric engineering at the Alan Turing Institute, UK. MG was supported by the EPSRC grants [EEP/P020720/1, EP/R018413/1, EP/R034710/1, EP/R004889/1] and a Royal Academy of Engineering Research Chair.

References

Abramowitz, M. and Stegun, I. A. *Handbook of Mathematical Functions*. Dover, 1972.

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

Baringhaus, L. and Franz, C. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1): 190–206, 2004.

Barp, A., Briol, F.-X., Kennedy, A. D., and Girolami, M. Geometry and dynamics for Markov chain Monte Carlo. *Annual Reviews in Statistics and its Applications*, 5:451–471, 2018a.

Barp, A., Oates, C., Porcu, E., and M, G. A Riemannian-Stein kernel method. *arXiv:1810.04946*, 2018b.

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, New York, 2004.

Calderhead, B. and Girolami, M. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics & Data Analysis*, 53(12):4028–4045, 2009.

Chen, C., Zhang, R., Wang, W., Li, B., and Chen, L. A unified particle-optimization framework for scalable Bayesian sampling. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018a.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. Stein points. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 843–852. PMLR, 2018b.

Chen, Y., Welling, M., and Smola, A. Super-samples from kernel herding. In *Proceedings of the 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*, pp. 2606–2615, 2016.

De Marchi, S., Schaback, R., and Wendland, H. Near-optimal data-independent point locations for radial basis function interpolation. *Advances in Computational Mathematics*, 23(3):317–330, 2005.

Detommaso, G., Cui, T., Spantini, A., Marzouk, Y., and Scheichl, R. A Stein variational Newton method. In *Advances in Neural Information Processing Systems 31*, pp. 9187–9197, 2018.

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

Eberle, A. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166 (3-4):851–886, 2016.

Freund, R. M., Grigas, P., and Mazumder, R. An extended Frank–Wolfe method with "in-face" directions, and its application to low-rank matrix completion. *SIAM Journal on Optimization*, 27(1):319–346, 2017.

Frogner, C. and Poggio, T. Approximate inference with Wasserstein gradient flows. *arXiv:1806.04542*, 2018.

Futami, F., Cui, Z., Sato, I., and Sugiyama, M. Frank-Wolfe Stein sampling. *arXiv:1805.07912*, 2018.

Girolami, M. and Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society. Series B*, 73(2):123–214, 2011.

Goodwin, B. C. Oscillatory behavior in enzymatic control process. *Advances in Enzyme Regulation*, 3:318–356, 1965.

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., Mackey, L., and Vollmer, S. Measuring sample quality with diffusions. *Annals of Applied Probability*, 2019. In press.

Graf, S. and Luschgy, H. *Foundations of Quantization for Probability Distributions*. Springer, 2007.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pp. 513–520, 2006.

Hu, T., Chen, Z., Sun, H., Bai, J., Ye, M., and Cheng, G. Stein neural sampler. *arXiv:1810.03545*, 2018.

Huggins, J. and Mackey, L. Random feature Stein discrepancies. In *Advances in Neural Information Processing Systems 31*, pp. 1903–1913. 2018.

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.

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

Joseph, V. R., Wang, D., Gu, L., Lyu, S., and Tuo, R. Deterministic sampling of expensive posteriors using minimum energy designs. *Technometrics*, 2018. To appear.

Lacoste-Julien, S. and Jaggi, M. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems*, pp. 496–504, 2015.

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.

Li, L., Li, Y., Liu, J., Liu, Z., and Lu, J. A stochastic version of Stein variational gradient descent for efficient sampling. *arXiv:1902.03394*, 2019.

Liu, C. and Zhu, J. Riemannian Stein variational gradient descent for Bayesian inference. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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. 2378–2386, 2016.

Liu, Q. and Wang, D. Stein variational gradient descent as moment matching. In *Advances in Neural Information Processing Systems*, pp. 8868–8877, 2018.

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

Lu, J., Lu, Y., and Nolen, J. Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 2018. To appear.

Mak, S. and Joseph, V. R. Support points. *Annals of Statistics*, 46(6A):2562–2592, 2018.

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

Meyn, S. and Tweedie, R. *Markov Chains and Stochastic Stability*. Springer Science & Business Media., 2012.

Muller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

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

Oates, C. J., Papamarkou, T., and Girolami, M. The controlled thermodynamic integral for Bayesian model evidence evaluation. *Journal of the American Statistical Association*, 111(514):634–645, 2016.

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

Parno, M. D. *Transport maps for accelerated Bayesian computation*. PhD thesis, Massachusetts Institute of Technology, 2015.

Robert, C. and Casella, G. *Monte Carlo Statistical Methods*. Springer, 2004.

Roberts, G. O. and Rosenthal, J. S. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

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

Santin, G. and Haasdonk, B. Convergence rate of the data-independent P-greedy algorithm in kernel-based approximation. *Dolomites Research Notes on Approximation*, 10, 2017.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *Annals of Statistics*, pp. 2263–2291, 2013.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of 6th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 583–602. University of California Press, 1972.

Székely, G. and Rizzo, M. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

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

Wang, D., Zeng, Z., and Liu, Q. Stein variational message passing for continuous graphical models. In *Proceedings of the 35th International Conference on Machine Learning, PMLR 80*, pp. 5219–5227, 2018.

Zhang, R., Chen, C., Li, C., and Carin, L. Policy optimization as Wasserstein gradient flows. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5737–5746, 2018.

Zhuo, J., Liu, C., Shi, J., Zhu, J., Chen, N., and Zhang, B. Message passing Stein variational gradient descent. In *Proceedings of the 35th International Conference on Machine Learning, PMLR 80:6013-6022*, 2018.