


# Optimal thinning of MCMC output

Marina Riabiz<sup>1,2</sup> | Wilson Ye Chen<sup>3</sup> | Jon Cockayne<sup>2</sup> |  
Pawel Swietach<sup>4</sup> | Steven A. Niederer<sup>1</sup> | Lester Mackey<sup>5</sup> |  
Chris. J. Oates<sup>2,6</sup> 

<sup>1</sup>King's College London, London, UK

<sup>2</sup>Alan Turing Institute, London, UK

<sup>3</sup>University of Sydney, Sydney, Australia

<sup>4</sup>Oxford University, Oxford, UK

<sup>5</sup>Microsoft Research, Cambridge, USA

<sup>6</sup>Newcastle University, Newcastle Upon Tyne, UK

## Correspondence

Chris. J. Oates, School of Mathematics, Statistics and Physics, Herschel Building, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.

Email: [chris.oates@ncl.ac.uk](mailto:chris.oates@ncl.ac.uk)

## Abstract

The use of heuristics to assess the convergence and compress the output of Markov chain Monte Carlo can be sub-optimal in terms of the empirical approximations that are produced. Typically a number of the initial states are attributed to 'burn in' and removed, while the remainder of the chain is 'thinned' if compression is also required. In this paper, we consider the problem of retrospectively selecting a subset of states, of fixed cardinality, from the sample path such that the approximation provided by their empirical distribution is close to optimal. A novel method is proposed, based on greedy minimisation of a kernel Stein discrepancy, that is suitable when the gradient of the log-target can be evaluated and approximation using a small number of states is required. Theoretical results guarantee consistency of the method and its effectiveness is demonstrated in the challenging context of parameter inference for ordinary differential equations. Software is available in the Stein Thinning package in Python, R and MATLAB.

## KEYWORDS

Bayesian computation, greedy optimisation, Markov chain Monte Carlo, reproducing kernel, Stein's method

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

## 1 | INTRODUCTION

The most popular computational tool for non-conjugate Bayesian inference is Markov chain Monte Carlo (MCMC). Introduced to statistics from the physics literature in Hastings (1970), Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990), an enormous amount of research effort has since been expended in the advancement of MCMC methodology. Such is the breadth of this topic that we do not attempt a survey here, but instead refer the reader to Robert and Casella (2013), Green et al. (2015) and the references therein to more advanced material. This paper is motivated by the fact that the approaches used for convergence assessment and to post-process the output of MCMC can strongly affect the estimates that are produced.

Let  $P$  be a distribution on a measurable space  $\mathcal{X}$  and let  $(X_i)_{i \in \mathbb{N}}$  be a Markov chain that is  $P$ -invariant. The Markov chain sample path provides an empirical approximation

$$\frac{1}{n} \sum_{i=1}^n \delta(X_i) \quad (1)$$

to  $P$ , where  $\delta(x)$  denotes a point mass centred at  $x \in \mathcal{X}$ . Our discussion supposes that a practitioner is prepared to simulate a Markov chain up to a maximum number of iterations,  $n$ , and that simulating further iterations is not practical; a scenario that is often encountered (e.g. see Section 4.3). In this setting, it is common (and indeed recommended) to replace Equation (1) with an alternative estimator

$$\frac{1}{m} \sum_{j=1}^m \delta(X_{\pi(j)}) \quad (2)$$

that is based on a subset of the total MCMC output. The  $m$  indices  $\pi(j) \in \{1, \dots, n\}$  indicate which states are retained and the identification of a suitable index set  $\pi$  is informed by the following considerations:

**Removal of Initial Bias:** The distribution of the initial states of the Markov chain may be quite different to  $P$ . To mitigate this, it is desirable to identify a ‘burn-in’  $(X_i)_{i=1}^b$  which is then discarded. The burn-in period  $b$  is typically selected using convergence diagnostics (Cowles & Carlin, 1996). These are primarily based on the empirical distribution of simple moment, quantile or density estimates across independent chains and making a judgement as to whether the ensemble of chains has converged to the distributional target. The main limitation of convergence diagnostics, as far as we are concerned in this work, is that in taking  $b$  large enough to make bias negligible, the number  $n - b$  of remaining samples may be rather small, such that the statistical efficiency of the estimator in Equation (2) is sub-optimal as an approximation of  $P$ . Nonetheless, a considerable portion of Bayesian pedagogy is devoted to the identification of the burn-in period, as facilitated using diagnostic tests that are built into commercial-grade software such as WinBUGS (Lunn et al., 2000), JAGS (Plummer, 2003), R (R Core Team, 2020), and Stan (Carpenter et al., 2017).

**Increased Statistical Efficiency:** It is often stated that discarding part of the MCMC output leads to a reduction in the statistical efficiency of the estimator (2) compared to Equation (1). This argument, made for example, in Geyer (1992), applies only when the procedure used to discard part of the MCMC output does not itself depend on the MCMC output and when the length  $n$  of the MCMC output is fixed. That estimation efficiency can be *improved* by discarding a portion of

the samples in a way that depends on the samples themselves is in fact well-established (see e.g. Dwivedi et al., 2019).

**Compression of MCMC Output:** A third motivation for estimators of the form Equation (2) is to control the cost of subsequent computation involving the MCMC output. Examples include approximating the expectation of a function  $f$ , where either evaluation of  $f$  or storage of its output is associated with a computational cost, and *Monte Carlo Maximum Likelihood*, where one constructs an approximate likelihood using MCMC, then performs optimisation on this approximate likelihood (Geyer & Thompson, 1992). In such situations one may want to control the cardinality  $m$  of the index set  $\pi$  and to use  $(X_{\pi(j)})_{j=1}^m$  as an experimental design on which  $f$  is evaluated. The most popular solution is to retain only every  $t$ th state visited by the Markov chain, a procedure known as ‘thinning’ of the MCMC output. See also the more sophisticated approach in Paige et al. (2016).

Taking these considerations into account, the most common approach used to select an index set  $\pi$  is based on the identification of a suitable burn-in period  $b$  and/or a suitable thinning frequency  $t$ , leading to an approximation of the form

$$\frac{1}{\lfloor (n-b)/t \rfloor} \sum_{i=1}^{\lfloor (n-b)/t \rfloor} \delta(X_{b+it}). \quad (3)$$

Here  $\lfloor r \rfloor$  denotes the integer part of  $r$ . This corresponds to a set of indices  $\pi$  in Equation (2) that discards the burn-in states and retains only every  $t$ th iteration from the remainder of the MCMC output. It includes the case where no states are removed when  $b = 0$  and  $t = 1$ . Despite their widespread usage, the interplay between the Markov chain sample path and the heuristics used to select  $b$  and  $t$  is not widely appreciated. In general it is unclear how much bias may be introduced by employing a post-processing heuristic that is itself based on the MCMC output. Indeed, even the basic question of when the post-processed estimator in Equation (3) is consistent when  $b$  and  $t$  are chosen based on the MCMC output appears not to have been studied.

In this paper we propose a novel method, called *Stein Thinning*, that selects an index set  $\pi$ , of specified cardinality  $m$ , such that the associated discrete approximation in Equation (2) is close to optimal among all approximations supported on the MCMC output. The method is designed to ensure that Equation (2) is a consistent approximation of  $P$ . This includes situations when the Markov chain on which it is based is not  $P$ -invariant, but we do of course require that the regions of high probability under  $P$  are explored. To achieve this we adopt a kernel Stein discrepancy as our optimality criterion. The minimisation of kernel Stein discrepancy is performed using a greedy sequential algorithm and the main contribution of our theoretical analysis is to study the interplay of the greedy algorithm with the randomness inherent to the MCMC output. The proposed *Stein Thinning* method is simple (see Algorithm 1), applicable to most problems where gradients of the log-posterior density can be computed, and implemented as convenient Python and MATLAB packages that require no additional user input other than the number  $m$  of states to be selected (see Appendix S1).

## 1.1 | Related work

Our work contributes to an active area of research that attempts to cast post-processing of MCMC as an optimisation problem. Mak and Joseph (2018) proposed a method, called *Support Points*, which selects a small number of states in order that an empirical measure supported

on those states minimises an ‘energy distance’ to  $P$ . However, computation of the energy distance requires access to  $P$ , and minimisation of energy distance requires a challenging non-convex optimisation problem to be solved, meaning that in practice approximations are required. Stein discrepancy provides a computable alternative, which was used in Liu and Lee (2017) to optimally weight an arbitrary set  $(X_i)_{i=1}^n \subset \mathbb{R}^d$  of states in a manner loosely analogous to importance sampling, at a computational cost of  $O(n^3)$ . The combined effect of applying the approach of Liu and Lee (2017) to MCMC output was analysed in Hodgkinson et al. (2020), who established situations in which the overall procedure will be consistent.

If a compressed representation of the posterior  $P$  is required, but one is not wedded to the use of MCMC for generation of candidate states, then several other methods can be used. Joseph et al. (2015, 2019) proposed a criterion to capture how well an empirical measure based on a point set approximates  $P$  and applied repeated numerical optimisation over  $\mathcal{X}$  to arrive at a suitable point set. A similar approach was taken in Chen et al. (2018), where a Stein discrepancy was numerically minimised. The reliance of both of these algorithms on non-convex numerical optimisation over  $\mathcal{X}$  renders their implementation and analysis difficult. Chen et al. (2019) considered using Markov chains to approximately perform numerical optimisation, allowing a tractable analytic treatment at the expense of a sub-optimal compression of  $P$ . An elegant alternative approach is to formulate a convex optimisation problem on the set of probability distributions on  $\mathcal{X}$ . In this spirit, Liu and Wang (2016) and Liu (2017) identified a gradient flow with  $P$  as a fixed point that can be approximately simulated using a particle method. At convergence, one obtains a compressed representation of  $P$ , however the theoretical analysis of this approach remains an open and active research topic (see e.g. Duncan et al., 2019).

The present paper differs from the contributions cited, in that (1) our algorithm requires only the output from one run of MCMC, which is a realistic requirement in many situations, and (2) we are able to provide a finite sample size error bound (Theorem 2) and a consistency guarantee (Theorem 3) for Stein Thinning, that cover precisely the algorithm that we implement.

## 1.2 | Outline of the paper

The paper proceeds, in Section 2, to recall the construction of a kernel Stein discrepancy and to present Stein Thinning. Then in Section 3 we establish a finite sample size error bound, as well as a widely-applicable consistency result that does not require the Markov chain to be  $P$ -invariant. In Section 4 we present an empirical assessment of Stein Thinning in the context of parameter inference for ordinary differential equation models. Conclusions are contained in Section 5.

## 2 | METHODS

In this section we introduce and analyse Stein Thinning. First, in Section 2.1, we recall the construction of a kernel Stein discrepancy and its theoretical properties. The Stein Thinning method is presented in Section 2.2, whilst Section 2.3 is devoted to implementational detail.

Before we proceed, we introduce a piece of notation that will often be used and recall the mathematical definition of a reproducing kernel:

**Notation:** Let  $\mathcal{P}$  denote the set of probability distributions  $P$  that admit a positive density  $p$  with Lipschitz gradient  $\nabla \log p$  on  $\mathbb{R}^d$ .

**Reproducing Kernel:** A reproducing kernel Hilbert space (RKHS) of functions on a set  $\mathcal{X}$  is a Hilbert space, denoted  $\mathcal{H}(k)$ , equipped with a function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , called a *kernel*, such that  $\forall x \in \mathcal{X}$  we have  $k(\cdot, x) \in \mathcal{H}(k)$  and  $\forall x \in \mathcal{X}, h \in \mathcal{H}(k)$  we have  $h(x) = \langle h, k(\cdot, x) \rangle_{\mathcal{H}(k)}$ . In this paper  $\langle \cdot, \cdot \rangle_{\mathcal{H}(k)}$  denotes the inner product in  $\mathcal{H}(k)$  and the induced norm will be denoted  $\| \cdot \|_{\mathcal{H}(k)}$ . For further details, see Berlinet and Thomas-Agnan (2004).

## 2.1 | Kernel Stein discrepancy

To construct a criterion for the selection of states from the MCMC output we require a notion of optimal approximation for probability distributions. To this end, recall that an *integral probability metric* (IPM) (Muller, 1997), based on a set  $\mathcal{F}$  of measure-determining functions on a measurable space  $\mathcal{X}$ , is defined as

$$D_{\mathcal{F}}(P, Q) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f dP - \int_{\mathcal{X}} f dQ \right|. \tag{4}$$

The fact that  $\mathcal{F}$  is measure-determining means that  $D_{\mathcal{F}}(P, Q) = 0$  if and only if  $P = Q$  is satisfied. Standard choices for  $\mathcal{F}$ , for example, that recover Wasserstein distance as the IPM, cannot be used in the Bayesian context due to the need to compute integrals with respect to  $P$  in Equation (4).

In the remainder of Section 2.1 we restrict attention to the setting  $P \in \mathcal{P}$ . To circumvent intractability of Equation (4), the notion of a *Stein discrepancy* was proposed in Gorham and Mackey (2015). This was based on Stein’s method (Stein, 1972), which consists of finding a set  $\mathcal{G}$  of sufficiently differentiable  $d$ -dimensional vector fields and a differential operator  $\mathcal{A}_P$ , depending on  $P$  and acting on elements of  $\mathcal{G}$ , such that  $\int_{\mathbb{R}^d} \mathcal{A}_P g \, dP = 0$  for all  $g \in \mathcal{G}$ . The proposal of Gorham and Mackey (2015) was to take  $\mathcal{F} = \mathcal{A}_P \mathcal{G}$  to be the image of  $\mathcal{G}$  under  $\mathcal{A}_P$  in Equation (4), leading to the *Stein discrepancy*

$$D_{\mathcal{A}_P \mathcal{G}}(P, Q) = \sup_{g \in \mathcal{G}} \left| \int_{\mathbb{R}^d} \mathcal{A}_P g \, dQ \right|. \tag{5}$$

Theoretical analysis had led to sufficient conditions for  $\mathcal{A}_P \mathcal{G}$  to be measure-determining (Gorham & Mackey, 2015). In this paper we focus on a particular form of Equation (5) due to Liu et al. (2016), Chwialkowski et al. (2016), Gorham and Mackey (2017), called a *kernel Stein discrepancy* (KSD). In this case,  $\mathcal{A}_P$  is the *Langevin Stein operator*  $\mathcal{A}_P g := p^{-1} \nabla \cdot (pg)$  derived in Gorham and Mackey (2015), where  $\nabla \cdot$  denotes the divergence operator in  $\mathbb{R}^d$  and  $\mathcal{G} := \{g : \mathbb{R}^d \rightarrow \mathbb{R}^d \mid \sum_{i=1}^d \|g_i\|_{\mathcal{H}(k)}^2 \leq 1\}$  is the unit ball in a Cartesian product of RKHS. It follows from construction that the set  $\mathcal{A}_P \mathcal{G}$  is the unit ball of another RKHS, denoted  $\mathcal{H}(k_P)$ , whose kernel is

$$k_P(x, y) := \nabla_x \cdot \nabla_y k(x, y) + \langle \nabla_x k(x, y), \nabla_y \log p(y) \rangle + \langle \nabla_y k(x, y), \nabla_x \log p(x) \rangle + k(x, y) \langle \nabla_x \log p(x), \nabla_y \log p(y) \rangle, \tag{6}$$

where  $\langle \cdot, \cdot \rangle$  denotes the standard Euclidean inner product,  $\nabla$  denotes the gradient operator and subscripts have been used to indicate the variables being acted on by the differential operators (Oates et al., 2017). Thus KSD is recognised as a maximum mean discrepancy in  $\mathcal{H}(k_P)$  (Song, 2008) and is fully characterised by the kernel  $k_P$ ; we therefore adopt the shorthand notation  $D_{k_P}(Q)$  for  $D_{\mathcal{A}_P \mathcal{G}}(P, Q)$ .

In the remainder of this section we recall the main properties of KSD. The first is a condition on the kernel  $k$  that guarantees elements of  $\mathcal{H}(k_p)$  have zero mean with respect to  $P$ . In what follows  $\|x\| = \langle x, x \rangle^{1/2}$  denotes the Euclidean norm on  $\mathbb{R}^d$ . It will be convenient to abuse operator notation, writing  $\nabla_x \nabla_y^\top k$  for the Hessian matrix of a bivariate function  $(x, y) \mapsto k(x, y)$ .

**Proposition 1** (Proposition 1 of Gorham & Mackey, 2017). *Let  $P \in \mathcal{P}$  and assume that  $\int_{\mathbb{R}^d} \|\nabla \log p\| dP < \infty$ . Let  $(x, y) \mapsto \nabla_x \nabla_y^\top k(x, y)$  be continuous and uniformly bounded on  $\mathbb{R}^d$ . Then  $\int_{\mathbb{R}^d} h dP = 0$  for all  $h \in \mathcal{H}(k_p)$ , where  $k_p$  is defined in Equation (6).*

The second main property of KSD that we will need is that it can be explicitly computed for an empirical measure  $Q = \frac{1}{n} \sum_{i=1}^n \delta(x_i)$ , supported on states  $x_i \in \mathbb{R}^d$ :

**Proposition 2** (Proposition 2 of Gorham & Mackey, 2017). *Let  $P \in \mathcal{P}$  and let  $(x, y) \mapsto \nabla_x \nabla_y^\top k(x, y)$  be continuous on  $\mathbb{R}^d$ . Then*

$$D_{k_p} \left( \frac{1}{n} \sum_{i=1}^n \delta(x_i) \right) = \sqrt{\frac{1}{n^2} \sum_{i,j=1}^n k_p(x_i, x_j)}, \quad (7)$$

where  $k_p$  was defined in Equation (6).

The third main property is that KSD provides convergence control. Let  $Q_n \Rightarrow P$  denote weak convergence of a sequence  $(Q_n)$  of measures to  $P$ . Theoretical analysis in Gorham and Mackey (2017), Chen et al. (2018), Huggins and Mackey (2018), Chen et al. (2019), Hodgkinson et al. (2020), and Gorham et al. (2020) established sufficient conditions for when convergence of Equation (7) to zero implies  $\frac{1}{n} \sum_{i=1}^n \delta(x_i) \Rightarrow P$ . For our purposes we present one such result, from Chen et al. (2019).

**Proposition 3** (Theorem 4 in Chen et al., 2019). *Let  $P \in \mathcal{P}$  be distantly dissipative, meaning that  $\liminf_{r \rightarrow \infty} \kappa(r) > 0$  where*

$$\kappa(r) : \inf \left\{ -2 \frac{\langle \nabla \log p(x) - \nabla \log p(y), x - y \rangle}{\|x - y\|^2} : \|x - y\| = r \right\}.$$

*Consider the kernel  $k(x, y) = (c^2 + \|\Gamma^{-1/2}(x - y)\|^2)^\beta$  for some fixed  $c > 0$ , a fixed positive definite matrix  $\Gamma$  and a fixed exponent  $\beta \in (-1, 0)$ . Then  $D_{k_p}(\frac{1}{n} \sum_{i=1}^n \delta(x_i)) \rightarrow 0$  implies  $\frac{1}{n} \sum_{i=1}^n \delta(x_i) \Rightarrow P$ , where  $k_p$  is defined in Equation (6).*

The properties just described ensure that KSD is a suitable optimality criterion to consider for the post-processing of MCMC output. However, all discrepancies are associated with finite sample size pathologies; see (Matsubara et al. 2021 Section 3.4) for a discussion of the pathologies of KSD. Our attention turns next to the development of algorithms for minimisation of KSD.

## 2.2 | Greedy minimisation of KSD

The convergence control afforded by Proposition 3 motivates the design of methods that select points  $(x_i)_{i=1}^n$  such that Equation (7) is approximately minimised. Continuous optimisation algorithms were proposed for this task in Chen et al. (2018) and Chen et al. (2019). In Chen et al. (2018), deterministic optimisation techniques were considered for low-dimensional problems,

whereas in Chen et al. (2019) a Markov chain was used to provide more a practical optimisation strategy when the state space is high-dimensional. In each case greedy sequential strategies were considered, wherein at iteration  $n$  a new state  $x_n$  is appended to the current sequence  $(x_1, \dots, x_{n-1})$ . Chen et al. (2018) also considered the use of conditional gradient algorithms (so-called *Frank-Wolfe*, or *kernel herding* algorithms) but found that greedy algorithms provided better performance across a range of experiments and therefore we focus on greedy algorithms in this manuscript.

The present paper is distinguished from earlier work in that we do not attempt to solve a continuous optimisation problem for selection of the next point  $x_n \in \mathcal{X}$ . Such optimisation problems are fundamentally difficult and can at best be approximately solved. Instead, we exactly solve the discrete optimisation problem of selecting a suitable element  $x_n$  from supplied MCMC output. In this sense, we expect our findings will be more widely applicable than previous work, since we are simply performing post-processing of MCMC output and there exists a variety of commercial-grade software for MCMC. The method that we propose, called `Stein Thinning`, is straight-forward to implement, and is stated in Algorithm 1 for a distribution  $P$  on a general measurable space  $\mathcal{X}$ . (The convention  $\sum_{i=1}^0 = 0$  is employed.)

**Data:** The output  $(x_i)_{i=1}^n$  from an MCMC method, a kernel  $k_P$  for which the conclusion of Proposition 3 holds, and a desired cardinality  $m \in \mathbb{N}$ .

**Result:** The indices  $\pi$  of a sequence  $(x_{\pi(j)})_{j=1}^m \subset \{x_i\}_{i=1}^n$  where the  $\pi(j)$  are elements of  $\{1, \dots, n\}$ .

**for**  $j = 1, \dots, m$  **do**

$$\left| \pi(j) \in \arg \min_{i=1, \dots, n} \frac{k_P(x_i, x_i)}{2} + \sum_{j'=1}^{j-1} k_P(x_{\pi(j')}, x_i); \right.$$

**end**

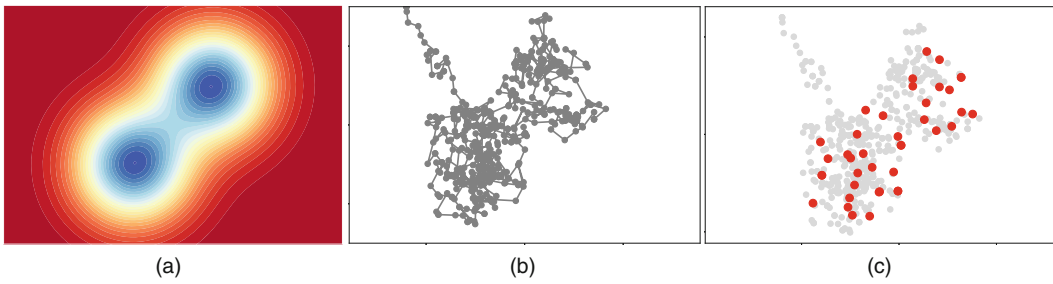
**Algorithm 1:** The proposed method; Stein Thinning.

The algorithm is illustrated on a simple bivariate Gaussian mixture in Figure 1. Observe in this figure that the points selected by the `Stein Thinning` do not belong to the burn-in period (which is visually clear), and that although the MCMC spent a disproportionate amount of time in one of the mixture components, the number of points selected by `Stein Thinning` is approximately equal across the two components of the target. The accuracy of the approximation produced by `Stein Thinning` is, nevertheless, gated by the quality of the MCMC output to which it is applied. A detailed empirical assessment is presented in Section 4.

*Remark 1* (Tie-breaking). In the event of a tie, a tie-breaking rule should be used to select the next index. For example, if the minimum in Algorithm 1 is realised by multiple candidate values  $\Pi(j) \subseteq \{1, \dots, n\}$ , one could adopt a tie-breaking rule that selects the smallest element of  $\Pi(j)$  as the value that is assigned to  $\pi(j)$ . The rule that is used has no bearing on our theoretical analysis in Section 3.

*Remark 2* (Complexity). The computation associated with iteration  $j$  of Algorithm 1 is  $O(nr_j)$  where  $r_j \leq \min(j, n)$  is the number of distinct indices in  $\{\pi(1), \dots, \pi(j-1)\}$ ; the computational complexity of Algorithm 1 is therefore  $O(n \sum_{j=1}^m r_j)$ . For typical MCMC algorithms the computational complexity is  $O(n)$ , so the complexity of `Stein Thinning` is equal to that for MCMC when  $m$  is fixed and higher when  $m$  is increasing with  $n$ , being at most  $O(nm^2)$ .





**FIGURE 1** Illustration of Stein Thinning: (a) Contours of the distributional target  $P$ . (b) Markov chain Monte Carlo (MCMC) output, limited to 500 iterations to mimic a challenging computational context, exhibiting burn-in and autocorrelation that must be identified and mitigated. (c) A subset of  $m = 40$  states from the MCMC output selected using Stein Thinning, which correctly ignores the burn-in period and stratifies states approximately equally across the two components of the target [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

**Remark 3** (Re-sampling). In general, the indices in  $\pi$  need not be distinct. That is, Algorithm 1 may prefer to include a duplicate state rather than to include a state which is not useful for representing  $P$ . Indeed, if  $m > n$  then the sequence  $(x_{\pi(j)})_{j=1}^m$  must contain duplicate entries. Theorem 1 in Section 3 clarifies this behaviour.

**Remark 4** (Finite sample error bound). The approximation produced by Stein Thinning satisfies a finite sample error bound

$$\left| \frac{1}{m} \sum_{j=1}^m f(x_{\pi(j)}) - \int_{\mathbb{R}^d} f(x) dP(x) \right| \leq D_{k_p} \left( \frac{1}{m} \sum_{j=1}^m \delta(x_{\pi(j)}) \right) \left\| f - \int_{\mathbb{R}^d} f(x) dP(x) \right\|_{\mathcal{H}(k_p)}$$

following Hickernell (1998) Equation (1.14) applied to  $f - \int f dP$ . This can be contrasted with the typically asymptotic analysis of MCMC. The practical estimation of the final term in this bound was discussed in Section 4 of South et al. (2021).

### 2.3 | Choice of Kernel

The suitability of KSD to quantify how well  $Q$  approximates  $P$  is determined by the choice of the kernel  $k$  in Equation (6). Several choices are possible and for  $P \in \mathcal{P}$ , based on Proposition 3 together with extensive empirical assessment, Chen et al. (2019) advocated the pre-conditioned inverse multi-quadric kernel  $k(x, y) := (1 + \|\Gamma^{-1/2}(x - y)\|^2)^{-1/2}$  where, compared to Proposition 3, we have fixed  $c = 1$  (without loss of generality) and  $\beta = -1/2$ . The suitability of these choices for Stein Thinning is verified in Appendix S5.1. The positive definite matrix  $\Gamma$  remains to be specified and it is natural to take a data-driven approach where the MCMC output is used to select  $\Gamma$ . Provided that a fixed number  $n_0 \in \mathbb{N}$  of the states  $(X_i)_{i=1}^{n_0}$  from the MCMC output are used in the construction of  $\Gamma$ , the consistency results for Stein Thinning that we establish in Section 3 are not affected. To explore different strategies for the selection of  $\Gamma$ , we focus on the following candidates:

- 1. Median (med):** The scaled identity matrix  $\Gamma = \ell^2 I$ , where  $\ell = \text{med} := \text{median}\{\|X_i - X_j\| : 1 \leq i < j \leq n_0\}$  is the median Euclidean distance between states (Garreau et al., 2018). In the



rare case that  $\text{med} = 0$ , an exception should be used, such as  $\ell = 1$ , to ensure a positive definite  $\Gamma$  is used.

2. **Scaled median** (`sclmed`): The scaled identity matrix  $\Gamma = \ell^2 I$ , where  $\ell = \text{med} / \sqrt{\log(m)}$ . This was proposed in Liu and Wang (2016) and can be motivated using the approximation  $\sum_{j'=1}^m k_P(x_{\pi(j)}, x_{\pi(j')}) \approx m \exp(-\ell^{-2} \text{med}^2) = 1$ . Note the dependence on  $m$  means that the preceding theoretical analysis does not apply when this heuristic is used.
3. **Sample covariance** (`smpcov`): The matrix  $\Gamma$  can be taken as a sample covariance matrix

$$\Gamma = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (X_i - \bar{X})(X_i - \bar{X})^\top, \quad \bar{X} := \frac{1}{n_0} \sum_{i=1}^{n_0} X_i,$$

provided that this matrix is non-singular.

The experiments in Section 4 shed light on which of these settings is the most effective, but we acknowledge that many other settings could also be considered. In what follows, we set  $n_0 = \min(n, 10^3)$  for the `med` and `sclmed` settings, to avoid an  $O(n^2)$  cost of computing  $\ell$ , and otherwise set  $n_0 = n$ , so that the whole of the MCMC output is used to select  $\Gamma$ . Python, R and MATLAB packages are provided and their usage is described in Appendix S1.

### 3 | THEORETICAL ASSESSMENT

The theoretical analysis in this section clarifies the limiting behaviour of Stein Thinning as  $m, n \rightarrow \infty$ . Our first main result concerns the behaviour of Stein Thinning on a fixed sequence  $(x_i)_{i=1}^n$ :

**Theorem 1** *Let  $\mathcal{X}$  be a measurable space and let  $P$  be a probability distribution on  $\mathcal{X}$ . Let  $k_P : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a reproducing kernel with  $\int_{\mathcal{X}} k_P(x, \cdot) dP = 0$  for all  $x \in \mathcal{X}$ . Let  $(x_i)_{i=1}^n \subset \mathcal{X}$  be fixed and consider an index sequence  $\pi$  of length  $m$  produced by Algorithm 1. Then we have the bound*

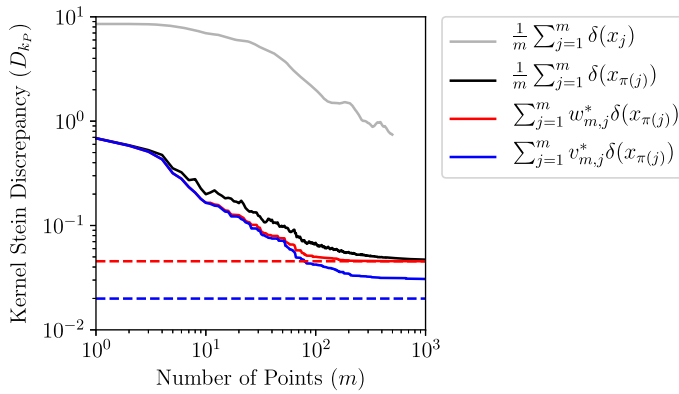
$$D_{k_P} \left( \frac{1}{m} \sum_{j=1}^m \delta(x_{\pi(j)}) \right)^2 \leq D_{k_P} \left( \sum_{i=1}^n w_i^* \delta(x_i) \right)^2 + \left( \frac{1 + \log(m)}{m} \right) \max_{i=1, \dots, n} k_P(x_i, x_i),$$

where the weights  $w^* = (w_1^*, \dots, w_n^*)$  in the first term satisfy

$$w^* \in \arg \min_{\substack{1_n^\top w = 1 \\ w \geq 0}} D_{k_P} \left( \sum_{i=1}^n w_i \delta(x_i) \right)$$

where  $1_n^\top = (1, \dots, 1)$  and  $w \geq 0$  indicates that  $w_i \geq 0$  for  $i = 1, \dots, n$ .

The proof of Theorem 1 is provided in Appendix S2.1. Its implication is that, given a sequence  $(x_i)_{i=1}^n$ , Stein Thinning produces an empirical distribution that converges in KSD to the optimal weighted empirical distribution  $\sum_{i=1}^n w_i^* \delta(x_i)$  based on that sequence. Properties of such optimally weighted empirical measures were studied in Liu and Lee (2017), Hodgkinson et al. (2020), and are not the focus of the present paper, where the case  $m \ll n$  is of principal interest.



**FIGURE 2** Illustration of Theorem 1: The gray curve represents the unprocessed output from MCMC in the example of Figure 1. The black curve represents Stein Thinning applied to this same output and, in addition, weighted output of Stein Thinning is shown for weights  $w_m^*$  subject subject to  $\sum w_{m,j}^* = 1$  and  $w_{m,j}^* \geq 0$  (solid red) and weights  $v_m^*$  subject only to  $\sum v_{m,j}^* = 1$  (solid blue). The dashed horizontal lines are the limiting values of their corresponding solid lines as the number  $m$  is increased [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

The role of Theorem 1 is to study the interaction between the greedy algorithm and a given sequence  $(x_i)_{i=1}^n$ , and this bound is central to our proof of Theorem 2 which deals with the case where  $(x_i)_{i=1}^n$  is replaced by MCMC output. Figure 2 illustrates the terms involved in Theorem 1. It is clear that a reduction in KSD is achieved by Stein Thinning of the MCMC output.

*Remark 5* (Optimal weights). To further improve the empirical approximation, we can consider an optimally-weighted sum  $\sum_{j=1}^m w_{m,j}^* \delta(x_{\pi(j)})$  where the  $w_{m,j}^*$  solve a convex optimisation problem analogous to Equation (8). Such weights minimise a quadratic function subject to a linear and a non-negativity constraint and can therefore be precisely computed. If the non-negativity constraint is removed and the indices in  $\pi$  are distinct then

$$v_m^* := \arg \min_{\mathbf{1}_m^\top v = 1} D_{k_P} \left( \sum_{j=1}^m v_j \delta(x_{\pi(j)}) \right) = \frac{K_P^{-1} \mathbf{1}_m}{\mathbf{1}_m^\top K_P^{-1} \mathbf{1}_m}, \quad (K_P)_{i,j} := k_P(x_{\pi(i)}, x_{\pi(j)}),$$

as derived in Oates et al. (2017). Figure 2 indicates that the benefit of applying weights  $w_m^*$  (red curve) to the output of Stein Thinning (black curve) is limited, likely because the  $x_{\pi(j)}$  were selected in a way that avoids redundancy in the point set. A larger improvement is provided by the weights  $v_m^*$  (blue curve), but in this case the associated empirical measure may not be a probability distribution.

*Remark 6* The use of a conditional gradient algorithm, instead of a greedy algorithm, in this context, amounts to simply removing the term  $k_P(x_{\pi(i)}, x_{\pi(j)})$  in Algorithm 1. As discussed in Chen et al. (2018), this term can be thought of as a regulariser that lends stability to the algorithm, avoiding selection of  $x_i$  that are far from the effective support of  $P$ .

*Remark 7* Theorem 1 is formulated at a high level of generality and can be applied on non-Euclidean domains  $\mathcal{X}$ . In Barp et al. (2021), Liu and Zhu (2018), Xu and Matsuda (2020) and Le et al. (2020) the authors proposed and discussed Stein operators  $\mathcal{A}_P$  for the non-Euclidean context.

Next, we consider the properties of Stein Thinning applied to MCMC output. Let  $V$  be a function  $V : \mathcal{X} \rightarrow [1, \infty)$  and, for a function  $f : \mathcal{X} \rightarrow \mathbb{R}$  and a measure  $\mu$  on  $\mathcal{X}$ , let  $\|f\|_V := \sup_{x \in \mathcal{X}} \frac{|f(x)|}{V(x)}$ ,  $\|\mu\|_V := \sup_{\|f\|_V \leq 1} |\int_{\mathcal{X}} f d\mu|$ . Recall that a  $\psi$ -irreducible and aperiodic Markov chain  $(X_i)_{i \in \mathbb{N}} \subset \mathcal{X}$  with  $n$ th step transition kernel  $P^n$  is  $V$ -uniformly ergodic (see Theorem 16.0.1 of Meyn & Tweedie, 2012) if and only if  $\exists R \in [0, \infty)$ ,  $\rho \in (0, 1)$  such that

$$\|P^n(x, \cdot) - P\|_V \leq RV(x)\rho^n \tag{9}$$

for all initial states  $x \in \mathcal{X}$  and all  $n \in \mathbb{N}$ . The notation  $\mathbb{E}$  will be used to denote expectation with respect to the law of the Markov chain in the sequel. Theorem 2 establishes a finite sample size error bound for Stein Thinning applied to MCMC output:

**Theorem 2** *Let  $\mathcal{X}$  be a measurable space and let  $P$  be a probability distribution on  $\mathcal{X}$ . Let  $k_P : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a reproducing kernel with  $\int_{\mathcal{X}} k_P(x, \cdot) dP = 0$  for all  $x \in \mathcal{X}$ . Consider a  $P$ -invariant, time-homogeneous Markov chain  $(X_i)_{i \in \mathbb{N}} \subset \mathcal{X}$  generated using a  $V$ -uniformly ergodic transition kernel, such that Equation (9) is satisfied with  $V(x) \geq \sqrt{k_P(x, x)}$  for all  $x \in \mathcal{X}$ . Suppose that, for some  $\gamma > 0$ ,*

$$b := \sup_{i \in \mathbb{N}} \mathbb{E} \left[ e^{\gamma k_P(X_i, X_i)} \right] < \infty, \quad M := \sup_{i \in \mathbb{N}} \mathbb{E} \left[ \sqrt{k_P(X_i, X_i)} V(X_i) \right] < \infty.$$

*Let  $\pi$  be an index sequence of length  $m$  produced by Algorithm 1 applied to the Markov chain output  $(X_i)_{i=1}^n$ . Then, with  $C = \frac{2k_P}{1-\rho}$ , we have that*

$$\mathbb{E} \left[ D_{k_P} \left( \frac{1}{m} \sum_{j=1}^m \delta(X_{\pi(j)}) \right)^2 \right] \leq \frac{\log(b)}{\gamma n} + \frac{CM}{n} + \left( \frac{1 + \log(m)}{m} \right) \frac{\log(nb)}{\gamma}. \tag{10}$$

The proof of Theorem 2 is provided in Appendix S2.2.

**Remark 8** The upper bound in Equation (10) is asymptotically minimised when (up to log factors)  $m$  is proportional to  $n$ . In practice we are interested in the case  $m \ll n$ , so we may for example set  $m = \lfloor \frac{n}{1000} \rfloor$  if we aim for substantial compression. It is not claimed that the bound in Equation (10) is tight and indeed empirical results in Section 4 endorse the use of Stein Thinning in the small  $m$  context.

**Remark 9** For  $P \in \mathcal{P}$  and  $k_P$  in Equation (6), based on a radial kernel  $k$ , meaning that  $k(x, y) = \phi(x - y)$  for some function  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$  satisfying  $\nabla \phi(0) = 0$ , we have that  $k_P(x, x) = -\Delta \phi(0) + \phi(0) \|\nabla \log p(x)\|^2$ . The function  $x \mapsto \sqrt{k_P(x, x)}$  appearing in the preconditions of Theorem 2 can therefore be understood in terms of  $\|\nabla \log p(x)\|$ . Further discussion of the preconditions of Theorem 2 is provided in Appendix S2.4.

Since convergence in mean-square does not in general imply almost sure convergence, we next strengthen the conclusions of Theorem 2. Our final result, Theorem 3, therefore establishes an almost sure convergence guarantee for Stein Thinning. Furthermore, the result that follows applies also in the ‘biased sampler’ case, where  $(X_i)_{i \in \mathbb{N}}$  is a  $Q$ -invariant Markov chain and  $Q$  need not equal  $P$ :

**Theorem 3** Let  $Q$  be a probability distribution on  $\mathcal{X}$  with  $P$  absolutely continuous with respect to  $Q$ . Consider a  $Q$ -invariant, time-homogeneous Markov chain  $(X_i)_{i \in \mathbb{N}} \subset \mathcal{X}$  generated using a  $V$ -uniformly ergodic transition kernel, such that  $V(x) \geq \frac{dP}{dQ}(x) \sqrt{k_P(x, x)}$ . Suppose that, for some  $\gamma > 0$ ,

$$b := \sup_{i \in \mathbb{N}} \mathbb{E} \left[ e^{\gamma \max\left(1, \frac{dP}{dQ}(X_i)^2\right) k_P(X_i, X_i)} \right] < \infty, \quad M := \sup_{i \in \mathbb{N}} \mathbb{E} \left[ \frac{dP}{dQ}(X_i) \sqrt{k_P(X_i, X_i)} V(X_i) \right] < \infty.$$

Let  $\pi$  be an index sequence of length  $m$  produced by Algorithm 1 applied to the Markov chain output  $(X_i)_{i=1}^n$ . If  $m \leq n$  and the growth of  $n$  is limited to at most  $\log(n) = O(m^{\beta/2})$  for some  $\beta < 1$ , then  $D_{k_P}(\frac{1}{m} \sum_{j=1}^m \delta(X_{\pi(j)})) \rightarrow 0$  almost surely as  $m, n \rightarrow \infty$ . Furthermore, if the preconditions of Proposition 3 are satisfied, then  $\frac{1}{m} \sum_{j=1}^m \delta(X_{\pi(j)}) \Rightarrow P$  almost surely as  $m, n \rightarrow \infty$ .

The proof of Theorem 3 is provided in Appendix S2.3. The interpretation of Theorem 3 is that one may sample states from a Markov chain that is not  $P$ -invariant and yet, under the stated assumptions (which ensure that regions of high probability under  $P$  are explored), one can use Stein Thinning to still obtain a consistent approximation of  $P$ . This can be contrasted, for example, with the Support Points method of Mak and Joseph (2018), which relies on  $P$  being well-approximated by the MCMC output. This completes our theoretical analysis of Stein Thinning.

## 4 | EMPIRICAL ASSESSMENT

In this section, we compare the performance of Stein Thinning with existing methods for post-processing MCMC output. Our motivation derives from a problem in which we must infer a 38-dimensional parameter in a calcium signalling model defined by a stiff system of 6 coupled ordinary differential equations (ODEs). Posterior uncertainty is required to be propagated through a high-fidelity simulation in a multi-scale and multi-physics model  $f$  of the human heart. Here, compression of the MCMC output can be used to construct an approximately optimal experimental design on which  $f$  can be evaluated. The calcium model is, however, unsuitable for conducting a thorough *in silico* assessment due to its associated computational cost. Therefore in Section 4.1 we first consider a simpler ODE model, where  $P$  can be accurately approximated. Then, as an intermediate example, in Section 4.2 we consider an ODE model that induces stronger correlations among the parameters in  $P$ , before addressing the calcium model in Section 4.3.

In Appendix S3 we describe the generic structure of a parameter inference problem for ODEs. In all instances the aim is to post-process the output from MCMC, in order to produce an accurate empirical approximation of the posterior supported on a small number  $m \ll n$  of the states that were visited. The following methods were compared:

1. The standard approach, which estimates a burn-in period using either the *GR diagnostic*  $\hat{b}^{\text{GR},L}$ ,  $L > 1$ , of Gelman and Rubin (1992), Brooks and Gelman (1998), Gelman et al. (2014) or the more sophisticated *VK diagnostic*  $\hat{b}^{\text{VK},L}$ ,  $L \geq 1$ , of Vats and Knudson (2018), in each case based on  $L$  independent chains as described in Appendix S4, followed by thinning as per Equation (3).

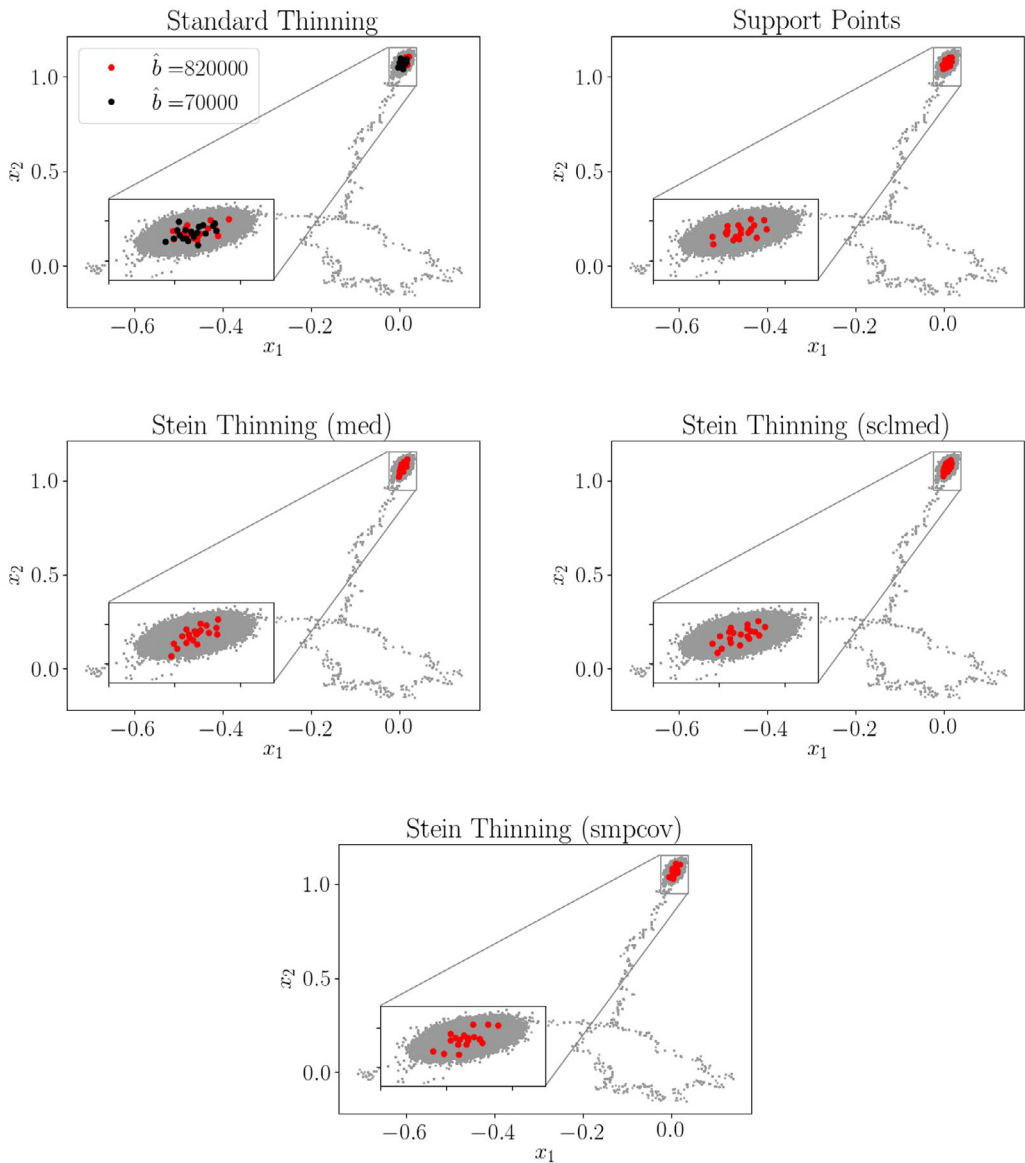
2. The `Support Points` algorithm proposed in Mak and Joseph (2018), implemented in the `R` package `support`.
3. The `Stein Thinning` algorithm that we have proposed, with each of the kernel choices described in Section 2.3.

To ensure that our empirical findings are not sensitive to the choice of MCMC method, we implemented four Metropolis–Hastings samplers that differ qualitatively according to the sophistication of their proposal. These were: (i) the Gaussian random walk (RW); (ii) the adaptive Gaussian random walk (ADA-RW), which uses an estimate of the covariance of the target (Haario et al., 1999); (iii) the Metropolis-adjusted Langevin algorithm (MALA), which takes a step in the direction of increasing Euclidean gradient, perturbed by Gaussian noise (Roberts & Tweedie, 1996); (iv) the preconditioned version of MALA (P-MALA), which employs a preconditioner based on the Fisher information matrix (Girolami & Calderhead, 2011). Full details are in Appendix S3. Metropolis–Hastings algorithms were selected on the basis that we were able to successfully implement them on the challenging calcium signalling model in Section 4.3, which required manually interfacing with the numerical integrator to produce reliable output.

#### 4.1 | Goodwin oscillator

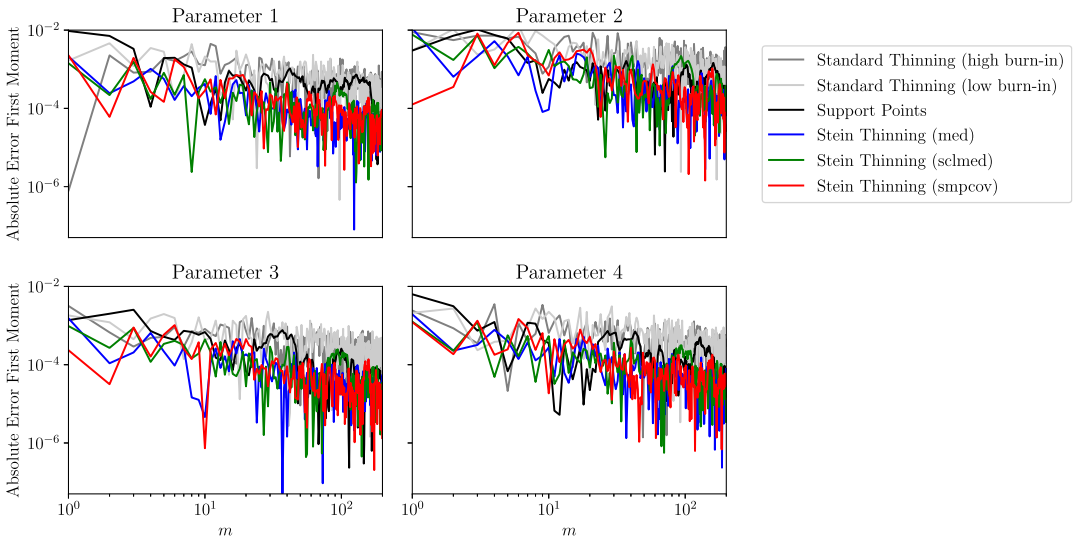
The first example that we consider is a negative feedback oscillator due to Goodwin (1965). The ODE model and the associated  $d = 4$  dimensional inference problem are described in Appendix S5.2, where one trace plot for each MCMC method, of length  $n = 2 \times 10^6$ , are presented in Figure S3.

First, we consider the standard approach to post-processing MCMC output, as per Equation (3). From the trace plots in Figure S3, it is clear that a burn-in period  $b > 0$  is required. For each method, we therefore computed the GR and VK diagnostics, to arrive at candidate values  $b$  for the burn-in period. Default settings were used for all diagnostics, which were computed both for the multivariate  $d$ -dimensional state vector and for the univariate marginals, as reported in Appendix S5.2. The GR diagnostics were computed using  $L = 6$  independent chains and the VK diagnostics were computed using both  $L = 1$  and  $L = 6$  independent chains; note that when  $L > 1$ , these diagnostics have access to more information in comparison with `Stein Thinning`, in terms of the number of samples that are available to the method. The estimated values for the burn-in period are reported in Appendix S5.2, Table S4. For all MCMC methods, neither the univariate nor the multivariate GR diagnostics were satisfied, so that  $\hat{b}^{\text{GR},6} > n$  and estimation using Equation (3) cannot proceed. The VK diagnostic produced values  $\hat{b}^{\text{VK},L} < n$ , which typically led to about half of the MCMC output being discarded. Although well-suited for their intended task of minimising bias in MCMC output, the smaller number of states left after burn-in removal may lead to inefficient approximation of  $P$  and derived quantities of interest, strikingly so in the case of the GR diagnostic. The use of an optimality criterion enables `Stein Thinning` to directly address this bias-variance trade-off. Of course, one can in principle run more iterations of MCMC to provide more diversity in the remainder of the sample path after burn-in is removed, but in applications such as the calcium model of Section 4.3 the computational cost associated with each iteration presents a practical limitation in running more iterations of an MCMC method. Effective methods to post-process limited output (or, equivalently, a long output from a poorly mixing Markov chain) are therefore important.



**FIGURE 3** Projections on the first two coordinates of the RW MCMC output from the Goodwin oscillator (gray dots), together with the first  $m = 20$  points selected using: the standard approach of discarding burn-in and thinning the remainder (the estimated burn in period is indicated in the legend); the `Support Points` method; Stein Thinning, for each of the settings `med`, `sclmed`, `smpcov` [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.com)]

Having identified a burn-in period, the standard approach thins the remainder of the sample path according to Equation (3). In the experiments that follow we focus on the VK diagnostic and consider both the smallest and largest estimates obtained for the burn-in period. The resulting index sets  $\pi$  are displayed, for  $m = 20$  and RW (the simplest MCMC method) in Figure 3 (top left panel), and in Appendix S5.2, Figures S6 (ADA-RW), S7 (MALA), S8 (P-MALA). In the same figures (top right panel) we show the set of `Support Points` obtained using algorithm proposed by Mak and Joseph (2018). The remaining panels display the output from Stein Thinning.



**FIGURE 4** Absolute error of estimates for the posterior mean of each parameter in the Goodwin oscillator, based on output from RW MCMC [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

Compared to the standard approach, Support Points and Stein Thinning produce sets that are more structured.

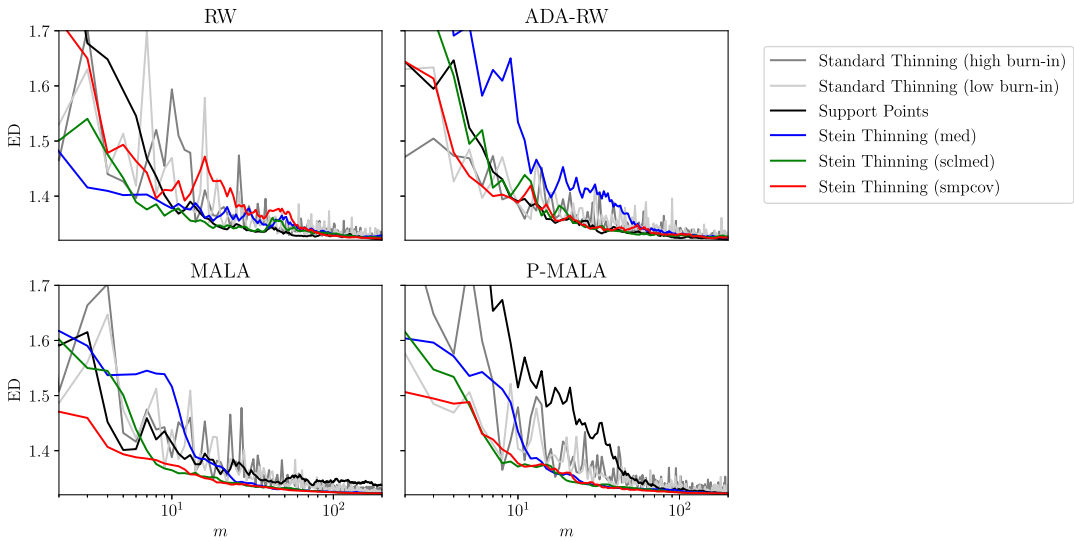
To assess the performance of these competing methods, we first considered the toy problem of approximating the posterior mean of each parameter in the Goodwin oscillator as an average of  $m$  points selected from the MCMC output. Figure 4 displays absolute errors for each method, based on RW; our ground truth was provided by an extended run of MCMC. Results for the other MCMC methods are provided in Appendix S5.2, Figures S9 (ADA-RW), S10 (MALA), S11 (P-MALA). Broadly speaking, Stein Thinning tends to provide more accurate estimators compared to the alternatives considered. From Figure 4 it is difficult to see any difference in performance between med, sclmed and smpcov. To gain more insight, in Appendix S5.2 we plot marginal density estimates in Figures S21 (RW), S13 (ADA-RW), S14 (MALA), S15 (P-MALA). It is apparent that Stein Thinning improves on the standard approach, whilst med and sclmed performed slightly better than smpcov. This may be because in smpcov there are more degrees of freedom in  $\Gamma$  that must be estimated. Support Points performed on a par with Stein Thinning based on smpcov.

To facilitate a more principled assessment, we computed two quantitative measures for how well the resulting empirical distributions approximate the posterior. These were (a) the energy distance (ED; Baringhaus & Franz, 2004; Székely & Rizzo, 2004), given up to an additive constant by

$$ED := \frac{2}{m} \sum_{j=1}^m \int \|x - x_{\pi(j)}\|_{\Sigma} dP(x) - \frac{1}{m^2} \sum_{j,j'=1}^m \|x_{\pi(j)} - x_{\pi(j')}\|_{\Sigma}, \tag{11}$$

where in this paper we used the norm  $\|x\|_{\Sigma} := \|\Sigma^{-1/2}x\|$  induced by the covariance matrix of  $P$ , with both  $\Sigma$  and Equation (11) being estimated from MCMC output, and (b) the KSD based on med, the simplest setting for  $\Gamma$ . ED serves as an objective performance measure, being closely related to the quantity that Support Points attempts to minimise (Mak & Joseph, 2018 used





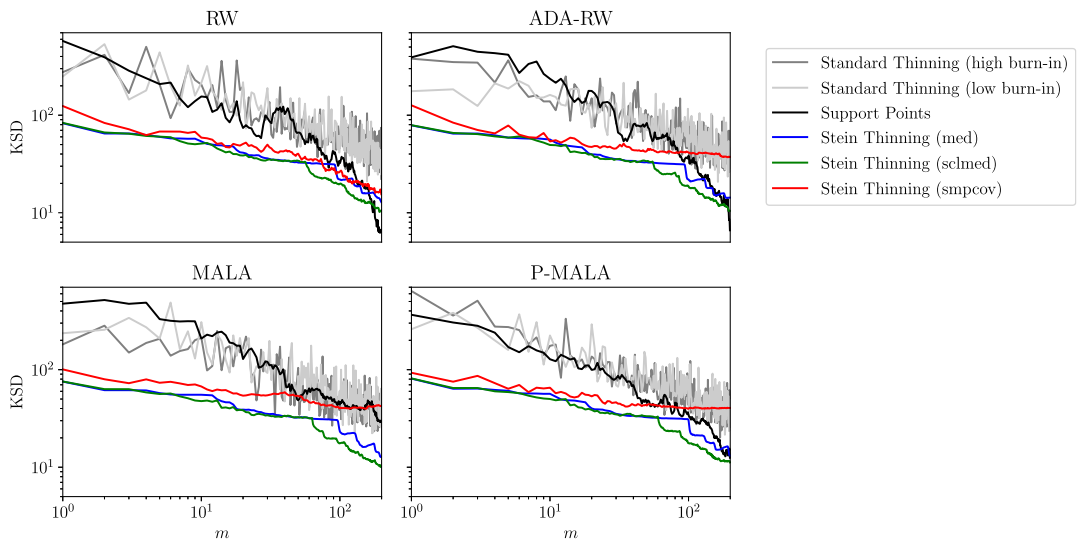
**FIGURE 5** Goodwin oscillator: Energy distance (ED) to the posterior, as per Equation (11), for empirical distributions obtained through traditional burn-in and thinning (grey lines), Support Points (black line) and Stein Thinning (colored lines), based on output from four different MCMC methods [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

the  $\|\cdot\|$  norm in place of  $\|\cdot\|_{\Sigma}$ ), while KSD is the performance measure that is being directly optimised in Stein Thinning. Our decision to include KSD in the assessment is motivated by three factors; (i) ED is somewhat insensitive to detail, making it difficult to rank competing methods; (ii) the empirical approximation of ED in Equation (11) relies on access to high-quality MCMC output, but this will not be available in Section 4.3; (iii) Stein discrepancies are the only computable performance measures in the Bayesian context, to the best of our knowledge, that have been proven to provide convergence control.

The results for ED are shown in Figure 5. Here Stein Thinning based on `sclmed` performed at least as well as the other methods considered and, surprisingly, out-performed Support Points when applied to MALA and P-MALA output. This may be because MALA and P-MALA provided worse approximations to  $P$  compared with RW and ADA-RW (recall that Support Points relies on the MCMC output providing an accurate approximation of  $P$ ). Note that neither ED nor KSD values will tend to 0 as  $m \rightarrow \infty$  in this experiment, since the number  $n$  of MCMC iterations was fixed. The corresponding results for KSD are presented in Figure 6 and show a clearer performance ordering of the competing methods, with Stein Thinning based on `med` and `sclmed` out-performing all other methods for all but the largest values of  $m$  considered. The `smpcov` setting performed well for small  $m$  but for large  $m$  its performance degraded. The performance ordering under KSD was identical across the different MCMC output.

## 4.2 | Lotka–Volterra

The second example that we consider is the predator-prey model of Lotka (1926) and Volterra (1926). A description of the  $d = 4$  dimensional inference task, the output from MCMC methods and the implementation of thinning procedures is reserved for Appendix S5.3. Compared to the Goodwin oscillator, the Lotka–Volterra posterior  $P$  exhibits stronger correlation among parameters. This has consequences for our assessment, since now all MCMC methods, and in particular



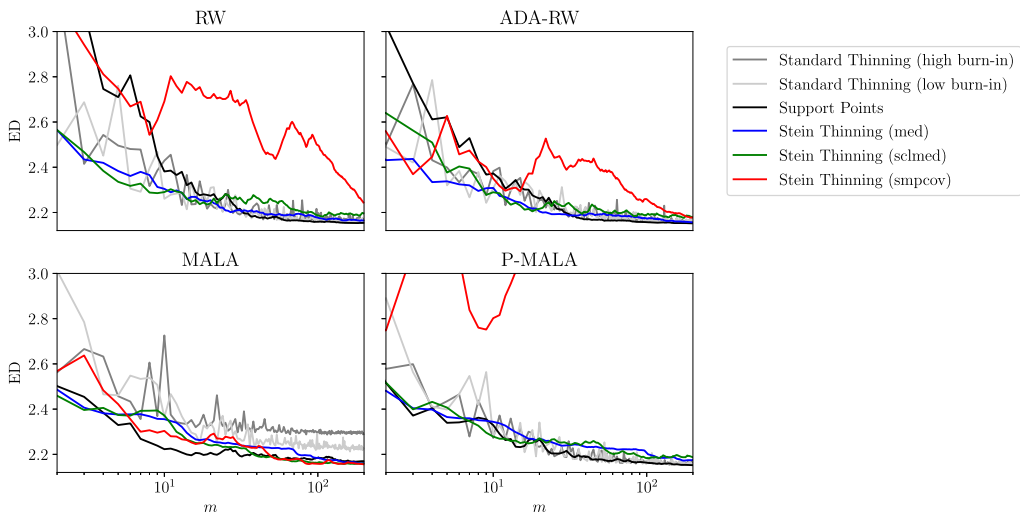
**FIGURE 6** Goodwin oscillator: Kernel Stein discrepancy (KSD) based on med, for empirical distributions obtained through traditional burn-in and thinning (grey lines), Support Points (black line) and Stein Thinning (colored lines), based on output from four different MCMC methods [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

MALA, mixed less well compared to corresponding results for the Goodwin oscillator, as can be seen from the trace plots in Appendix S5.3, Figure S17. Results are reported for ED in Figure 7. It can be seen that Stein Thinning based on med and sclmed performed comparably with Support Points, being better for small  $m$  in the case of RW and ADA-RW and marginally worse for large  $m$  in RW, ADA-RW and P-MALA. Interestingly, the setting smpcov was associated with poor performance on output from RW, ADA-RW and especially P-MALA. This may be because, when  $\Gamma$  is poorly conditioned, any error in an estimate for  $\Gamma$  will be amplified when computing  $\Gamma^{-1}$ . However, in the case of MALA, which mixed poorly, the standard approach of burn-in removal and thinning performed poorly and all settings of Stein Thinning provided an improvement.

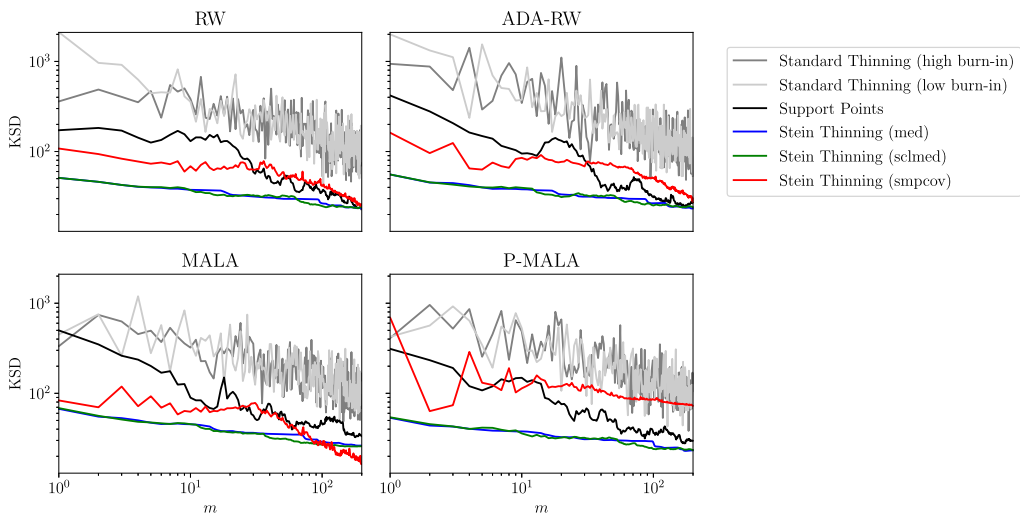
Results for KSD are reported in Figure 8. The performance ordering of competing methods under KSD is similar to that reported in Section 4.1, except for the smpcov setting which appears to improve the performance of Stein Thinning for larger values of  $m$  in the context of MALA. This may be because smpcov serves to ‘whiten’ the correlation structure in  $P$ , such that the resulting geometry is more favourable for the construction of an empirical approximation. However, this improved performance was not seen on P-MALA. In all cases Stein Thinning out-performed Support Points.

### 4.3 | Calcium signalling model

Our final example is a model for calcium signalling in cardiac cells, illustrated in Appendix S5.4, Figure S24. The model describes an electrically activated intracellular calcium signal that in turn activates the sub-cellular sarcomere, causing the muscle cell to contract and the heart to beat. The intracellular calcium signal is crucial for healthy cardiac function. However, under pathological conditions, dysregulation of this intra-cellular signal can play a central role in the initiation and sustenance of life-threatening arrhythmias. Computational models are increasingly being



**FIGURE 7** Lotka–Volterra model: Energy distance (ED) to the posterior, as per Equation (11), for empirical distributions obtained through traditional burn-in and thinning (grey lines), Support Points (black line) and Stein Thinning (colored lines), based on output from four different MCMC methods [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



**FIGURE 8** Lotka–Volterra model: Kernel Stein discrepancy (KSD) based on med, for empirical distributions obtained through traditional burn-in and thinning (grey lines), Support Points (black line) and Stein Thinning (colored lines), based on output from four different MCMC methods [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

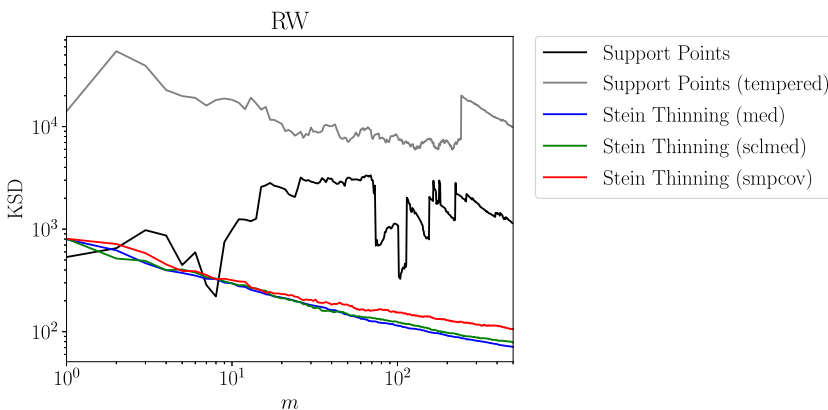
applied to study this highly orchestrated multi-scale signalling cascade to determine how changes in cell-scale calcium regulation, encoded in calcium model parameters, impact whole-organ cardiac function (Campos et al., 2015; Colman, 2019; Niederer et al., 2019). The computational cost of simulating from tissue-scale and organ-scale models is high, with single simulations taking thousands of CPU hours (Augustin et al., 2016; Niederer et al., 2011; Strocchi et al., 2020). This limits the capacity to propagate uncertainty in calcium signalling model parameters up to organ-scale simulations, so that at present it remains unclear how uncertainty in calcium signalling

parameters impacts the predictions made by a whole-organ model. Our motivation for developing Stein Thinning was to obtain a compressed representation of the posterior distribution for the  $d = 38$  dimensional parameter of a calcium signalling model, based on a cell-scale experimental dataset, which can subsequently be used as an experimental design to propagate uncertainty through a whole-organ model.

This motivating problem entails a second complication in that, compared to the example in Section 4.1 and even the example in Section 4.2, the development of an efficient MCMC method appears to be difficult. Thus, in the experiment that follows, we cannot rely on any of the MCMC methods that we described at the start of Section 4 to provide anything more than a crude approximation of the posterior, at best. This is evidenced by the non-overlapping approximations to the posterior marginals produced when different random seeds are used; see Figures S26 to S29. (Of course, it is possible that a more sophisticated sampling method may be designed for this task, but our aim here is not to develop a new sampling method.) Tempering of the likelihood provides a straightforward route to improve the mixing of MCMC, but the invariant distribution  $Q$  will then no longer equal  $P$ . Here we explore the potential for Stein Thinning to perform bias-correction for such  $Q$ -invariant MCMC output, in the spirit of Theorem 3.

Our focus in the remainder is on output from the RW MCMC method. This MCMC method was selected since (a) gradient-free methods can be easier to tune when the posterior is concentrated (Livingstone & Zanella, 2020), and (b) once the sample path has been computed, the associated gradients can be computed in parallel. Both standard and tempered MCMC were performed; in the latter case the likelihood was tempered so that the (biased) target  $Q$  was just about tractable for MCMC (see Appendix S5.4). In each case a total of  $n = 4 \times 10^6$  iterations of MCMC were performed, representing two weeks' CPU time.

Figure 9 reports the KSD based on `med`, for index sets of cardinality up to  $m = 500$ ; see Appendix S5.4 for results for KSD based on `sclmed` (Figure S30) and `smpcov` (Figure S31). Considering first the tempered MCMC output, the lower values of KSD achieved by Stein Thinning are consistent with fact that Stein Thinning corrects for bias due to tempering, while Support Points does not. Furthermore, Stein Thinning of tempered MCMC results in lower values of KSD compared to Support Points applied to standard MCMC output, with the latter being negatively affected by the non-convergence of the MCMC.



**FIGURE 9** Calcium signalling model. Kernel Stein discrepancy (KSD) based on `med`, for empirical distributions obtained using Support Points and Stein Thinning, based on output from RW MCMC applied to either  $P$  or a tempered version of  $P$  [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

Inspection of the univariate marginals demonstrates that the combination of tempering and Stein Thinning produces approximations that are robust to changes in the random seed, while the approximations produced by standard MCMC with an equivalent computational budget are not; see Figures S26 to S29.

## 5 | CONCLUSION

In this paper, standard approaches used to post-process and compress output from MCMC were identified as being sub-optimal when one considers the approximation quality of the empirical distribution that is produced. A novel method, Stein Thinning, was proposed that seeks a subset of the MCMC output, of fixed cardinality, such that the associated empirical approximation is close to optimal. The theoretical analysis that we have provided for Stein Thinning handles the effect of the post-processing procedure jointly with the randomness involved in simulating from the Markov chain, such that consistency of the overall estimator is established.

Although we focused on MCMC, the proposed method can be applied to any computational method that provides a collection of states as output. These include approximate (biased) MCMC methods, where Stein Thinning may be able to provide bias correction in the spirit of Theorem 3. However, the main limitation of Stein Thinning is that it requires gradients of the log-target to be computed, which is not always practical.

Our research was motivated by challenging parameter inference problems that arise in ODEs, in particular in cardiac modelling where one is interested in propagating calcium signalling parameter uncertainty through a whole-organ simulation—a task that would naïvely be impractical or impossible using the full MCMC output. Our ongoing research is exploiting Stein Thinning in this context and is enabling us to perform scientific investigations that were not feasible beforehand. Furthermore, in a sequel we demonstrate that approximate implementations of Stein Thinning can massively reduced its implementation cost (Teymur et al., 2021).

## ACKNOWLEDGEMENTS

The authors are grateful for support from the Lloyd's Register Foundation programme on data-centric engineering and the programme on health and medical sciences at the Alan Turing Institute. MR, SN and CJO were supported by the British Heart Foundation (BHF; SP/18/6/33805). JC was supported by the UKRI Strategic Priorities Fund (EP/T001569/1). PS was supported by the BHF (RG/15/9/31534). SN was supported by the EPSRC (EP/P01268X/1, NS/A000049/1, EP/M012492/1), the BHF (PG/15/91/31812, FS/18/27/33543), the NIHR (II-LB-1116-20001) and the Wellcome Trust (WT 203148/Z/16/Z). The authors thank Matthew Graham, Liam Hodgkinson, Rob Salomone, and the anonymous Editor, Associate Editor, and Reviewers, for helpful comments on the manuscript.

## ORCID

Chris. J. Oates  <https://orcid.org/0000-0002-4444-8603>

## REFERENCES

Augustin, C.M., Neic, A., Liebmann, M., Prassl, A.J., Niederer, S.A., Haase, G. et al. (2016) Anatomically accurate high resolution modeling of human whole heart electromechanics: a strongly scalable algebraic multigrid solver method for nonlinear deformation. *Journal of Computational Physics*, 305, 622–646.

- Baringhaus, L. & Franz, C. (2004) On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1), 190–206.
- Barp, A., Oates, C., Porcu, E. & Girolami, M. (2021) A Riemann–Stein kernel method. *Bernoulli*, to appear.
- Berlinet, A. & Thomas-Agnan, C. (2004) *Reproducing Kernel Hilbert spaces in probability and statistics*. New York: Springer Science & Business Media.
- Brooks, S.P. & Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434–455.
- Campos, F.O., Shiferaw, Y., Prassl, A.J., Boyle, P.M., Vigmond, E.J. & Plank, G. (2015) Stochastic spontaneous calcium release events trigger premature ventricular complexes by overcoming electrotonic load. *Cardiovascular Research*, 107(1), 175–183.
- Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M. et al. (2017) Stan: a probabilistic programming language. *Journal of Statistical Software*, 76(1).
- Chen, W.Y., Mackey, L., Gorham, J., Briol, F.-X. & Oates, C.J. (2018) Stein points. In: *Proceedings of the 35th international conference on machine learning*.
- Chen, W.Y., Barp, A., Briol, F.-X., Gorham, J., Mackey, L., Girolami, M. et al. (2019) Stein points Markov chain Monte Carlo. In: *Proceedings of the 36th international conference on machine learning*.
- Chwialkowski, K., Strathmann, H. & Gretton, A. (2016) A kernel test of goodness of fit. In: *Proceedings of the 33rd international conference on machine learning*.
- Colman, M.A. (2019) Arrhythmia mechanisms and spontaneous calcium release: bi-directional coupling between re-entrant and focal excitation. *PLoS Computational Biology*, 15(8), e1007260.
- Cowles, M.K. & Carlin, B.P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434), 883–904.
- Duncan, A., Nüsken, N. & Szpruch, L. (2019) On the geometry of Stein variational gradient descent. *arXiv:1912.00894*.
- Dwivedi, R., Feldheim, O.N., Gurel-Gurevich, O. & Ramdas, A. (2019) The power of online thinning in reducing discrepancy. *Probability Theory and Related Fields*, 174(1–2), 103–131.
- Garreau, D., Jitkrittum, W. & Kanagawa, M. (2018) Large sample analysis of the median heuristic. *arXiv:1707.07269*.
- Gelfand, A.E. & Smith, A.F. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398–409.
- Gelman, A. & Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2014) *Bayesian data analysis*, vol. 2. Boca Raton, FL: CRC Press.
- Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.
- Geyer, C.J. (1992) Practical Markov chain Monte Carlo. *Statistical Science*, 7(4), 473–483.
- Geyer, C.J. & Thompson, E.A. (1992) Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*, 54(3), 657–683.
- Girolami, M. & Calderhead, B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 73(2), 123–214.
- Goodwin, B.C. (1965) Oscillatory behavior in enzymatic control process. *Advances in Enzyme Regulation*, 3, 318–356.
- Gorham, J. & Mackey, L. (2015) Measuring sample quality with Stein’s method. In: *Proceedings of the 29th conference on neural information processing systems*.
- Gorham, J. & Mackey, L. (2017) Measuring sample quality with kernels. In: *Proceedings of the 34th international conference on machine learning*.
- Gorham, J., Raj, A. & Mackey, L. (2020) Stochastic Stein discrepancies. In: *Proceedings of the 34th annual conference on neural information processing systems*.
- Green, P.J., Latuszyński, K., Pereyra, M. & Robert, C.P. (2015) Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4), 835–862.
- Haario, H., Saksman, E. & Tamminen, J. (1999) Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3), 375–396.

- Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications.
- Hickernell, F. (1998) A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, 67(221), 299–322.
- Hodgkinson, L., Salomone, R. & Roosta, F. (2020) The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv:2001.09266*.
- Huggins, J. & Mackey, L. (2018) Random feature Stein discrepancies. In: *Proceedings of the 31st conference on neural information processing systems*.
- Joseph, V.R., Dasgupta, T., Tuo, R. & Wu, C. (2015) Sequential exploration of complex surfaces using minimum energy designs. *Technometrics*, 57(1), 64–74.
- Joseph, V.R., Wang, D., Gu, L., Lyu, S. & Tuo, R. (2019) Deterministic sampling of expensive posteriors using minimum energy designs. *Technometrics*, 61(3), 297–308.
- Le, H., Lewis, A., Bharath, K. & Fallaize, C. (2020) A diffusion approach to Stein's method on Riemannian manifolds. *arXiv:2003.11497*.
- Liu, Q. (2017) Stein variational gradient descent as gradient flow. In: *Proceedings of the 31st conference on neural information processing systems*.
- Liu, Q. & Lee, J.D. (2017) Black-box importance sampling. In: *Proceedings of the 20th international conference on artificial intelligence and statistics*.
- Liu, Q. & Wang, D. (2016) Stein variational gradient descent: a general purpose Bayesian inference algorithm. In: *Proceedings of the 30th conference on neural information processing systems*.
- Liu, C. & Zhu, J. (2018) Riemannian Stein variational gradient descent for Bayesian inference. In: *Proceedings of the 32nd AAAI conference on artificial intelligence*.
- Liu, Q., Lee, J.D. & Jordan, M.I. (2016) A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In: *Proceedings of the 33rd international conference on machine learning*.
- Livingstone, S. & Zanella, G. (2020) The Barker proposal: combining robustness and efficiency in gradient-based MCMC. *arXiv:1908.11812*.
- Lotka, A.J. (1926) Elements of physical biology. *Science Progress in the Twentieth Century (1919–1933)*, 21(82), 341–343.
- Lunn, D.J., Thomas, A., Best, N. & Spiegelhalter, D. (2000) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4), 325–337.
- Mak, S. & Joseph, V.R. (2018) Support points. *The Annals of Statistics*, 46(6A), 2562–2592.
- Matsubara, T., Knoblauch, J., Briol, F.-X. & Oates, C.J. (2022) Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. <https://doi.org/10.48550/arXiv.2104.07359>
- Meyn, S. & Tweedie, R. (2012) *Markov Chains and stochastic stability*. New York: Springer Science & Business Media.
- Muller, A. (1997) Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2), 429–443.
- Niederer, S.A., Mitchell, L., Smith, N. & Plank, G. (2011) Simulating human cardiac electrophysiology on clinical time-scales. *Frontiers in Physiology*, 2, 14.
- Niederer, S.A., Lumens, J. & Trayanova, N.A. (2019) Computational models in cardiology. *Nature Reviews Cardiology*, 16(2), 100–111.
- Oates, C.J., Girolami, M. & Chopin, N. (2017) Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society, Series B*, 79(3), 695–718.
- Paige, B., Sejdinovic, D. & Wood, F.D. (2016) Super-sampling with a reservoir. In: *Proceedings of the 32nd conference on uncertainty in artificial intelligence*.
- Plummer, M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd international workshop on distributed statistical computing*.
- R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing. Available from: <https://www.R-project.org>
- Robert, C. & Casella, G. (2013) *Monte Carlo statistical methods*. New York: Springer Science & Business Media.
- Roberts, G.O. & Tweedie, R.L. (1996) Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4), 341–363.



- Song, L. (2008) Learning via Hilbert space embedding of distributions. PhD thesis, School of Information Technologies, University of Sydney.
- South, L.F., Karvonen, T., Nemeth, C., Girolami, M. & Oates, C. (2021) Semi-exact control functionals from Sard's method. *Biometrika*, to appear.
- Stein, C. (1972) 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*. University of California Press, pp. 583–602.
- Strocchi, M., Gsell, M.A., Augustin, C.M., Razeghi, O., Roney, C.H., Prassl, A.J. et al. (2020) Simulating ventricular systolic motion in a four-chamber heart model with spatially varying robin boundary conditions to model the effect of the pericardium. *Journal of Biomechanics*, 101, 109645.
- Székely, G.J. & Rizzo, M.L. (2004) Testing for equal distributions in high dimension. *InterStat*, 5(16.10), 1249–1272.
- Tanner, M.A. & Wong, W.H. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.
- Teymur, O., Gorham, J., Riabiz, M. & Oates, C.J. (2021) Optimal quantisation of probability measures using maximum mean discrepancy. In: *Proceedings of the 24th international conference on artificial intelligence and statistics*.
- Vats, D. & Knudson, C. (2018) Revisiting the Gelman-Rubin diagnostic. *arXiv:1812.09384*.
- Volterra, V. (1926) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Memoria della Reale Accademia Nazionale dei Lincei*, 6, 31–113.
- Xu, W. & Matsuda, T. (2020) A Stein goodness-of-fit test for directional distributions. In: *Proceedings of the 23rd international conference on artificial intelligence and statistics*.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Riabiz, M., Chen, W.Y., Cockayne, J., Swietach, P., Niederer, S.A., Mackey, L. et al. (2022) Optimal thinning of MCMC output. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(4), 1059–1081. Available from: <https://doi.org/10.1111/rssb.12503>