ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)

Unbiased Markov chain Monte Carlo methods with couplings

Pierre E. Jacob,

Harvard University, Cambridge, USA

John O'Leary

Harvard University, Cambridge, and Acadian Asset Management, Boston, USA

and Yves F. Atchadé

Boston University, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, December 11th, 2019*, Professor A. Doucet *in the Chair*]

Summary. Markov chain Monte Carlo (MCMC) methods provide consistent approximations of integrals as the number of iterations goes to ∞. MCMC estimators are generally biased after any fixed number of iterations. We propose to remove this bias by using couplings of Markov chains together with a telescopic sum argument of Glynn and Rhee. The resulting unbiased estimators can be computed independently in parallel. We discuss practical couplings for popular MCMC algorithms. We establish the theoretical validity of the estimators proposed and study their efficiency relative to the underlying MCMC algorithms. Finally, we illustrate the performance and limitations of the method on toy examples, on an Ising model around its critical temperature, on a high dimensional variable-selection problem, and on an approximation of the cut distribution arising in Bayesian inference for models made of multiple modules.

Keywords: Coupling; Markov chain Monte Carlo methods; Parallel computing; Unbiased estimation

1. Introduction

Markov chain Monte Carlo (MCMC) methods constitute a popular class of algorithms to approximate high dimensional integrals arising in statistics and other fields (Liu, 2008; Robert and Casella, 2004; Brooks *et al.*, 2011; Green *et al.*, 2015). These iterative methods provide estimators that are consistent as the number of iterations grows large but potentially biased for any fixed number of iterations, which discourages the parallel execution of many short chains (Rosenthal, 2000). Consequently, efforts have focused on exploiting parallel processors within each iteration (Tjelmeland, 2004; Brockwell, 2006; Lee *et al.*, 2010; Jacob *et al.*, 2011; Calderhead, 2014; Goudie *et al.*, 2017; Yang *et al.*, 2017) and on the design of parallel chains targeting different distributions (Altekar *et al.*, 2004; Wang *et al.*, 2015; Srivastava *et al.*, 2015). Still, MCMC estimators are ultimately justified by asymptotics in the number of iterations, which is discordant with current trends in computing hardware, characterized by increasing parallelism but stagnating clock speeds.

Address for correspondence: Pierre E. Jacob, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA.
E-mail: pjacob@fas.harvard.edu

1369–7412/20/82000

In this paper we propose a general construction to produce unbiased estimators of integrals with respect to a target probability distribution from MCMC kernels. The lack of bias means that these estimators can be implemented on parallel processors in the framework of Glynn and Heidelberger (1991), without communication between processors. Confidence intervals can be constructed with asymptotic guarantees in the number of processors, in contrast with standard MCMC confidence intervals that are justified asymptotically in the number of iterations (e.g. Flegal *et al.* (2008), Gong and Flegal (2016), Atchadé (2016) and Vats *et al.* (2018)). The lack of bias has additional benefits, as discussed in Section 5.5 in which we make use of its interplay with the law of iterated expectations to perform modular inference; see also the discussion in Section 6.

Our contribution follows the path breaking work of Glynn and Rhee (2014), which uses couplings to construct unbiased estimators of integrals with respect to an invariant distribution. They illustrated their construction on Markov chains represented by iterated random functions, leveraging the contraction properties of such functions. Glynn and Rhee (2014) also considered Harris recurrent chains for which an explicit minorization condition holds. Previously, McLeish (2011) employed similar debiasing techniques to obtain 'nearly unbiased' estimators from a single MCMC chain. More recently Jacob *et al.* (2019) removed the bias from conditional particle filters (Andrieu *et al.*, 2010) by coupling chains so that they meet in finite time. The present paper brings this type of 'Rhee–Glynn' construction to generic MCMC algorithms, with a novel analysis of estimator efficiency and a variety of examples. Our proposed construction involves couplings of MCMC algorithms, which we discuss for generic Metropolis–Hastings and Gibbs samplers.

Couplings have been used to study the convergence properties of MCMC algorithms from both theoretical and practical points of view (e.g. Reutter and Johnson (1995), Johnson (1996), Rosenthal (1997), Johnson (1998, 2013), Neal (1999), Roberts and Rosenthal (2004) and Johndrow and Mattingly (2017)). Couplings also underpin perfect samplers (Propp and Wilson, 1996; Murdoch and Green, 1998; Casella *et al.*, 2001; Flegal and Herbei, 2012; Lee *et al.*, 2014; Huber, 2016). A notable aspect of the approach of Glynn and Rhee (2014) that is preserved in our method is that only two chains must be coupled for the proposed estimator to be unbiased, without further assumptions on the state space or target distribution. Thus the approach applies more broadly than perfect samplers (see Glynn (2016)) while yielding unbiased estimators rather than exact samples. Coupling pairs of Markov chains also forms the basis of the approach of Neal (1999), with a similar motivation for parallel computation. The proposed estimation technique also shares aims with regeneration methods (e.g. Mykland *et al.* (1995) and Brockwell and Kadane (2005)), and we propose a numerical comparison in Section 5.2.

In Section 2 we introduce our estimators and present a coupling of random-walk Metropolis–Hastings (MH) chains as an illustration. In Section 3 we establish the efficiency properties of these estimators, discuss the verification of key assumptions and describe the use of the proposed estimators on parallel processors in light of results from for example Glynn and Heidelberger (1991). In Section 4 we describe how to couple some important MCMC algorithms and illustrate the effect of dimension on algorithms' performance with a multivariate normal distributions target. Section 5 contains more challenging examples including a multimodal target, a comparison with regeneration methods, sampling problems in large dimensional discrete spaces arising in Bayesian variable selection and Ising models, and an application to modular inference. We discuss our findings in Section 6. Scripts in R (R Core Team, 2015) are available from `https://github.com/pierrejacob/unbiasedmcmc` and supplementary materials are available on line.

2. Unbiased estimation from coupled chains

2.1. Rhee–Glynn estimator

Given a target probability distribution π on a Polish space $\mathcal{X}$ and a measurable real-valued test function h that is integrable with respect to π, we want to estimate the expectation $\mathbb{E}_\pi[h(X)] = \int h(x)\pi(\mathrm{d}x)$. Let P denote a Markov transition kernel on $\mathcal{X}$ that leaves π invariant, and let π_0 be some initial probability distribution on $\mathcal{X}$. Our estimators are based on a coupled pair of Markov chains $(X_t)_{t\geqslant 0}$ and $(Y_t)_{t\geqslant 0}$, which marginally start from π_0 and evolve according to P. In particular, we suppose that $\bar{P}$ is a transition kernel on the joint space $\mathcal{X}\times\mathcal{X}$ such that $\bar{P}\{(x,y),A\times\mathcal{X}\} = P(x,A)$ and $\bar{P}\{(x,y),\mathcal{X}\times A\} = P(y,A)$ for any $x,y\in\mathcal{X}$ and any measurable set A. We then construct the coupled Markov chain $(X_t,Y_t)_{t\geqslant 0}$ as follows. We draw (X_0,Y_0) such that $X_0\sim\pi_0$ and $Y_0\sim\pi_0$. Given (X_0,Y_0) we draw $X_1\sim P(X_0,\cdot)$, and then for any $t\geqslant 1$, given $X_0,(X_1,Y_0),\ldots,(X_t,Y_{t-1})$, we draw $(X_{t+1},Y_t)\sim\bar{P}\{(X_t,Y_{t-1}),\cdot\}$. We consider the following assumptions.

Assumption 1. As $t\to\infty$, $\mathbb{E}[h(X_t)]\to\mathbb{E}_\pi[h(X)]$. Furthermore, there is an $\eta>0$ and $D<\infty$ such that $\mathbb{E}[|h(X_t)|^{2+\eta}]\leqslant D$ for all $t\geqslant 0$.

Assumption 2. The chains are such that the meeting time $\tau := \inf\{t\geqslant 1 : X_t = Y_{t-1}\}$ satisfies $\mathbb{P}(\tau>t)\leqslant C\delta^t$ for all $t\geqslant 0$, for some constants $C<\infty$ and $\delta\in(0,1)$.

Assumption 3. The chains stay together after meeting, i.e. $X_t = Y_{t-1}$ for all $t\geqslant\tau$.

By construction, each of the marginal chains $(X_t)_{t\geqslant 0}$ and $(Y_t)_{t\geqslant 0}$ has initial distribution π_0 and transition kernel P. Assumption 1 requires these chains to result in a uniformly bounded $(2+\eta)$-moment of h; more discussion on moments of Markov chains can be found in Tweedie (1983). Since X_0 and Y_0 may be drawn from any coupling of π_0 with itself, it is possible to set $X_0 = Y_0$. However, X_1 is then generated from $P(X_0,\cdot)$, so that $X_1\neq Y_0$ in general. Thus one cannot force the meeting time to be small by setting $X_0 = Y_0$. Assumption 2 puts a condition on the coupling that is operated by $\bar{P}$ and would not in general be satisfied for an independent coupling. Coupled kernels must be carefully designed, using for example common random numbers and maximal couplings, for assumption 2 to be satisfied. We present a simple case in Section 2.2 and further examples in Section 4. We stress that the state space is not assumed to be discrete, and that the constants D and η of assumption 1 and C and δ of assumption 2 do not need to be known to implement the approach proposed. Assumption 3 typically holds by design; coupled chains that stay identical after meeting were termed 'faithful' in Rosenthal (1997).

Under these assumptions we introduce the following motivation for an unbiased estimator of $\mathbb{E}_\pi[h(X)]$, following Glynn and Rhee (2014). We begin by writing $\mathbb{E}_\pi[h(X)]$ as $\lim_{t\to\infty}\mathbb{E}[h(X_t)]$. Then, for any fixed $k\geqslant 0$,

$$
\begin{aligned}
\mathbb{E}_\pi[h(X)] &= \mathbb{E}[h(X_k)] + \sum_{t=k+1}^{\infty}\{\mathbb{E}[h(X_t)] - \mathbb{E}[h(X_{t-1})]\} && \text{(expanding the limit as a telescoping sum)}\\
&= \mathbb{E}[h(X_k)] + \sum_{t=k+1}^{\infty}\{\mathbb{E}[h(X_t)] - \mathbb{E}[h(Y_{t-1})]\} && \text{(since the chains have the same marginals)}\\
&= \mathbb{E}[h(X_k) + \sum_{t=k+1}^{\infty}\{h(X_t) - h(Y_{t-1})\}] && \text{(swapping the expectations and limit)}\\
&= \mathbb{E}[h(X_k) + \sum_{t=k+1}^{\tau-1}\{h(X_t) - h(Y_{t-1})\}] && \text{(by assumption 3).}
\end{aligned}
$$

We note that the sum in the last equation is 0 if $k+1>\tau-1$. The heuristic argument above suggests that the estimator $H_k(X,Y)=h(X_k)+\Sigma_{t=k+1}^{\tau-1}\{h(X_t)-h(Y_{t-1})\}$ should have expectation $\mathbb{E}_\pi[h(X)]$. We observe that this estimator requires $\tau-1$ calls to $\bar{P}$ and $\max(1,k+1-\tau)$ calls to P; thus under assumption 2 its cost has a finite expectation.

In Section 3 we establish the validity of the estimator under the three conditions above; this formally justifies the swap of expectation and limit. The estimator can be viewed as a debiased version of $h(X_k)$. Unbiasedness is guaranteed for any choice of $k\geqslant 0$, but both the cost and the variance of $H_k(X,Y)$ are sensitive to k; see Section 3.1. Thanks to this unbiasedness property, we can sample $R\in\mathbb{N}$ independent copies of $H_k(X,Y)$ in parallel and average the results to estimate $\mathbb{E}_\pi[h(X)]$ consistently as $R\to\infty$; we defer further considerations on the use of unbiased estimators on parallel processors to Section 3.3.

Before presenting examples and enhancements to the estimator above, we discuss the relationship between our approach and existing work. There is a rich literature applying forward couplings to study Markov chain convergence (Johnson, 1996, 1998, 2013; Thorisson, 2000; Lindvall, 2002; Rosenthal, 2002; Douc *et al.*, 2004; Nikooienejad *et al.*, 2016), and to obtain new algorithms such as perfect samplers (Huber, 2016) and the methods of Neal (1999) and Neal and Pinto (2001). Our approach is closely related to Glynn and Rhee (2014), who employed pairs of Markov chains to obtain unbiased estimators. The present work combines similar arguments with couplings of MCMC algorithms and proposes further improvements to remove bias at a reduced loss of efficiency.

Indeed Glynn and Rhee (2014) did not apply their methodology to the MCMC setting. They considered chains that are associated with contractive iterated random functions (see also Diaconis and Freedman (1999)), and Harris recurrent chains with an explicit minorization condition. A minorization condition refers to a small set $\mathcal{C}$, $\lambda>0$, an integer $m\geqslant 1$ and a probability measure ν such that, for all $x\in\mathcal{C}$ and some measurable set A, $P^m(x,A)\geqslant\lambda\nu(A)$. Such a condition is said to be explicit if the set, constant and probability measure are known by the user. Finding explicit small sets that are useful in practice can present a technical challenge, even for MCMC experts (see the discussion and references in Cowles and Rosenthal (1998)). When available, explicit minorization conditions can also be employed to identify regeneration times, yielding estimators that are amenable to parallel computation in the framework of Mykland *et al.* (1995) and Brockwell and Kadane (2005). By contrast Johnson (1996, 1998) and Neal (1999) addressed the question of coupling MCMC algorithms so that pairs of chains meet exactly, without analytical knowledge on the target distribution. The present paper focuses on the use of couplings of this type in the framework of Glynn and Rhee (2014).

2.2. Coupled Metropolis–Hastings example

Before further examination of our estimator and its properties, we present a coupling of MH chains that will typically satisfy assumptions 1–3 in realistic settings; this coupling was proposed in Johnson (1998) as part of a method to diagnose convergence. We postpone discussion of other couplings of MCMC algorithms to Section 4. We recall that each iteration t of the MH algorithm (Hastings, 1970) begins by drawing a proposal X^* from a Markov kernel $q(X_t,\cdot)$, where X_t is the current state. The next state is set to $X_{t+1}=X^*$ if $U\leqslant\pi(X^*)q(X^*,X_t)/\{\pi(X_t)q(X_t,X^*)\}$, where U denotes a uniform random variable on $[0,1]$, and $X_{t+1}=X_t$ otherwise.

We define a pair of chains so that each proceeds marginally according to the MH algorithm and jointly so that the chains will meet exactly after a random number of steps. We suppose that the pair of chains are in states X_t and Y_{t-1}, and we consider how to generate X_{t+1} and Y_t so that $\{X_{t+1}=Y_t\}$ might occur.

If $X_t\neq Y_{t-1}$, the event $\{X_{t+1}=Y_t\}$ cannot occur if both chains reject their respective propos-

als, X^* and Y^*. Meeting will occur if these proposals are identical and if both are accepted. Marginally, the proposals follow $X^*|X_t \sim q(X_t, \cdot)$ and $Y^*|Y_{t-1} \sim q(Y_{t-1}, \cdot)$. If $q(x, x^*)$ can be evaluated for all x and x^*, then we can sample from a maximal coupling between the two proposal distributions, which is a coupling of $q(X_t, \cdot)$ and $q(Y_{t-1}, \cdot)$ maximizing the probability of the event $\{X^* = Y^*\}$. How to sample from maximal couplings of continuous distributions is described in Thorisson (2000) and in Section 4.1. One can accept or reject the two proposals by using a common uniform random variable U. The chains will stay together after they meet: at each step after meeting, the proposals will be identical with probability 1, and jointly accepted or rejected with a common uniform variable. This coupling requires neither explicit minorization conditions nor contractive properties of a random-function representation of the chain.

2.3. Time-averaged estimator

To motivate our next estimator, we note that we can compute $H_k(X, Y)$ for several values of k from the same realization of the coupled chains, and that the average of these is unbiased as well. For any fixed integer m with $m \geqslant k$, we can run coupled chains for $\max(m, \tau)$ iterations, compute the estimator $H_l(X, Y)$ for each $l \in \{k, \ldots, m\}$ and take the average $H_{k:m}(X, Y) = (m - k + 1)^{-1} \Sigma_{l=k}^{m} H_l(X, Y)$, as we summarize in algorithm 1 in Table 1. We refer to $H_{k:m}(X, Y)$ as the *time-averaged estimator*; the estimator $H_k(X, Y)$ is retrieved when $m = k$. Alternatively we could average the estimators $H_l(X, Y)$ by using weights $w_l \in \mathbb{R}$ for $l \in \{k, \ldots, m\}$, to obtain $\Sigma_{l=k}^{m} w_l H_l(X, Y)$. This will be unbiased if $\Sigma_{l=k}^{m} w_l = 1$.

Rearranging terms in $(m - k + 1)^{-1} \Sigma_{l=k}^{m} H_l(X, Y)$, we can write the time-averaged estimator as

$$H_{k:m}(X, Y) = \frac{1}{m - k + 1} \sum_{l=k}^{m} h(X_l) + \sum_{l=k+1}^{\tau-1} \min\left(1, \frac{l - k}{m - k + 1}\right) \{h(X_l) - h(Y_{l-1})\}. \tag{2.1}$$

The term $(m - k + 1)^{-1} \Sigma_{l=k}^{m} h(X_l)$ corresponds to a standard MCMC average with m total iterations and a burn-in period of $k - 1$ iterations. We can interpret the other term as a bias correction. If $\tau \leqslant k + 1$, then the correction term equals 0. This provides some intuition for the choice of k and m: large k-values lead to the bias correction being equal to 0 with large probability, and large values of m result in $H_{k:m}(X, Y)$ being similar to an estimator obtained from a long MCMC run. Thus we expect the variance of $H_{k:m}(X, Y)$ to be similar to that of MCMC estimators for appropriate choices of k and m.

The estimator $H_{k:m}(X, Y)$ requires $\tau - 1$ calls to $\bar{P}$ and $\max(1, m + 1 - \tau)$ calls to P, which are overall comparable with m calls to P when m is large. Indeed, for the proposed couplings, calls to $\bar{P}$ are approximately twice as expensive as calls to P. Therefore, the cost of $H_{k:m}(X, Y)$ is

Table 1. Algorithm 1: unbiased 'time-averaged' estimator $H_{k:m}(X, Y)$ of $\mathbb{E}_\pi[h(X)]$

Step 1: draw X_0 and Y_0 from an initial distribution π_0 and draw $X_1 \sim P(X_0, \cdot)$
Step 2: set $t = 1$: while $t < \max(m, \tau)$, where $\tau = \inf\{t \geqslant 1 : X_t = Y_{t-1}\}$,

(a) draw $(X_{t+1}, Y_t) \sim \bar{P}\{(X_t, Y_{t-1}), \cdot\}$,
(b) set $t \leftarrow t + 1$

Step 3: for each $l \in \{k, \ldots, m\}$, compute $H_l(X, Y) = h(X_l) + \Sigma_{t=l+1}^{\tau-1} \{h(X_t) - h(Y_{t-1})\}$;
return $H_{k:m}(X, Y) = (m - k + 1)^{-1} \Sigma_{l=k}^{m} H_l(X, Y)$, or compute $H_{k:m}(X, Y)$ with equation (2.1)

comparable with $2(\tau-1)+\max(1, m+1-\tau)$ iterations of the underlying MCMC algorithm. Thus both the variance and the cost of $H_{k:m}(X, Y)$ will approach those of MCMC estimators for large values of k and m. This motivates the use of the estimator $H_{k:m}(X, Y)$ with $m > k$, which enables us to control the loss of efficiency that is associated with the removal of burn-in bias in contrast with the basic estimator $H_k(X, Y)$ of Section 2.1. We discuss the choice of k and m in further detail in Section 3 and in the subsequent experiments. A variant of estimator (2.1) can be obtained by considering a time lag that is greater than 1 between the two chains $(X_t)_{t\geqslant 0}$ and $(Y_t)_{t\geqslant 0}$, with the meeting time defined as the first time t for which $\{X_t = Y_{t-\text{lag}}\}$ occurs. This introduces another tuning parameter but was found to be fruitful in Biswas *et al.* (2019).

We conclude this section with a few remarks on practical implementations. First, the test function h does not have to be specified at run time in algorithm 1 in Table 1. One can store the coupled chains and choose the test function later. Also, one typically resorts to thinning the output of an MCMC sampler if the memory cost of storing chains is prohibitive, or if the cost of evaluating the test function of interest is significant compared with the cost of each MCMC iteration (e.g. Owen (2017)). This is feasible in the framework proposed: one could consider a variation of algorithm 1 where each call to the Markov kernels P and $\bar{P}$ would be replaced by multiple calls to them. We also observe that the estimators proposed can take values outside the range of the test function h; for instance they can take negative values even if the range of the test function contains only non-negative values.

Finally, we stress the difficulty that is inherent in choosing an initial distribution π_0. The estimators are unbiased for any choice of π_0, including point masses, but this choice has an effect on both the computing cost and the variance. There is also a choice about whether to draw X_0 and Y_0 independently from π_0 or not; in our experiments we use independent draws. We shall see in Section 5.1 that unfortunate choices of initial distributions can severely affect the performance of the estimators proposed. This suggests trying more than one choice of initialization, especially in the setting of multimodal targets. Overall the choice of π_0 and its relative importance compared with standard MCMC sampling are open questions.

2.4. Signed measure estimator

We can formulate the proposed estimation procedure in terms of a signed measure $\hat{\pi}$ defined by

$$\hat{\pi} = \frac{1}{m-k+1}\sum_{l=k}^{m}\delta_{X_l} + \sum_{l=k+1}^{\tau-1}\min\left(1, \frac{l-k}{m-k+1}\right)(\delta_{X_l} - \delta_{Y_{l-1}}), \tag{2.2}$$

which is obtained by replacing test function evaluations by delta masses in equation (2.1), as in Section 4 of Glynn and Rhee (2014). The measure $\hat{\pi}$ is of the form $\hat{\pi} = \sum_{l=1}^{N}\omega_l\delta_{Z_l}$ where the weights satisfy $\sum_{l=1}^{N}\omega_l = 1$ and where the atoms (Z_l) are values among the history of the coupled chains. Some of the weights (ω_l) may be negative, making $\hat{\pi}$ a signed empirical measure. In this view the unbiasedness property states that $\mathbb{E}[\sum_{l=1}^{N}\omega_l h(Z_l)] = \mathbb{E}_\pi[h(X)]$ for a test function h.

We can consider the convergence behaviour of $\hat{\pi}^R = R^{-1}\sum_{r=1}^{R}\hat{\pi}^{(r)}$ towards π, where $(\hat{\pi}^{(r)})$ for $r \in \{1, \ldots, R\}$ are independent replications of $\hat{\pi}$. Glynn and Rhee (2014) obtained a Glivenko–Cantelli result for a similar measure related to their estimator. In the current setting, assume for simplicity that π is univariate or else consider only one of its marginals. To emphasize the importance of the number of replications R, we rewrite the weights and atoms as $\hat{\pi}^R = \sum_{l=1}^{N_R}\omega_l\delta_{Z_l}$. Introduce the function $s \mapsto \hat{F}^R(s) = \sum_{l=1}^{N_R}\omega_l\mathbb{1}(Z_l \leqslant s)$ on $\mathbb{R}$. Proposition 2 in Section 3 states that $\hat{F}^R$ converges to F as $R \to \infty$ uniformly with probability 1, where F is the cumulative distribution function of π.

The function $s \mapsto \hat{F}^R(s)$ is not monotonically increasing because of negative weights among (ω_l), which motivates the following comments regarding the estimation of quantiles of π. Assume from now on that the pairs (ω_l, Z_l) are ordered such that $Z_l \leqslant Z_{l+1}$. For any $q \in (0,1)$ there might be more than one index l such that $\Sigma_{i=1}^{l-1}\omega_i \leqslant q$ and $\Sigma_{i=1}^{l}\omega_i > q$; the quantile estimate might be defined as Z_l for any such l. The convergence of $\hat{F}^R$ to F indicates that all such estimates are expected to converge to the qth quantile of π. Therefore the signed measure representation leads to a way of estimating quantiles of the target distribution in a consistent way as $R \to \infty$. The construction of confidence intervals for these quantiles, perhaps by bootstrapping the R independent copies, stands as an interesting area for future research. Another route to estimate quantiles of π would be to project marginals of $\hat{\pi}^R$ onto the space of probability measures, for instance by using a generalization of the Wasserstein metric to signed measures (Mainini, 2012). One could also estimate F by using isotonic regression (Chatterjee *et al.*, 2015), considering $\hat{F}^R(s)$ for various values s as noisy measurements of $F(s)$.

3. Properties and parallel implementation

The proofs of the results of this section are in the on-line supplementary materials. Our first result establishes the basic validity of the estimators proposed.

Proposition 1. Under assumptions 1–3, for all $k \geqslant 0$ and $m \geqslant k$, the estimator $H_{k:m}(X, Y)$ has expectation $\mathbb{E}_\pi[h(X)]$, a finite variance and a finite expected computing time.

A direct consequence of proposition 1 is that an average of R independent copies of $H_{k:m}(X, Y)$ converges to $\mathbb{E}_\pi[h(X)]$ as $R \to \infty$. We discuss more sophisticated results on unbiased estimators and parallel processing in Section 3.3 and other uses of such estimators in Sections 5.5 and 6. Following Glynn and Rhee (2014), we provide proposition 2 on the signed measure estimator (2.2). We recall that such estimators apply to univariate target distributions or to the marginal distributions of a multivariate target.

Proposition 2. Under assumptions 2 and 3, for all $m \geqslant k \geqslant 0$, and assuming that $(X_t)_{t \geqslant 0}$ converges to π in total variation, introduce the function $s \mapsto \hat{F}^R(s) = \Sigma_{l=1}^{N_R} \omega_l \mathbb{1}(Z_l \leqslant s)$, where $(\omega_l, Z_l)_{l=1}^{N_R}$ are weighted atoms obtained from R independent copies of $\hat{\pi}$ in equation (2.2). Denote by F the cumulative distribution function of π. Then

$$\sup_{s \in \mathbb{R}} |\hat{F}^R(s) - F(s)| \underset{R \to \infty}{\longrightarrow} 0 \qquad \text{almost surely.}$$

Section 3.1 studies the variance and efficiency of $H_{k:m}(X, Y)$, Section 3.2 concerns the verification of assumption 2 by using drift conditions and Section 3.2 discusses estimation on parallel processors in the presence of a budget constraint.

3.1. Variance and efficiency

We consider the effect of k and m on the efficiency of the estimators proposed, which will then suggest guidelines for the choice of these tuning parameters. Estimators $H_{k:m}^{(r)}(X, Y)$, for $r = 1, \ldots, R$, can be generated independently and averaged. More estimators can be produced in a given computing budget if each estimator is cheaper to produce. The trade-off can be understood in the framework of Glynn and Whitt (1992) (see also Rhee and Glynn (2012) and Glynn and Rhee (2014)), by defining the asymptotic inefficiency as the product of the variance and expected cost of the estimator. That product is the asymptotic variance of $R^{-1}\Sigma_{r=1}^{R} H_{k:m}^{(r)}(X, Y)$ as the computational budget, as opposed to the number of estimators R, goes to ∞ (Glynn and Whitt,

1992). Of primary interest is the comparison of this asymptotic inefficiency with the asymptotic variance of standard MCMC estimators. We start by writing the time-averaged estimator (2.1) as

$$H_{k:m}(X,Y) = \mathrm{MCMC}_{k:m} + \mathrm{BC}_{k:m},$$

where $\mathrm{MCMC}_{k:m}$ is the MCMC average $(m-k+1)^{-1}\Sigma_{l=k}^{m} h(X_l)$ and $\mathrm{BC}_{k:m}$ is the bias correction term. The variance of $H_{k:m}(X,Y)$ can be written

$$\mathbb{V}\{H_{k:m}(X,Y)\} = \mathbb{E}[\{\mathrm{MCMC}_{k:m} - \mathbb{E}_\pi[h(X)]\}^2] + 2\mathbb{E}[\{\mathrm{MCMC}_{k:m} - \mathbb{E}_\pi[h(X)]\}\mathrm{BC}_{k:m}] + \mathbb{E}[\mathrm{BC}_{k:m}^2].$$

Defining the mean-squared error of the MCMC estimator as $\mathrm{MSE}_{k:m} = \mathbb{E}[\{\mathrm{MCMC}_{k:m} - \mathbb{E}_\pi[h(X)]\}^2]$, the Cauchy–Schwarz inequality yields

$$\mathbb{V}\{H_{k:m}(X,Y)\} \leqslant \mathrm{MSE}_{k:m} + 2\sqrt{\mathrm{MSE}_{k:m}}\sqrt{\mathbb{E}[\mathrm{BC}_{k:m}^2]} + \mathbb{E}[\mathrm{BC}_{k:m}^2]. \tag{3.1}$$

To bound $\mathbb{E}[\mathrm{BC}_{k:m}^2]$, we introduce a geometric drift condition on the Markov kernel P.

Assumption 4. The Markov kernel P is π invariant, φ irreducible and aperiodic, and there is a measurable function $V : \mathcal{X} \to [1,\infty)$, $\lambda \in (0,1)$, $b < \infty$, and a small set $\mathcal{C}$ such that, for all $x \in \mathcal{X}$,

$$\int P(x, \mathrm{d}y) V(y) \leqslant \lambda V(x) + b\mathbb{1}(x \in \mathcal{C}).$$

We refer the reader to Meyn and Tweedie (2009) for the definitions and core theoretical tools for working with Markov chains on a general state space: in particular chapter 5 for aperiodicity, φ-irreducibility and small sets, and chapter 15 for geometric drift conditions; see also Roberts and Rosenthal (2004). Geometric drift conditions are known to hold for various MCMC algorithms (e.g. Roberts and Tweedie (1996a, b), Jarner and Hansen (2000), Atchade (2006), Khare and Hobert (2013), Choi and Hobert (2013) and Pal and Khare (2014)). Assumption 4 often plays a central role in establishing geometric ergodicity (e.g. theorem 9 in Roberts and Rosenthal (2004)). We show next that this assumption enables an informative bound on $\mathbb{E}[\mathrm{BC}_{k:m}^2]$.

Proposition 3. Suppose that assumptions 2–4 hold, with a function V for which the integral $\int V(x)\pi_0(\mathrm{d}x)$ is finite. If the function h is such that $\sup_{x\in\mathcal{X}} |h(x)|/V(x)^\beta < \infty$ for some $\beta \in [0, \frac{1}{2})$, then for all $m \geqslant k \geqslant 0$ we have

$$\mathbb{E}[\mathrm{BC}_{k:m}^2] \leqslant \frac{C_{\delta,\beta}\delta_\beta^k}{(m-k+1)^2},$$

for some constants $C_{\delta,\beta} < \infty$, and $\delta_\beta = \delta^{1-2\beta} \in (0,1)$, with $\delta \in (0,1)$ as in assumption 2.

Using proposition 3, inequality (3.1) becomes

$$\mathbb{V}[H_{k:m}(X,Y)] \leqslant \mathrm{MSE}_{k:m} + 2\sqrt{\mathrm{MSE}_{k:m}}\frac{\sqrt{(C_{\delta,\beta}\delta_\beta^k)}}{m-k+1} + \frac{C_{\delta,\beta}\delta_\beta^k}{(m-k+1)^2}. \tag{3.2}$$

The variance of $H_{k:m}(X,Y)$ is thus bounded by the mean-squared error MSE of an MCMC estimator plus additive terms that vanish geometrically in k and polynomially in $m-k$.

To facilitate the comparison between the efficiency of $H_{k:m}(X,Y)$ and that of MCMC estimators, we add simplifying assumptions. First, the rightmost terms of inequality (3.2) decrease geometrically with k, at a rate that is driven by $\delta_\beta = \delta^{1-2\beta}$ where δ is as in assumption 2. This motivates a choice of k depending on the distribution of the meeting time τ. In practice, we can

sample independent realizations of the meeting time and choose k such that $\mathbb{P}(\tau > k)$ is small, i.e. we choose k as a large quantile of the meeting times.

Dropping the third term on the right-hand side of inequality (3.2), which is smaller than the second term, assuming that $\mathrm{MSE}_{k:m} > 0$ and that $m > \tau$ with large probability, we obtain the approximate inequality

$$\mathbb{E}[2(\tau-1)+\max(1,m+1-\tau)]\mathbb{V}\{H_{k:m}(X,Y)\} \lessapprox \{m+\mathbb{E}(\tau)\}\mathbb{V}\{H_{k:m}(X,Y)\}$$
$$\lessapprox (m-k+1)\mathrm{MSE}_{k:m}\left\{1+\frac{k+\mathbb{E}(\tau)}{m-k+1}\right\}\left[1+\frac{2}{\sqrt{(m-k+1)}}\sqrt{\left\{\frac{C_{\delta,\beta}\delta_\beta^k}{(m-k+1)\mathrm{MSE}_{k:m}}\right\}}\right].$$

As k increases we expect $(m-k+1)\mathrm{MSE}_{k:m}$ to converge to $\mathbb{V}\{(m-k+1)^{-1/2}\Sigma_{t=k}^m h(X_t)\}$, where X_k would be distributed according to π. Denote this variance by $V_{k,m}$. The limit of $V_{k,m}$ as $m \to \infty$ is the asymptotic variance of the MCMC estimator, denoted by V_∞. Hence, for k and $m-k$ both large, the loss of efficiency of the method compared with standard MCMC methods is approximately $1+(k+\mathbb{E}[\tau])/(m-k)$.

This informal series of approximations suggests that we can retrieve an asymptotic efficiency that is comparable with the underlying MCMC estimators with appropriate choices of k and m that depend on the distribution of the meeting time τ. These choices are thus sensitive to the coupling of the chains and not only to the performance of the underlying MCMC algorithm. Choosing m as a multiple of k, such as $5k$ or $10k$, makes intuitive sense when considering that k/m is the proportion of iterations that are simply discarded in the event that $\tau < k$. In other words, the bias of the MCMC algorithm can be removed at the cost of an increased variance, which can in turn be reduced by choosing sufficiently large values of k and m. This results in a trade-off with the desired level of parallelism: one might prefer to keep k and m small, yielding a suboptimal efficiency for $H_{k:m}(X,Y)$, but enabling more independent copies to be generated in a given computing time.

3.2. Verifying assumption 2

We discuss how assumption 4 on the Markov kernel P can be used to verify assumption 2, on the shape of the meeting time distribution. Informally, assumption 4 guarantees that the bivariate chain $\{(X_t, Y_{t-1}), t \geqslant 1\}$ visits $\mathcal{C} \times \mathcal{C}$ infinitely often, where $\mathcal{C}$ is a small set. If there is a positive probability of the event $\{X_{t+1} = Y_t\}$ for every t such that $(X_t, Y_{t-1}) \in \mathcal{C} \times \mathcal{C}$, then we expect assumption 2 to hold. The next result formalizes that intuition. The proof is based on a modification of an argument by Douc *et al.* (2004). We introduce $\mathcal{D} = \{(x,y) \in \mathcal{X} \times \mathcal{X} : x = y\}$. Then assumption 3 reads $\bar{P}\{(x,x), \mathcal{D}\} = 1$ for all $x \in \mathcal{X}$.

Proposition 4. Suppose that P satisfies assumption 4 with a small set $\mathcal{C}$ of the form $\mathcal{C} = \{x : V(x) \leqslant L\}$ where $\lambda + b/(1+L) < 1$. Suppose also that there exists $\epsilon \in (0,1)$ such that

$$\inf_{(x,y)\in\mathcal{C}\times\mathcal{C}} \bar{P}\{(x,y), \mathcal{D}\} \geqslant \epsilon. \tag{3.3}$$

Then there is a finite constant C' and a $\kappa \in (0,1)$, such that, for all $n \geqslant 1$,

$$\mathbb{P}(\tau > n) \leqslant C'\pi_0(V)\kappa^n,$$

where $\pi_0(V) = \int V(x)\pi_0(\mathrm{d}x)$. Hence assumption 2 holds as long as $\pi_0(V) < \infty$.

If assumption 4 holds with a small set of the form $\mathcal{C} = \{x : V(x) \leqslant L\}$ for some $L > 0$, then it holds also for $\mathcal{C} = \{x : V(x) \leqslant L'\}$ for all $L' \geqslant L$. In that case we can always choose L sufficiently large that $\lambda + b/(1+L) < 1$. Hence the main restriction in proposition 4 is the assumption that

the small sets in assumption 4 are of the form $\{x : V(x) \leqslant L\}$, i.e. level sets of V. This is known to be true in some cases. For instance it is known from theorem 2.2 of Roberts and Tweedie (1996b) that, for a large class of MH algorithms, any non-empty compact set is a small set, and therefore for these algorithms it suffices to check that the level sets of the drift function V are compact. Common examples of drift functions include $V(x) = c/\sqrt{\pi(x)}$ (Roberts and Tweedie, 1996b; Jarner and Hansen, 2000; Atchade, 2006), $V(x) = c\exp(b|x|)$ (Roberts and Tweedie, 1996a) or the example in Pal and Khare (2014), which all have compact level sets under mild regularity conditions.

The work of Middleton *et al.* (2018) contains results that generalize propositions 3 and 4 to Markov chains satisfying polynomial drift conditions (e.g. Andrieu and Vihola (2015)), leading to polynomial tails for the associated meeting times.

3.3. Parallel implementation under budget constraints

Our main motivation for unbiased estimators comes from parallel processing; see Sections 5.5 and 6 for other motivations. Independent unbiased estimators with finite variance can be generated on separate machines and combined into consistent and asymptotically normal estimators. If the number of estimators is prespecified, this follows from the central limit theorem for independent and identically distributed variables. We might prefer to specify a time budget, and to generate as many estimators as possible within the budget. The lack of bias allows the application of a variety of results on budget-constrained parallel simulations, which we briefly review here, following Glynn and Heidelberger (1990, 1991).

We denote the proposed estimator by H and its expectation, which is the object of interest here, by $\pi(h)$. Generating H takes a random time C. We write $N(t)$ for the number of independent copies of H that can be produced by time t. The sequence $(H_n, C_n)_{n\in\mathbb{N}}$ refers to independent and identically distributed copies of (H, C), so we can write $N(t) = \sup\{n \geqslant 0 : C_1 + \ldots + C_n \leqslant t\}$, with $N(t) = 0$ if $t < C_1$. We add the subscript p to refer to objects that are associated with processor $p \in \{1, \ldots, P\}$.

The first result is that the estimator $\bar{H}_p(t)$, defined for all $1 \leqslant p \leqslant P$ as 0 if $N_p(t) = 0$, and by the sample average of $H_{p1}, \ldots, H_{pN_p(t)}$ otherwise, is biased: $\mathbb{E}[\bar{H}_p(t)] = \mathbb{E}[H] - \mathbb{E}[H\mathbb{1}(C > t)]$. Corollary 6 of Glynn and Heidelberger (1990) states that, if $\mathbb{E}[|H\exp(\alpha C)|] < \infty$ for some $\alpha > 0$, then the bias is negligible compared with $\exp(-\alpha t)$ as $t \to \infty$. By the Cauchy–Schwarz inequality, $\mathbb{E}[|H\exp(\alpha C)|]^2$ is less than the product of $\mathbb{E}[H^2]$ and $\mathbb{E}[\exp(2\alpha C)|]$. In our context, $\mathbb{E}[H^2]$ is finite under proposition 1, and $\mathbb{E}[\exp(2\alpha C)|]$ is finite for a range of values of α that depends on the value of δ in assumption 2.

We can define an unbiased estimator of $\pi(h)$ with a slight modification of $\bar{H}_p(t)$. For all $p \in \{1, \ldots, P\}$, set $\tilde{H}_p(t) = \bar{H}_p(t)$ if $N_p(t) > 0$ and $\tilde{H}_p(t) = H_{p1}$ if $N_p(t) = 0$. With $\tilde{N}_p(t) = \max\{1, N_p(t)\}$, then $\tilde{H}_p(t)$ is the sample average of $H_{p1}, \ldots, H_{p\tilde{N}_p(t)}$. The computation of $\tilde{N}_p(t)$ requires the completion of H_{p1}, and thus we cannot necessarily return $\tilde{H}_p(t)$ at time t, in contrast with $\bar{H}_p(t)$. In contrast, we have $\mathbb{E}[\tilde{H}_p(t)] = \mathbb{E}[H] = \pi(h)$, i.e. the estimator is unbiased, provided that $\mathbb{E}[|H|] < \infty$ (corollary 7 of Glynn and Heidelberger (1990)). We denote the average of $\tilde{H}_p(t)$ over P processors by $\tilde{H}(P, t) = P^{-1}\Sigma_{p=1}^{P}\tilde{H}_p(t)$, which is unbiased for $\pi(h)$.

Asymptotic results on $\tilde{H}(P, t)$ can be found in Glynn and Heidelberger (1991) and are summarized below. We first have the consistency results: $\lim_{t\to\infty}\tilde{H}(P, t) = \lim_{P\to\infty}\tilde{H}(P, t) = \pi(h)$ almost surely for all t and P, and, if $\mathbb{E}[|H|^{1+\delta}]$ for some $\delta > 0$ and if $\{t_P\}$ is a sequence such that $\lim_{P\to\infty} t_P = \infty$, then $\tilde{H}(P, t_P)$ converges to $\pi(h)$ in probability as $P \to \infty$. Next, we can construct confidence intervals for $\pi(h)$ based on $\tilde{H}(P, t)$, following the end of section 3 in Glynn and Heidelberger (1991). Indeed, define

$$\hat{\sigma}_1^2(P,t) = \frac{1}{P-1}\sum_{p=1}^{P}\{\tilde{H}_p(t) - \tilde{H}(P,t)\}^2,$$

$$\tilde{\tau}(P,t) = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{\tilde{N}_p(t)}\sum_{n=1}^{\tilde{N}_p(t)} C_{pn},$$

$$\hat{\sigma}_2^2(P,t) = \tilde{\tau}(P,t)\left\{\frac{1}{P}\sum_{p=1}^{P}\frac{1}{\tilde{N}_p(t)}\sum_{n=1}^{\tilde{N}_p(t)} H_{pn}^2 - \tilde{H}(P,t)^2\right\},$$

where $\tilde{N}_p(t) = \max\{1, N_p(t)\}$. Then we have the three following central limit theorems: for fixed t and $P \to \infty$,

$$\frac{\sqrt{P}}{\hat{\sigma}_1(P,t)}\{\tilde{H}(P,t) - \pi(h)\} \to \mathcal{N}(0,1); \tag{3.4}$$

for fixed P and $t \to \infty$,

$$\frac{\sqrt{(Pt)}}{\hat{\sigma}_2(P,t)}\{\tilde{H}(P,t) - \pi(h)\} \to \mathcal{N}(0,1); \tag{3.5}$$

if $t_P \to \infty$ as $P \to \infty$,

$$\frac{\sqrt{(Pt_P)}}{\hat{\sigma}_2(P,t_P)}\{\tilde{H}(P,t_P) - \pi(h)\} \to \mathcal{N}(0,1). \tag{3.6}$$

These results require moment conditions such as $\mathbb{E}[\tilde{H}_p(t)^2] < \infty$. The central limit theorem (3.4) will be used to construct confidence intervals in Sections 5.3 and 5.4.

We conclude this section with a remark on the setting where t is fixed and the number of processors P goes to ∞. There, the time to obtain $\tilde{H}(P,t)$ would typically increase with P. Indeed at least one estimator needs to be completed on each processor for $\tilde{H}(P,t)$ to be available. The completion time behaves as the maximum of independent copies of the cost C. Under assumption 2, the completion time for $\tilde{H}(P,t)$ has expectation behaving as $\log(P)$ when $P \to \infty$, for fixed t. Other tail assumptions (Middleton *et al.*, 2018) would lead to different behaviour for the completion time that is associated with $\tilde{H}(P,t)$.

4. Couplings of Markov chain Monte Carlo algorithms

We consider couplings of various MCMC algorithms that satisfy assumptions 2 and 3. These couplings are widely applicable and do not require extensive analytical knowledge of the target distribution. We stress that they are not optimal in general, and we expect that other constructions would yield more efficient estimators. We begin in Section 4.1 by reviewing maximal couplings.

4.1. Sampling from maximal couplings

A maximal coupling between two distributions p and q on a space $\mathcal{X}$ is a distribution of a pair of random variables (X, Y) that maximizes $\mathbb{P}(X = Y)$, subject to the marginal constraints $X \sim p$ and $Y \sim q$. We write p and q both for these distributions and for their probability density functions with respect to a common dominating measure, and we refer to the uniform distribution on the interval $[a, b]$ by $\mathcal{U}([a, b])$. A procedure to sample from a maximal coupling is described in algorithm 2 in Table 2; see for example section 4.5 of chapter 1 of Thorisson (2000), and Johnson (1998) where it is termed γ-coupling.

We justify algorithm 2 and compute its cost. Denote by (X, Y) the output of the algorithm. First, X follows p from step 1. To prove that Y follows q, introduce a measurable set A. We write

Table 2. Algorithm 2: sampling from a maximal coupling of p and q

Step 1: sample $X \sim p$ and $W|X \sim \mathcal{U}\{[0, p(X)]\}$: if $W \leqslant q(X)$, output (X, X) *Step 2*: otherwise, sample $Y^* \sim q$ and $W^*|Y^* \sim \mathcal{U}\{[0, q(Y^*)]\}$ until $W^* > p(Y^*)$, and output (X, Y^*)

$\mathbb{P}(Y \in A) = \mathbb{P}(Y \in A, \text{step 1}) + \mathbb{P}(Y \in A, \text{step 2})$, where the events $\{\text{step 1}\}$ and $\{\text{step 2}\}$ refer to the algorithm terminating at step 1 or 2. We compute

$$\mathbb{P}(Y \in A, \text{ step 1}) = \int_A \int_0^\infty \mathbb{1}\{w \leqslant q(x)\} \frac{\mathbb{1}\{0 \leqslant w \leqslant p(x)\}}{p(x)} p(x) \mathrm{d}w \, \mathrm{d}x = \int_A \min\{p(x), q(x)\} \mathrm{d}x.$$

We can deduce from this that $\mathbb{P}(\text{step 1}) = \int_{\mathcal{X}} \min\{p(x), q(x)\} \mathrm{d}x$. For $\mathbb{P}(Y \in A, \text{ step 2})$ to be equal to $\int_A [q(x) - \min\{p(x), q(x)\}] \mathrm{d}x$, we need

$$\int_A [q(x) - \min\{p(x), q(x)\}] \mathrm{d}x = \mathbb{P}(Y \in A | \text{step 2}) \left[1 - \int_{\mathcal{X}} \min\{p(x), q(x)\} \mathrm{d}x\right],$$

and we conclude that the distribution of Y given $\{\text{step 2}\}$ should for all x have a density $\tilde{q}(x)$ equal to $[q(x) - \min\{p(x), q(x)\}]/[1 - \int \min\{p(x'), q(x')\} \mathrm{d}x']$. Step 2 is a standard rejection sampler using q as a proposal distribution to target $\tilde{q}$, which concludes the proof that $Y \sim q$. We also confirm that algorithm 2 maximizes the probability of $\{X = Y\}$. Under the algorithm,

$$\mathbb{P}(X = Y) = \mathbb{P}(\text{step 1}) = \int_{\mathcal{X}} \min\{p(x), q(x)\} \mathrm{d}x = 1 - d_{\mathrm{TV}}(p, q),$$

where $d_{\mathrm{TV}}(p, q) = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \mathrm{d}x$ is the total variation distance. By the coupling inequality (Lindvall, 2002), this proves that the algorithm implements a maximal coupling.

To assess the cost of algorithm 2, note that step 1 costs one draw from p, one evaluation from p and one from q. Each attempt in the rejection sampler of step 2 costs one draw from q, one evaluation from p and one from q. Hereafter we refer to the cost of one draw and two evaluations by 'one unit'. Observe that the probability of acceptance in step 2 is given by $\mathbb{P}\{W^* \geqslant p(Y^*)\} = 1 - \int_{\mathcal{X}} \min\{p(y), q(y)\} \mathrm{d}y$. Then, the number of attempts in step 2 has a geometric distribution with mean $[1 - \int_{\mathcal{X}} \min\{p(y), q(y)\} \mathrm{d}y]^{-1}$, and step 2 itself occurs with probability $1 - \int_{\mathcal{X}} \min\{p(y), q(y)\} \mathrm{d}y$. Therefore the overall expected cost is two units. The expectation of the cost is the same for all distributions p and q, whereas the variance of the cost depends on $d_{\mathrm{TV}}(p, q)$, and in fact goes to ∞ as this distance goes to 0.

In algorithm 2, the value of X is not used in the generation of Y^* within step 2. In other words, conditionally on $\{X \neq Y\}$, the two output variables are independent. We might prefer to correlate the outputs in the event $\{X \neq Y\}$, e.g. in random-walk MH steps as in the next section. We describe a maximal coupling presented in Bou-Rabee *et al.* (2018). It applies to distributions p and q on $\mathbb{R}^d$ such that $X \sim p$ can be represented as $X = \mu_1 + \Sigma^{1/2} \dot{X}$, and $Y \sim q$ as $Y = \mu_2 + \Sigma^{1/2} \dot{Y}$, where the pair $(\dot{X}, \dot{Y})$ follows a coupling of some distribution s with itself. The construction requires that s is spherically symmetrical: $s(x) = s(y)$ for all $x, y \in \mathbb{R}^d$ such that $\|x\| = \|y\|$, where $\|\cdot\|$ denotes the Euclidean norm. For instance, if s is a standard multivariate normal distribution, then $X \sim \mathcal{N}(\mu_1, \Sigma)$ and $Y \sim \mathcal{N}(\mu_2, \Sigma)$.

Let $z = \Sigma^{-1/2}(\mu_1 - \mu_2)$ and $e = z/\|z\|$. We independently draw $\dot{X} \sim s$ and $U \sim \mathcal{U}([0, 1])$ and let

$$\dot{Y} = \begin{cases} \dot{X} + z & \text{if } U \leqslant \min\left\{1, \dfrac{s(\dot{X} + z)}{s(\dot{X})}\right\}, \\ \dot{X} - 2(e'\dot{X})e & \text{otherwise.} \end{cases}$$

This procedure outputs a pair $(\dot{X}, \dot{Y})$ that follows a coupling of s with itself. We then define $(X, Y) = (\mu_1 + \Sigma^{1/2}\dot{X}, \mu_2 + \Sigma^{1/2}\dot{Y})$. On the event $\{\dot{Y} = \dot{X} + z\}$, we have $X = Y$. On the event $\{\dot{Y} \neq \dot{X} + z\}$, the vector $\dot{X} - 2(e'\dot{X})e$ is the reflection of $\dot{X}$ through the hyperplane orthogonal to e that passes through the origin. We show that the output (X, Y) follows a maximal coupling of p and q, which we refer to as a maximal coupling with reflection on the residuals, or a 'reflection maximal coupling'. First we show that $\dot{Y}$ follows s, closely following the argument in Bou-Rabee *et al.* (2018). For a measurable set B, we compute

$$\mathbb{P}(\dot{Y} \in B) = \int \mathbb{1}_B(x+z) \min\left\{1, \frac{s(x+z)}{s(x)}\right\} s(x)\mathrm{d}x + \int \mathbb{1}_B\{x - 2(e'x)e\} \max\left\{0, 1 - \frac{s(x+z)}{s(x)}\right\} s(x)\mathrm{d}x.$$

The first integral above becomes $\int \mathbb{1}_B(w) \min\{s(w-z), s(w)\}\mathrm{d}w$, after a change of variables $w := x + z$. To simplify the second integral we make the change of variables $w := x - 2(e'x)e$. Since this corresponds to a reflection with respect to a plane orthogonal to e, we have $\mathrm{d}w = \mathrm{d}x$, and $x = w - 2(e'w)e$; thus

$$\int \mathbb{1}_B\{x - 2(e'x)e\} \max\{0, s(x) - s(x+z)\}\mathrm{d}x = \int \mathbb{1}_B(w) \max\{0, s(w) - s(w-z)\}\mathrm{d}w,$$

where we have used $s\{w - 2(e'w)e\} = s(w)$ and $s\{w - 2(e'w)e + z\} = s(w - z)$, because $\|w - 2(e'w)e\| = \|w\|$ and $\|w - 2(e'w)e + z\| = \|w - z\|$. Summing the two integrals we obtain $\mathbb{P}(\dot{Y} \in B) = \int_B s(w)\mathrm{d}w$, so $\dot{Y} \sim s$.

To verify that the procedure corresponds to a maximal coupling of p and q, we observe that

$$\begin{aligned}\mathbb{P}(X \neq Y) = \mathbb{P}(\dot{Y} \neq \dot{X} + z) &= 1 - \int \min\{s(x), s(x+z)\}\mathrm{d}x \\ &= 1 - \int \min[s\{\Sigma^{-1/2}(\tilde{x} - \mu_1)\}, s\{\Sigma^{-1/2}(\tilde{x} - \mu_2)\}]|\Sigma^{-1/2}|\mathrm{d}\tilde{x},\end{aligned}$$

with the change of variable $\tilde{x} := \mu_1 + \Sigma^{1/2}x$. This is precisely the total variation distance between p and q, on writing their densities in terms of the density of s. Note that the computational cost that is associated with the above sampling technique is deterministic, in contrast with the cost of algorithm 2.

Finally, for discrete distributions with common finite support, a procedure for sampling from a maximal coupling is described in Section 5.4, with a cost that is also deterministic.

4.2. Metropolis–Hastings steps

In Section 2.2 we described a coupling of MH chains due to Johnson (1998); we summarize the coupled kernel $\bar{P}\{(X_t, Y_{t-1}), \cdot\}$ in the following procedure.

Step 1: sample $(X^*, Y^*)|(X_t, Y_{t-1})$ from a maximal coupling of $q(X_t, \cdot)$ and $q(Y_{t-1}, \cdot)$.
Step 2: sample $U \sim \mathcal{U}([0, 1])$.
Step 3: if $U \leqslant \min\{1, \pi(X^*)q(X^*, X_t)/\pi(X_t)q(X_t, X^*)\}$, then $X_{t+1} = X^*$; otherwise $X_{t+1} = X_t$.
Step 4: if $U \leqslant \min\{1, \pi(Y^*)q(Y^*, Y_{t-1})/\pi(Y_{t-1})q(Y_{t-1}, Y^*)\}$, then $Y_t = Y^*$; otherwise $Y_t = Y_{t-1}$.

Here we address the verification of assumptions 1–3 for this algorithm. Assumption 1 can be verified for MH chains under conditions on the target and the proposal (Nummelin, 2002; Roberts and Rosenthal, 2004). In some settings the explicit drift function that is given in theorem 3.2 of Roberts and Tweedie (1996b) may be used to verify assumption 2 as in Section 3.2. The probability of coupling at the next step given that the chains are in X_t and Y_{t-1} can be controlled

as follows. First, the probability of proposing the same value X^* depends on the total variation distance between $q(X_t, \cdot)$ and $q(Y_{t-1}, \cdot)$, which is typically strictly positive if X_t and Y_{t-1} are in bounded subsets of $\mathcal{X}$. Furthermore, the probability of accepting X^* is often strictly positive on bounded subsets of $\mathcal{X}$, for instance when $\pi(x) > 0$ for all $x \in \mathcal{X}$. Assumption 3 is satisfied by design thanks to the use of maximal couplings and common uniform variable U in the above procedure.

Different considerations drive the choice of proposal distribution in standard MCMC sampling and in our proposed estimators. In the case of random-walk proposals with variance Σ, larger variances lead to smaller total variation distances between $q(X_t, \cdot)$ and $q(Y_{t-1}, \cdot)$ and thus larger probabilities of proposing identical values. However, meeting events only occur if proposals are accepted, which is unlikely if Σ is too large. This trade-off could lead to a different choice of Σ than the optima known for the marginal chains (Roberts *et al.*, 1997), and deserves further investigation.

We perform experiments with a d-dimensional normal target distribution $\mathcal{N}(0, V)$, where V is the inverse of a matrix drawn from a Wishart distribution with identity scale matrix and d degrees of freedom. This setting, which has been borrowed from Hoffman and Gelman (2014), yields normal distribution targets with strong correlations and a dense precision matrix. Below, each independent run is performed with an independent draw of V. We consider normal random-walk proposals with variance Σ set to V/d. The division by d heuristically follows from the scaling results of Roberts *et al.* (1997). We initialize the chains either from the target distribution, or from a normal distribution centred at $(1, \ldots, 1)$ with identity covariance matrix. We first couple the proposals with a maximal coupling given by algorithm 2. The resulting average meeting times, based on 1000 independent runs, are given in Fig. 1(a). The plot indicates an exponential increase of the average meeting times with the dimension, under both initialization strategies. In passing, this illustrates that meeting times can be large even if the chains marginally start at stationarity, i.e. in a setting where there is no burn-in bias.

Next we perform the same experiments with the reflection maximum coupling that was described in the previous section. The results are shown in Fig. 1(b). The average meeting times now increase at a rate that appears closer to linear in the dimension. This is to be compared with established theoretical results on the linear performance of standard MH estimators with respect to the dimension (Roberts *et al.*, 1997). A formal justification of the scaling that is observed in Fig. 1(b) is an open question, and so is the design of more effective coupling strategies.

4.3. Gibbs sampling

Gibbs sampling is another popular class of MCMC algorithms, in which components of a Markov chain are updated alternately by sampling from the target conditional distributions (chapter 10 of Robert and Casella (2004)), implemented for example in the software packages JAGS (Plummer, 2003). In Bayesian statistics, these conditional distributions sometimes belong to a standard family such as the normal, gamma or inverse gamma distributions. Otherwise, the conditional updates might require MH steps. We can introduce couplings in each conditional update, using either maximal couplings of the target conditionals, if these are standard distributions, or maximal couplings of the proposal distributions in MH steps targeting the target conditionals. Controlling the probability of meeting at the next step over a set, as required for the application of proposition 4, can be done case by case. Drift conditions for Gibbs samplers also tend to rely on case-by-case arguments (see for example Rosenthal (1996)).

Gibbs samplers tend to perform well for targets with weak correlations between the components being updated; otherwise Gibbs chains are expected to mix poorly. We perform numerical experiments on normal target distributions in varying dimensions to observe the effect of correlations on the meeting times of coupled Gibbs chains. For each target $\mathcal{N}(0, V)$, we introduce

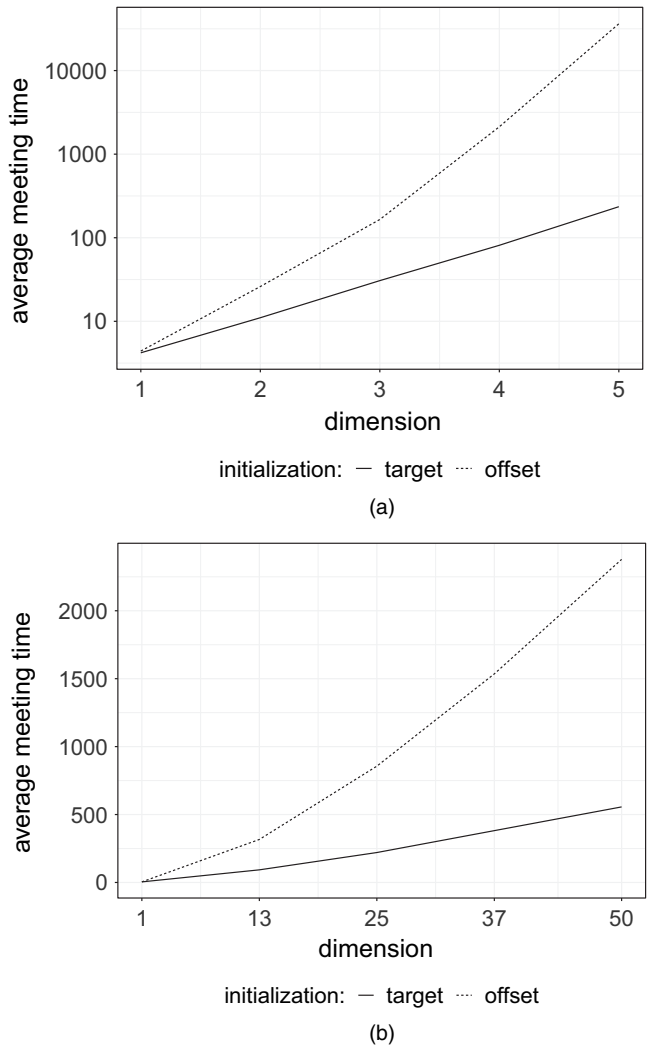

Fig. 1. Scaling of the average meeting time of a coupled MH algorithm with the dimension of the target $\mathcal{N}(0, V)$, where V is the inverse of a Wishart draw, as described in Section 4.2 (the chains are either initialized from the target, or from a normal $\mathcal{N}(\mathbf{1}_d, I_d)$ distribution, where $\mathbf{1}_d$ is a vector of 1s ('offset' in the legend)): (a) using maximal coupling of algorithm 2; (b) using reflection maximal coupling described in Section 4.1

an MH-within-Gibbs sampler, where each univariate component i is updated with a single Metropolis step, using normal distribution proposals with variance $V_{i,i}$. Here an iteration of the sampler refers to a complete scan of the components. Fig. 2(a) presents the median meeting times as a function of the dimension, when V is the inverse of a Wishart draw as in the previous section. In this highly correlated setting, the meeting times scale poorly with the dimension. The plot presents the median instead of the average, because we have stopped the runs after 500000 iterations; the median is robust to this truncation, but not the average. We remark that shorter meeting times are obtained when initializing the chains away from the target distribution.

Next we consider a normal distribution target with covariance matrix V defined by $V_{i,j} = 0.5^{-|i-j|}$, which induces weak correlations between components; the inverse of V is tridiagonal. In that case, the same Gibbs sampler performs much more favourably, as we can see from Fig.

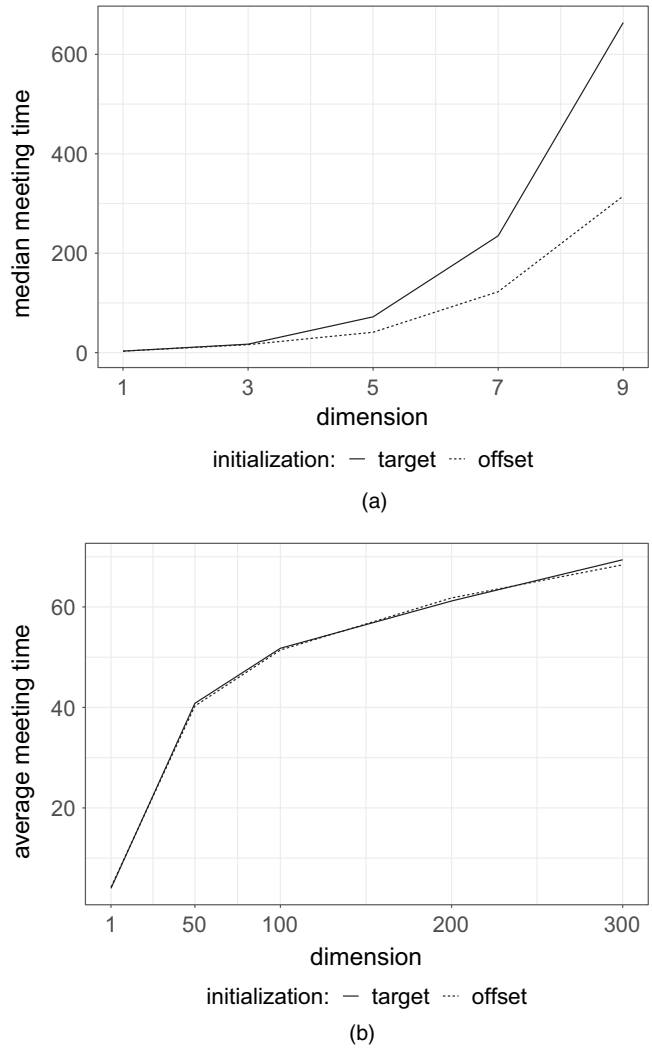

Fig. 2. (a) Scaling of the median and (b) average meeting time of a coupled Gibbs algorithm with the dimension of the target $\mathcal{N}(0, V)$ (the chains are either initialized from the target, or from a normal $\mathcal{N}(\mathbf{1}_d, I_d)$ distribution, where $\mathbf{1}_d$ is a vector of 1s ('offset' in the legend)): (a) the covariance V of the target is generated as the inverse of a Wishart sample, inducing strong correlations; (b) the covariance V is defined as $V_{ij} = 0.5^{-|i-j|}$, inducing weak correlations

2(b). The average meeting times seem to scale sublinearly with the dimension, under both choices of initializations π_0. Couplings of other Gibbs samplers will be encountered in the numerical experiments of Section 5.

4.4. Coupling of other Markov chain Monte Carlo algorithms

Among extensions of the MH algorithm, Metropolis-adjusted Langevin algorithms (e.g. Roberts and Tweedie (1996a)) are characterized by the use of a proposal distribution given current state X_t that is normal with mean $X_t + h\nabla \log\{\pi(X_t)\}/2$ and variance $h\Sigma$, with tuning parameter $h > 0$ and covariance matrix Σ. Maximal couplings or reflection maximal couplings of the proposals

could be readily implemented to obtain faithful chains. Going further in the use of gradient information, the Hamiltonian or hybrid Monte Carlo algorithm (Duane *et al.*, 1987; Neal, 1993, 2011) is a popular MCMC algorithm for large dimensional targets. In Heng and Jacob (2019), the framework of the present paper is applied to pairs of Hamiltonian Monte Carlo chains, with a focus of the verification of assumptions 1–3 in that context. Such couplings were analysed in detail in Mangoubi and Smith (2017) and Bou-Rabee *et al.* (2018) to obtain convergence rates for the underlying chains. We refer to Heng and Jacob (2019) for more details and provide for completeness some experiments on the normal distribution target that was described above in the on-line supplementary materials.

The present paper generalizes unbiased estimators that are obtained by coupling conditional particle filters in Jacob *et al.* (2019). These algorithms, which were introduced in Andrieu *et al.* (2010), target the distribution of latent processes given observations and fixed parameters for non-linear state space models. The couplings of conditional particle filters in Jacob *et al.* (2019) involve a combination of common random numbers and maximal couplings. Couplings of particle independent MH sampling, which is a particular case of MH sampling with an independent proposal distribution, are simpler to design and are considered in Middleton *et al.* (2019).

The design of generic and efficient MCMC kernels is a topic of active on-going research (see for example Murray *et al.* (2010), Goodman and Weare (2010), Pollock *et al.* (2016), Vanetti *et al.* (2017) and Titsias and Yau (2017) and references therein). Any new kernel could lead to unbiased estimators with the framework proposed, as long as appropriate couplings can be implemented.

5. Illustrations

Section 5.1 illustrates the effect of k, m and the initial distribution π_0, identifying a situation where some care is required. Section 5.2 considers the removal of the bias from a Gibbs sampler previously considered for perfect sampling and regeneration methods. Section 5.3 introduces an Ising model and a coupling of a replica exchange algorithm, and we present experiments performed on parallel processors. Section 5.4 considers a high dimensional variable-selection example, with an MH algorithm that has previously been shown to scale linearly with the number of variables. Finally, Section 5.5 focuses on the problem of approximating the cut distribution arising in modular inference, which illustrates the appeal of unbiased estimators beyond parallel computing.

5.1. Bimodal target

We use a bimodal target distribution and a random-walk MH algorithm to illustrate our method and to highlight some of its limitations. In particular, we consider a mixture of univariate normal distributions with density $\pi(x) = 0.5\mathcal{N}(x; -4, 1) + 0.5\mathcal{N}(x; 4, 1)$, which we sample from using random-walk MH with normal proposal distributions of variance $\sigma_q^2 = 9$. This enables regular jumps between the modes of π. We set the initial distribution π_0 to $\mathcal{N}(10, 10^2)$, so that chains are likely to start closer to the mode at 4 than the mode at -4. Over 1000 independent runs, we find that the meeting time τ has an average of 20 and a 99% quantile of 105.

We consider the task of estimating $\int \mathbb{1}(x > 3)\pi(\mathrm{d}x) \approx 0.421$. First, we consider the choice of k and m. Over 1000 independent experiments, we approximate the expected cost $\mathbb{E}[2(\tau - 1) + \max(1, m - \tau + 1)]$ and the variance $\mathbb{V}\{H_{k:m}(X, Y)\}$, and compute the inefficiency as the product of the two (as in Section 3.1). We then divide the inefficiency by the asymptotic variance of the MCMC estimator, which is denoted by V_∞, which we obtain from 10^6 iterations and a burn-in period of 10^4 by using the R package `CODA` (Plummer *et al.*, 2006).

We present the results in Table 3. First, we see that the inefficiency is sensitive to the choice of k and m. Second, we see that when k and m are sufficiently large we can retrieve an inefficiency that is comparable with that of the underlying MCMC algorithm. The ideal choice of k and m will depend on trade-offs between inefficiency, the desired level of parallelism and the number of processors that are available. We present a histogram of the target distribution, obtained by using $k = 200$ and $m = 2000$, in Fig. 3(a). These histograms are produced by averaging unbiased estimators of expectations of indicator functions, corresponding to consecutive intervals. Confidence intervals at level 95% are obtained from the central limit theorem and are represented as grey boxes, with vertical bars showing the point estimates.

Next, we consider a more challenging case by setting $\sigma_q^2 = 1$, again with $\pi_0 = \mathcal{N}(10, 10^2)$. These values make it difficult for the chains to jump between the modes of π. Over $R = 1000$ runs we find an average meeting time of 769, with a 99% quantile of 9186. When the chains start in different modes, the meeting times are often dramatically larger than when the chains start by the same mode. One can still recover accurate estimates of the target distribution, but k and m must be set to larger values. With $k = 20000$ and $m = 30000$, we obtain the 95% confidence interval $[0.397, 0.430]$ for $\int \mathbb{1}(x > 3)\pi(\mathrm{d}x) \approx 0.421$. We show a histogram of π in Fig. 3(b).

Table 3. Cost, variance and inefficiency divided by MCMC asymptotic variance V_∞, for various choices of k and m, for the test function $h : x \longmapsto \mathbb{1}(x > 3)$, in the bimodal target example of Section 5.1

k	m	*Cost*	*Variance*	*Inefficiency*/V_∞
1	k	37	4.1×10^{2}	1878.4
1	$10k$	39	3.6×10^{2}	1703.5
1	$20k$	45	3.0×10^{2}	1624.8
100	k	119	9.0	130.6
100	$10k$	1019	2.3×10^{-2}	2.9
100	$20k$	2019	7.9×10^{-3}	1.9
200	k	219	2.4×10^{-1}	6.5
200	$10k$	2019	5.3×10^{-3}	1.3
200	$20k$	4019	2.4×10^{-3}	1.2

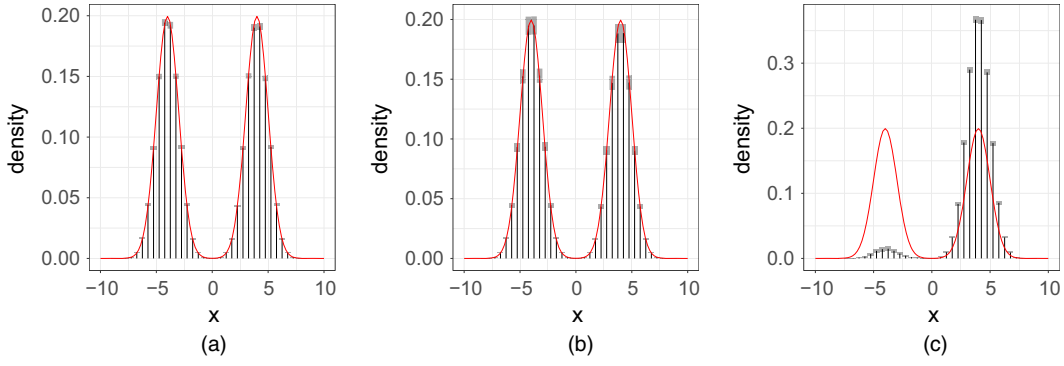

Fig. 3. Histograms of the mixture target distribution of Section 5.1, obtained with the proposed unbiased estimators, based on a normal random-walk MH algorithm, with a proposal variance σ_q^2 and an initial distribution π_0, over $R = 1000$ experiments (———, target density function): (a) $\sigma_q^2 = 3^2$ and $\pi_0 = \mathcal{N}(10, 10^2)$; (b) $\sigma_q^2 = 1^2$ and $\pi_0 = \mathcal{N}(10, 10^2)$; (c) $\sigma_q^2 = 1^2$ and $\pi_0 = \mathcal{N}(10, 1^2)$

Finally we consider a third case, with $\sigma_q^2 = 1$ as before but now with π_0 set to $\mathcal{N}(10, 1)$. This initialization makes it unlikely for a chain to start near the mode at -4. The pair of chains typically converge around the mode at 4 and meet in a small number of iterations. Over $R = 1000$ replications, we find an average meeting time of 9 and a 99% quantile of 35. A 95% confidence interval on $\int \mathbb{1}(x > 3)\pi(\mathrm{d}x)$ obtained from the estimators with $k = 50$ and $m = 500$ is [0.799, 0.816], which is far from the true value of 0.421. The associated histogram of π is shown in Fig. 3(c).

Sampling 9000 additional estimators yields a 95% confidence interval $[-0.353, 1.595]$, again using $k = 50$ and $m = 500$. Among these extra 9000 values, a few correspond to cases where one chain jumped to the leftmost mode before meeting the other. This resulted in large meeting times and thus a large empirical variance for $H_{k:m}$. On noting a large empirical variance one can then decide to use larger values of k and m. We conclude that, although our estimators are unbiased and are consistent in the limit as $R \to \infty$, poor performance of the underlying Markov chains combined with ill-chosen initializations can still produce misleading results for any finite R, such as 1000 in this example.

5.2. *Gibbs sampler for nuclear pump failure data*

Next we consider a classic Gibbs sampler for a model of pump failure counts, which was used for example in Murdoch and Green (1998) to illustrate perfect samplers for continuous distributions, and in Mykland *et al.* (1995) to illustrate their regeneration approach. Here we focus on a comparison with the regeneration approach, which was motivated by similar practical concerns to those in this paper, in particular to avoid an arbitrary choice of burn-in, to construct confidence intervals on the expectations of interest and to make principled use of parallel processors. Mykland *et al.* (1995) showed how to construct regeneration times—random times between which the chain forms independent and identically distributed 'tours'. They defined a consistent estimator for arbitrary test functions, whose asymptotic variance takes a simple form. The estimator is then obtained by aggregating over these independent tours.

The data consist of operating times $(t_n)_{n=1}^K$ and failure counts $(s_n)_{n=1}^K$ for $K = 10$ pumps at the Farley-1 nuclear power station, as first described in Gaver and O'Muircheartaigh (1987). The model specifies $s_n \sim \text{Poisson}(\lambda_n t_n)$ and $\lambda_n \sim \text{gamma}(\alpha, \beta)$, where $\alpha = 1.802$, $\beta \sim \text{gamma}(\gamma, \delta)$, $\gamma = 0.01$ and $\delta = 1$. The Gibbs sampler for this model consists of the following update steps:

$$\lambda_n \mid \text{rest} \sim \text{gamma}(\alpha + s_n, \beta + t_n) \qquad \text{for } n = 1, \ldots, K,$$

$$\beta \mid \text{rest} \sim \text{gamma}\Bigg(\gamma + 10\alpha, \delta + \sum_{n=1}^{K} \lambda_n\Bigg).$$

Here $\text{gamma}(\alpha, \beta)$ refers to the distribution with density $x \mapsto \Gamma(\alpha)^{-1}\beta^\alpha x^{\alpha-1}\exp(-\beta x)$. We initialize all parameter values to 1 (the initialization was not specified in Mykland *et al.* (1995)). To form our estimator we apply maximal couplings at each conditional update of the Gibbs sampler, as described in Section 4.3.

We begin by drawing 1000 meeting times independently. Following the guidelines of Section 3.1, we set $k = 7$, corresponding to the 99% quantile of τ and $m = 10k = 70$. For the regeneration approach, Mykland *et al.* (1995) gave a set of tuning parameters which we adopt below. Applying the regeneration approach to 1000 Gibbs sampler runs of 5000 iterations each, we observe on average 1996 complete tours per run with an average length of 2.50 iterations per tour. These values agree with the count of 1967 tours of average length 2.56 reported in Mykland *et al.* (1995). We observe a posterior mean estimate for β of 2.47 with a variance of 1.89×10^{-4} over the 1000 independent runs, which implies an efficiency value of $(5000 \times 1.89 \times 10^{-4})^{-1} = 1.06$.

This exceeds the efficiency of 0.94 that was achieved by our estimator with the choice of $k = 7$ and $m = 70$. In contrast, the regeneration approach often requires more extensive analytical work with the underlying Markov chain; we refer to Mykland *et al.* (1995) for a detailed description. For reference, the underlying Gibbs sampler achieves an efficiency of 1.08, based on a long run of 5×10^5 iterations and a burn-in of 10^3 iterations. More extensive comparisons with other regeneration approaches such as that of Brockwell and Kadane (2005) would deserve investigation.

5.3. Ising model

We consider an Ising model on a 32×32 square lattice with periodic boundaries. This provides a setting where a basic MCMC sampler can mix slowly depending on an inverse temperature parameter θ, and where a replica exchange strategy as in Geyer (1991) can be helpful. We also use this example to illustrate the use of our estimators on a large computing cluster, with the considerations that were reviewed in Section 3.3. For i and j in $\{1, \ldots, 32\}^2$ we write $i \sim j$ if i and j are neighbours in the square lattice with periodic boundaries. We write $x_i \in \{-1, 1\}$ for the spin at location i, and $x = \{x_i\}$ for the full grid. We write $t(x)$ for the 'natural statistic' $t(x) = 0.5 \Sigma_{i \in \{1,\ldots,32\}^2} \Sigma_{j \sim i} x_i x_j$ summing the products of pairs of neighbours. The 0.5-multiplier here results in each pair of neighbouring sites being counted only once. Under the model, the probability that is associated with a grid x is $\pi_\theta(x) \propto \exp\{\theta t(x)\}$, where $\theta > 0$ denotes an inverse temperature parameter that calibrates the degree of correlation between neighbouring sites.

We consider a single-site Gibbs sampler, called a heat bath algorithm in this context, to approximate the distribution π_θ given a value of θ. One iteration of the algorithm consists of a sweep through all the locations $i \in \{1, \ldots, 32\}^2$. For each i we draw x_i from its conditional distribution under π_θ given all the other spins. It can be checked that the conditional probability of $\{x_i = 1\}$ given the other spins equals $\exp(\theta s_i)/\{\exp(\theta s_i) + \exp(-\theta s_i)\}$, where s_i denotes the sum of spins over the four neighbours of i. We initialize the chains by drawing spins uniformly in $\{-1, 1\}$ at each site, independently across sites.

A simple strategy to couple heat bath chains consists of sampling from the maximal coupling of each conditional distribution. For a grid of θ-values from 0.3 and 0.48, we run 100 pairs of chains until they meet. We then plot the average meeting time as a function of θ in Fig. 4(a), noting that the average meeting time increases sharply to values above 10^6 as θ approaches its critical value (see the related discussion in Propp and Wilson (1996)). We conclude that it would be expensive to produce unbiased estimators based on the heat bath algorithm for values of θ above 0.48, for reasons related to the behaviour of the underlying algorithm.

There are several ways to address the degeneracy of the heat bath algorithm as θ increases. Specialized algorithms have been proposed to update groups of spins jointly (Swendsen and Wang, 1987; Wolff, 1989). Here, we consider an approach based on an ensemble of N chains that regularly exchange their states: a technique often termed replica exchange or parallel tempering. Following for example Geyer (1991), we introduce N chains, $x^{(1)}, \ldots, x^{(N)}$, with each $x^{(n)}$ targeting $\pi_{\theta^{(n)}}$ with different values of $\theta^{(n)}$ ordered as $\theta^{(1)} < \ldots < \theta^{(N)}$. Each iteration of the algorithm proceeds as follows. With probability $p_{\text{swap}} \in (0, 1)$, for $n \in \{1, \ldots, N-1\}$ (sequentially), we propose exchanging the states $x^{(n)}$ and $x^{(n+1)}$ corresponding to $\theta^{(n)}$ and $\theta^{(n+1)}$. We accept this swap with probability $\min[1, \pi_{\theta^{(n)}}(x^{(n+1)})\pi_{\theta^{(n+1)}}(x^{(n)})/\{\pi_{\theta^{(n)}}(x^{(n)})\pi_{\theta^{(n+1)}}(x^{(n+1)})\}]$, which simplifies to $\min[1, \exp\{(\theta^{(n)} - \theta^{(n+1)})\{t(x^{(n+1)}) - t(x^{(n)})\}]$. Otherwise we perform a full sweep of single-site Gibbs updates, independently across chains.

A coupling of this algorithm involves a pair of ensembles with N chains each; the two ensembles are identical if chain n in the first ensemble equals chain n in the second ensemble, for all

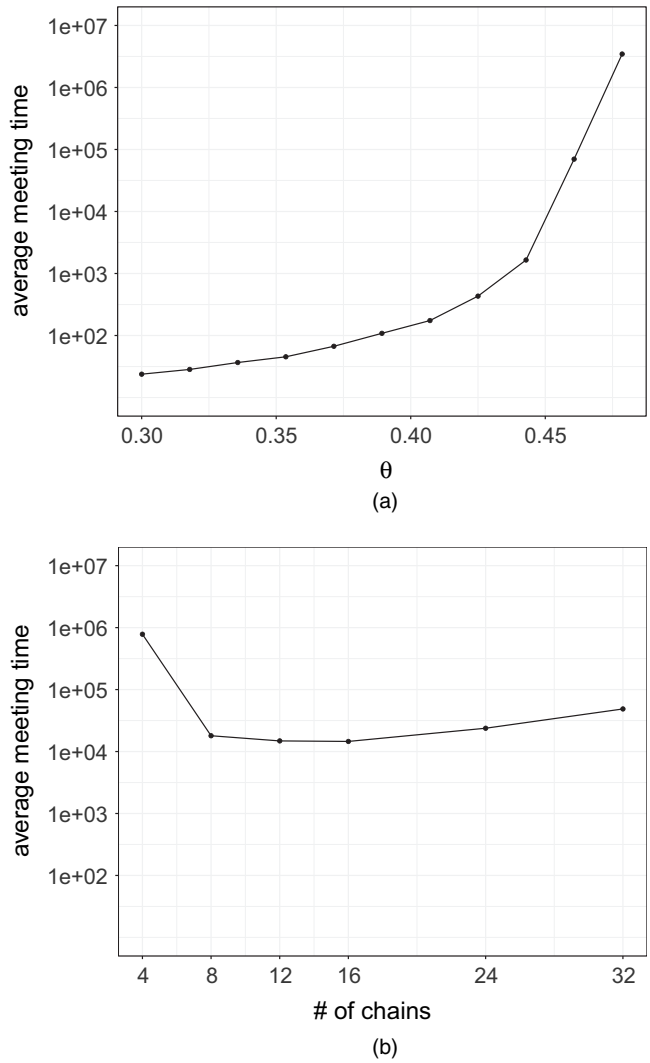

Fig. 4. For the Ising model of Section 5.3 on a 32×32 grid, average meeting times corresponding to different coupled Markov chains: (a) coupled single-site Gibbs sampler, for different inverse temperatures θ; (b) coupled replica exchange algorithm with N chains, on a grid of values $\theta^{(1)} < \ldots < \theta^{(N)}$ with $\theta^{(1)} = 0.3$ and $\theta^{(N)} = 0.55$, for various values of N

$n \in \{1, \ldots, N\}$. We use common random numbers to decide whether to perform swap moves or single-site Gibbs moves, and whether to accept the proposed states in the event of a swap move. In the event of a single-site Gibbs move, we maximally couple each conditional update.

Throughout the following experiments we use $p_{\text{swap}} = 0.01$ and introduce an equally spaced grid of θ-values from $\theta^{(1)} = 0.3$ to $\theta^{(N)} = 0.55$ for several choices of N. We note that these grids includes θ-values at which we have seen that the single-site Gibbs sampler mixes poorly. Fig. 4(b) shows the resulting average meeting times over 100 independent runs, as a function of the number of chains N. The average meeting time first decreases with the number of chains but then increases again. A possible explanation is that the mixing of the chains first improves as N increases and then stabilizes; however, it becomes more difficult for the ensembles to meet when

N increases since all chains in the ensembles must meet. The minimum average meeting time is here attained for $N = 16$ chains per ensemble.

Setting $N = 16$, $k = 10^5$ and $m = 2 \times 10^5$ we now illustrate the use of the proposed unbiased estimators on a cluster. The test function is taken as $x \mapsto t(x)$ defined above, so we estimate $\Sigma_x t(x)\pi_\theta(x)$ for various values of θ. We use 500 processors to generate unbiased estimates with a time budget of 30 min. Within that time, each processor generated between one and seven estimators, with an average of 3.7 estimators per processor and a total of 1858 estimators. The chronology of the generation of these estimates is illustrated in Fig. 5(a). For each processor, horizontal segments of different colours indicate the duration that is associated with each estimator. The final estimates with standard errors are shown in Fig. 5(b), where we can see that the standard errors are very small compared with the values of the estimates, for each value of θ. These standard errors were computed as $\hat{\sigma}_1(P,t)/\sqrt{P}$ following equation (3.4), the central limit theorem corresponding to the large processor count limit.

5.4. Variable selection

For our next example we consider a variable-selection problem following Yang *et al.* (2016) to illustrate the scaling of our proposed method on high dimensional discrete state spaces. For integers p and n, let $Y \in \mathbb{R}^n$ represent a response variable depending on covariates $X_1, \ldots, X_p \in \mathbb{R}^n$. We consider the task of inferring a binary vector $\gamma \in \{0,1\}^p$ representing which covariates to select as predictors of Y, with the convention that X_i is selected if $\gamma_i = 1$. For any γ, we write $|\gamma| = \Sigma_{i=1}^p \gamma_i$ for the number of selected covariates and X_γ for the $n \times |\gamma|$ matrix of covariates chosen by γ. Inference on γ proceeds by way of a linear regression model relating Y to X_γ, namely $Y = X_\gamma \beta_\gamma + w$ with $w \sim \mathcal{N}(0, \sigma^2 I_n)$.

We assume a prior on γ of $\pi(\gamma) \propto p^{-\kappa|\gamma|}\mathbb{1}(|\gamma| \leqslant s_0)$. This distribution puts mass only on vectors γ with fewer than s_0 1s, imposing a degree of sparsity. Given γ we assume a normal prior for the regression coefficient vector $\beta_\gamma \in \mathbb{R}^{|\gamma|}$ with zero mean and variance $g\sigma^2(X_\gamma' X_\gamma)^{-1}$. Finally, we give the precision σ^{-2} an improper prior $\pi(\sigma^{-2}) \propto 1/\sigma^{-2}$. This leads to the marginal likelihood

$$\pi(Y|X,\gamma) \propto \frac{(1+g)^{-|\gamma|/2}}{\{1+g(1-R_\gamma^2)\}^{n/2}}, \qquad R_\gamma^2 = \frac{Y' X_\gamma (X_\gamma' X_\gamma)^{-1} X_\gamma' Y}{Y' Y}.$$

To approximate the distribution $\pi(\gamma|X,Y)$, Yang *et al.* (2016) employed an MCMC algorithm whose kernel P is a mixture of two Metropolis kernels. The first component $P_1(\gamma,\cdot)$ selects a co-ordinate $i \in \{1,\ldots,p\}$ uniformly at random and flips γ_i to $1-\gamma_i$. The resulting vector γ^* is then accepted with probability $1 \wedge \pi(\gamma^*|X,Y)/\pi(\gamma|X,Y)$, where $a \wedge b$ denotes $\min(a,b)$ for $a, b \in \mathbb{R}$. Sampling a vector γ' from the second kernel $P_2(\gamma,\cdot)$ proceeds as follows. If $|\gamma|$ equals 0 or p, then γ' is set to γ. Otherwise, co-ordinates i_0 and i_1 are drawn uniformly among $\{j : \gamma_j = 0\}$ and $\{j : \gamma_j = 1\}$ respectively. The proposal γ^* has $\gamma^*_{i_0} = \gamma_{i_1}$, $\gamma^*_{i_1} = \gamma_{i_0}$ and $\gamma^*_j = \gamma_j$ for the other components. Then γ' is set to γ^* with probability $1 \wedge \pi(\gamma^*|X,Y)/\pi(\gamma|X,Y)$, and to γ otherwise. The MCMC kernel $P(\gamma,\cdot)$ targets $\pi(\gamma|X,Y)$ by sampling from $P_1(\gamma,\cdot)$ or from $P_2(\gamma,\cdot)$ with equal probability. Note that each MCMC iteration can only benefit from parallel processors to a limited extent, since $|\gamma|$ is always less than s_0, itself chosen to be a small value; thus the calculation of R_γ^2 involves only linear algebra of small matrices.

We consider the following strategy to couple the above MCMC algorithm. To sample a pair of states $(\gamma', \tilde{\gamma}')$ given $(\gamma, \tilde{\gamma})$, we first use a common uniform random variable to decide whether to sample from a coupling $\bar{P}_1$ of P_1 to itself or a coupling $\bar{P}_2$ of P_2 to itself. The coupled kernel $\bar{P}_1\{(\gamma,\tilde{\gamma}),\cdot\}$ proposes flipping the same co-ordinate for both vectors γ and $\tilde{\gamma}$ and then uses a common uniform random variable in the acceptance step. For the coupled kernel $\bar{P}_2\{(\gamma,\tilde{\gamma}),\cdot\}$,

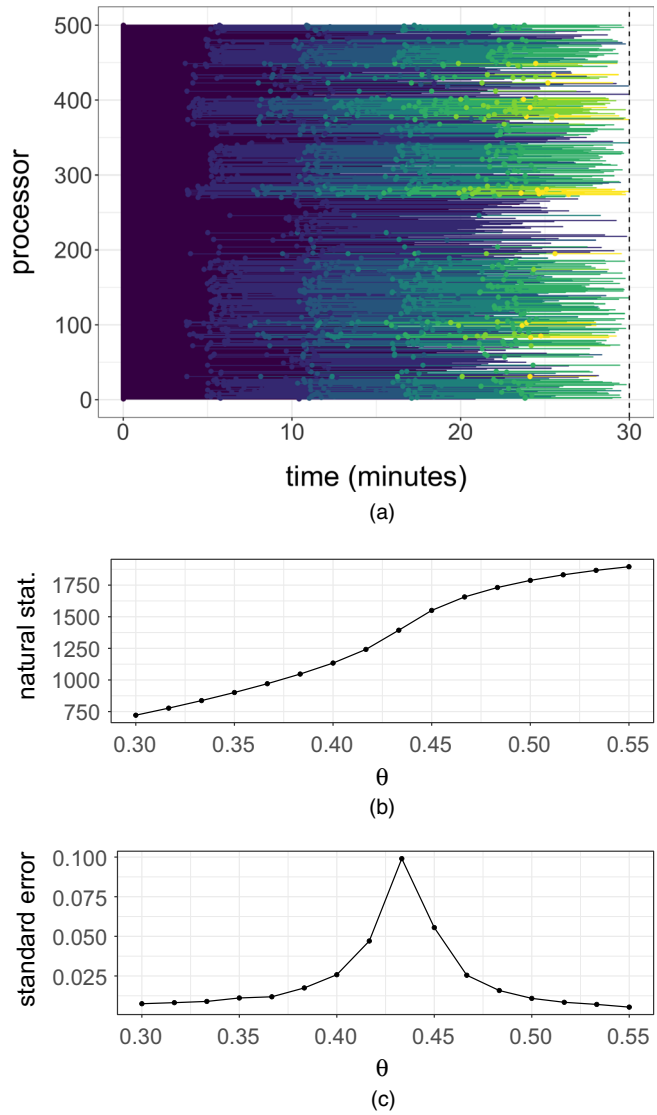

Fig. 5. For the Ising model of Section 5.3 on a 32 × 32 grid, (a) chronology of the generation of unbiased estimators on 500 processors over 30 min, and (b) estimates of the expected natural statistic $\Sigma_x t(x)\pi_\theta(x)$, and (c) standard errors, for a grid of 16 values $0.3 = \theta^{(1)} < \ldots < \theta^{(N)} = 0.55$: results obtained by coupling a replica exchange algorithm with 16 chains

we need to select two pairs of indices: $(i_0, \tilde{i}_0)$ and $(i_1, \tilde{i}_1)$. We obtain the first pair by sampling from a maximal coupling of the discrete uniform distributions on $\{j : \gamma_j = 0\}$ and $\{j : \tilde{\gamma}_j = 0\}$. This yields indices $(i_0, \tilde{i}_0)$ with the greatest possible probability that $i_0 = \tilde{i}_0$. We use the same approach to sample a pair $(i_1, \tilde{i}_1)$ to maximize the probability that $i_1 = \tilde{i}_1$. Finally we use a common uniform variable to accept or reject the proposals. If either vector γ or $\tilde{\gamma}$ has no 0s or no 1s, then it is kept unchanged.

We recall that one can sample from a maximal coupling of two discrete probability distributions $q = (q_1, \ldots, q_N)$ and $\tilde{q} = (\tilde{q}_1, \ldots, \tilde{q}_N)$ as follows. First, let $c = (c_1, \ldots, c_N)$ be the distribution with probabilities $c_n = (q_n \wedge \tilde{q}_n)/\alpha$ for $\alpha = \Sigma_{n=1}^N q_n \wedge \tilde{q}_n$ and define residual distributions q' and

$\tilde{q}'$ with probabilities $q'_n = (q_n - \alpha c_n)/(1-\alpha)$ and $\tilde{q}'_n = (\tilde{q}_n - \alpha c_n)/(1-\alpha)$. Then, with probability α, draw $i \sim c$ and output (i, i). Otherwise draw $i \sim q'$ and $\tilde{i} \sim \tilde{q}'$ and output $(i, \tilde{i})$. The resulting pair follows a maximal coupling of q and $\tilde{q}$, since $\mathbb{P}(i = \tilde{i}) = \alpha = 1 - d_{\mathrm{TV}}(q, \tilde{q})$, and marginally $\mathbb{P}(i = n) = \alpha c_n + (1-\alpha) q'_n = q_n$, and likewise for $\mathbb{P}(\tilde{i} = n)$, for all $n \in \{1, \ldots, N\}$. The procedure involves $\mathcal{O}(N)$ operations for N the size of the state space.

We now consider an experiment like those of Yang *et al.* (2016). We define

$$\beta^* = \mathrm{SNR} \sqrt{\left\{ \sigma_0^2 \frac{\log(p)}{n} \right\}} (2, -3, 2, 2, -3, 3, -2, 3, -2, 3, 0, \ldots, 0)' \in \mathbb{R}^p,$$

and generate Y given X and β^* from the model with $\sigma^2 = 1$, $\sigma_0^2 = 1$, $n \in \{500, 1000\}$, $p \in \{1000, 5000\}$ and signal-to-noise parameter $\mathrm{SNR} \in \{0.5, 1, 2\}$. We also set $s_0 = 100$, $g = p^3$ and $\kappa = 2$ (exactly as in Yang *et al.* (2016); the value of κ was obtained by personal communication) and generate the covariates X by using a multivariate normal distribution with covariance matrix Σ either equal to a unit diagonal matrix or with entries $\Sigma_{ij} = \exp(-|i-j|)$. We refer to these two cases as the independent design and correlated design cases respectively. We draw from the initial distribution π_0 by creating a vector of p 0s, sampling s_0 co-ordinates uniformly from $\{1, \ldots, p\}$ without replacement, and setting the corresponding entries to 1 with probability 0.5.

For various values of n, p and SNR, and the two types of design, we run coupled chains 100 times independently until they meet. We report the average meeting times in Tables 4 and 5. The average meeting times are of the order of 10^4–10^5, depending on the problem; the maximum is attained in the correlated design at $n = 500$, $p = 1000$ and $\mathrm{SNR} = 2$. In contrast with this, the experiments in Yang *et al.* (2016) identify the scenario $n = 500$, $p = 5000$ and $\mathrm{SNR} = 1$ as the most challenging. This discrepancy deserves further study; it could be due to variations from a synthetic data set to another, or to differences in the criteria being reported.

To illustrate the effect of dimension, we focus on the independent design setting with $n = 500$ and $\mathrm{SNR} = 1$, and we consider values of p between 100 and 1000. For each value of p, we run

Table 4. Average meeting times in the independent design

n	*p*	*Results for SNR = 0.5*	*Results for SNR = 1*	*Results for SNR = 2*
500	1000	4937	7586	6031
500	5000	24634	25602	38083
1000	1000	4729	5893	4892
1000	5000	23407	46398	24712

Table 5. Average meeting times in the correlated design

n	*p*	*Results for SNR = 0.5*	*Results for SNR = 1*	*Results for SNR = 2*
500	1000	5536	5485	216996
500	5000	27535	28756	29083
1000	1000	4921	5451	5613
1000	5000	24101	29215	23043

coupled chains 1000 times independently until they meet. We present violin plots representing the distributions of meeting times divided by p in Fig. 6(a). The distribution of scaled meeting times appears to be approximately constant as a function of p, suggesting that meeting times increase linearly in p. This is consistent with the findings of Yang *et al.* (2016), where mixing times are shown to increase linearly in p.

Focusing now on the independent design case with $n = 500$, $p = 1000$ and $\mathrm{SNR} = 1$ we consider various values of the prior hyperparameter κ in $\{0.1, 1, 2\}$. We set $k = 75000$ and $m = 150000$ and generate unbiased estimators on a cluster for 120 min, using 200 processors for each value of κ, and so 600 processors in total. The test function is chosen so that the estimand $\pi(h)$ is the vector of inclusion probabilities $\mathbb{P}(\gamma_i = 1|X, Y)$ for $i \in \{1, \ldots, 20\}$. Within the time budget, 39282 estimates were produced, with each processor producing between eight and 181 of these. The largest observed meeting time was 81423. The meeting times were similar across the three values of κ.

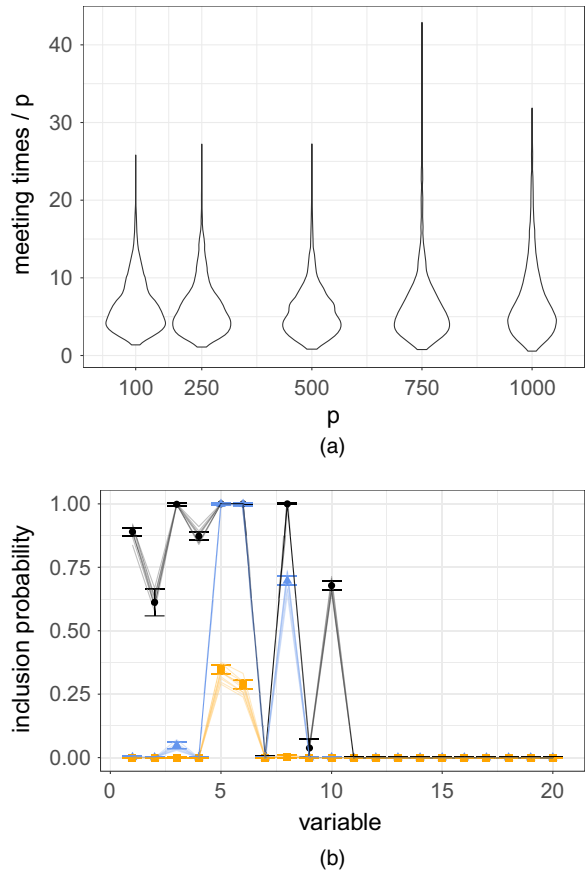

Fig. 6. (a) Meeting times divided by p for $p \in \{100, 250, 500, 750, 1000\}$ and $n = 500$, SNR = 1, in the variable-selection example of Section 5.4 with independent design, based on $R = 1000$ independent repeats (the violins represent the distributions of scaled meeting times for different p) and (b) posterior probabilities of inclusion for the first 20 variables, in the setting $n = 500$, $p = 1000$, SNR = 1, for three values of the prior hyperparameter κ (the error bars representing 95% confidence intervals were obtained after 120 min of calculation on 600 processors, using $k = 75000$ and $m = 150000$) (———, ———, ———, standard MCMC estimates based on 10 independent chains of length 10^6; ●, $\kappa = 0.1$; ▲, $\kappa = 1$; ■, $\kappa = 2$)

Fig. 6(b) shows the results in the form of 95% confidence intervals shown as error bars, using expression (3.4), the central limit theorem relevant when the time budget is fixed and the number of processors grows large. We observe that κ has a strong effect on the probability of including the first 10 variables in this setting, and that the most satisfactory results are obtained for $\kappa = 0.1$ rather than for $\kappa = 2$, recalling that β^* has non-zero entries in its first 10 components. Note that the error bars are narrow but still noticeable, particularly for $\kappa = 0.1$. On Fig. 6(b), the full lines represent estimates that were obtained with 10 independent MCMC runs with 10^6 iterations each, discarding the first 10^5 iterations as burn-in. These MCMC estimates present noticeable variability in spite of the large number of iterations. In a standard MCMC setting, we might run chains for more iterations until the estimates agree across independent runs. In the framework proposed, we increase the precision by generating more independent unbiased estimators without necessarily modifying k or m.

Fig. 6(b) suggests that the variable selection procedure that is considered here is sensitive to the prior hyperparameter κ; we refer to Yang *et al.* (2016), and to Johnson (2013) and Nikooienejad *et al.* (2016) for related discussions on Bayesian variable selection in high dimension and convergence of MCMC algorithms.

5.5. Cut distribution

Finally, our proposed estimator can be used to approximate the cut distribution, which poses a significant challenge for existing MCMC methods (Plummer, 2014; Jacob *et al.*, 2017). This illustrates another appeal of the unbiasedness property, beyond the motivation for parallel computation.

Consider two models: one with parameters θ_1 and data Y_1 and another with parameters θ_2 and data Y_2, where the likelihood of Y_2 might depend on both θ_1 and θ_2. For instance the first model could be a regression with data Y_1 and coefficients θ_1, and the second model could be another regression whose covariates are the residuals, coefficients or fitted values of the first regression (Pagan, 1984; Murphy and Topel, 2002). In principle one could introduce an encompassing model and conduct joint inference on θ_1 and θ_2 via the posterior distribution. In that case, misspecification of either model would lead to misspecification of the ensemble and thus to a misleading quantification of uncertainty, as noted in several studies (e.g. Liu *et al.* (2009), Plummer (2014), Lunn *et al.* (2009), McCandless *et al.* (2010), Zigler (2016) and Blangiardo *et al.* (2011)).

The cut distribution (Spiegelhalter *et al.*, 2003; Plummer, 2014) allows the propagation of uncertainty about θ_1 to inference on θ_2 while preventing misspecification in the second model from affecting estimation in the first. The cut distribution is defined as

$$\pi_{\text{cut}}(\theta_1, \theta_2) = \pi_1(\theta_1)\pi_2(\theta_2|\theta_1).$$

Here $\pi_1(\theta_1)$ refers to the distribution of θ_1 given Y_1 in the first model alone, and $\pi_2(\theta_2|\theta_1)$ refers to the distribution of θ_2 given Y_2 and θ_1 in the second model. Often, the density $\pi_2(\theta_2|\theta_1)$ can be evaluated up to only a constant in θ_2, which may vary with θ_1. This makes the cut distribution difficult to approximate with MCMC algorithms (Plummer, 2014).

A naive approach consists of first running an MCMC algorithm targeting $\pi_1(\theta_1)$ to obtain a sample $(\theta_1^n)_{n=1}^{N_1}$, perhaps after discarding a burn-in period and thinning the chain. Then, for each θ_1^n, one can run an MCMC algorithm targeting $\pi_2(\theta_2|\theta_1^n)$, yielding N_2 samples. One might again discard some burn-in and thin the chains, or just keep the final state of each chain. The resulting joint samples approximate the cut distribution. However, the validity of this approach relies on a double limit in N_1 and N_2. Diagnosing convergence may also be difficult given the number of chains in the second stage, each of which targets a different distribution $\pi_2(\theta_2|\theta_1^n)$.

If we could sample $\theta_1 \sim \pi_1$ and $\theta_2|\theta_1 \sim \pi_2(\theta_2|\theta_1)$, then the pair (θ_1, θ_2) would follow the cut distribution. The same two-stage rationale can be applied in the framework proposed. Consider a test function $(\theta_1, \theta_2) \mapsto h(\theta_1, \theta_2)$. Writing $\mathbb{E}_{\text{cut}}$ for expectations with respect to π_{cut}, the law of iterated expectations yields

$$\mathbb{E}_{\text{cut}}[h(\theta_1, \theta_2)] = \int \left\{ \int h(\theta_1, \theta_2)\pi_2(\mathrm{d}\theta_2|\theta_1) \right\} \pi_1(\mathrm{d}\theta_1) = \int \bar{h}(\theta_1)\pi_1(\mathrm{d}\theta_1).$$

Here $\bar{h}(\theta_1) = \int h(\theta_1, \theta_2)\pi_2(\mathrm{d}\theta_2|\theta_1)$. In the framework proposed, we can make an unbiased estimator of $\bar{h}(\theta_1)$ for all θ_1 and then plug these estimators into an unbiased estimator of the integral $\int h(\theta_1)\pi_1(\mathrm{d}\theta_1)$. This is perhaps clearer by using the signed measure representation of Section 2.4: one can obtain a signed measure $\hat{\pi}_1 = \Sigma_{l=1}^{N} \omega_l \delta_{\theta_{1,l}}$ approximating π_1, and then obtain an unbiased estimator of $\bar{h}(\theta_{1,l})$ for all l, denoted by $\bar{H}_l$. Then the weighted average $\Sigma_{l=1}^{N} \omega_l \bar{H}_l$ is an unbiased estimator of $\mathbb{E}_{\text{cut}}[h(\theta_1, \theta_2)]$ by the law of iterated expectations. Such estimators can be generated independently in parallel, and their average provides a consistent approximation of an expectation with respect to the cut distribution.

We consider the example that was described in Plummer (2014), inspired by an investigation of the international correlation between human papilloma virus (HPV) prevalence and cervical cancer incidence (Maucort-Boulch *et al.*, 2008). The first module concerns HPV prevalence, with data independently collected in 13 countries. The parameter $\theta_1 = (\theta_{1,1}, \ldots, \theta_{1,13})$ receives a beta(1,1) prior distribution independently for each component. The data $(Y_1, \ldots, Y_{13})$ consist of 13 pairs of integers. The first represents the number of women who were infected with high-risk HPV, and the second represents population sizes. The likelihood specifies a binomial model for Y_i, independently for each component i. The posterior for this model is given by a product of beta distributions.

The second module concerns the relationship between HPV prevalence and cancer incidence, and posits a Poisson regression. The parameters are $\theta_2 = (\theta_{2,1}, \theta_{2,2}) \in \mathbb{R}^2$ and receive a normal prior with zero mean and variance 10^3 per component. The likelihood in this module is given by

$$Z_{1,i} \sim \text{Poisson}\{\exp(\theta_{2,1} + \theta_{1,i}\theta_{2,2} + Z_{2,i})\} \qquad \text{for } i \in \{1, \ldots, 13\},$$

where the data $(Z_{1,i}, Z_{2,i})_{i=1}^{13}$ are pairs of integers. The first component represents numbers of cancer cases, whereas the second is the number of woman-years of follow-up. The Poisson regression model might be misspecified, motivating departures from inference based on the joint model (Plummer, 2014).

Here we can draw directly from the first posterior, denoted by $\pi_1(\theta_1)$, and obtain a sample $(\theta_1^n)_{n=1}^{N}$. For each θ_1^n we consider an MH algorithm targeting $\pi_2(\theta_2|\theta_1^n)$, using a normal random-walk proposal with variance Σ. We couple this algorithm by using reflection maximal couplings of the proposals as in Section 4.1. In preliminary runs, starting with a standard bivariate normal distribution as an initial distribution and a proposal covariance matrix set to identity, we estimate the first two moments of the cut distribution, and we use them to refine the initial distribution π_0 and the proposal covariance matrix Σ. With these settings we obtain the distribution of meeting times that is shown in Fig. 7(a). We then set $k = 100$ and $m = 10k$, and obtain approximations of the cut distribution represented by histograms in Figs 7(b) and 7(c), using $N = 10000$ unbiased estimators. The overlaid curves correspond to a kernel density estimate obtained by running $m = 1000$ steps of MCMC targeting $\pi_2(\theta_2|\theta_1^n)$ with θ_1^n drawn from $\pi_1(\theta_1)$, for $n \in \{1, \ldots, N\}$, and keeping the final mth state of each chain. The proposed estimators can be refined by increasing the number N of independent replications, whereas the MCMC estimators would converge only in the double limit of N and m going to ∞.

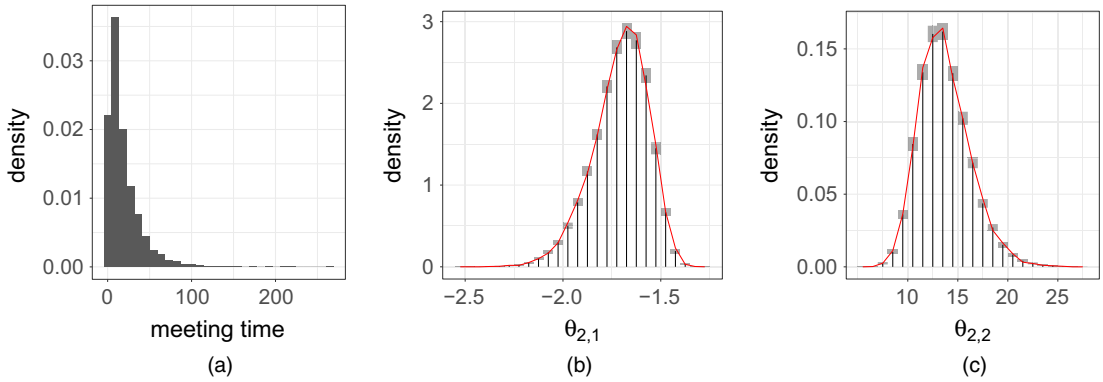

Fig. 7. (a) Meeting times and histograms of (b) $\theta_{2,1}$ and (c) $\theta_{2,2}$, in the example of Section 5.5, computed from 10000 unbiased estimators, with $k = 100$ and $m = 1000$ (▒, 95% confidence intervals on the histogram estimates; ———, marginal probability density functions of the cut distribution, obtained by running $N = 10000$ MCMC chains for 1000 steps, each targeting $\pi_2(\theta_2|\theta_1^n)$ for θ_1^n drawn from $\pi_1(\theta_1)$, and retaining the last state only)

6. Discussion

By combining the powerful technique of Glynn and Rhee (2014) with couplings of MCMC algorithms, unbiased estimators of integrals with respect to the target distribution can be constructed. Their efficiency can be controlled with tuning parameters k and m, for which we have proposed guidelines: k can be chosen as a large quantile of the meeting time τ, and m as a multiple of k. Improving on these simple guidelines stands as a subject for future research. In numerical experiments we have argued that the estimators proposed yield a practical way of parallelizing MCMC computations in a range of settings. We stress that coupling pairs of Markov chains does not improve their marginal mixing properties, and that poor mixing of the underlying chains can lead to poor performance of the resulting estimator. The choice of initial distribution π_0 can have undesirable effects on the estimators, as in the multimodal example of Section 5.1. Unreliable estimators would also result from stopping the chains before their meeting time.

Couplings of MCMC algorithms can be devised by using maximal couplings reflection couplings and common random numbers. We have focused on couplings that can be implemented without further analytical knowledge about the target distribution or about the MCMC kernels. However, these couplings might result in prohibitively large meeting times, either because the marginal chains mix slowly, as in Section 5.1, or because the coupling strategy is ineffective, as in Section 4.2.

Regarding convergence diagnostics, the framework proposed yields the following representation for the total variation between π_k and π, where π_k denotes the marginal distribution of X_k:

$$d_{\mathrm{TV}}(\pi_k, \pi) = \frac{1}{2} \sup_{h:\,|h|\leqslant 1} |\mathbb{E}[h(X_k)] - \mathbb{E}_\pi[h(X)]|$$
$$= \frac{1}{2} \sup_{h:\,|h|\leqslant 1} \left| \mathbb{E}\left[\sum_{t=k+1}^{\tau-1} \{h(X_t) - h(Y_{t-1})\} \right] \right|,$$

Here the supremum ranges over all bounded measurable functions under the assumptions stated. The equality above has several consequences. For instance, the triangle inequality implies that $d_{\mathrm{TV}}(\pi_k, \pi) \leqslant \min(1, \mathbb{E}[\max\{0, (\tau - k - 1)\}])$, and we can approximate $\mathbb{E}[\max\{0, (\tau - k - 1)\}]$ by Monte Carlo sampling for a range of k-values. This is pursued in Biswas *et al.* (2019), where

the construction proposed is extended to allow for arbitrary time lags between the coupled chains.

Thanks to its potential for parallelization, the framework proposed can facilitate a consideration of MCMC kernels that might be too expensive for serial implementation. For instance, one can improve MH-within-Gibbs samplers by performing more MH steps per component update, Hamiltonian Monte Carlo sampling by using smaller step sizes in the numerical integrator (Heng and Jacob, 2019) and particle MCMC sampling by using more particles in the particle filters (Andrieu *et al.*, 2010; Jacob *et al.*, 2019). We expect the optimal tuning of MCMC kernels to be different in the proposed framework from when used marginally.

On top of enabling the application of the results of Glynn and Heidelberger (1991) to accommodate budget constraints, the lack of bias of the estimators proposed can be beneficial in combination with the law of total expectation, to implement modular inference procedures as in Section 5.5. In Rischard *et al.* (2018) the lack of bias was exploited in new estimators of Bayesian cross-validation criteria. In Chen *et al.* (2018) similar unbiased estimators were used in the expectation step of an expectation–maximization algorithm. There may be other settings where the lack of bias is appealing, for instance in gradient estimation for stochastic gradient descents (Tadić and Doucet, 2017).

Acknowledgements

The authors are grateful to Jeremy Heng and Luc Vincent-Genod for useful discussions. The authors gratefully acknowledge support by the National Science Foundation through grants DMS-1712872 (Pierre E. Jacob) and DMS-1513040 (Yves F. Atchadé).

References

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. and Ronquist, F. (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, **20**, 407–415.

Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc.* B, **72**, 269–342.

Andrieu, C. and Vihola, M. (2015) Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.*, **25**, 1030–1077.

Atchade, Y. F. (2006) An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodol. Comput. Appl. Probab.*, **8**, 235–254.

Atchadé, Y. F. (2016) Markov chain Monte Carlo confidence intervals. *Bernoulli*, **22**, 1808–1838.

Biswas, N., Jacob, P. E. and Vanetti, P. (2019) Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems*, pp. 7389–7399.

Blangiardo, M., Hansell, A. and Richardson, S. (2011) A Bayesian model of time activity data to investigate health effect of air pollution in time series studies. *Atmosph. Environ.*, **45**, 379–386.

Bou-Rabee, N., Eberle, A. and Zimmer, R. (2018) Coupling and convergence for Hamiltonian Monte Carlo. *Preprint arXiv:1805.00452*. Department of Mathematical Sciences, Rutgers University, Camden.

Brockwell, A. E. (2006) Parallel Markov chain Monte Carlo simulation by pre-fetching. *J. Computnl Graph. Statist.*, **15**, 246–261.

Brockwell, A. E. and Kadane, J. B. (2005) Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *J. Computnl Graph. Statist.*, **14**, 436–458.

Brooks, S. P., Gelman, A., Jones, G. and Meng, X.-L. (eds) (2011) *Handbook of Markov Chain Monte Carlo*. Boca Raton: CRC Press.

Calderhead, B. (2014) A general construction for parallelizing Metropolis–Hastings algorithms. *Proc. Natn. Acad. Sci. USA*, **111**, 17408–17413.

Casella, G., Lavine, M. and Robert, C. P. (2001) Explaining the perfect sampler. *Am. Statistn*, **55**, 299–305.

Chatterjee, S., Guntuboyina, A. and Sen, B. (2015) On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, **43**, 1774–1800.

Chen, W., Ma, L. and Liang, X. (2018) Blind identification based on expectation-maximization algorithm coupled with blocked Rhee–Glynn smoothing estimator. *IEEE Communs Lett.*, **22**, 1838–1841.

Choi, H. M. and Hobert, J. P. (2013) The Pólya–Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electron. J. Statist.*, **7**, 2054–2064.

Cowles, M. K. and Rosenthal, J. S. (1998) A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. *Statist. Comput.*, **8**, 115–124.
Diaconis, P. and Freedman, D. (1999) Iterated random functions. *SIAM Rev.*, **41**, 45–76.
Douc, R., Moulines, E. and Rosenthal, J. S. (2004) Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Ann. Appl. Probab.*, **14**, 1643–1665.
Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987) Hybrid Monte Carlo. *Phys. Lett.* B, **195**, 216–222.
Flegal, J. M., Haran, M. and Jones, G. L. (2008) Markov chain Monte Carlo: can we trust the third significant figure? *Statist. Sci.*, **23**, 250–260.
Flegal, J. M. and Herbei, R. (2012) Exact sampling for intractable probability distributions via a Bernoulli factory. *Electron. J. Statist.*, **6**, 10–37.
Gaver, D. P. and O'Muircheartaigh, I. G. (1987) Robust empirical Bayes analyses of event rates. *Technometrics*, **29**, 1–15.
Geyer, C. J. (1991) Markov chain Monte Carlo maximum likelihood. *Technical Report*. School of Statistics, University of Minnesota, Minneapolis.
Glynn, P. W. (2016) Exact simulation versus exact estimation. In *Proc. Winter Simulation Conf,* pp. 193–205. New York: Institute of Electrical and Electronics Engineers.
Glynn, P. W. and Heidelberger, P. (1990) Bias properties of budget constraint simulations. *Ops Res.*, **38**, 801–814.
Glynn, P. W. and Heidelberger, P. (1991) Analysis of parallel replicated simulations under a completion time constraint. *ACM Trans. Modlng Comput. Simulns*, **1**, 3–23.
Glynn, P. W. and Rhee, C.-H. (2014) Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.* A, **51**, 377–389.
Glynn, P. W. and Whitt, W. (1992) The asymptotic efficiency of simulation estimators. *Ops Res.*, **40**, 505–520.
Gong, L. and Flegal, J. M. (2016) A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *J. Computnl Graph. Statist.*, **25**, 684–700.
Goodman, J. and Weare, J. (2010) Ensemble samplers with affine invariance. *Communs Appl. Math. Computnl Sci.*, **5**, 65–80.
Goudie, R. J. B., Turner, R. M., De Angelis, D. and Thomas, A. (2017) MultiBUGS: a parallel implementation of the BUGS modelling framework for faster Bayesian inference. *J. Statist. Softwr.*, to be published.
Green, P. J., Łatuszyński, K., Pereyra, M. and Robert, C. P. (2015) Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statist. Comput.*, **25**, 835–862.
Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
Heng, J. and Jacob, P. E. (2019) Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, **106**, 287–302.
Hoffman, M. D. and Gelman, A. (2014) The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, **15**, 1593–1623.
Huber, M. (2016) *Perfect Simulation*. Boca Raton: CRC Press.
Jacob, P. E., Lindsten, F. and Schön, T. B. (2019) Smoothing with couplings of conditional particle filters. *J. Am. Statist. Ass.*, to be published, doi 10.1080/01621459.2018.1548856.
Jacob, P. E., Murray, L. M., Holmes, C. C. and Robert, C. P. (2017) Better together?: Statistical learning in models made of modules. *Preprint arXiv:1708.08719*. Department of Statistics, Harvard University, Cambridge.
Jacob, P. E., Robert, C. P. and Smith, M. H. (2011) Using parallel computation to improve independent Metropolis–Hastings based estimation. *J. Computnl Graph. Statist.*, **20**, 616–635.
Jarner, S. F. and Hansen, E. (2000) Geometric ergodicity of Metropolis algorithms. *Stoch. Processes Appl.*, **85**, 341–361.
Johndrow, J. E. and Mattingly, J. C. (2017) Coupling and decoupling to bound an approximating Markov chain. *Preprint arXiv:1706.02040*. Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia.
Johnson, V. E. (1996) Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *J. Am. Statist. Ass.*, **91**, 154–166.
Johnson, V. E. (1998) A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *J. Am. Statist. Ass.*, **93**, 238–248.
Johnson, V. E. (2013) On numerical aspects of Bayesian model selection in high and ultrahigh-dimensional settings. *Baysn Anal.*, **8**, 741–758.
Khare, K. and Hobert, J. P. (2013) Geometric ergodicity of the Bayesian lasso. *Electron. J. Statist.*, **7**, 2150–2163.
Lee, A., Doucet, A. and Łatuszyński, K. (2014) Perfect simulation using atomic regeneration with application to sequential Monte Carlo. *Preprint arXiv:1407.5770*. University of Bristol, Bristol.
Lee, A., Yau, C., Giles, M. B., Doucet, A. and Holmes, C. C. (2010) On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *J. Computnl Graph. Statist.*, **19**, 769–789.
Lindvall, T. (2002) *Lectures on the Coupling Method*. New York: Dover Publications.
Liu, F., Bayarri, M. and Berger, J. (2009) Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Baysn Anal.*, **4**, 119–150.

Liu, J. S. (2008) *Monte Carlo Strategies in Scientific Computing*. New York: Springer Science and Business Media.
Lunn, D., Best, N., Spiegelhalter, D., Graham, G. and Neuenschwander, B. (2009) Combining MCMC with sequential PKPD modelling. *J. Pharmkinet. Pharmdynam.*, **36**, 19–38.
Mainini, E. (2012) A description of transport cost for signed measures. *J. Math. Sci.*, **181**, 837–855.
Mangoubi, O. and Smith, A. (2017) Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *Preprint arXiv:1708.07114*. Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester.
Maucort-Boulch, D., Franceschi, S. and Plummer, M. (2008) International correlation between human papillomavirus prevalence and cervical cancer incidence. *Cancer Epidem. Biomark. Prevn*, **17**, 717–720.
McCandless, L. C., Douglas, I. J., Evans, S. J. and Smeeth, L. (2010) Cutting feedback in Bayesian regression adjustment for the propensity score. *Int. J. Biostatist.*, **6**, no. 2, 1–22.
McLeish, D. (2011) A general method for debiasing a Monte Carlo estimator. *Monte Carlo Meth. Appl.*, **17**, 301–315.
Meyn, S. and Tweedie, R. (2009) *Markov Chains and Stochastic Stability*, 2nd edn. Cambridge: Cambridge University Press.
Middleton, L., Deligiannidis, G., Doucet, A. and Jacob, P. E. (2018) Unbiased Markov chain Monte Carlo for intractable target distributions. *Preprint arXiv:1807.08691*. Department of Statistics, University of Oxford, Oxford.
Middleton, L., Deligiannidis, G., Doucet, A. and Jacob, P. E. (2019) Unbiased smoothing using particle independent Metropolis-Hastings. *Proc. Mach. Learn. Res.*, **89**, 2378–2387.
Murdoch, D. J. and Green, P. J. (1998) Exact sampling from a continuous state space. *Scand. J. Statist.*, **25**, 483–502.
Murphy, K. M. and Topel, R. H. (2002) Estimation and inference in two-step econometric models. *J. Bus. Econ. Statist.*, **20**, 88–97.
Murray, I., Adams, R. P. and MacKay, D. J. C. (2010) Elliptical slice sampling. In *Proc. 13th Int. Conf. Artificial Intelligence and Statistics*, *Chia Laguna*, pp. 541–548.
Mykland, P., Tierney, L. and Yu, B. (1995) Regeneration in Markov chain samplers. *J. Am. Statist. Ass.*, **90**, 233–241.
Neal, R. M. (1993) Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems* (eds S. J. Hanson, J. D. Cowan and C. L. Giles), pp. 475–482.
Neal, R. M. (1999) Circularly-coupled Markov chain sampling. *Preprint arXiv:1711.04399*. Department of Statistics, University of Toronto, Toronto.
Neal, R. M. (2011) MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, vol. 2 (eds S. P. Brooks, A. Gelman, G. Jones and X.-L. Meng). Boca Raton: CRC Press.
Neal, R. M. and Pinto, R. L. (2001) Improving Markov chain Monte Carlo estimators by coupling to an approximating chain. *Technical Report*. Department of Statistics, University of Toronto, Toronto.
Nikooienejad, A., Wang, W. and Johnson, V. E. (2016) Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics*, **32**, 1338–1345.
Nummelin, E. (2002) MC's for MCMC'ists. *Int. Statist. Rev.*, **70**, 215–240.
Owen, A. B. (2017) Statistically efficient thinning of a Markov chain sampler. *J. Computnl Graph. Statist.*, **26**, 738–744.
Pagan, A. (1984) Econometric issues in the analysis of regressions with generated regressors. *Int. Econ. Rev.*, **25**, 221–247.
Pal, S. and Khare, K. (2014) Geometric ergodicity for Bayesian shrinkage models. *Electron. J. Statist.*, **8**, 604–645.
Plummer, M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proc. 3rd Int. Wrkshp Distributed Statistical Computing, Vienna* (eds K. Hornik, F. Leisch and A. Zeileis).
Plummer, M. (2014) Cuts in Bayesian graphical models. *Statist. Comput.*, **25**, 37–43.
Plummer, M., Best, N., Cowles, K. and Vines, K. (2006) CODA: convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11.
Pollock, M., Fearnhead, P., Johansen, A. M. and Roberts, G. O. (2016) The scalable Langevin exact algorithm: Bayesian inference for big data. *Preprint arXiv:1609.03436*.
Propp, J. G. and Wilson, D. B. (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Rand. Struct. Algs*, **9**, 223–252.
R Core Team (2015) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Reutter, A. and Johnson, V. E. (1995) General strategies for assessing convergence of MCMC algorithms using coupled sample paths. Institute of Statistics and Decision Sciences, Duke University, Durham.
Rhee, C.-H. and Glynn, P. W. (2012) A new approach to unbiased estimation for SDE's. In *Proc. Winter Simulation Conf.*, article 17.
Rischard, M., Jacob, P. E. and Pillai, N. (2018) Unbiased estimation of log normalizing constants with applications to Bayesian cross-validation. *Preprint arXiv:1810.01382*.
Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer.

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, **7**, 110–120.
Roberts, G. O. and Rosenthal, J. S. (2004) General state space Markov chains and MCMC algorithms. *Probab. Surv.*, **1**, 20–71.
Roberts, G. O. and Tweedie, R. L. (1996a) Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, **2**, 341–363.
Roberts, G. O. and Tweedie, R. L. (1996b) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–110.
Rosenthal, J. S. (1996) Analysis of the Gibbs sampler for a model related to James-Stein estimators. *Statist. Comput.*, **6**, 269–275.
Rosenthal, J. S. (1997) Faithful couplings of Markov chains: now equals forever. *Adv. Appl. Math.*, **18**, 372–381.
Rosenthal, J. S. (2000) Parallel computing and Monte Carlo algorithms. *Far East J. Theoret. Statist.*, **4**, 207–236.
Rosenthal, J. S. (2002) Quantitative convergence rates of Markov chains: a simple account. *Electron. Communs Probab.*, **7**, 123–128.
Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003) *WinBUGS User Manual*. Cambridge: Medical Research Council Biostatistics Unit.
Srivastava, S., Cevher, V., Dinh, Q. and Dunson, D. (2015) WASP: scalable Bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics* (eds G. Lebanon and S. V. N. Vishwanathan), pp. 912–920.
Swendsen, R. H. and Wang, J.-S. (1987) Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.*, **58**, 86–88.
Tadić, V. B. and Doucet, A. (2017) Asymptotic bias of stochastic gradient search. *Ann. Appl. Probab.*, **27**, 3255–3304.
Thorisson, H. (2000) *Coupling, Stationarity, and Regeneration*. New York: Springer.
Titsias, M. K. and Yau, C. (2017) The Hamming ball sampler. *J. Am. Statist. Ass.*, **112**, 1598–1611.
Tjelmeland, H. (2004) Using all Metropolis–Hastings proposals to estimate mean values. *Technical Report*. Department of Mathematical Sciences, Norwegian University of Science and Technology.
Tweedie, R. (1983) The existence of moments for stationary Markov chains. *J. Appl. Probab.*, **20**, 191–196.
Vanetti, P., Bouchard-Côté, A., Deligiannidis, G. and Doucet, A. (2017) Piecewise deterministic Markov chain Monte Carlo. *Preprint arXiv:1707.05296*. Department of Statistics, University of Oxford, Oxford.
Vats, D., Flegal, J. M. and Jones, G. L. (2018) Strong consistency of multivariate spectral variance estimators in Markov chain Monte Carlo. *Bernoulli*, **24**, 1860–1909.
Wang, X., Guo, F., Heller, K. A. and Dunson, D. B. (2015) Parallelizing MCMC with random partition trees. In *Advances in Neural Information Processing Systems 28* (eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett), pp. 451–459.
Wolff, U. (1989) Comparison between cluster Monte Carlo algorithms in the Ising model. *Phys. Lett.* B, **228**, 379–382.
Yang, S., Chen, Y., Bernton, E. and Liu, J. S. (2017) On parallelizable Markov chain Monte Carlo algorithms with waste-recycling. *Statist. Comput.*, **28**, 1073–1081.
Yang, Y., Wainwright, M. J. and Jordan, M. I. (2016) On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist.*, **44**, 2497–2532.
Zigler, C. M. (2016) The central role of Bayes theorem for joint estimation of causal effects and propensity scores. *Am. Statistn*, **70**, 47–54.

Discussion on the paper by Jacob, O'Leary and Atchadé

Darren J. Wilkinson (*Newcastle University, Newcastle upon Tyne*)

Jacob, O'Leary and Atchadé are to be congratulated for this interesting and useful contribution to the Markov chain Monte Carlo (MCMC) literature. Since in general we cannot initialize an MCMC chain with an exact sample from the target, we rely on asymptotic convergence to equilibrium, but lack of convergence leads to biased estimates. The method outlined in the paper provides a method for removing the bias from any estimates that are produced from the algorithm output, by using a pair of coupled chains. The approach is quite general and can potentially be applied to many different kinds of MCMC algorithms—this paper concentrates on Metropolis–Hastings and Gibbs sampling, but there are several related papers exploring applications to other MCMC schemes.

It is important to keep in mind that coalescence of the coupled chains and convergence to equilibrium are different. Practical application of the method relies on the time-averaged estimator (equation (2.1)). This is just the regular MCMC estimate with a correction term that terminates once the chains have coupled.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Unbiased Markov chain Monte Carlo with couplings'.

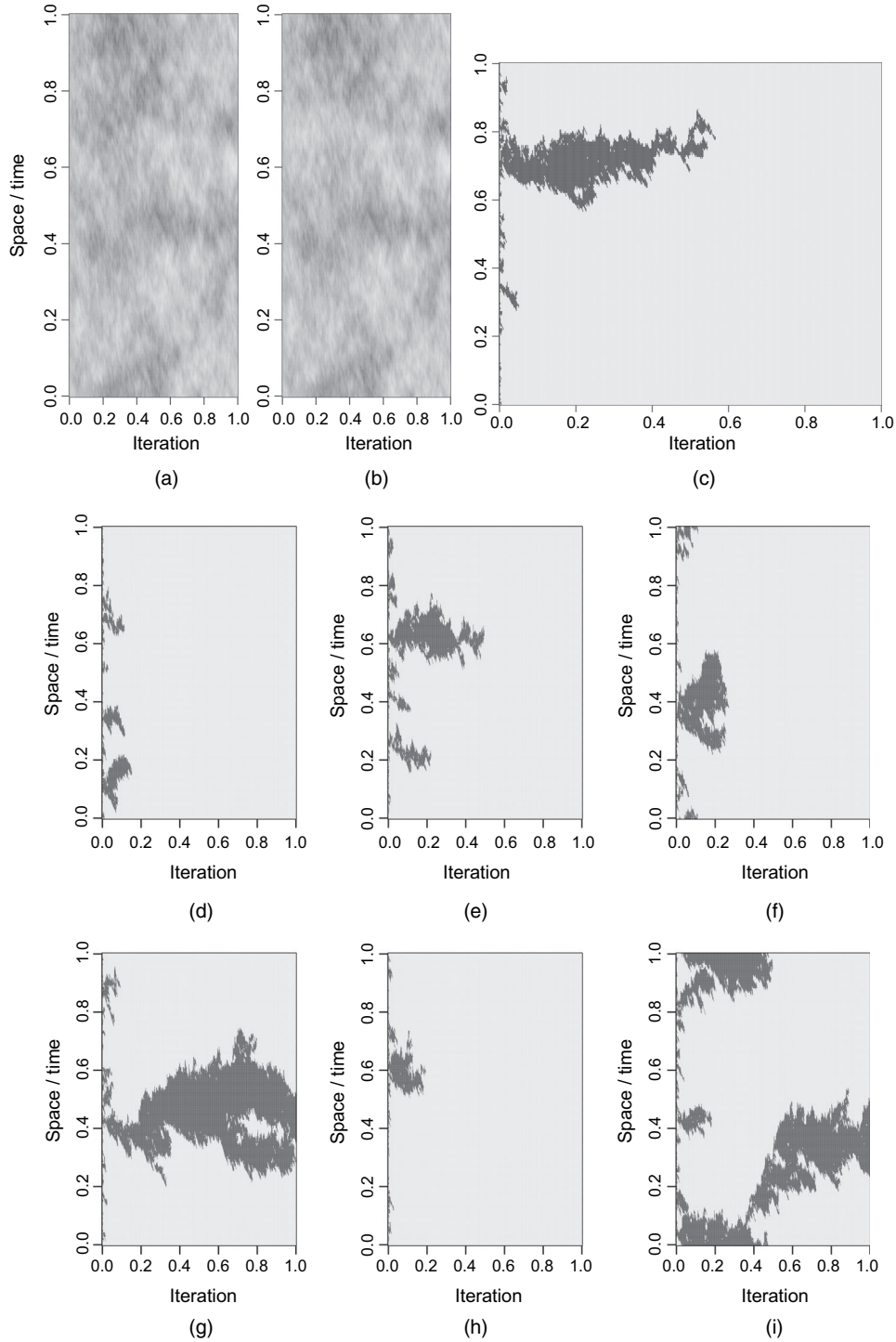

Fig. 8. Coupled AR(1) chains (a) 1 and (b) 2 and (c) a plot showing the uncoupled AR(1) variables and (d)–(i) for additional replicates 1–6: two of these six replicates failed to coalesce within 500 iterations

Choosing k (the 'burn-in') to be a high quantile of the coupling time distribution will ensure that most estimates contain no correction at all. But, since k represents 'wasted' computation, it seems desirable to encourage the chains to couple rapidly. Although it is possible to engineer early coupling (via identical initial conditions and a low acceptance rate Metropolis–Hastings kernel, for example), this might be a dangerous strategy (see the example in Section 5.1). When the correction term is present, it can sometimes be large, so developing a better understanding of the distributional properties of the correction term is likely to be useful.

The technique proposed for coupling Metropolis–Hastings kernels works for multivariate as well as univariate samplers, though reflection maximal coupling works much better than a simple independent maximal coupling in high dimensions. However, it also makes sense to want to couple Gibbs samplers and other componentwise update algorithms. The paper proposes that, for each component update of a Gibbs sampler, the full conditionals for the two chains should be coupled. It is clear that, once all components have coalesced, the full state of the chain will match, and the coupling of the two chains will be faithful. There is little theory providing reassurance that this will happen in reasonable time for challenging problems, though it does seem to work reasonably well, empirically.

Consider a linear Gaussian auto-regressive AR(1) process defined by

$$X_t = \alpha X_{t-1} + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2),$$

for $t = 1, 2, \ldots, T$, with periodic boundaries. The full conditional for each variable X_t is of the form

$$X_t | X_{t-1}, X_{t+1} \sim N\left\{ \frac{\alpha}{1+\alpha^2}(X_{t-1} + X_{t+1}), \frac{\sigma}{1+\alpha^2} \right\}.$$

We can run two chains by cycling through each variable in turn and sampling from the coupled full conditionals. Here we use $T = 200$ and $\alpha = 0.99$, and run initially for $n = 500$ iterations (Fig. 8). By studying the coupling of variables in the Gibbs sampler, we see that coupling follows a stochastic process, somewhat reminiscent of the annealing of a one-dimensional Ising model, and that it is not always rapid. We can study the coupling time distribution for this model with simple Monte Carlo sampling, to find that the mean coupling time is around 2.6×10^3, with a median of around 1×10^3. The distribution is long tailed, with a maximum observed coupling time over 1000 replicates of 67000. Choosing k to be a very high quantile of this distribution will therefore be computationally demanding.

One of the main motivations for obtaining unbiased estimates of posterior expectations is to fix one of the issues with parallel chains MCMC sampling. We would like to run many chains independently on different processors and then to pool results in some way. In this case any bias in the chains will not 'average out' across multiple processors, since there will be non-negligible bias in every chain. Unbiased estimators can be safely averaged to obtain new unbiased estimators with reduced variance. However, the debiasing does not eliminate burn-in—it just corrects for it, so there is still repeated burn-in, limiting the scalability of the parallel chains approach. The paper recommends choosing k as a high quantile of the coupling time distribution, and m a multiple of k. Any non-zero k limits speed-up of parallel coupled chains relative to one long run, via *Amdahl's law* for parallel chains MCMC sampling (Wilkinson, 2005):

$$\text{SpeedUp}(N) = \frac{b+n}{b+n/N} \overset{N}{\underset{\infty}{\to}} \frac{b+n}{b},$$

for burn-in b and monitoring run n on N processors (e.g. an asymptotic limit of 10 for $n = 9b$). This suggests that parallel chains may be useful in the context of a small number of available processors (of the order of 10), but possibly an increasingly inefficient proposition for large numbers of processors (greater than of the order of 10^2).

The method proposed in the paper is often straightforward to implement and seems to work on a broad range of samplers. It is potentially useful in the context of parallelization of MCMC algorithms, since averaging unbiased estimators is relatively safe, but it is not clear that it fundamentally solves parallel MCMC sampling, since some kind of burn-in must still be repeated for every pair of chains. Many open questions remain regarding the codevelopment of MCMC and coupling algorithms to optimize efficiency. Overall, this represents an interesting and thought-provoking contribution to the literature, and so it gives me great pleasure to propose the vote of thanks.

Chris Sherlock (*Lancaster University*)

I congratulate Jacob, O'Leary and Atchadé for this methodological development which is relatively straightforward to implement yet has the potential to broaden the use of 'embarrassingly parallel' (Dean

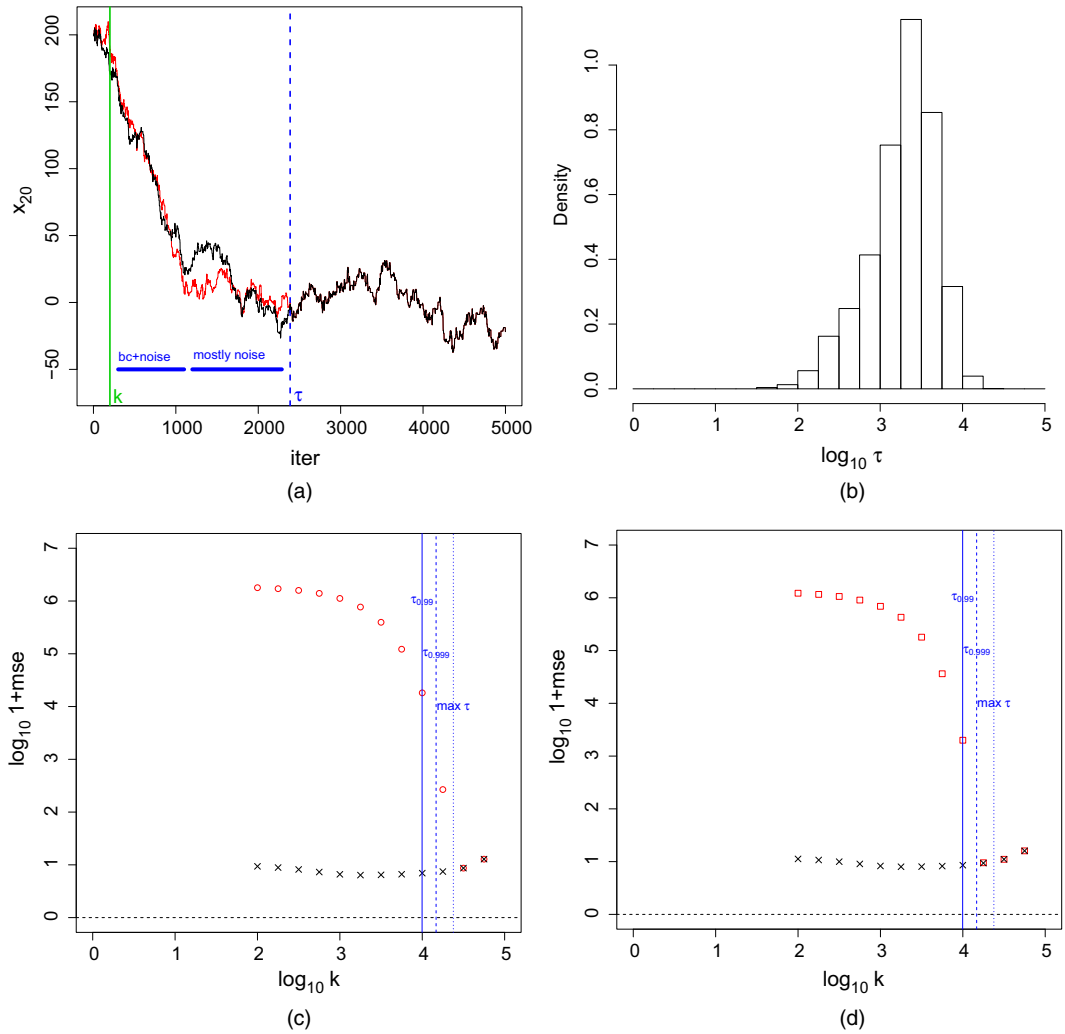

Fig. 9. Plots from the experiment described in the text: (a) first 5000 iterations of the first simulation (⋮, τ; |, putative k); (b) histogram of all 5000 values of $\log_{10}(\tau)$; (c) $\log_{10}(1 + \text{MSE})$ (mean taken over the 5000 simulations) of unbiased MCMC (○) and MCMC (×) sampling as a function of k (|, $\tau_{0.99}$; ⋮, $\tau_{0.999}$; ⋮, largest τ (not $\tau_{\max}$) value over the 5000 simulations); (d) $\log_{10}(1 + \text{MSE})$ of Rao–Blackwellized unbiased MCMC (□) and MCMC (×) as a function of k

and Ghemewat, 2008) Markov chain Monte Carlo (MCMC) sampling to a much more general class of problem. It gives me great pleasure, therefore, to second the vote of thanks.

Reverse engineering the key construction on the third page, we see that $H_k(X, Y)$ is unbiased because *on average* X_t is one step ahead of Y_{t-1}; moreover, as is clear from Fig. 9(a), most of the bias correction occurs during the relatively brief phase that in standard MCMC sampling is called 'burn-in'. Also, from Fig. 9(a), the value of the bias correction term $\Sigma_{t=k+1}^{\tau-1}\{h(X_t) - h(Y_{t-1})\}$ depends on the meanderings of the two chains until their coupling time and is potentially very noisy. This is mitigated by averaging to form $H_{k:m}$ in equation (2.1), but it still depends on the meanderings of the two chains. Now $\text{var}(\text{BC}_{k:m}) \geqslant \mathbb{E}[\text{var}(\text{BC}_{k:m}|\tau)]$, but a simple back of the envelope calculation suggests that $\text{var}(\text{BC}_{k:m}|\tau) = \mathcal{O}\{(\tau-k)^3/(m-k)^2\}\mathbf{1}_{(\tau>k+1)}$, motivating setting k to a high quantile of τ and choosing m so that, when $\tau > k + l$, the denomator dominates the numerator. This also underlines the importance of reflection maximal coupling where, for random-walk proposals, τ is approximately linear in dimension rather than

exponential. See, for example, Bladt *et al.* (2014) and references therein for the background to reflection coupling and other similar methods such as projection coupling.

A target of $N\{0, \mathrm{diag}(1^2, 2^2, \ldots, 20^2)\}$ was explored using a proposal of $X^*|x \sim N(x, \lambda^2 I_{20})$ for 10^5 iterations starting from $N\{(10, 20, \ldots, 200)^{\mathrm{T}}, I_{20}\}$, and Y^* was obtained by using the reflection maximal coupling. Fig. 9(b) shows the distribution of $\log_{10}(\tau)$ over 5000 repeat simulations. In standard MCMC sampling, the burn-in is decided from the trace plots of a few runs, with the post-burn-in values used for estimation. Thus we keep m fixed at 10^5 and examine the effect of increasing k on estimation of $h(X) = X_{20}$, the least well-explored component. Fig. 9(c) shows that the mean-squared error (MSE) of unbiased MCMC is prohibitively large until beyond the 0.999-quantile of τ, yet the MSE of the MCMC value might be deemed adequate and is stable throughout.

As mentioned, the noise is due to the meanderings of the two chains and the large variability in τ; Fig. 9 and the argument above show that the dominant contribution is from the most extreme τ-values. I now provide a Rao–Blackwellization for Metropolis–Hastings MCMC sampling, integrating over the distribution of coupling events, which brings the most extreme τ-values under control and leads to the MSEs in Fig. 9(d). These show a modest reduction for low and medium values of k but an order of magnitude at the 0.99-quantile and a factor of 30 at $k = 10^{4.25}$, at a cost of less than twice that of unbiased MCMC sampling.

Conditionally on (X_t, Y_{t-1}) reflection coupling proposes $X^* = Y^*$ with probability α_{rc}; otherwise the proposal $Y^{**} \neq X^*$ is a deterministic function of (X_t, Y_{t-1}) and X^*. Let $\alpha(a, b)$ denote the appropriate Metropolis–Hastings acceptance probability from a current value a to a proposed value b. If $Y^* = X^*$ then define $\alpha_\wedge := \alpha(X_t, X^*) \wedge \alpha(Y_{t-1}, X^*)$, $\alpha_\vee := \alpha(X_t, X^*) \vee \alpha(Y_{t-1}, X^*)$, $\alpha_{X\setminus Y} := 0 \vee [\alpha(X_t, X^*) - \alpha(Y_{t-1}, X^*)]$ and $\alpha_{Y\setminus X} := 0 \vee [\alpha(Y_{t-1}, X^*) - \alpha(X_t, X^*)]$; with probability $\alpha_\wedge$, X^* is accepted by both chains and coupling occurs. Instead of sampling whether or not coupling occurs, we ascertain its probability: $p_{\mathrm{coup}} := \alpha_{rc} \times \alpha_\wedge$. If $p_{\mathrm{coup}} = 1$ then coupling is inevitable; we set $\tau_{\max} = t$ and sample the X-chain from t onwards. Otherwise, we know that, if coupling occurred at $\tau = t$, $X_{t:m}$ would be sufficient additional information to calculate $H_{k:m}$, so we may continue sampling the Y-chain *conditionally on its not coupling* with X at time t and then simply ignore these later Y-values when conditioning on $\tau = t$. Specifically, we track and store a variable p_t^{nc}: the probability that the chains are not coupled by and including time t. At iteration $t+1$ we perform the following steps.

Step 1: $p_{t+1}^{\mathrm{nc}} = p_t^{\mathrm{nc}}(1 - p_{\mathrm{coup}})$.
Step 2: if $p_{\mathrm{coup}} < 1$ then sample an action from {'perform accept–reject steps on X^* and Y^{**}', 'accept X^* and reject Y^*', 'reject X^* and accept Y^*', 'reject X^* and Y^*'} according to the probability vector

$$\left\{ \frac{1 - \alpha_{rc}}{1 - p_{\mathrm{coup}}}, \frac{\alpha_{rc}\alpha_{X\setminus Y}}{1 - p_{\mathrm{coup}}}, \frac{\alpha_{rc}\alpha_{Y\setminus X}}{1 - p_{\mathrm{coup}}}, \frac{\alpha_{rc}(1 - \alpha_\vee)}{1 - p_{\mathrm{coup}}} \right\};$$

otherwise set $\tau_{\max} = t + 1$.

The Rao–Blackwellized bias correction is then

$$\sum_{\tau^*=1}^{\tau_{\max}} \mathbb{P}(\tau = \tau^*) \mathrm{BC}_{k:m}\{X, Y(\tau^*)\} = \sum_{t=k+1}^{\tau_{\max}-1} \mathbb{P}(\tau > t) \min\left(1, \frac{t-k}{m-k+1}\right) \{h(X_t) - h(Y_{t-1})\},$$

where $Y(\tau^*) := (Y_{1:(\tau^*-1)}, X_{\tau^*:(m-1)})$. In the above experiments, always, $\tau_{\max} \ll m$; moreover $\mathrm{var}(\mathbb{E}[\tau^3]) \approx 2.4\mathbb{E}[\mathrm{var}(\tau^3)]$, suggesting that variability *between* simulations is more important; however, a few very influential simulations had extremely large values for τ, from the tail of their distribution, and it is control of these which so reduces the MSE.

Certain two-block Gibbs samplers, such as for exploring an auto-regressive AR(1) process via a block of odd-numbered components and a block of even-numbered components, could be similarly Rao–Blackwellized since, conditionally on one of the blocks coupling, the other block would be guaranteed to couple. The Rao–Blackwellization could also be applied to the estimate of d_{TV} from Section 6.

The vote of thanks was passed by acclamation.

Radu Craiu (*University of Toronto*) **and Xiao-Li Meng** (*Harvard University, Cambridge*)
We thank Jacob, O'Leary and Atchadé for their tantalizing paper. As developers of Markov chain Monte Carlo (MCMC) methodology our curiosity is triggered, and as users of MCMC sampling our hopes are boosted. Indeed, we are inspired to continue our exploration of improving MCMC samplers and estimators

Table 6. Steps for implementing the proposed control variate estimators based on N (independent) paralleled coupled chains

Step 1: run N coupled chains $\{X^{(i)}, Y^{(i)}\}_{i=1}^{N}$ in parallel; record $\{\tau_i\}_{i=1}^{N}$ and draw independent and identically distributed Bernoulli(0.5) $\{\xi_i\}_{i=1}^{N}$
Step 2: compute $\{H^*_{k:m}(X^{(i)}, Y^{(i)}; \eta_i)\}_{i=1}^{N}$, where η_i is an integer median of $\{\tau_j - \xi_j, j \neq i\}$
Step 3: compute the final estimator $\bar{H}^*_{k:m}$ as the (sample) average of $\{H^*_{k:m}(X^{(i)}, Y^{(i)}; \eta_i)\}_{i=1}^{N}$

Table 7. Pump failure data: estimated standard errors for $\bar{H}_{7:7}$ and $\bar{H}^*_{7:7}$ when $h(\theta) = \beta$, $h(\theta) = \beta^2$ and $h(\theta) = \beta^3$ with $\theta = (\lambda_1, \ldots, \lambda_{10}, \beta)$

Estimator	*Number of chains N*	$h(\theta) = \beta$		$h(\theta) = \beta^2$		$h(\theta) = \beta^3$	
		Mean	*Standard deviation*	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*
$\bar{H}_{7:7}$	2	2.47	0.82	6.50	4.62	18.66	21.89
$\bar{H}^*_{7:7}$	2	2.46	0.67	6.49	3.70	18.64	17.32
$\bar{H}_{7:7}$	5	2.45	0.52	6.48	2.96	18.66	14.16
$\bar{H}^*_{7:7}$	5	2.47	0.48	6.50	2.73	18.64	12.55
$\bar{H}_{7:7}$	10	2.47	0.38	6.51	2.17	18.68	10.39
$\bar{H}^*_{7:7}$	10	2.47	0.35	6.52	1.96	18.66	9.14

Table 8. Same set-up as in Table 7 except here $m = 70$, and hence there is a time averaging over $m - k + 1 = 66$ iterations

Estimator	*Number of chains N*	$h(\theta) = \beta$		$h(\theta) = \beta^2$		$h(\theta) = \beta^3$	
		Mean	*Standard deviation*	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*
$\bar{H}_{7:70}$	2	2.47	0.09	6.61	0.47	19.11	2.19
$\bar{H}^*_{7:70}$	2	2.47	0.09	6.61	0.47	19.13	2.18
$\bar{H}_{7:70}$	5	2.47	0.05	6.61	0.30	19.11	1.38
$\bar{H}^*_{7:70}$	5	2.47	0.05	6.61	0.30	19.13	1.37
$\bar{H}_{7:70}$	10	2.47	0.05	6.61	0.20	19.11	0.95
$\bar{H}^*_{7:70}$	10	2.47	0.04	6.61	0.19	19.12	0.93

Table 9. Simulated data set from the same model for the pump failure data example in the paper

Times	0.95	1.55	1.56	1.84	2.14	2.18	2.67	3.06	3.07	3.40	3.41	3.82
Counts	2	0	2	0	1	0	0	0	0	0	0	2

via efficiency swindles (e.g. Van Dyk and Meng (2001), Craiu and Meng (2001, 2005), Craiu and Lemieux (2007) and Yu and Meng (2011)), among which figures prominently the *control variates* method.

We start by noting that $H_k(X, Y)$, defined in Section 2, can be expressed in two ways:

$$H_k(X,Y) = h(X_k) + \sum_{j=k+1}^{\tau-1} \{h(X_j) - h(Y_{j-1})\} = h(X_{(\tau-1)\vee k}) + \sum_{j=k}^{\tau-2} \{h(X_j) - h(Y_j)\}, \tag{1}$$

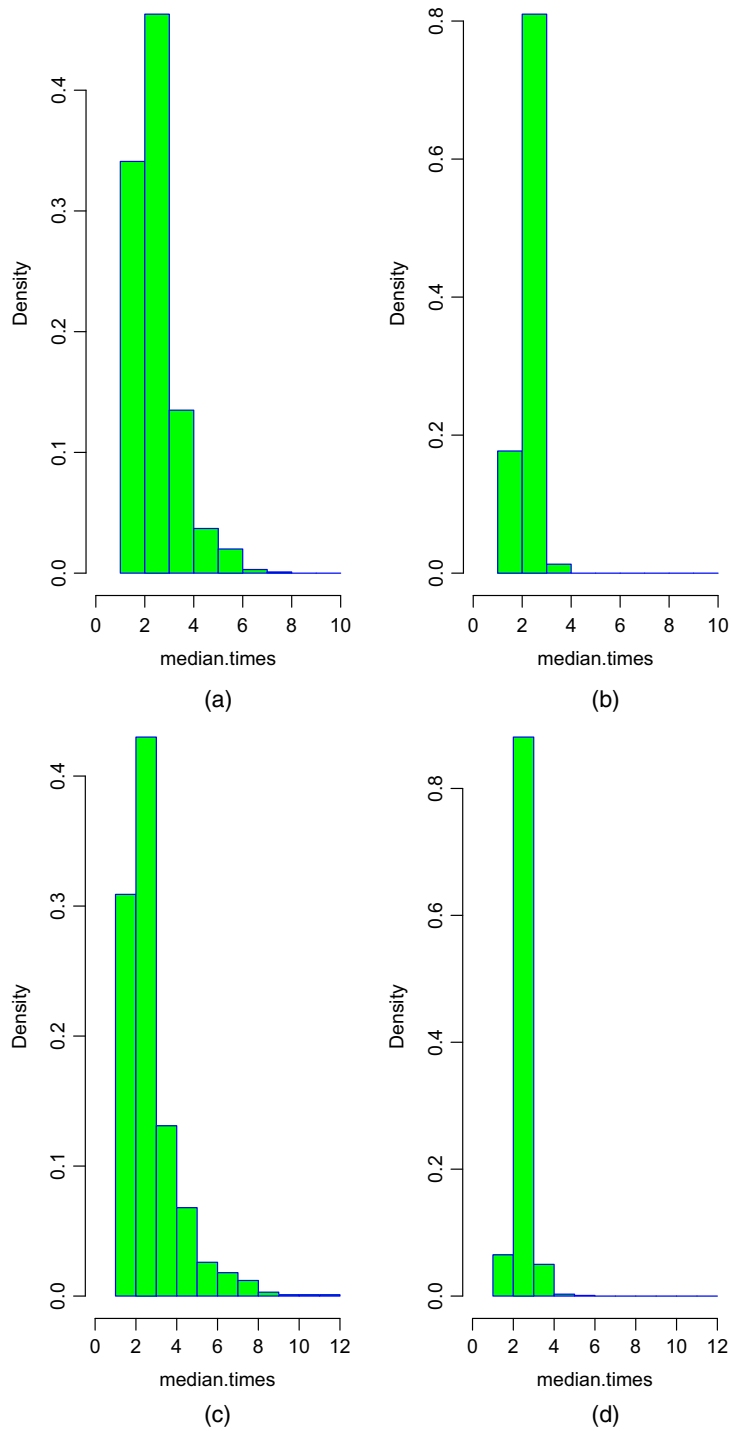

Fig. 10. Histogram of coupling times for (a), (b) the pump data and (c), (d) the simulated data for (a), (c) two (b), (d) 10 parallel chains

Table 10. Same as Table 8 with $k=3$, $m=70$ and the new data given in Table 9

Estimator	*Number of chains N*	$h(\theta)=\beta$		$h(\theta)=\beta^2$		$h(\theta)=\beta^3$	
		Mean	*Standard deviation*	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*
$\bar{H}_{3:70}$	2	2.64	0.24	7.75	1.54	22.28	8.68
$\bar{H}^*_{3:70}$	2	2.64	0.20	7.73	1.38	22.19	7.44
$\bar{H}_{3:70}$	5	2.65	0.14	7.77	0.89	22.31	4.87
$\bar{H}^*_{3:70}$	5	2.65	0.11	7.76	0.85	22.29	4.59
$\bar{H}_{3:70}$	10	2.64	0.11	7.74	0.63	22.27	3.38
$\bar{H}^*_{3:70}$	10	2.64	0.10	7.74	0.62	22.26	3.36

where $A \vee B = \max\{A, B\}$. The first expression is the authors' which rendered the insight that $H_k(X, Y)$ includes a *time forward correction* of $h(X_k)$'s bias, when $\tau > k+1$. The second expression indicates that $H_k(X, Y)$ is also a *time backward correction* for the bias in $h(X_{\tau-1})$ when $\tau > k+1$. (No correction is needed when $\tau \leqslant k+1$.)

The most important insight from the second expression, however, is that $\Delta_j \equiv h(X_j) - h(Y_j)$ has zero expectation for any j. This implies that, for any integer-valued η independent of $\{(X_j, Y_j), j = 1, \ldots\}$, we can use $C_k(\eta) = \Sigma_{j=k}^{\eta-2} \Delta_j$ as a *control variate* for $H_k(X, Y)$, i.e.

$$H_k^*(X, Y; \eta) = H_k(X, Y) + C_k(\eta) = h(X_{(\tau-1)\vee k}) + \sum_{j=k}^{\tau-2} \Delta_j - \sum_{j=k}^{\eta-2} \Delta_j, \tag{2}$$

always shares the mean of $H_k(X, Y)$ but can have a smaller variance with a judicious choice of η.

Our preliminary theoretical investigations (Craiu and Meng, 2020) show that a good choice of η is the median of $\tau - \xi$, where ξ is an independent Bernoulli(0.5) variable. This suggests immediately the method given in Table 6, where $H_{k:m}$ is the time averaging extension of H_k, defined by the authors' expression (2.1), and similarly $H^*_{k:m}$ extends H^*_k. Table 7 compares $\bar{H}^*_{k:m}$ with $\bar{H}_{k:m}$, the average of the authors' N estimators $\{H_{k:m}(X^{(i)}, Y^{(i)})\}_{i=1}^N$, using their pump failure data, but without time averaging (hence $m = k = 7$). We see a decrease of standard deviations for all cases. The reductions are small because τ_is are small (e.g. τ_is vary mostly from 1 to 6), and hence C_k cannot contribute much. Indeed the gains vanish with time averaging using $m = 70$, as seen in Table 8.

To confirm that this vanishing is due to fast coupling, we simulated a new data set (Table 9) by using the same model for the pump failure data, for which the τ_is are somewhat larger (e.g. mostly $\{\tau_i = 1 \sim 9\}$). The distributions of coupling times for the pump data and the simulated examples are shown in Fig. 10 respectively for two and 10 parallel chains. Table 10 repeats Table 8 with these new data, where we see the gains appear again, albeit quite small. More extensive investigations are reported in Craiu and Meng (2020), especially on the use of equation (2) for improving the theoretical bounds given in Biswas *et al.* (2019).

Andi Q. Wang (*University of Bristol*), **Murray Pollock and Gareth O. Roberts** (*University of Warwick, Coventry, and Alan Turing Institute, London*) **and David Steinsaltz** (*University of Oxford*)

We congratulate Jacob, O'Leary and Atchadé for a stimulating and interesting paper. Exploiting coupling constructions within Markov chain Monte Carlo (MCMC) methods can be very powerful; in this work they have exhibited a range of coupling constructions, which can be adopted for certain commonly used MCMC samplers to debias the output. An alternative approach to removing bias is to build global couplings into the design of new MCMC algorithms. Section 5.2 briefly reviews one traditional version of this approach, using *regenerations*, following a method proposed by Mykland *et al.* (1995).

We wish to call attention to recent advances in regenerative algorithms, that build on two decades of development in the theory of quasi-stationarity in Markov processes. One recent example is the 'Restore' sampler, proposed in Wang *et al.* (2019), a continuous time regenerative sampler, constructed from three components: a continuous time Markov process Y taking values in the state space $\mathcal{X}$, a regeneration rate $\kappa : \mathcal{X} \to [0, \infty)$ and a probability density function μ on $\mathcal{X}$, the regeneration distribution.

From an algorithmic perspective, the process X_t runs according to the law of the Markov process Y but also randomly *regenerates* at rate $t \mapsto \kappa(X_t)$: a new location is drawn (independently) from the regeneration distribution μ. Under appropriate conditions on the components Y, κ and μ, the resulting process X has a given target probability distribution π as its invariant measure. Fixing Y, μ and the target density π, the corresponding choice of regeneration rate κ, to ensure π-stationarity, is

$$\kappa(x) = \frac{L^*\pi(x)}{\pi(x)} + C\frac{\mu(x)}{\pi(x)}, \qquad x \in \mathcal{X},$$

where L is the infinitesimal generator of the Markov process Y, and the constant C is chosen such that $\kappa \geqslant 0$.

The process's killing and subsequent regeneration mechanism enables us to construct a natural coupling on the entire state space. Suppose that a killing time can be identified that would have occurred regardless of the state of the process. (This can be done straightforwardly if, say, the regeneration rate κ is uniformly bounded away from 0.) In this case, regardless of the prior evolution of the process, all sample paths can be coupled—because of the fixed regeneration distribution—to a common point in the state space.

By appealing to the *coupling from the past* principle (Propp and Wilson, 1996), it suffices then to simulate the restore process from the last such occurrence in the past forwards to time 0 to achieve a perfect draw from the target distribution. For further details, we refer the reader to Wang *et al.* (2019).

Anthony Lee (*University of Bristol*), **Sumeetpal S. Singh** (*University of Cambridge*) **and Matti Vihola** (*University of Jyväskylä*)

We congratulate Jacob, O'Leary and Atchadé for their inspiring work, which elaborates on the general Markov chain debiasing method of Glynn and Rhee (2014) and demonstrates how it can be applied in practice in various settings. We would like to highlight the importance of understanding the scalability properties of the related couplings, in particular with respect to increasing model dimension, to assess the applicability of the method for challenging problems for which inference is difficult with existing techniques.

We focus here on the context of hidden Markov model smoothing, where debiasing was suggested earlier (Jacob *et al.*, 2020) based on coupled conditional particle filters (Chopin and Singh, 2015). Our follow-up work (Lee *et al.*, 2020) presented another coupling based on the backward sampling conditional particle filter (Whiteley, 2010). The latter approach, termed the coupled conditional backward sampling particle filter, has provably stable behaviour in increasing data record length T (Lee *et al.*, 2020). More specifically, under suitabie 'strong mixing' conditions, the coupling time τ_T of this filter—with fixed number of particles—satisfies assumption 2 of the present paper such that $\mathbb{P}(\tau_T > t) \leqslant C_T\delta^t$, with constants $\log(C_T) = O(T)$ and $\delta \in (0, 1)$, and these orders appear tight (Lee *et al.*, 2020).

Since the Markov chain is uniformly ergodic, an easy modification of the proof of proposition 3 for bounded functions yields the following upper bound for the 'bias correction' term:

$$\mathbb{E}[\mathrm{BC}_{k:m}^2] \leqslant 4C_T\delta^k\|h\|_\infty^2(m-k+1)^{-2}(1-\delta^{1/2})^{-4}.$$

Clearly, for any $\epsilon > 0$, there exists $c \in (0, \infty)$ such that for all $k \geqslant k_T = cT$ it holds that $\mathbb{E}[\mathrm{BC}_{k:m}^2] \leqslant \epsilon\|h\|_\infty^2(m-k+1)^{-2}$. Furthermore, c may be chosen so that $\log(C_T\delta^{k_T}) = O(-T)$, demonstrating that the bias correction part quickly becomes negligible for large T, as long as $k \geqslant k_T$.

If we consider the suggested unbiased parallelization with P processing units and aim for minimal wall time, the required run time is $\kappa\tau_P^*$, where κ is the cost of a single iteration and τ_P^* is the maximum of P independent coupling times. It is easy to see that in our context $\mathbb{E}[\tau_P^*] = O\{T + \log(P)\}$ and, since $\kappa = O(T)$, we conclude that the minimal time of the maximum of P independent realizations is $O\{T^2 + T\log(P)\}$. The use of a fixed number of particles implies that the space complexity is, however, $O(TP)$.

The following contributions were received in writing after the meeting.

Christine P. Chai (*Microsoft Corporation, Redmond*)

Bayesian methodology is known to provide biased estimates with a lower mean-squared error (Hoff, 2009), so I am excited to see improvements on removing the bias from Markov chain Monte Carlo (MCMC) kernels. This paper provides excellent theoretical methodology, and I have several questions about the description of the unbiased MCMC estimators.

(a) Why does the paper use only confidence intervals, not credible intervals? I understand that the central limit theorem is a frequentist concept (Gray *et al.*, 2015) but, since this paper is about Bayesian methods, I wonder where the term 'credible interval' applies.
(b) What is the contextual definition of the 'meeting time' of the chains? The mathematical definition is provided, but I think that stating the definition in plain text would also be helpful.
(c) Why is the convergence speed measured as the 'average meeting time' in the figures? I believe that one time unit refers to the cost of one draw and two evaluations in the sampling algorithm, so the concept makes sense. But the word 'time' makes me relate this term to the actual computation time. Since the details of the platform (e.g. RStudio) is not mentioned in the paper, I wonder whether there is a better term to describe the computation cost of MCMC sampling.
(d) I agree that parallel computing is not a panacea, and that each MCMC iteration can receive only limited benefits from parallel processors. Given the potential for parallelization of the framework proposed, how is the scalability related to the asymptotically exact, 'embarrassingly parallel' MCMC algorithm described in Neiswanger *et al.* (2013)?

(The opinions and views expressed here are those of the author and do not necessarily state or reflect those of Microsoft.)

Martin Chak, Nikolas Kantas and Grigorios A. Pavliotis (*Imperial College London*)

We congratulate Jacob, O'Leary and Atchadé for their major and inspiring contribution. In our discussion we shall sketch our ideas for extending the use of couplings to include variance reduction in addition to bias removal. Our approach is based on extending the framework presented in Nüsken and Pavliotis (2019). There, couplings of n π-invariant continuous time Markov processes $(Z^i; i = 1, \ldots, n)$ were considered for variance reduction (with respect to asymptotic variance). Sections 3 and 5 of Nüsken and Pavliotis (2019) present non-trivial constructions for such couplings of Markov processes that can be used in sampling contexts. Typically one should expect that these couplings enhance the exploration of the state space, which seems to be in contrast with constructing Markov processes that meet as required for unbiased Markov chain Monte Carlo sampling. We believe that these two opposing behaviours can be combined. Instead of using independent samples of coupled faithful pairs of chains, one could consider correlating these pairs, $Z^i = (X^i, Y^i)$, such that each of X^i and Y^i have marginally in i the same law and then use an unbiased estimator. Given X^i one could design Y^i so that they eventually meet and the X^is could be jointly propagated by using correlated noise inputs designed for improved joint asymptotic variance or better exploration of the state space as in Nüsken and Pavliotis (2019). This can be achieved in discrete time (or in continuous time) by extending the work of Nüsken and Pavliotis (2019) to discrete time Markov processes and Markov chain Monte Carlo sampling (or by extending the work in this paper for continuous Markov processes). Finally we note that in the most simple case moving from $\mathcal{X}$ to work on a joint state space $\mathcal{X}^n$ requires modifying the test function $h(x)$ to $(1/n)\Sigma_{i=1}^n h(x^i)$ or a normalized weighted average, so it would be interesting to include variance reduction techniques such as control variates.

Nicolas Chopin (*École Nationale de la Statistique et de l'Administration Economique, Institut Polytechnique de Paris*)

I congratulate Jacob, O'Leary and Atchadé on their very elegant, coupling-based, approach to making Markov chain Monte Carlo algorithms parallelizable. To illustrate the points I would like to make, I consider the Gibbs sampler of Albert and Chib (1993), for a probit model $(y_i = \mathrm{sgn}(z_i),\ z_i = \beta^{\mathrm{T}} x_i + \epsilon_i,\ \epsilon_i \sim N(0, 1))$ where one simulates alternately

step 1, $\beta | y, z$ (Gaussian), and
step 2, $z | \beta, y$ (independently, $z_i \sim N_{>0}(\beta^{\mathrm{T}} x_i, 1)$ if $y_i = 1$).

Even if we restrict our attention to maximum coupling for step 1, several options are available to couple at step 1:

(a) maximum coupling for the z_is;
(b) maximum coupling for the residuals ϵ_i (in this specific case, coupling is particularly simple (and fast); to couple two variates from distributions $N_{>c_1}(0, 1)$ and $N_{>c_2}(0, 1)$, assuming $c_1 < c_2$, sample $X \sim N_{>c_1}(0, 1)$, and take $Y = X$ if $X > c_2$; otherwise sample $Y \sim N_{[c_1, c_2]}(0, 1)$);
(c) optimal transport for the z_is (i.e. use the inverse cumulative distribution function algorithm with the same input). In that case, the z_is become increasing close but do not become equal. But note that this still helps to increase the probability that the β-chains become coupled.

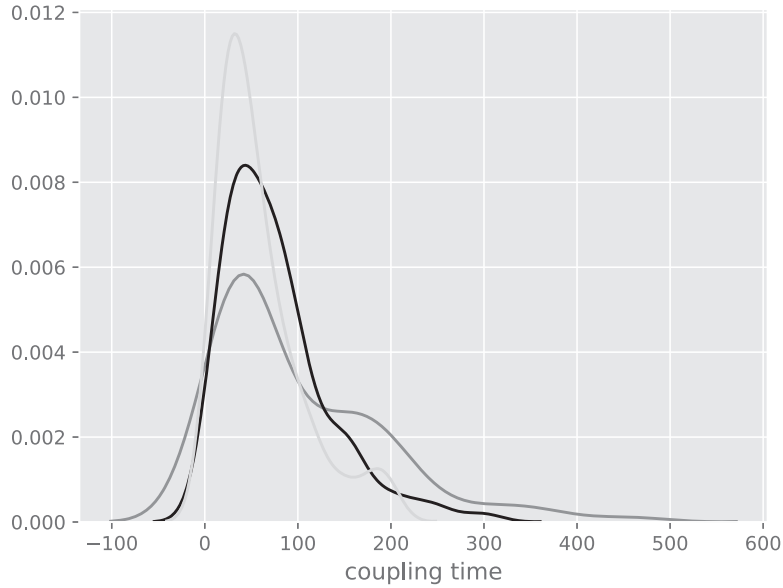

Fig. 11. Kernel density estimates of the coupling times for the three coupling approaches: ——, latent coupling; ——, residuals coupling; ——, latent transport

Fig. 11 shows kernel density estimates for the coupling times of the three approaches (100 runs; data, German credit, with standardized predictors; prior, $N(0, 10^2)$ for each β_j). If computational time is taken into account, approach (b) is the clear winner, as it seems 20 times faster (at least on this particular data set, and with the usual *caveats* regarding implementation). But it is also interesting to see that approach (c) (transport) may be competitive without forcing the z_i's to be equal.

I would be glad to hear the authors' views, more generally, on

(a) the effect of parameterization on coupling performance and
(b) the opportunity to use transport coupling (or any coupling that does not generate *equal* outputs with positive probability) for certain components of the chain.

Mathieu Gerber and Anthony Lee (*University of Bristol*)

This is a very interesting paper that demonstrates how couplings can be used in practical applications to obtain unbiased approximations.

Controlled variance of cost of Thorisson's algorithm

We consider a modification of Thorisson's algorithm for simulating from a coupling $\Gamma_\phi(\mu, \nu)$ of two distributions μ and ν. Let $\nu = \nu_0 + \nu_1$ where $\nu_0 \ll \mu$, $\nu_1 \perp \mu$ and

$$w(x) := \begin{cases} \mathrm{d}\nu_0/d\mu & x \in S^{\mathrm{C}}, \\ \infty & x \in S, \end{cases}$$

where $S \subseteq \mathcal{X}$ satisfies $\mu(S) = \nu_1(S^{\mathrm{C}}) = 0$. Let $\phi : \mathcal{X} \to [0, 1]$ satisfy $\phi \leqslant w$ pointwise. The algorithm is as follows.

Step 1: sample $X \sim \mu$. If $1 = B \sim \text{Bernoulli}\{\phi(X)\}$, output (X, X).
Step 2: repeat sample $Y \sim \nu$; with probability $1 - \phi(Y)/w(Y)$ output (X, Y).

Letting (X, Z) denote the output of the algorithm, we have $\mathbb{P}(X \in A) = \mu(A)$ and, using $\mathbb{P}(B = 1) = 1 - \mu(\phi)$ and $\nu(\phi/w) = \mu(\phi)$,

$$\begin{aligned} \mathbb{P}(Z \in A) &= \mathbb{P}(Z \in A, B = 1) + \mathbb{P}(Z \in A, B = 0) \\ &= \mu(\phi \cdot \mathbf{1}_A) + \nu(\mathbf{1}_A \cdot \{1 - \phi/w\}) \end{aligned}$$

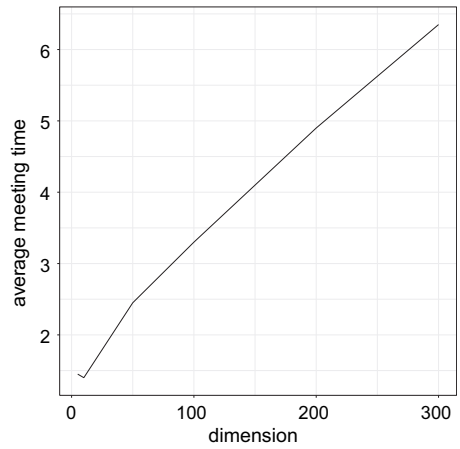

Fig. 12. Scaling of the average meeting time of a reparameterized, coupled Metropolis-within-Gibbs algorithm with the dimension of the posterior distribution of a Bayesian logistic regression model with d parameters and $n = 100d$ observations

$$= \mu(\phi \cdot \mathbf{1}_A) + \nu(A) - \mu(\mathbf{1}_A \cdot \phi)$$
$$= \nu(A),$$

so that (X, Z) has the right marginals. The choice $\phi(x) = \min\{1, w(x)\}$ gives $\mathbb{P}(X \neq Z) = 1 - \mu(\phi) = \|\mu - \nu\|_{\mathrm{TV}}$, which corresponds to Thorisson's algorithm.

The number of samples from ν is $N = (1 - B)G$, where $G \sim \text{geometric}\{1 - \mu(\phi)\}$ independent of B. The expectation and variance of N are 1 and $2\mu(\phi)/\{1 - \mu(\phi)\}$ respectively. Hence, one can reduce the variance of N by increasing $\mathbb{P}(X \neq Z) = 1 - \mu(\phi)$, i.e. by making $\Gamma_\phi(\mu, \nu)$ non-maximal. For example, one could choose $\phi(x) = \min\{C, w(x)\}$ for some $C < 1$, so that the variance is upper bounded by $2C/(1 - C)$.

Reparameterization

Section 4.3 of the paper discusses the coupling of Metropolis-within-Gibbs samplers, and the numerical results suggest that the average meeting time scales (sub)linearly with the dimension d of the target distribution π, provided that its components are weakly dependent. When this condition is not verified one can find a diffeomorphism g such that the components of $\pi_g = \pi \circ g^{-1}$, the image of π by g, are weakly dependent. Although finding such a diffeomorphism can be difficult in general, when π is a posterior distribution the Bernstein–von Mises theorem suggests taking $g(x) = A^{-1}(x - \mu)$ with $\mu \approx \mathbb{E}_\pi[X]$ and A such that $AA^{\mathrm{T}} \approx \mathrm{cov}_\pi(X)$. In this case, $\pi_g \approx \mathcal{N}(0, I_d)$ and we can follow the approach used in Fig. 2 (with $V = I_d$) to approximate π_g, and therefore also π.

To illustrate, we consider a logistic regression model, with an $\mathcal{N}(0, I_d)$ prior distribution and $n = 100d$ covariates, and report in Fig. 12 the average meeting time (over 20 replications) as a function of d, which appears to scale (sub)linearly with d.

Jiaqi Gu (*University of Hong Kong*)

I congratulate Jacob, O'Leary and Atchadé for their thought-provoking paper. In the paper, I am interested in two aspects:

(a) adapting couplings to substitute a meeting time with finite expected value for traditional convergence diagnosis methods in Markov chain Monte Carlo (MCMC) algorithms and
(b) removing the bias of MCMC estimators with telescopic sums at finite expected computational cost.

MCMC methods (Metropolis *et al.*, 1953; Hastings, 1970; Geman and Geman, 1984) have been widely used for Bayesian inference. Although the asymptotic convergence of MCMC estimators has been proved in theory (Tierney, 1994),

$$\frac{1}{n}\sum_{t=1}^{n} h(X_t) \to \mathbb{E}_\pi[h(X)], \qquad \text{almost surely,}$$

the bias of MCMC estimators still exists after an arbitrary number of iterations. In this paper, with two elaborately designed chains $\{X_t\}$ and $\{Y_t\}$ which have a meeting time $\tau(E\tau < \infty)$, $\mathbb{E}_\pi[h(X)]$ is decomposed as the expectation of finite summation

$$\mathbb{E}\left[h(X_k) + \sum_{t=k+1}^{\tau-1} h(X_t) - h(Y_{t-1})\right].$$

Consequently, unbiased estimator $H_{k:m}(X, Y)$ could be obtained at finite expected computation cost $O(\mathbb{E}[\tau])$. By investigating the performance of couplings in numerical experiments under various settings, the authors propose simple guidelines to implement their unbiased Bayesian inference method as efficiently as possible.

There are still some possible extensions of their method. One example is missing data. As illustrated in Sections 4.2 and 4.3, the average meeting time would increase as the dimension grows. This would make it more difficult to adapt such a strategy to latent variable models. Like the Ising model in Section 5.3, latent variable models

$$f(\mathbf{X}_{\text{obs}}; \boldsymbol{\theta}) = \int f(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}; \boldsymbol{\theta}) \mathrm{d}\mathbf{X}_{\text{mis}}$$

usually consist of high dimensional latent variables $\mathbf{X}_{\text{mis}}$ which need updates but are irrelevant to the target $\mathbb{E}_\pi[h(\boldsymbol{\theta})]$. Given that the dimension of $\mathbf{X}_{\text{mis}}$ is far greater than parameters $\boldsymbol{\theta}$, the expected meeting time $\mathbb{E}[\tau]$ will be extremely large and the unbiased estimators could not be obtained within available computational time if the definition of meeting time is kept unchanged. An extension of the proposed method to deal with missing data problems is needed.

Ajay Jasra (*King Abdullah University of Science and Technology, Thuwal*), **Kody Law** (*University of Manchester*) **and Jeremy Heng** (*École Supérieure des Sciences et Commerciales Business School, Singapore*)

We thank Jacob, O'Leary and Atchadé for a most interesting paper. Our discussion centres on a very promising extension of the method for unbiased estimation as we now explain. Consider a probability measure π on measurable space $(\mathbf{X}, \mathscr{X})$ for which we want to compute $\pi(\varphi) := \int_{\mathbf{X}} \varphi(x)\pi(\mathrm{d}x)$ with $\varphi : \mathbf{X} \to \mathbb{R}$, π integrable and measurable. Suppose that one can deal with only a sequence of *biased* probability measures $(\pi_l)_{l \geqslant 0}$ on $(\mathbf{X}, \mathscr{X})$, with $\pi_l(|\varphi|) < \infty \ \forall\, l \in \mathbb{Z}^+$, such that $\lim_{l \to \infty} \pi_l(\varphi) = \pi(\varphi)$; examples include partially observed diffusion processes (e.g. Jasra *et al.* (2017)) or inverse problems (e.g. Beskos *et al.* (2017)). Consider a positive probability mass function $\mathbb{P}_L$ on $\mathbb{Z}^+$. It is known (Rhee and Glynn, 2015; Vihola, 2018) that if one can find a sequence of independent random variables $(\xi_l)_{l \geqslant 0}$ that are independent of $L \sim \mathbb{P}_L$ such that

$$\begin{aligned} \mathbb{E}[\xi_0] &= \pi_0(\varphi), \\ \mathbb{E}[\xi_l] &= \pi_l(\varphi) - \pi_{l-1}(\varphi) \qquad \forall\, l \in \mathbb{N}, \\ \sum_{l \in \mathbb{Z}^+} \frac{\mathbb{E}[\xi_l^2]}{\mathbb{P}_L(l)} &< \infty \end{aligned} \tag{3}$$

then $\xi_L / \mathbb{P}_L(L)$ is an unbiased estimator of $\pi(\varphi)$. This is the 'single-term' estimator as discussed by Rhee and Glynn (2015) and Vihola (2018), but alternatives such as the 'coupled sum' estimator are also possible. If l is a parameter induced by an Euler discretization of a stochastic differential equation or a finite element approximation of a solution of a partial differential equation one can completely remove the discretization bias, while only working with biased versions of π. The main idea is now to utilize the authors' approach to obtain the sequence of $(\xi_l)_{l \geqslant 0}$. If $l = 0$ then one can apply the coupled Metropolis–Hastings (MH) method associated with π_0. For $l \in \mathbb{N}$ one must be more careful as, to satisfy condition (3), it will typically not be sufficient to run two *independent* unbiased Markov chain Monte Carlo algorithms for-estimating $\pi_l(\varphi)$ and $\pi_{l-1}(\varphi)$ respectively. The scheme which we are developing is based on constructing a four-way coupling of MH kernels. This is possible, for example, by sampling a four-way 'maximally' coupled MH proposal, i.e., if $Q : \mathbf{X} \to \mathscr{P}(\mathbf{X})$ ($\mathscr{P}(\mathbf{X})$ are the probability measures on $(\mathbf{X}, \mathscr{X})$) is a Markov kernel and $x_{1:4} \in \mathbf{X}^4$, we sample a proposal kernel $\check{Q} : \mathbf{X}^4 \to \mathscr{P}(\mathbf{X}^4)$, such that for any $j \in \{1, \ldots, 4\}$ $\int_{x'_{-j} \in \mathbf{X}^3} \check{Q}(x_{1:4}, \mathrm{d}x'_{1:4}) = Q(x_j, \mathrm{d}x'_j)$ (x_{-j} is $x_{1:4}$ without the jth co-ordinate) and

$$\int_{\{x_{1:4}\in\mathbf{X}^4:x_1=x_2=x_3=x_4\}} \check{Q}(x_{1:4}, \mathrm{d}x'_{1:4}) = \int_{\mathbf{X}^4} \bigwedge_{j=1}^{4} Q(x_j, \mathrm{d}x'_j).$$

Acceptance or rejection of proposals for these four chains can be conducted in a similar manner to the authors' but, when $l \in \mathbb{N}$, care is required to ensure faithfulness of the chains targeting π_l and π_{l-1}. We believe that this *doubly randomized* scheme will have multiple applications and moreover have finite variance and expected cost, as well as providing a competitive alternative to Agapiou *et al.* (2018). In addition, the benefits of such unbiased multilevel Monte Carlo methods (Giles, 2008), relative to advanced multilevel Monte Carlo methods such as in Beskos *et al.* (2017) and Jasra *et al.* (2017), would be the lack of bias, amenability to parallelization and comparable or reduced computational cost for a given error.

Lawrence Middleton, George Deligiannidis and Arnaud Doucet (*University of Oxford*)
We congratulate Jacob, O'Leary and Atchadé for this seminal work for both its elegance and its importance. The implementation of this methodology requires being able to couple two Markov chain Monte Carlo (MCMC) chains and is highly dependent on the specific MCMC algorithm used. We review here the scheme proposed in Middleton *et al.* (2019) to couple two independent Metropolis–Hastings (MH) chains and explain how this idea can be combined with sequential Monte Carlo (SMC) methods to estimate unbiasedly expectations with respect to high dimensional probability measures.

Given a target density $\pi(x)$ and proposal density $q(x)$, we can couple two MH chains X and Y to ensure that $X_t = Y_{t-1}$ for all $t \geqslant \tau$ for some meeting time τ by using the same proposal and uniform random variables for X at iteration i and Y at iteration $i-1$. Furthermore, by also using $q(x)$ as initial distribution for both chains and selecting the proposal at the first iteration for X to be equal to the initial distribution of Y, we can ensure that $\mathbb{P}(\tau = 1) \geqslant \frac{1}{2}$. Additionally, conditionally on X_0, the meeting time follows a geometric distribution; see Middleton *et al.* (2019) for details.

The independent MH algorithm is of limited interest for high dimensional targets. However, the particle independent MH (PIMH) method introduced in Andrieu *et al.* (2010) empirically scales much better. This algorithm is nothing but an independent MH algorithm targeting an artificial target $\pi_N(x, z)$ admitting the desired target $\pi(x)$ as a marginal by using an SMC proposal $q_N(x, z)$, where N is the number of particles and z some auxiliary variables. We can thus couple PIMH chains exactly like IMH chains and, under weak regularity assumptions, $\mathbb{P}(\tau = 1) \to 1$ as $N \to \infty$. In particular, we can use PIMH steps to estimate unbiasedly expectations with respect to the smoothing distribution of non-linear non-Gaussian state space models by using a particle filter proposal. To estimate unbiasedly expectations with respect to a general probability measure, we can simply use SMC samplers (Del Moral *et al.*, 2006) or annealed importance sampling (Neal, 2001) proposals within PIMH sampling and couple the resulting chains (Middleton *et al.* (2019), appendix B).

From a practical point of view, the coupling approach of Jacob and his colleagues is easier to put into practice than the random-truncation technique of Glynn and Rhee (2014) in this context. Indeed, as long as the likelihood of the observations given the latent states is bounded for state space models or the likelihood is bounded for general posterior targets, then the PIMH algorithm is uniformly ergodic for any N and, under very weak additional assumptions, the coupling procedure returns finite variance estimates in finite expected time. To obtain similar guarantees for the methodology of Glynn and Rhee (2014), quantitative convergence results for the PIMH algorithm would need to be obtained to select an appropriate distribution for the truncation time.

Daniel Paulin (*University of Edinburgh*)
Congratulations go to Jacob, O'Leary and Atchadé for their very interesting paper and the many follow-ups. They have opened up a new field of unbiased Markov chain Monte Carlo sampling that is widely applicable and helps convergence diagnostics and parallelization.

A question for future research is whether the expected coupling times scale well in high dimensional settings. In recent years, considerable progress has been made in the understanding of high dimensional scaling of the Metropolis-adjusted Langevin algorithm (Dwivedi *et al.*, 2019) and Hamiltonian Monte Carlo sampling (Chen *et al.*, 2019) and Gibbs sampling (Qin and Hobert, 2018). It could be interesting to see whether couplings where the expected coupling time has similarly good dimension dependence can be constructed for such practically relevant algorithms.

Emilia Pompe and Chris Holmes (*University of Oxford*)
We congratulate Jacob, O'Leary and Atchadé on their impressive contribution, which has already stimu-

lated further research in the area and will certainly inspire many other developments. The idea of running parallel computations for Markov chain Monte Carlo (MCMC) algorithms, which is sequential by nature, has recently been attracting much attention in the Bayesian community, e.g. Scott *et al.* (2016), Dai *et al.* (2019) and Rendell *et al.* (2018). We believe that the authors achieved a major breakthrough in this field

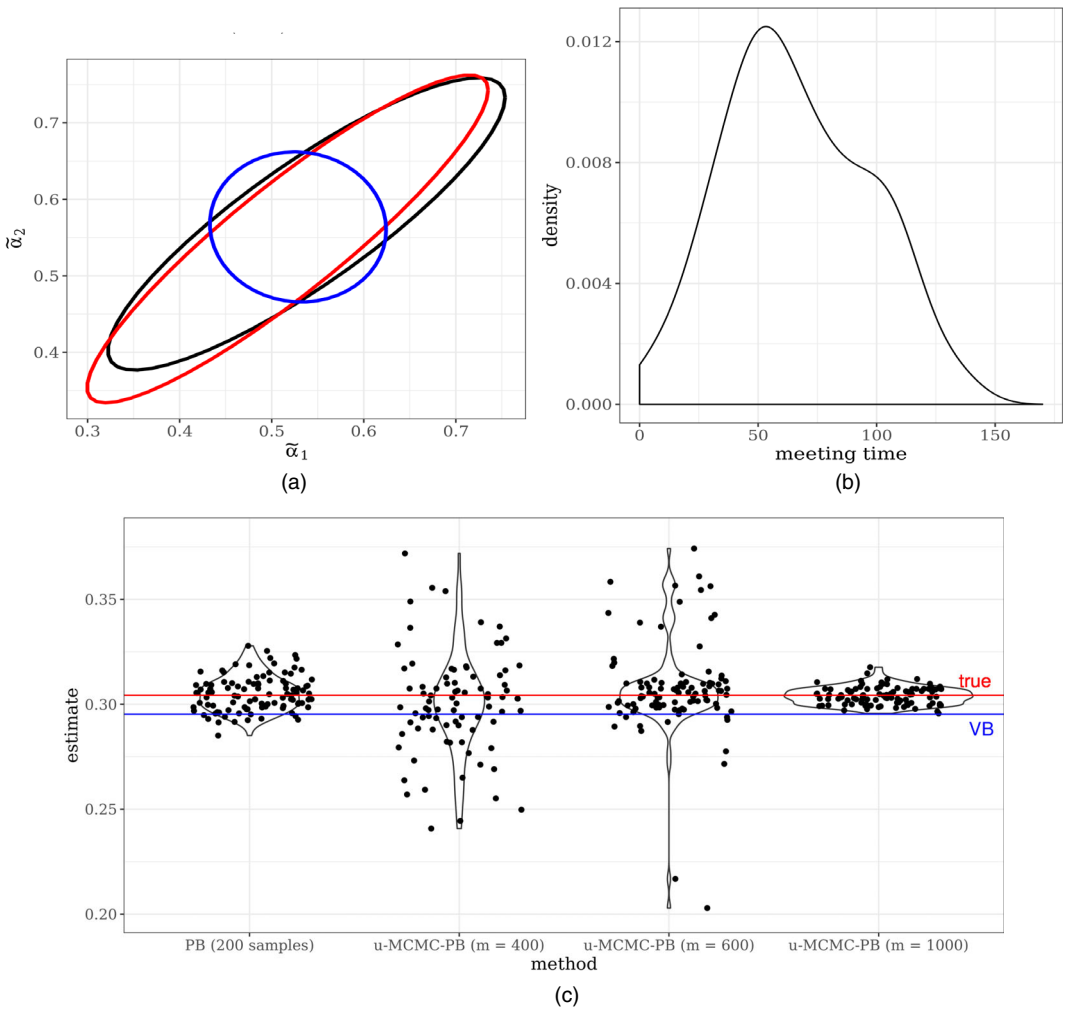

Fig. 13. Reproduction of the toy example of Lyddon *et al.* (2018) (see section 3.1 and Fig. 1 of Lyddon *et al.* (2018) for details) and comparison of these results with a sampling scheme where, instead of drawing samples γ_i from the variational Bayes (VB) approximation, we follow the idea of unbiased MCMC sampling and construct a pair of coupled Markov chains targeting this approximation: (a) 95% probability contour of the true posterior (——) and different approximations (VB (——) and VB corrected through the posterior bootstrap (PB) (——); we set the parameter of the PB as $T = 10000$ and $c = 1$); (b) density of the coupling times of the Markov chains; (c) violin plots obtained for a function g (see expression (4)) defined as $\tilde{\alpha}_1\tilde{\alpha}_2$ (a product of co-ordinates) where $\tilde{\alpha} := \arg\max_\gamma \Sigma_{i=1}^{n} w_i \log\{f_\gamma(x_i)\} + \Sigma_{i=1}^{T} w_{n+i} \log\{f_\gamma(x_{n+i})\}$ and $\log(f_\gamma)$ denotes the log-likelihood function (we compare results obtained with 200 samples drawn from the PB and where unbiased MCMC sampling is used at the first sampling stage, for various values of k and m used for defining the unbiased estimator (see Section 2.3); the parameter k was each time set to $0.1m$; each experiment was repeated 100 times; we can see that the variance decreases dramatically as the values of k and m increase; it can also be observed that using the PB or the PB combined with unbiased MCMC sampling indeed corrects the VB approximation—after applying any of these procedures the results are closer to the true value of $\alpha_1\alpha_2$ calculated for the posterior)

by proposing a general framework for parallelizing computations, applicable to many MCMC algorithms (Heng and Jacob, 2019; Middleton *et al.*, 2018).

Moreover, an important advantage of this method is applicability to models composed of modules (Liu *et al.*, 2009; Plummer, 2015; Zigler, 2016; Jacob *et al.*, 2017), where because of misspecification of the full model it may be beneficial to estimate parameters of these modules sequentially, as discussed in Section 5.5. The law of total expectation ensures that the estimators of integrals with respect to the 'cut distribution' in such models are unbiased. This can be useful more generally, when one wants to use different Monte Carlo techniques preserving unbiasedness, e.g. (standard) importance sampling at subsequent stages of inference. This remarkable property of the unbiased MCMC technique could also be useful beyond cut models as we now discuss.

We can extend unbiased MCMC to the area of Bayesian non-parametrics and the posterior bootstrap proposed by Lyddon *et al.* (2018) to address issues arising in model misspecification. Following the notation of algorithm 1 of Lyddon *et al.* (2018), suppose that we want to obtain an unbiased estimator of $g(\tilde{\alpha})$. In the current version of algorithm 1 this can be obtained via averaging $g(\tilde{\alpha}^{(i)})$ since it is assumed that we can draw independent and identically distributed samples from $\gamma^{(i)} \sim \pi(\gamma|x_{1:n})$ and $x_k^{(i)} \sim f_{\gamma^{(i)}}$ for $k = n+1, \ldots, n+T$. In many settings, however, sampling directly from one of those distributions (or both) is impossible, e.g. when the prior for γ is non-conjugate.

To construct an unbiased estimator of $g(\tilde{\alpha})$, we can follow the procedure described in Section 5.5 for

$$\theta_1 := \gamma$$

and

$$\pi_1(\gamma) \sim \pi(\gamma|x_{1:n}),$$

$$\theta_2 := (x_{(n+1):(n+T)}, w)$$

and

$$\pi_2\{(x_{(n+1):(n+T)}, w)|\gamma\} \sim f_\gamma(x_{n+1}) \ldots f_\gamma(x_{n+T}) \mathrm{Dir}\left(w; \underbrace{1, \ldots, 1}_{n}, \underbrace{c/T, \ldots, c/T}_{T}\right),$$

$$h(\theta_1, \theta_2) = h(\gamma, x_{(n+1):(n+T)}, w) := g\left\{\arg\max_\alpha \sum_{j=1}^{n} w_j u(x_j, \alpha) + \sum_{j=1}^{T} w_{n+j} u(x_{n+j}, \alpha)\right\}. \quad (4)$$

Fig. 13 illustrates this on a toy example from Lyddon *et al.* (2018).

An interesting line of research would be developing similar methodology for adaptive MCMC sampling where the proposal or target distributions are updated as the algorithm runs (Roberts and Rosenthal, 2009; Pompe *et al.*, 2018). This is particularly important in the case of two-stage estimation procedures, where after obtaining $\theta_1^{(1)}, \ldots, \theta_1^{(N_1)}$ in the second stage we run N_1 chains, possibly in parallel. Since each chain targets a different distribution $\pi_2(\theta_2|\theta_1^{(i)})$, there are N_1 transition kernels to be tuned—this is why the efficiency of the algorithm could potentially be significantly improved by allowing for updating the proposal distributions on the fly.

C. P. Robert, G. Clarté, A. Hairault, C. Lawless and R. Ryder (*Université Paris Dauphine, Université Paris Sciences et Lettres and Ceremade, Paris*)

We congratulate Jacob, O'Leary and Atchadé for this excellent paper.

Associating Markov chain Monte Carlo (MCMC) sampling with unbiasedness is indeed quite a challenge since MCMC algorithms very rarely produce simulations from the exact target, unless specific tools like renewal control can be exploited in an efficient manner. Although renewal events are often easy to produce, renewal control is too rare an occurrence to consider it as a generic convergence assessment method (Guihenneuc-Jouyaux and Robert, 1998; Robert, 1998). Although coupling is also at the base of perfect simulation methods (Kendall and Møller, 2000; Huber, 2004), the analogy between this debiasing technique and perfect sampling is difficult to fathom, since the coupling of two chains is not a perfect sampling instant, occurring instead much earlier. When discussing the implementation of coupling in Metropolis and Gibbs settings, the authors propose a most simple optimal coupling algorithm which we were not aware of: a form of accept–reject sampling also found in a similar way in perfect sampling. (Something that is only obvious in retrospect is that the variance of the resulting unbiased estimator is at least the variance of the original MCMC estimator.)

We would, however, stress that unbiasedness is not the ultimate goal one should seek when running a Monte Carlo approximation, since the paper may otherwise give the impression of achieving a 'free-lunch' result. Reaching (exact) stationarity and exploring the posterior target within an allocated number of iterations was and remains the primary goal of an MCMC algorithm. Assessing both features is a considerable challenge that has not been solved by a foolproof technique, but coupling techniques have been proposed in the early days of MCMC methods towards this goal. In addition to Mykland *et al.* (1995), see, for example, Robert (1995, 1998), Guihenneuc-Jouyaux and Robert (1998), Hobert *et al.* (2002), Roberts and Rosenthal (2001), Hobert and Robert (2004) or Qin and Hobert (2019). Although renewal approaches include a burn-in period, this part of the simulation can be recycled thanks to the authors' technique and added to the independent and identically distributed errors terms in the renewal representation of ergodic averages. However, coupling is clearly working in the settings examined therein, whereas renewal apparently does not. In toy examples like the Efron and Morris (1973) baseball data and the Gelfand and Smith (1990) pump failure data, the parameters k and m of the algorithm can even be optimized against the variance of the averaged averages, which proves a remarkable feat.

Robin J. Ryder, Grégoire Clarté, Adrien Hairault, Caroline Lawless and Christian P. Robert (*Ceremade, Paris, Université Paris-Dauphine and Université Paris Sciences et Lettres*)

We congratulate Jacob, O'Leary and Atchadé for this excellent paper.

In 'traditional' MCMC methods, it is standard to check that stationarity has been attained by running a small number of parallel chains, initiated at different starting points, to verify that the final distribution is independent of the initialization—even though the single- *versus* multiple-chain(s) debate erupted initially with Gelman and Rubin (1992) *versus* Geyer (1992).

As noted by the authors, a bad choice of the initial distribution p_0 can lead to poor properties. In essence, this occurs and remains undetected for the current proposal because the coupling of the chains occurs long before the chain reaches stationarity. We make two suggestions to alleviate this issue, and hence add a stationarity check as a by-product of the run.

(a) The chains X and Y need to have the same initial distribution, but different pairs of chains on different parallel cores can afford different initial distributions. The resulting estimator remains unbiased. We would therefore suggest that parallel chains be initiated from distributions which put weight on different parts of the parameter space. Ideas from the quasi-Monte-Carlo literature (see Gerber and Chopin (2015)) could be used here.

(b) We also note that, although the marginal distributions of X and Y need to be identical, any joint distribution on (X, Y) produces an unbiased algorithm. We would suggest that it is preferable that X and Y meet (shortly) after the chains have reached stationarity. Here is one possible strategy for this: let p and p' be two distributions which put weight on different parts of the space, and $Z \sim \text{Bernoulli}(\frac{1}{2})$. If $Z = 0$, take $X_0 \sim p$ and $Y_0 \sim p'$; otherwise take $X_0 \sim p'$ and $Y_0 \sim p$. The marginal distribution of both X_0 and Y_0 is $\frac{1}{2}(p + p')$, but the two chains will start in different parts of the parameter space and are likely to meet after they have both reached stationarity.

The ideal algorithm is one which gives a correct answer when it has converged, and a warning or error when it has not. MCMC chains which have not yet reached stationarity (e.g. because they have not found all modes of a multimodal distribution) can be difficult to detect. Here, this issue is more likely to be detected since it would lead to the coupling not occurring: $\mathbb{E}[\tau]$ is large, and this is a feature, since it warns the practitioner that their kernel is ill fitted to the target density.

Leah F. South and Chris Nemeth (*Lancaster University*) **and Chris. J. Oates** (*Newcastle University, Newcastle upon Tyne, and Alan Turing Institute, London*)

Jacob, O'Leary and Atchadé are to be congratulated on an impressive and thought-provoking contribution to the field. Traditional Markov chain Monte Carlo (MCMC) sampling has benefitted from the development of gradient-based control variates (Assaraf and Caffarel, 1999; Barp *et al.*, 2018; Mira *et al.*, 2013; Oates *et al.*, 2017; South *et al.*, 2018), but it may be more difficult to design gradient-based control variates for unbiased MCMC sampling. Following the notation in the paper, let $\hat{\pi}^R(h) := (1/R)\Sigma_{r=1}^R H_{k:m}^{(r)}(X, Y)$ and $\pi(h) := \mathbb{E}[h(X)]$. Under assumptions 1–3, proposition 1 establishes that $\sigma(h)^2 := \mathbb{V}\{H_{k:m}(X, Y)\} < \infty$, so

$$\sqrt{R}\{\hat{\pi}^R(h) - \pi(h)\} \xrightarrow{\text{d}} N\{0, \sigma(h)^2\}$$

as $R \to \infty$. A control variate g should therefore be selected such that $\pi(g) = 0$ and $\sigma(h - g) \ll \sigma(h)$. In inequality (3.2) it was demonstrated that, in the large m and k limit, the quantity $\sigma(h)^2$ is just the asymptotic

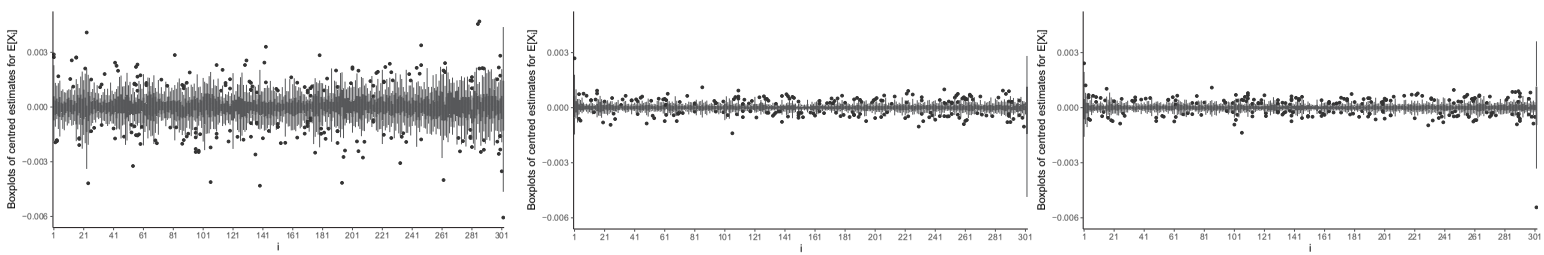

Fig. 14. Boxplots of 30 estimates of the marginal posterior expectations $\mathbb{E}[X_i]$ for $i = 1, \ldots, 302$ in the logistic regression example of Heng and Jacob (2019) (the following strategies were compared: (i) direct minimization of $\hat{\sigma}(h - g)$ and (ii) minimization of the bound (5) on $\sigma(h - g)$, with $\eta, \lambda \ll 1$, $\gamma \gg 1$ and π approximated with MCMC output; in each case g was a first-order Stein control variate (Mira *et al.*, 2013; South *et al.*, 2018) estimated by using the first $\lfloor R/2 \rfloor$ chains, whereas $\pi(h)$ was estimated by using the remaining $R - \lfloor R/2 \rfloor$ chains so that the estimators remain unbiased; runs are based on $k = 330$, $m = 3300$ and $R = 32$; the empirical means from approach (ii) are subtracted for visualization; the median variance reduction factor by using approaches (i) and (ii) is approximately 20; however, approach (i) depended strongly on the numerical approach that was used to minimize the non-convex objective function $\hat{\sigma}(h - g)^2$; code to reproduce the experiment is provided at https://github.com/LeahPrice/debiasedhmc): (a) unbiased MCMC sampling; (b) approach (i); (c) approach (ii)

variance from traditional MCMC sampling; existing gradient-based control variates can therefore be used (Belomestny *et al.*, 2019; Mijatović and Vogrinc, 2018). However, at finite m and k the dependence of $\sigma(h)$ on h is far from explicit. One could use sample splitting to construct an approximation of the form

$$\hat{\sigma}(h)^2 = \frac{1}{\lfloor R/2 \rfloor} \sum_{r=1}^{\lfloor R/2 \rfloor} \left\{ H_{k:m}^{(r)}(X, Y) - \frac{1}{\lfloor R/2 \rfloor} \sum_{r'=1}^{\lfloor R/2 \rfloor} H_{k:m}^{(r')}(X, Y) \right\}^2$$

and attempt to minimize $\hat{\sigma}(h-g)$. Alternatively, one could bound $\sigma(h-g)$ in terms of quantities that are independent of the Markov chain and then minimize the bound. One such bound is provided in the following result, stated for $k=m=0$ for simplicity, which we do not claim to be in any sense optimal.

Let assumptions 1–3 be satisfied, with η as in assumption 1 and C and δ as in assumption 2. Let π_t be the law of X_t and assume that $\lambda := \sup_{t \geqslant 0} d_{\mathrm{TV}}(\pi, \pi_t) < \infty$. Let $\mathcal{H}$ be a reproducing kernel Hilbert space, with norm denoted '$\|\cdot\|_{\mathcal{H}}$' and with kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying $K(x, x) \leqslant 1$ for all $x \in \mathcal{X}$. If $|h|^{2+\eta} \in \mathcal{H}$ then

$$\begin{aligned} \sigma(h) &\leqslant \gamma \{ \pi(|h|^{2+\eta}) + \lambda \| |h|^{2+\eta} \|_{\mathcal{H}} \}^{1/(2+\eta)} + \mathbb{E}[h(X_0)^2]^{1/2}, \\ \gamma^2 &= 4C^{\eta/(2+\eta)} \frac{\delta^{\eta/(2+\eta)}}{(1-\delta^{\eta/(4+2\eta)})^2}, \end{aligned} \tag{5}$$

where the positive constants γ and λ are h independent. The proof can be found in South *et al.* (2019). None of the three h-dependent quantities in the bound depend on the law of the Markov chain and thus minimization of the bound may be practical. The, in practice unknown, values of $\gamma(\delta)$ and λ determine which of the three terms dominate the bound.

Fig. 14 displays the variance reduction that is achieved in the 302-dimensional logistic regression example of Heng and Jacob (2019). These results are encouraging, but more work is required to develop an understanding of gradient-based control variates for unbiased MCMC sampling.

Paul Vanetti and Arnaud Doucet (*University of Oxford*)

Bias correction for Markov chain Monte Carlo sampling is a long-standing problem, for which this paper represents a breakthrough. We propose two generalizations to derive new unbiased estimators.

First we propose the *lagged estimator*, obtained by starting chain Y after L steps of the chain X and coupling those two chains such that $X_t = Y_{t-L}$ for all $t \geqslant \tau$ where τ is the meeting time. For some N, k such that $N \geqslant k + L$, we can then exploit the identity

$$\frac{1}{L} \sum_{n=N-L+1}^{N} \mathbb{E}[h(X_n)] = \frac{1}{L} \left[\sum_{i=k}^{k+L-1} \mathbb{E}[h(X_i)] + \sum_{j=k+L}^{N} \{ \mathbb{E}[h(X_j)] - \mathbb{E}[h(Y_{j-L})] \} \right]. \tag{6}$$

Lags greater than 1 were used in Biswas *et al.* (2019) to improve probability metric bounds. The estimator (6) is similar to the time-averaged estimator in the paper under discussion when $L = m - k$, but the bias correction term that is incurred is not inflated by a large coefficient. Our empirical results (Table 11)

Table 11. Standard deviation for (a) the time-averaged estimator (with lag 1) and (b) the lagged estimator (with $L = m - k$)†

k	m	σ	k	L	σ
(*a*)			(*b*)		
1	10	430	1	9	66.7
10	100	34.2	10	90	11.9
100	1000	0.119	100	900	0.119

†The underlying Markov chain Monte Carlo chain is a Metropolis random walk with target $\pi(x) = \mathcal{N}(x; 0, 1)$, initialization $\pi_0(x_0) = \mathcal{N}(x_0; 0, 5^2)$, proposal $q(x, x^*) = \mathcal{N}(x^*; x, 1)$ and test function $h(x) = x^2$. σ is the standard deviation of a single estimate.

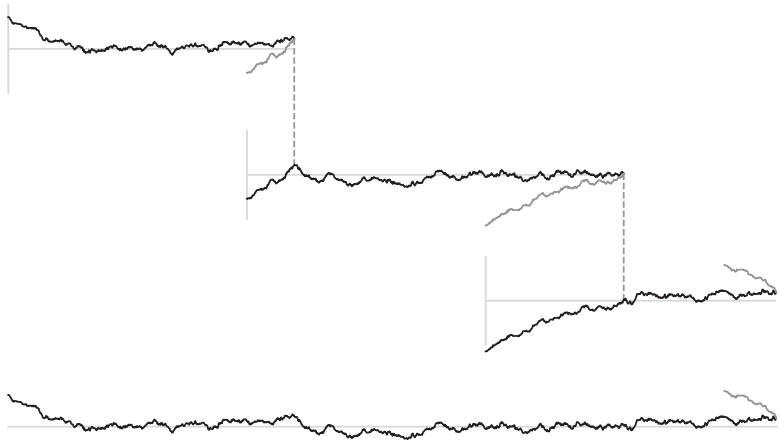

Fig. 15. Sharing chains between pairs (the top three curves are three pairs; the X-chain of the second is used as the Y-chain for the first, and similarly the X-chain of the third for the Y-chain of the second; the final curve is the aggregate chain): ———, positive contributions to the total estimate; ———, negative contributions

in a simple scenario suggest that estimator (8) can outperform the time-averaged estimator at a similar computational cost.

For L sufficiently large, where X_L is approximately stationary, the bias correction term of estimator (8) can be interpreted as a removal of the burn-in, representing the difference between a stationary chain and the first iterations of a new chain. This motivates our second innovation, which is to use the X-chain of each pair of coupled chains as the Y-chain for another pair.

If we simulate R pairs $(X^{(r)}, Y^{(r)})$, using the chain $Y_i^{(r)} = X_i^{(r+1)}$ for $r \in \{1, \ldots, R-1\}$, and for $r = R$ we take a novel chain (and not the X for any other pair), then averaging the estimates over the R pairs yields an estimator in which the negative bias correction terms for the first $R-1$ pairs cancel with the first samples from the next pair. Thus the estimator is equivalent to that obtained by a single long chain with a lag choice of RL; see Fig. 15 for an illustration.

If we had used $Y_i^{(R)} = X_i^{(1)}$ for the Rth pair, the resulting scheme would very closely match the parallel implementation of circular coupling (Neal, 1999). The framework that is provided in this discussion gives a new interpretation to this scheme, which was designed to provide 'states that all have close to the equilibrium distribution'. A natural question to ask is whether it yields unbiased estimates and, if this is not so, whether it can be modified to achieve exact unbiasedness.

Dootika Vats (*Indian Institute of Technology Kanpur*) **and Galin L. Jones** (*University of Minnesota, Twin Cities*)

We congratulate Jacob, O'Leary and Atchadé on considerably furthering the cause of unbiased estimation in Markov chain Monte Carlo sampling. Although there is often a need for the methods discussed in the paper, unbiased estimation is trivially achieved by successfully starting from stationarity. The authors mention coupling-based perfect simulation techniques as a means to draw exact samples from the target distribution, but we note that simple accept–reject samplers are useful in surprisingly many examples. Accept–reject samplers can be inefficient for high dimensional target distributions but, even there, in conjunction with *linchpin* variables (Archila, 2016; Huber, 2016), reasonably efficient accept–reject samplers can be used to produce one exact draw from the target.

For example, consider the model in Section 5.2 used to study the nuclear pump failure data. The joint posterior density of $\lambda = (\lambda_1, \ldots, \lambda_K)$ and β can factorized as

$$f(\lambda, \beta) = f(\lambda|\beta)\, f(\beta).$$

Since it is easy to draw samples from $f(\lambda|\beta)$, β is a linchpin variable. Accept–reject samples from the univariate marginal posterior of β are easier to obtain than accept–reject samples from the $(K+1)$-dimensional joint posterior of (λ, β). In fact, a simple and efficient accept–reject sampler that targets $f(\beta)$ is easy to construct. Samples from the joint posterior can then be obtained via sequential sampling.

Based on a similar linchpin construction, Jones *et al.* (2006) presented an efficient accept–reject sampler for the baseball batting averages data and model presented in section S3. If the number of covariates is small, efficient accept–reject samplers are also possible for the Bayesian lasso posterior in section S5 and the variable selection model in section 5.2. Archila (2016) presented a few other hierarchical models where efficient accept–reject samplers can be implemented via the linchpin variable approach.

In many situations, accept–reject sampling is too computationally burdensome for a full Monte Carlo procedure, but it may be able to provide *one* draw at a reasonable cost. When this draw is used as a starting value in a Markov chain, the usual Monte Carlo estimator of *any* expectation is trivially unbiased. This procedure is easily parallelizable and requires no burn-in. However, in settings where direct sampling is inefficient, the unbiased estimation techniques of Jacob, O'Leary and Atchadé are promising.

The **authors** replied later, in writing, as follows.

We are grateful to the discussants for their stimulating, creative and insightful comments. In light of the many methodological enhancements and new applications presented here, we join Craiu and Meng in feeling hopeful about the future of Markov chain Monte Carlo (MCMC) methods.

Ease of use

We were glad to read some confirmation that the method proposed is simple to implement. Various discussions (Wilkinson; Sherlock; Chopin; Gerber and Lee; Pompe and Holmes; South, Nemeth and Oates; Vanetti and Doucet) involved original numerical experiments illustrating different aspects of the method, possible improvements and new applications.

Middleton, Deligiannidis and Doucet emphasize that the approach proposed does not involve random truncation variables, in contrast with for example Glynn and Rhee (2014) and Agapiou *et al.* (2018). The construction of such variables can require hard-to-determine quantities, such as the contraction rate associated with a given coupling. Interested readers might consult Agapiou *et al.* (2018) for an example of estimation of such a contraction rate, in the case of a coupled Metropolis–Hastings algorithm on a logistic regression with three covariates. Removing the need for truncation variables by considering chains that meet exactly appears both as a straightforward modification of Glynn and Rhee (2014) and as a step forward in terms of applicability.

Parallelization goals and limitations

One motivation for unbiased estimation is its amenability to parallel computing: a goal shared by regeneration techniques (Mykland *et al.*, 1995; Brockwell and Kadane, 2005), circularly coupled MCMC sampling (Neal, 1999) and perfect samplers (e.g. Lee *et al.* (2014)).

The detailed discussion of Wilkinson illustrates the different goals one might have in combining parallel computing and Monte Carlo methods. At first, *Amdahl's law* and the results summarized in Section 3.3 of the paper, extracted from for example Glynn and Heidelberger (1991), may seem to contradict one another. Whereas the former suggests diminishing returns in the number of processors P, the central limit theorem in equation (3.4) indicates steady improvement at the standard Monte Carlo rate as $P \to \infty$. This apparent paradox can be resolved by introducing a certain amount of context and by distinguishing between different goals.

Let us take the perspective of a user who can run a Monte Carlo experiment on P machines, yielding an estimator with precision α, in run time T. We summarize the performance by the pair (α, T). Next, suppose that the user has access to $2P$ machines. What are our user's hopes? One option would be to achieve twice the precision while keeping the run time close to T, for performance $(2\alpha, T)$. Alternatively, the user may hope to produce an equally accurate result but in half the time, achieving performance $(\alpha, T/2)$.

The class of estimators that is considered in the paper achieves the first type of improvement in most settings, but it achieves the second type of improvement only if the time budget is sufficiently large to begin with. We believe that, irrespective of the Monte Carlo method under consideration, the second type of gain is realistic only when starting from a moderate or large run time T. The presence of incompressible computing costs and the intuition that any Monte Carlo algorithm must entail a baseline amount of serial computation would seem to limit the potential for speed-ups when T is small to begin with.

To illustrate this discussion with a set of simple experiments, consider pairs of variables (H, C) defined as $H \sim \text{Exp}(1)$ and $C = 1 + H$. Suppose that we are interested in the estimation of the expectation of H, which we think of as some unbiased estimator, and that C represents the time cost of computing H. We consider the estimator $\tilde{H}_p(T)$ generated on each processor p by averaging independent copies of H within the time budget T and waiting for at least one copy to be completed, as in Section 3.3 of the paper. The estimator

Table 12. For various specified budgets T and number of processors P, each cell contains the pair (precision, expected completion time)

T	$P=128$	$P=256$	$P=512$	$P=1024$
6	187.2, 6.77	376.5, 7.18	694.2, 7.8	1486.9, 8.52
12	547.7, 12.00	1025.3, 12.00	2190.4, 12.01	3960.9, 12.01
24	1312.5, 24.00	2649.1, 24.00	5013.6, 24.00	10295.9, 24.00

on processor p has completion time $C_p(T) = \max(T, C_{p,1})$ where $C_{p,1}$ is the cost of the first estimator. The final estimator is the average $P^{-1}\Sigma_{p=1}^{P}\tilde{H}_p(T)$. Its expected completion time $\mathcal{C}(P,T)$ is the expectation of $\max\{C_1(T),\ldots,C_P(T)\}$, and we take its precision $\alpha(P,T)$ to be the inverse of its variance.

In Table 12 we report the precision $\alpha(P,T)$ and completion time $\mathcal{C}(P,T)$ for various $P \in \{128, 256, 512, 1024\}$ and $T \in \{6, 12, 24\}$, based on 10^3 independent runs. We have the following observations.

(a) When we set the time budget T to 6 s, the estimators take on average between 6 and 9 s to complete, with a visible increase with P. If we had set the budget to 0, we would have still observed approximately the same expected completion times. Thus, in that regime increasing the number of processors increases the run time. When we set the budget to 12 or 24 s, the expected completion times become nearly constant and equal to the budget; the actual increase is logarithmic in P.

(b) The precision increases linearly with P, on every row of the table, but there are some noticeable Monte Carlo variations. By comparing two consecutive cells of any row the precision approximately doubles whereas the completion time increases very slowly. When T is sufficiently large, doubling the number of processors doubles the precision while keeping the expected completion time essentially fixed.

(c) In contrast, if we compare for example $T=24$ and $P=512$, and $T=12$ and $P=1024$, we see that the run time is indeed halved, but that some precision is lost. We believe that this corresponds to Amdahl's law in Wilkinson's discussion. This effect would eventually disappear as the budget T increases.

Similarly, we believe that comparisons between unbiased MCMC estimators and 'one long run' of an MCMC algorithm would be most informative with adequate amounts of context, including the number of processors, budget constraints, the desired precision, the choice of k and m for the proposed estimators and the choice of burn-in for the MCMC algorithm.

Theory on meeting times

Several discussants highlight the importance of understanding the behaviour of the meeting time in the estimators proposed. Indeed the meeting time drives both the cost of the estimator and its variance via the bias correction term.

A few results on the meeting times or closely related objects already exist. In their comment, Lee, Singh and Vihola summarize their analysis of couplings of conditional particle filters, including precise results on the effect of the number of observations and particles on the meeting times. Paulin's comment recalls that much progress has been made on the understanding of standard MCMC methods in recent years; we would like to add that many of these works have employed coupling strategies, such as Diaconis *et al.* (2010) and Qin and Hobert (2019) for Gibbs samplers, or Mangoubi and Smith (2017) and Bou-Rabee *et al.* (2018) for Hamiltonian Monte Carlo sampling.

We hope that similar techniques will provide a precise understanding of the cost and variance of the proposed estimators in the future. More discussion in the context of Hamiltonian Monte Carlo sampling can be found in Heng and Jacob (2019). We also note that the rich literature around coupling from the past (Propp and Wilson, 1996) contains theoretical results on meeting times. For example, Collevecchio *et al.* (2018) found the asymptotic distribution of a rescaled meeting time to be Gumbel in the context of a Fortuin–Kasteleyn model.

Variance reduction

It is encouraging to see the enthusiasm and creativity of the discussants around variance reduction tech-

niques. Indeed these techniques directly improve the estimators proposed and the discussants have managed to come up with control variates, Rao–Blackwellization and antithetic variables.

Vanetti and Doucet propose a generalization of the proposed estimators by using a lag $L \geqslant 1$ between the chains. In this approach, one samples $X_0, X_1, \ldots, X_L$ by using the original MCMC algorithm with kernel P, Y_0 from π_0, and then recursively X_{t+1} and Y_{t-L+1} given X_t and Y_{t-L} from $\bar{P}$. The meeting time τ can be defined as $\inf\{t \geqslant 1 : X_t = Y_{t-L}\}$. Generalizing equation (2.1) in the paper, we write the lagged, time-averaged estimator as

$$H_{k:m}^{(L)} = \frac{1}{m-k+1} \sum_{l=k}^{m} h(X_l) + \sum_{l=k+L}^{\tau-1} \min\left\{1, \frac{\lceil (l-k)/L \rceil}{m-k+1}\right\} \{h(X_l) - h(Y_{l-L})\}.$$

Using a value of L that is larger than 1 is always possible and can yield a large improvement in efficiency. We thank Vanetti and Doucet for this important generalization. We encourage users to implement such lagged estimators and to experiment with the choice of L.

Craiu and Meng propose an equally general variance reduction technique in the form of a control variate exploiting the fact that X_t and Y_t have the same marginal distribution for any fixed t. We look forward to their upcoming paper on the topic. Sherlock proposes a Rao–Blackwellization method for coupled Metropolis–Hastings chains, which leverages an estimate of the probability of coupling at the next step conditionally on the current states. South, Nemeth and Oates propose gradient-based control variates—adapted from those recently developed for standard MCMC estimators (South *et al.*, 2018)—with promising preliminary results in a logistic regression setting. Finally Chak, Kantas and Pavliotis envision an approach that is reminiscent of antithetic variables, where different pairs of Markov chains would be constructed to produce a negative correlation in the resulting unbiased estimators. Such couplings might be generated by adapting techniques from Nüsken and Pavliotis (2019).

Cost reduction

With bias out of the picture, the variance and average cost of the estimators proposed constitute their most notable performance characteristics. Thus, it is important to develop a better understanding of how to reduce the cost as well as the variance. Selecting the most efficient MCMC kernel available for the problem at hand is obviously recommended. Then, for a fixed MCMC kernel P, reducing the cost amounts to improving the coupling strategy.

Chopin provides numerical results in the case of the Gibbs sampler of Albert and Chib (1993), emphasizing that various strategies are possible for each update of the Gibbs sampler. In more complicated Gibbs samplers, the possibilities quickly multiply, and thus it would be useful to develop general guidelines for users. In our experience, 'common random numbers' provide a good starting point, since these are simple to implement and often yield chains that contract towards one another to at least some extent. Generating enough contraction to bring the chains close together is the main challenge, since other techniques, such as the maximal couplings that were used extensively in this paper, can be used to obtain exact meetings when the chains are already close; see Heng and Jacob (2019) for example.

The parameterization of a Gibbs sampler will also certainly impact not only its marginal performance (Papaspiliopoulos and Roberts, 2008) but also the implementation and performance of coupling strategies. The comment of Gerber and Lee regarding parameterization with Bernstein–von Mises heuristics is helpful in that regard, as well as the comment of Vats and Jones regarding the potential presence of low dimensional *linchpin* variables.

New applications

Some comments pointed out the potential use of the proposed estimators in new applications, which we find exciting.

We look forward to reading more about the work described by Jasra, Law and Heng, and note that it will be an interesting methodological application of multimarginal couplings. There might be connections with the comment of Chak, Kantas and Pavliotis where anticorrelations between pairs of chains would result from four-way couplings.

The application described in the comment by Pompe and Holmes, following up on the 'cut distribution' of Section 5.5, is very exciting to us as it leverages another appeal of unbiased estimators, beyond parallel computation. We agree with Pompe and Holmes that it would be useful to obtain unbiased estimators from adaptive MCMC methods, but at this point this appears to be a challenging open question.

The discussion of Gu raises the question of the applicability of the proposed framework to missing data and latent variable models. We refer to the comments of Lee, Singh and Vihola and of Middleton, Deligiannidis and Doucet, as well as references therein, which concern the smoothing problem in hidden

Markov models or state space models. These provide examples of unbiased estimation strategies that have proved effective in high dimensional latent variable models.

Bayesian computation

In response to the comments of Chai and Gu, we emphasize that the methods that are considered in the paper are not specifically Bayesian, in the sense that they do not require the conceptualization of unknown quantities as random variables on which prior distributions would be specified. For Bayesian methods in numerical analysis, see Diaconis (1988), Bernardo *et al.* (1992) or more recently Cockayne *et al.* (2019). The term 'Bayesian computation' can refer either to computational methods that are used in Bayesian statistics or to Bayesian methods to perform computation, which seems to foster some confusion. This paper concerns non-Bayesian computational methods which might be useful in Bayesian statistics and elsewhere.

Initialization

The discussion of Ryder, Clarté, Hairault, Lawless and Robert proposes an interesting strategy for initializing the chains. We share their view that large meeting times can sometimes indicate convergence issues with the underlying Markov kernel, which we have experienced ourselves, and can indeed be seen as a feature rather than a bug.

We note that using different initial distributions for different runs result in estimators that are unbiased but not identically distributed, and thus care would be advised for the construction of confidence intervals. In contrast, π_0 can be chosen to be a mixture of various components and X_0 and Y_0 can indeed be drawn from any coupling of π_0 with itself. Thus one could for instance maximize the probability that X_0 and Y_0 are drawn from different components of the same mixture.

The lagged estimator that is proposed in the comment of Vanetti and Doucet has implications for the selection of an initialization strategy. As the lag increases, the issue of small meeting times that can occur when two chains start nearby will tend to disappear. Thus users might find lagged estimators with $L > 1$ to be safer in general, as was found for the estimation of total variation upper bounds in Biswas *et al.* (2019).

Perfect sampling

Various discussants noted the connections and distinctions between our method and perfect samplers, which often rely on couplings of Markov chains (Propp and Wilson, 1996; Glynn, 2016). As noted in Robert, Clarté, Hairault, Lawless and Ryder the relationship between perfect sampling and unbiased estimation is indeed elusive. It is unclear whether the machinery of this paper bring us any closer to perfect sampling.

Vats and Jones remind us that perfect samplers might be easier to implement than commonly thought. Similarly, we can wonder why regeneration approaches such as Brockwell and Kadane (2005) are not used more often. There might remain an important gap between the generic applicability of MCMC algorithms such as Metropolis–Hastings algorithms and that of current perfect samplers. Perhaps the work advertised by Wang, Pollock, Roberts and Steinsaltz will help to close that gap. We speculate that, even if perfect samplers were to become fully generic, there would still be room for MCMC strategies, unbiased or not, if they come at a smaller computational cost.

References in the discussion

Agapiou, S., Roberts, G. O. and Vollmer, S. J. (2018) Unbiased Monte Carlo: posterior estimation for intractable/infinite-dimensional models. *Bernoulli*, **24**, 1726–1786.

Albert J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 699–679.

Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc.* B, **72**, 269–342.

Archila, F. H. A. (2016) Markov chain Monte Carlo for linear mixed models. *PhD Thesis*. University of Minnesota.

Assaraf, R. and Caffarel, M. (1999) Zero-variance principle for Monte Carlo algorithms. *Phys. Rev. Lett.*, **83**, 46–82.

Barp, A., Oates, C., Porcu, E. and Girolami, M. (2018) A Riemannian-Stein kernel method. *Preprint arXiv: 1810.04946*. Imperial College London, London.

Belomestny, D., losipoi, L., Moulines, E., Naumov, A. and Samsonov, S. (2019) Variance reduction for Markov chains with application to MCMC. *Preprint arXiv:1910.03643*. Duisburg–Essen University.

Bernardo, J., Berger, J., Dawid, A. and Smith, A. (1992) Some Bayesian numerical analysis. *Baysn Statist.*, **4**, 345–363.

Beskos, A., Jasra, A., Law, K. J., Tempone, R. and Zhou, Y. (2017) Multilevel sequential Monte Carlo samplers. *Stoch. Processes Appl.*, **127**, 1417–1440.
Biswas, N., Jacob, P. E. and Vanetti, P. (2019) Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems* (eds H. Wallach, H. Larochelle, A. Beygelzimer, F. Atchadé-Buc, E. Fox and R. Garnett), pp. 7389–7399. Red Hook: Curran Associates.
Bladt, M., Finch, S. and Sørensen, M. (2014) Simulation of multivariate diffusion bridge. *J. R. Statist. Soc.* B, **78**, 343–369.
Bou-Rabee, N., Eberle, A. and Zimmer, R. (2018) Coupling and convergence for Hamiltonian Monte Carlo. *Preprint arXiv:1805.00452*. Department of Mathematical Sciences, Rutgers University, Camden.
Brockwell, A. E. and Kadane, J. B. (2005) Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *J. Computnl Graph. Statist.*, **14**, 436–458.
Chen, Y., Dwivedi, R., Wainwright, M. J. and Yu, B. (2019) Fast mixing of Metropolized Hamiltonian Monte Carlo: benefits of multi-step gradients. *Preprint arXiv:1905.12247*.
Chopin, N. and Singh, S. S. (2015) On particle Gibbs sampling. *Bernoulli*, **21**, 1855–1883.
Cockayne, J., Oates, C. J., Sullivan, T. and Girolami, M. (2019) Bayesian probabilistic numerical methods. *SIAM Rev.*, **61**, 756–789.
Collevecchio, A., Elçi, E. M., Garoni, T. M. and Weigel, M. (2018) On the coupling time of the heat-bath process for the Fortuin–Kasteleyn random-cluster model. *J. Statist. Phys.*, **170**, 22–61.
Craiu, R. V. and Lemieux, C. (2007) Acceleration of the multiple-try Metropolis algorithm using antithetic and stratified sampling. *Statist. Comput.*, **17**, 109.
Craiu, R. V. and Meng, X. L. (2001) Antithetic coupling for perfect sampling. In *Baysesian Methods, with Applications to Science, Policy and Official Statistics—Selected Papers from ISBA 2000* (ed. E. I. George).
Craiu, R. V. and Meng, X. L. (2005) Multiprocess parallel antithetic coupling for backward and forward Markov chain Monte Carlo. *Ann. Statist.*, **33**, 661–697.
Craiu, R. V. and Meng, X. L. (2020) Double Happiness: Enhancing the coupled gains of *L*-lag coupling via control variates. To be published.
Dai, H., Pollock, M. and Roberts, G. (2019) Monte Carlo fusion. *J. Appl. Probab.*, **56**, 174–191.
Dean, J. and Ghemewat, S. (2008) Mapreduce: simplified data processing on large clusters. *Communs ACM*, **51**, 107–113.
Del Moral, P., Doucet, A. and Jasra, A. (2006) Sequential Monte Carlo samplers. *J. R. Statist. Soc.* B, **68**, 411–436.
Diaconis, P. (1988) Bayesian numerical analysis. In *Statistical Decision Theory and Related Topics IV*, vol. 1 (eds S. S. Gupta and J. O. Berger), pp. 163–175. New York: Springer.
Diaconis, P., Khare, K. and Saloff-Coste, L. (2010) Gibbs sampling, conjugate priors and coupling. *Sankhya* A, **72**, 136–169.
Dwivedi, R., Chen, Y., Wainwright, M. J. and Yu, B. (2019) Log-concave sampling: Metropolis-Hastings algorithms are fast. *J. Mach. Learn. Res.*, **20**, 1–42.
Efron, B. and Morris, C. (1973) Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Am. Statist. Ass.*, **68**, 117–130.
Gelfand, A. and Smith, A. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.
Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.
Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, no. 6, 721–741.
Gerber, M. and Chopin, N. (2015) Sequential quasi Monte Carlo (with discussion). *J. R. Statist. Soc.* B, **77**, 509–579.
Geyer, C. J. (1992) Practical Markov chain Monte Carlo. *Statist. Sci.*, **7**, 473–483.
Giles, M. B. (2008) Multilevel Monte Carlo path simulation. *Ops Res.*, **56**, 607–617.
Glynn, P. W. (2016) Exact simulation versus exact estimation. In *Proc. Winter Simulation Conf.* (eds T. Huschka and S. E. Chick), pp. 193–205. New York: Institute of Electrical and Electronics Engineers.
Glynn, P. W. and Heidelberger, P. (1991) Analysis of parallel replicated simulations under a completion time constraint. *ACM Trans. Modlng Comput. Simulns*, **1**, 3–23.
Glynn, P. W. and Rhee, C.-H. (2014) Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.* A, **51**, 377–389.
Gray, K., Hampton, B., Silveti-Falls, T., McConnell, A. and Bausell, C. (2015) Comparison of Bayesian credible intervals to frequentist confidence intervals. *J. Mod. Appl. Statist. Meth.*, **14**, 8.
Guihenneuc-Jouyaux, C. and Robert, C. P. (1998) Finite Markov chain convergence results and MCMC convergence assessments. *J. Am. Statist. Ass.*, **93**, 1055–1067.
Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
Heng, J. and Jacob, P. E. (2019) Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, **106**, 287–302.
Hoff, P. D. (2009) *A First Course in Bayesian Statistical Methods*. New York: Springer.
Hobert, J. P., Jones, G. L., Presnell, B. and Rosenthal, J. S. (2002) On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika*, **89**, 731–743.

Hobert, J. P. and Robert, C. P. (2004) A mixture representation of π with applications in Markov chain Monte Carlo and perfect sampling. *Ann. Appl. Probab.*, **14**, 1295–1305.
Huber, M. (2004) Perfect sampling using bounding chains. *Ann. Appl. Probab.*, **14**, 734–753.
Huber, M. L. (2016) *Perfect Simulation*. Boca Raton: Chapman and Hall–CRC.
Jacob, P. E., Lindsten, F. and Schön, T. B. (2020) Smoothing with couplings of conditional particle filters. *J. Am. Statist. Ass.*, to be published.
Jacob, P. E., Murray, L. M., Holmes, C. C. and Robert, C. P. (2017) Better together?: Statistical learning in models made of modules. *Preprint arXiv:1708.08719*. Department of Statistics, Harvard University, Cambridge.
Jasra, A., Kamatani, K., Law, K. J. H. and Zhou, Y. (2017) Multilevel particle filters. *SIAM J. Numer. Anal.*, **55**, 3068–3096.
Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006) Fixed-width output analysis for Markov chain Monte Carlo. *J. Am. Statist. Ass.*, **101**, 1537–1547.
Kendall, W. S. and Møller, J. (2000) Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Adv. Appl. Probab.*, **32**, 844–865.
Lee, A., Doucet, A. and Łatuszyński, K. (2014) Perfect simulation using atomic regeneration with application to sequential Monte Carlo. *Preprint arXiv:1407.5770*. University of Bristol, Bristol.
Lee, A., Singh, S. S. and Vihola, M. (2020) Coupled conditional backward sampling particle filter. *Ann. Statist.*, to be published.
Liu, F., Bayarri, M. J. and Berger, J. O. (2009) Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Baysn Anal.*, **4**, 119–150.
Lyddon, S., Walker, S. and Holmes, C. C. (2018) Nonparametric learning from Bayesian models with randomized objective functions. In *Advances in Neural Information Processing Systems* (eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett), pp. 2071–2081. Red Hook: Curran Associates.
Mangoubi, O. and Smith, A. (2017) Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *Preprint arXiv:1708.07114*. Worcester Polytechnic Institute, Worcester.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.
Middleton, L., Deligiannidis, G., Doucet, A. and Jacob, P. E. (2018) Unbiased Markov chain Monte Carlo for intractable target distributions. *Preprint arXiv:1807.08691*. Department of Statistics, University of Oxford, Oxford.
Middleton, L., Deligiannidis, G., Doucet, A. and Jacob, P. E. (2019) Unbiased smoothing using particle independent Metropolis-Hastings. In *Proc. 22nd Int. Conf. Artificial Intelligence and Statistics* (eds K. Chaudhuri and M. Saugiyama), pp. 2378–2387.
Mijatović, A. and Vogrinc, J. (2018) On the Poisson equation for Metropolis–Hastings chains. *Bernoulli*, **24**, 2401–2428.
Mira, A., Solgi, R. and Imparato, D. (2013) Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statist. Comput.*, **23**, 653–662.
Mykland, P., Tierney, L. and Yu, B. (1995) Regeneration in Markov chain samplers. *J. Am. Statist. Ass.*, **90**, 233–241.
Neal, R. M. (1999) Circularly-coupled Markov chain sampling. *Preprint arXiv:1711.04399*. Department of Statistics, University of Toronto, Toronto.
Neal, R. M. (2001) Annealed importance sampling. *Statist. Comput.*, **11**, 125–139.
Neiswanger, W., Wang, C. and Xing, E. (2013) Asymptotically exact, embarrassingly parallel MCMC. *Preprint arXiv:1311.4780*.
Nüsken, N. and Pavliotis, G. (2019) Constructing sampling schemes via coupling: Markov semigroups and optimal transport. *SIAM J. Uncertnty Quant.*, **7**, 324–382.
Oates, C. J., Girolami, M. and Chopin, N. (2017) Control functionals for Monte Carlo integration. *J. R. Statist. Soc.* B, **79**, 695–718.
Papaspiliopoulos, O. and Roberts, G. (2008) Stability of the Gibbs sampler for Bayesian hierarchical models. *Ann. Statist.*, **38**, 95–117.
Plummer, M. (2015) Cuts in Bayesian graphical models. *Statist. Comput.*, **25**, 37–43.
Pompe, E., Holmes, C. and Łatuszyński, K. (2018) A framework for adaptive MCMC targeting multimodal distributions. *Preprint arXiv:1812.02609*. University of Oxford, Oxford.
Propp, J. G. and Wilson, D. B. (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Rand. Struct. Algs*, **9**, 223–252.
Qin, Q. and Hobert, J. P. (2018) Wasserstein-based methods for convergence complexity analysis of MCMC with application to Albert and Chib's algorithm. *Preprint arXiv:1810.08826*.
Qin, Q. and Hobert, J. P. (2019) Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression. *Ann. Statist.*, **47**, 2320–2347.
Rendell, L. J., Johansen, A. M., Lee, A. and Whiteley, N. (2018) Global consensus Monte Carlo. *Preprint arXiv:1807.09288*. University of Warwick, Coventry.
Rhee, C. H. and Glynn, P. W. (2015) Unbiased estimation with square root convergence for SDE models. *Ops Res.*, **63**, 1026–1043.

Robert, C. P. (1995) Convergence control techniques for MCMC algorithms. *Statist. Sci.*, **10**, 231–253.
Robert, C. P. (1998) *Discretization and MCMC Convergence Assessment*. New York: Springer.
Roberts, G. and Rosenthal, J. S. (2009) Examples of adaptive MCMC. *J. Computnl Graph. Statist.*, **18**, 349–367.
Roberts, G. O. and Rosenthal, J. S. (2001) Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, **16**, 351–367.
Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E. (2016) Bayes and big data: the consensus Monte Carlo algorithm. *Int. J. Mangmnt Sci. Engng Mangmnt*, **11**, 78–88.
South, L. F., Nemeth, C. and Oates, C. J. (2019) Discussion of "Unbiased Markov chain Monte Carlo with couplings" by Pierre E. Jacob, John O'Leary and Yves F. Atchadé. *Preprint arXiv:1912.10496*. Lancaster University, Lancaster.
South, L. F., Oates, C. J., Mira, A. and Drovandi, C. (2018) Regularised zero-variance control variates for high-dimensional variance reduction. *Preprint arXiv:1811.05073*. Queensland University of Technology, Brisbane.
Tierney, L. (1994) Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701–1728.
Van Dyk, D. A. and Meng, X. L. (2001) The art of data augmentation (with discussion). *J. Computnl Graph. Statist.*, **10**, 1–50.
Vihola, M. (2018) Unbiased estimators and multilevel Monte Carlo. *Ops Res.*, **66**, 448–462.
Wang, A. Q., Pollock, M., Roberts, G. O. and Steinsaltz, D. (2019) Regeneration-enriched Markov processes with application to Monte Carlo. *Preprint*. (Available from `http://arxiv.org/abs/1910.05037`.)
Whiteley, N. (2010) Discussion on 'Particle Markov chain Monte Carlo methods', by C. Andrieu, A. Doucet and R. Holenstein. *J. R. Statist. Soc.* B, **72**, 306–307.
Wilkinson, D. J. (2005) Parallel Bayesian computation. In *Handbook of Parallel Computing and Statistics* (ed. E. J. Kontoghiorghes), ch. 16, pp. 481–512. Boca Raton: CRC Press.
Yu, Y. and Meng, X. L. (2011) To center or not to center: That is not the question—an Ancillarity–Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency (with discussion). *J. Computnl Graph. Statist.*, **20**, 531–570.
Zigler, C. M. (2016) The central role of Bayes' theorem for joint estimation of causal effects and propensity scores. *Am. Statistn*, **70**, 47–54.