Rao-Blackwellised parallel MCMC

Tobias Schwedes*
Department of Mathematics,
Imperial College London,

Ben Calderhead
Department of Mathematics,
Imperial College London

Abstract

Multiple proposal Markov chain Monte Carlo (MP-MCMC) as introduced by Calderhead (2014) allow for computationally efficient and parallelisable inference, whereby multiple states are proposed and computed simultaneously. In this paper, we improve the resulting integral estimators by sequentially using the multiple states within a Rao-Blackwellised estimator. We further propose a novel adaptive Rao-Blackwellised MP-MCMC algorithm, which generalises the adaptive MCMC algorithm introduced by Haario et al. (2001) to allow for multiple proposals. We prove its asymptotic unbiasedness, and demonstrate significant improvements in sampling efficiency through numerical studies.

1 Introduction

Markov chain Monte Carlo methods still act in practice as the workhorse for performing Bayesian inference over sophisticated mathematical models (Foreman-Mackey et al., 2013; Martin et al., 2011). As model complexity and data volume increases, so it becomes more important to develop scalable and computationally efficient approaches for asymptotically exact inference. Multiple proposal MCMC offers an enticing way forward by allowing for a parallelisable MCMC framework, whereby multiple proposals may be made and computed in parallel before being subsampled in such a way that the correct stationary distribution is targeted (Calderhead, 2014). A straightfoward extension of this involves the construction of a weighted estimator that makes use of all proposed states; this may simply be considered as a Rao-Blackwellised version.

Rao-Blackwellisation can be seen as a version of the Waste-Recycling method by Frenkel (2006) or similar (Ceperley et al., 1977; Tjelmeland, 2004; Frenkel, 2004; Delmas and Jourdain, 2009; Yang et al., 2018); its name arising due to the fact that every proposal is used, including the ones rejected by MCMC. For instance, Tjelmeland (2004) proposes to take a weighted mean between several proposals in Metropolis-Hastings which otherwise would have been discarded except the single accepted proposal. Delmas and Jourdain (2009) propose a control variates approach, in which the correction term uses all proposed states of a Metropolis algorithm. Yang et al. (2018) suggest an alternative, slightly more complex algorithm inspired by that presented in (Calderhead, 2014) and apply Waste-Recycling to construct a "locally-weighted" estimator.

In this paper, we take the original algorithm by Calderhead (2014) as our starting point. We develop the theoretical justification for such a construction and carefully prove that it targets the correct stationary distribution; noting that this is not straightforward as we are essentially defining a Markov chain that operates on a product space, and thus the balance and detailed balance conditions must be defined accordingly.

We then proceed to elucidate the application of Rao-Blackwellisation to such an algorithm in Section 3, using a different heuristic to that in the Waste-Recycling literature. We clarify its relationship to the original algorithm in (Calderhead, 2014) and to previously presented approaches, making explicit the variables that are Rao-Blackwellised. We provide a formal derivation and proof of its correctness in terms of unbiasedness of the resulting estimator for integrals with respect to the target, and give conditions under which the estimate exhibits lower asymptotic variance than the original.

Finally, in Section 4, we develop this approach further by proposing a non-trivial adaptive version of this parallelisable and Rao-Blackwellised MCMC method and prove asymptotic unbiasedness for a selection of common classes of proposal distributions. We demonstrate the resulting improvements in computational efficiency and in particular consider the comparison with multiple single Markov chains running in parallel.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. *t.schwedes14@imperial.ac.uk

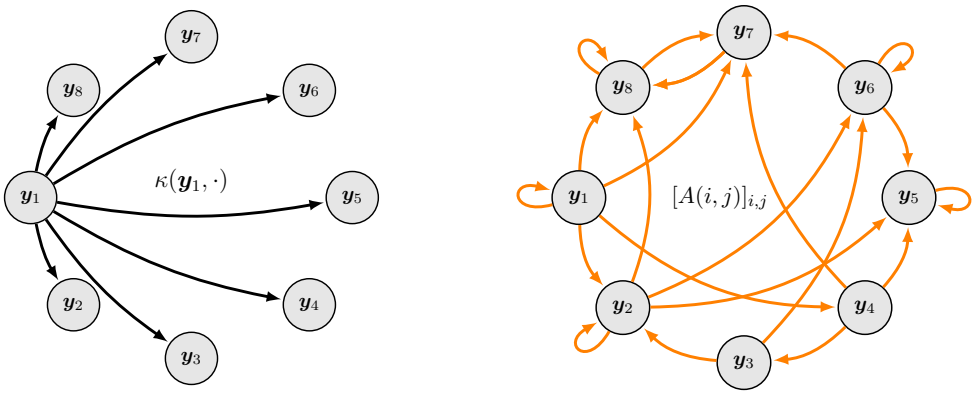


Figure 1: MP-MCMC in two steps: Step 1, based on a current proposal state $\boldsymbol{y}_I$ for say, $I = 1$, multiple states $\boldsymbol{y}_2, ..., \boldsymbol{y}_{N+1}$ are proposed in parallel via a kernel κ (*left*). Step 2, consider the collection of proposals $\boldsymbol{y}_{1:N+1}$ as states of a Markov chain with transition probabilities $[A(i,j)]_{i,j}$ and sample from it (*right*)

2 MP-MCMC

2.1 Deriving MP-MCMC

During one iteration of MP-MCMC (Calderhead, 2014) we subsample from states $\boldsymbol{y}_1, ..., \boldsymbol{y}_{N+1}$, which comprise the current state $\boldsymbol{y}_i$ and N new i.i.d. proposed states $\boldsymbol{y}_{\setminus i}$, where $\boldsymbol{y}_{\setminus i} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_{i-1}, \boldsymbol{y}_{i+1}, ..., \boldsymbol{y}_{N+1})$. A subsample $\boldsymbol{y}_I$ is collected according to an auxiliary random variable $I = 1, ..., N+1$, that is, we collect $\boldsymbol{y}_i$ as a sample if $I = i$, while I is sampled in sequence M times per iteration. Here, $\boldsymbol{y}_1, ..., \boldsymbol{y}_{N+1}$ and I are related via $p(\boldsymbol{y}_1, ..., \boldsymbol{y}_{N+1}, I) = \frac{1}{N+1} p_I(\boldsymbol{y}_1, ..., \boldsymbol{y}_{N+1})$ and

$$p_I(\boldsymbol{y}_{1:N+1}) = \pi(\boldsymbol{y}_I)\tilde{\kappa}(\boldsymbol{y}_I, \boldsymbol{y}_{\setminus I}), \tag{1}$$

where $\tilde{\kappa}(\boldsymbol{y}_I, \boldsymbol{y}_{\setminus I}) = \prod_{i \neq I} \kappa(\boldsymbol{y}_I, \boldsymbol{y}_i)$ and κ is the proposal kernel. In other words, I determines the factorisation of the joint distribution of $(\boldsymbol{y}_{1:N+1}, I)$.

We iterate between sampling I given $\boldsymbol{y}_1, ..., \boldsymbol{y}_{N+1}$, collecting M samples of $\boldsymbol{y}_I$, and then sampling N new states $\boldsymbol{y}_{\setminus I}$ via the kernel $\kappa(\boldsymbol{y}_I, \cdot)$ evaluated at the current $\boldsymbol{y}_I$. The algorithm thus takes on a structure similar to Gibbs sampling, which mirrors the accept/reject step (i.e. sampling I) and proposal steps (i.e. sampling $\boldsymbol{y}_{\setminus I}$) in standard Metropolis-Hastings.

Samples of I are generated according to a finite state Markov chain. That is, given $I = i$ and $\boldsymbol{y}_1, ..., \boldsymbol{y}_{N+1}$ we sample $I = j$, and hence collect $\boldsymbol{y}_j$ as a new state, according to a transition probability $A(i,j)$, which satisfies,

$$\sum_{j=1}^{N+1} \pi(\boldsymbol{y}_j)\tilde{\kappa}(\boldsymbol{y}_j, \boldsymbol{y}_{\setminus j})A(j,i) = \pi(\boldsymbol{y}_i)\tilde{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\setminus i}). \tag{2}$$

A single iteration of this algorithm is visualised in Figure 1. Concretely, Calderhead (2014) suggests

$$A(i,j) = \begin{cases} \frac{1}{N}\min(1, R(i,j)) & \text{if } j \neq i \\ 1 - \sum_{j \neq i} A(i,j) & \text{otherwise,} \end{cases} \tag{3}$$

where $R(i,j) = \pi(\boldsymbol{y}_j)\tilde{\kappa}(\boldsymbol{y}_j, \boldsymbol{y}_{\setminus j})/[\pi(\boldsymbol{y}_i)\tilde{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\setminus i})]$. For $N = 1$, A reduces to the acceptance probability is Metropolis-Hastings. Another option is $A(\cdot, j) = w_j \propto \pi(\boldsymbol{y}_j)\tilde{\kappa}(\boldsymbol{y}_j, \boldsymbol{y}_{\setminus j})$, where the normalising constant is easily found summing over j. For $N = 1$, A reduces to Barker's acceptance probability (Barker, 1965).

Whereas in Metropolis-Hastings or Barker's method a single sample is drawn per iteration, we sample $M > 1$ times due to the increased coverage of the state space and degree of freedom from multiple proposals. The last sample per iteration becomes the initial state of the subsequent iteration. The procedure is repeated until a required number of samples is achieved, see Algorithm 1 for details.

Algorithm 1: Multiple proposal MCMC (MP-MCMC)

Input: Starting point $\boldsymbol{x}_0 = \boldsymbol{y}_1 \in \mathbb{R}^d$, number of proposals N, number of accepted samples per iteration M, auxiliary variable $I = 1$ and counter $n = 0$;

for each MCMC iteration $\ell = 1, 2, ...$ **do**
  Draw N new points $\boldsymbol{y}_{\setminus I}$, conditioned on I, independently from the proposal kernel $\kappa(\boldsymbol{y}_I, \cdot)$;
  Calculate the transition probabilities $A(i,j)$ for $i, j = 1, ..., N+1$ satisfying the balance condition (2), which can be done in parallel;
  for $m = 1, ..., M$ **do**
    Sample I' via $A(I, \cdot)$;
    Set $I = I'$;
    Set new sample $\boldsymbol{x}_{[(\ell-1)M+m]} = \boldsymbol{y}_I$;
  end
end

Likelihood calculations occur only when sampling new states, and therefore can be straightforwardly parallelised. We note that MP-MCMC versions of Metropolis-adjusted Langevin, Riemann manifold Metropolis-adjusted Langevin as well as Hamiltonian MCMC are implied in the formulation of the algorithm due to a free choice of the proposal kernel κ.

2.2 Invariance of the stationary distribution

We now prove that the stationary distribution π is indeed preserved under MP-MCMC updates. This is a new result as Calderhead (2014) did not provide any formal proof for this.

Lemma 2.1. *If the balance condition in equation* (2) *is fulfilled, then updates according to Algorithm 1 preserve* π*.*

Proof. See Appendix A.1. □

3 Rao-Blackwellised MP-MCMC

Estimating integrals of the form $\int f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ can readily be achieved by MP-MCMC when averaging $f(\boldsymbol{x}_i)$ of samples x_i. We now introduce a provably more efficient estimate to that, based exactly on the same iterations of MP-MCMC. The improvement we achieve is hence for free. More precisely, instead of subsampling proposals, we use Rao-Blackwellisation which incorporates all proposals per iteration and assigns each one a suitable weight so that the resulting estimator is asymptotically unbiased. For clarity, an estimator $(\hat{\vartheta}_L)_L$ for ϑ is asymptotically unbiased if $\mathbb{E}(\hat{\vartheta}_L) \to \vartheta$ for $L \to \infty$.

We first provide some intuition for this approach, position it with regards to existing literature, then give a more formal derivation, followed by a brief numerical study.

3.1 Intuition

Assuming stationarity in the underlying MP-MCMC, the relative frequency of subsamples $\boldsymbol{y}_i$ among all $N+1$ proposed points in one iteration approaches the stationary probability $w_i \propto \pi(\boldsymbol{y}_i)\tilde{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\setminus i})$ for M becoming large. In the limit, $M \to \infty$, subsampling from proposed points is equivalent to accepting each $\boldsymbol{y}_i$ and weighting it according to w_i. In particular,

$$\frac{1}{M}\sum_{m=1}^{M} f\left(\boldsymbol{x}_m\right) \xrightarrow{M\to\infty} \sum_{i=1}^{N+1} w_i f\left(\boldsymbol{y}_i\right). \tag{4}$$

In other words, the arithmetic mean of collected subsamples from MP-MCMC converges to a Rao-Blackwellised estimate. We formalise this limiting case as *Rao-Blackwellised multiple proposal MCMC* (RB-MP-MCMC) in Algorithm 2. A single iteration of this algorithm is visualised in Figure 2.

Algorithm 2: Rao-Blackwellised multiple proposal MCMC (RB-MP-MCMC)

Input: Starting point (proposal) $\boldsymbol{y}_1 \in \mathbb{R}^d$, number of proposals N, auxiliary variable $I = 1$, integrand f;

1 **for** each MCMC iteration $\ell = 1, 2, ...$ **do**

2 Draw N new points $\boldsymbol{y}_{\setminus I}$, conditioned on I, independently from the proposal kernel $\tilde{\kappa}(\boldsymbol{y}_I, \cdot)$;

3 Calculate the stationary distribution of I conditioned on $\boldsymbol{y}_{1:N+1}$, i.e. $\forall$ $i = 1, ..., N+1$, $p(I = i|\boldsymbol{y}_{1:N+1}) = \pi(\boldsymbol{y}_i)\tilde{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\setminus i}) / \sum_j \pi(\boldsymbol{y}_j)\tilde{\kappa}(\boldsymbol{y}_j, \boldsymbol{y}_{\setminus j})$, which can be done in parallel;

4 Compute $\tilde{\mu}^{(f)}_{\ell,N} = \sum_i p(I = i|\boldsymbol{y}_{1:N+1}) f(\boldsymbol{y}_i)$;

5 Update Rao-Blackwell estimate $\hat{\mu}^{(f)}_{\ell,N} = \hat{\mu}^{(f)}_{\ell-1,N} + \frac{1}{\ell}(\tilde{\mu}^{(f)}_{\ell,N} - \hat{\mu}^{(f)}_{\ell-1,N})$, where $\hat{\mu}^{(f)}_{0,N} = 0$;

6 Sample new I via the stationary distribution $p(\cdot|\boldsymbol{y}_{1:N+1})$;

7 **end**

Note that Algorithm 2 implicitly generates a Markov chain: in every iteration, we sample from the $N+1$ proposals (line 6), conditioned on which N new proposals are drawn in the subsequent iteration. The chain defined by the subsamples from each iteration corresponds to Algorithm 1 with $M = 1$ and $A(\cdot, j) \propto w_j$. This is the Markov chain we refer to when speaking of the chain underlying RB-MP-MCMC.

3.2 Relation to Waste-Recycling

We note that similar methods can be found in the literature in the context of Waste-Recycling (Frenkel, 2006; Ceperley et al., 1977; Tjelmeland, 2004; Frenkel, 2004; Delmas and Jourdain, 2009; Yang et al., 2018). For instance, Tjelmeland (2004) suggests a Waste-Recycling estimator based on Metropolis-Hastings, where out of multiple proposed states per iteration only one is accepted. He proposes to take the weighted mean of both accepted and rejected states according to a general weight matrix function, of which the weights associated with the Rao-Blackwell estimate discussed in this work are a special case. Delmas and Jourdain (2009) consider a control variates approach: their estimator is the sum of the classic arithmetic mean of accepted samples in Metropolis-Hastings and a correction term depending on some function Ψ. Using multiple proposals in a special case of Metropolis-Hastings they call

Boltzmann algorithm, and the specific choice of $\Psi = f$ in the correction term, their control variate estimator reduces to our Rao-Blackwell estimator.

Our work adds to the existing results by using a heuristic based on Rao-Blackwellisation, by extending Algorithm 2 to allow for adaptivity in proposed samples, and by proving theoretical statements, e.g., on the improvement on MP-MCMC by Rao-Blackwellisation, Lemma 3.1, or the asymptotic unbiasedness of the adaptive Rao-Blackwellised estimate, see Theorem 4.1.

3.3 Formal derivation

In what follows we formally derive Algorithm 2 as Rao-Blackwellision of MP-MCMC. Let a superscript (ℓ) denote variables associated to the ℓth out of L iterations, e.g., $\boldsymbol{x}_m^{(\ell)}$ denote the mth collected sample in the ℓth iteration in MP-MCMC. The arithmetic mean follows as

$$\hat{\mu}_{L,M,N}^{(f)} = \frac{1}{LM} \sum_{\ell=1}^{L} \sum_{m=1}^{M} f\left(\boldsymbol{x}_m^{(\ell)}\right). \tag{5}$$

If $u_m^{(\ell)} \sim \mathcal{U}(0,1)$, for $m = 1, ..., M$, and $\eta_i^{(\ell)} = \sum_{j=1}^{i} w_j^{(\ell)}$, for $i = 1, ..., N+1$, and $\eta_0 := 0$, then the samples $\boldsymbol{x}_m^{(\ell)}$ can be written as

$$\boldsymbol{x}_m^{(\ell)} = \sum_{i=1}^{N+1} \mathbb{1}_{\left(\eta_{i-1}^{(\ell)}, \eta_i^{(\ell)}\right]}\left(u_m^{(\ell)}\right) \boldsymbol{y}_i^{(\ell)}, \tag{6}$$

where we assume that for the underling MP-MCMC acceptance probability in iteration ℓ, $A(\cdot, j) \propto w_j^{(\ell)}$. The resulting Rao-Blackwellised estimator can be seen as Waste-Recycling for the Boltzmann algorithm in (Delmas and Jourdain, 2009).

While in stationarity, $\boldsymbol{x}_m^{(\ell)} \sim \pi$, and hence

$$\mathbb{E}_\pi\left[\hat{\mu}_{L,M,N}^{(f)}\right] = \mathbb{E}_\pi\left[f(\boldsymbol{x})\right]. \tag{7}$$

In any iteration ℓ, $\boldsymbol{y}_{1:N+1}^{(\ell)}$ is a sufficient statistic for $\mathbb{E}_\pi[\boldsymbol{x}]$, since the conditional distribution of the samples $\boldsymbol{x}_m^{(\ell)} = \sum_{i=1}^{N+1} \mathbb{1}_{(\eta_{i-1}^{(\ell)}, \eta_i^{(\ell)}]}(u_m^{(\ell)}) \boldsymbol{y}_i^{(\ell)}$, where $m = 1, ..., M$, depends on $\mathbb{E}_\pi[\boldsymbol{x}]$ only through $\boldsymbol{y}_{1:N+1}^{(\ell)}$. The i.i.d. random variates $u_m^{(\ell)} \sim \mathcal{U}(0,1)$ for $m = 1, ..., M$, are independent of $\mathbb{E}_\pi[\boldsymbol{x}]$. Therefore, the conditional expectation of $\hat{\mu}_{L,M,N}^{(f)}$ given the proposed states $\boldsymbol{y}_{1:N+1}^{(\ell)}$ produces a Rao-Blackwellisation of $\hat{\mu}_{L,M,N}^{(f)}$.

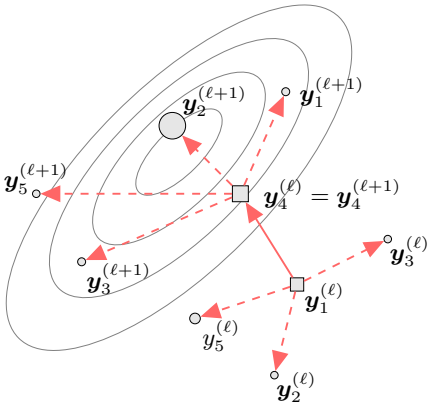


Figure 2: In every iteration of RB-MP-MCMC, proposals $\boldsymbol{y}_i^{(\ell)}$ for $i = 1, ..., N+1$ are generated analogously to MP-MCMC, but are then associated with the $w_i^{(\ell)} = p(I^{(\ell)} = i | \boldsymbol{y}_{1:N+1}^{(\ell)})$, thereby prioritising proposals that are most informative about the target, to form the weighted estimate $\sum_{i=1}^{N+1} w_i^{(\ell)} f(\boldsymbol{y}_i^{(\ell)})$

Note that,

$$\mathbb{E}_\pi\left[\sum_{i=1}^{N+1} \mathbb{1}_{\left(\eta_{i-1}^{(\ell)}, \eta_i^{(\ell)}\right]}\left(u_m^{(\ell)}\right) f\left(\boldsymbol{y}_i^{(\ell)}\right) \middle| \boldsymbol{y}_{1:N+1}^{(\ell)}\right] \tag{8}$$

$$= \int_{(0,1)} \sum_{i=1}^{N+1} \mathbb{1}_{\left(\eta_{i-1}^{(\ell)}, \eta_i^{(\ell)}\right]}\left(u_m^{(\ell)}\right) f\left(\boldsymbol{y}_i^{(\ell)}\right) \mathrm{d}u_m^{(\ell)} \tag{9}$$

$$= \sum_{i=1}^{N+1} w_i^{(\ell)} f\left(\boldsymbol{y}_i^{(\ell)}\right). \tag{10}$$

Hence, the Rao-Blackwell estimate $\hat{\mu}_{L,N}^{(f)}$ can be written as,

$$\begin{aligned} \hat{\mu}_{L,N}^{(f)} &= \mathbb{E}_\pi\left[\hat{\mu}_{L,M,N}^{(f)} \middle| \boldsymbol{y}_{1:N+1}^{(\ell)}, \ell = 1, ..., L\right] \\ &= \frac{1}{L} \sum_{\ell=1}^{L} \sum_{i=1}^{N+1} w_i^{(\ell)} f\left(\boldsymbol{y}_i^{(\ell)}\right). \end{aligned} \tag{11}$$

The weights $w_i^{(\ell)}$ present a special case of the general weight matrix function introduced in the Waste-Recycling algorithm by Tjelmeland (2004).

The following result states that $\hat{\mu}_{L,N}^{(f)}$ is never a worse estimate than the original MP-MCMC estimate $\hat{\mu}_{L,M,N}^{(f)}$.

Lemma 3.1. *Let $\hat{\mu}_{L,N}^{(f)}$ denote the Rao-Blackwell estimate based on the MP-MCMC mean estimate $\hat{\mu}_{L,M,N}^{(f)}$. Then,*

$$\mathrm{Var}\left(\hat{\mu}_{L,N}^{(f)}\right) \leq \mathrm{Var}\left(\hat{\mu}_{L,M,N}^{(f)}\right). \tag{12}$$

Proof. See Appendix A.2. □

Note that Delmas and Jourdain (2009) prove in their Proposition 2.3 that Waste-Recycling can degrade Metropolis-Hastings, using the usual acceptance probability. However, in the Boltzmann case they show this never occurs (Proposition 2.5), hence there is no contradiction between the above result and theirs. However, their results on boundedness of variance only consider the asymptotic case; their main result, Theorem 3.1, derives asymptotic normality and a formula for the asymptotic variance. In contrast, we consider the non-asymptotic case to derive the boundedness of variance compared to Calderhead's algorithm, which implies the result we get from (Delmas and Jourdain, 2009).

Lemma 3.2 (Asymptotic unbiasedness of RB-MP-MCMC estimates)**.** *Let the underlying Markov chain described by Algorithm 2 be positive Harris. Then,* $\hat{\mu}^{(f)}_{L,N}$ *for* $L \geq 1$ *defined in* (11) *is an asymptotically unbiased estimate for* $\mathbb{E}_\pi[f(\boldsymbol{y})]$.

Proof. See Appendix A.3. □

Hence, RB-MP-MCMC produces unbiased estimates for $\mu^{(f)}$ when in equilibrium, which includes posterior mean estimates. We now show that the same holds true for posterior covariance estimates.

Proposition 3.3. *Let the underlying Markov chain described by Algorithm 2 be positive Harris. Then, the covariance estimate* $\hat{\Sigma}_{L,N}$ *for* $L \geq 1$ *given by*

$$\hat{\Sigma}_{L,N} = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{i=1}^{N+1} w_i^{(\ell)}\left(\boldsymbol{y}_i^{(\ell)} - \hat{\boldsymbol{\mu}}_{L,N}\right)\left(\boldsymbol{y}_i^{(\ell)} - \hat{\boldsymbol{\mu}}_{L,N}\right)^T,$$

where $\hat{\boldsymbol{\mu}}_{L,N} = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{i=1}^{N+1} w_i^{(\ell)}\boldsymbol{y}_i^{(\ell)}$*, is an asymptotically unbiased estimate for the target covariance.*

Proof. For a proof, we refer to Appendix A.4. □

3.4 Numerical effect of increasing M in multiple proposal MCMC

In what follows we analyse the impact of an increasing number of subsamples M per iteration in MP-MCMC numerically on the MSE of resulting estimates. Due to a decrease in empirical variance with increasing M, a reduction in the MSE for increasing M is expected, with a lower bound given by the limiting case $M \to \infty$ represented by Rao-Blackwellised MP-MCMC.

Our experiments consist of sampling a standard Gaussian $\mathcal{N}(0,1)$ target, using MP-MCMC with the independence proposal kernel $\kappa = \mathcal{N}(0, 2.4^2)$. In each iteration $M = \alpha N$ states are generated by subsampling from proposals, which leads to a total number of $n_\alpha = \alpha\hat{n} = \alpha LN$ samples per simulation; here we choose $L = 511$. The target mean is estimated via the arithmetic mean of samples. Further, we apply RB-MP-MCMC to estimate the target mean, using the same κ as before, and compare both outcomes.

According to Figure 3, increasing the number of subsamples per iteration leads to an improvement in the MSE convergence rate. More precisely, for increasing α the constant $c(\alpha)$ in the MSE rate $c(\alpha)N^{-1}$ is reduced and approaches the limiting lowest value $c(\infty) \approx \frac{1}{2}c(1)$ associated with RB-MP-MCMC.

Algorithm 3: Adaptive Rao-Blackwellised multiple proposal MCMC (RB-MP-MCMC)
All code altered compared to RB-MP-MCMC, Algorithm 2, is highlighted

Input: Initialise starting point (proposal) $\boldsymbol{y}_1 \in \mathbb{R}^d$, number of proposals N, auxiliary variable $I = 1$, integrand f and adaptation parameter Υ_1;

for each MCMC iteration $\ell = 1, 2, \ldots$ **do**
- Draw N new points $\boldsymbol{y}_{\setminus I}$, conditioned on I and Υ_ℓ, independently from the proposal kernel $\kappa_{\Upsilon_\ell}(\boldsymbol{y}_I, \cdot)$;
- Calculate the stationary distribution of I conditioned on $\boldsymbol{y}_{1:N+1}$ and Υ_ℓ, i.e. $\forall$ $i = 1, \ldots, N+1$, $p(I = i|\boldsymbol{y}_{1:N+1}, \Upsilon_\ell) = \pi(\boldsymbol{y}_i)\tilde{\kappa}_{\Upsilon_\ell}(\boldsymbol{y}_i, \boldsymbol{y}_{\setminus i}) / \sum_j \pi(\boldsymbol{y}_j)\tilde{\kappa}_{\Upsilon_\ell}(\boldsymbol{y}_j, \boldsymbol{y}_{\setminus j})$, which can be done in parallel;
- Compute $\tilde{\mu}^{(f)}_{\ell,N} = \sum_i p(I = i|\boldsymbol{y}_{1:N+1}, \Upsilon_\ell) f(\boldsymbol{y}_i)$;
- Update Rao-Blackwell estimate $\hat{\mu}^{(f)}_{\ell,N} = \hat{\mu}^{(f)}_{\ell-1,N} + \frac{1}{\ell}(\tilde{\mu}^{(f)}_{\ell,N} - \hat{\mu}^{(f)}_{\ell-1,N})$, where $\hat{\mu}^{(f)}_{0,N} = 0$;
- Sample new I via the stationary distribution $p(\cdot|\boldsymbol{y}_{1:N+1}, \Upsilon_\ell)$;
- Update adaptation parameter $\Upsilon_{\ell+1} = G(\Upsilon_\ell, \boldsymbol{y}_{1:N+1})$;

end

4 Adaptive Rao-Blackwellised multiple proposal MCMC

It is widely known that adaptively learning the proposal distribution may significantly increase the performance of an underlying sampling method, see (Haario et al., 2001; Atchadé and Rosenthal, 2005; Haario et al., 2006; Giordani and Kohn, 2008; Roberts and Rosenthal, 2009) for MCMC, and (Au and Beck, 1999; Cappé et al., 2008; Cornuet et al., 2012) for importance sampling. In what follows, we extend Algorithm 2 to incorporate adaptivity and we prove asymptotic unbiasedness for a selection of common classes of proposal distributions.

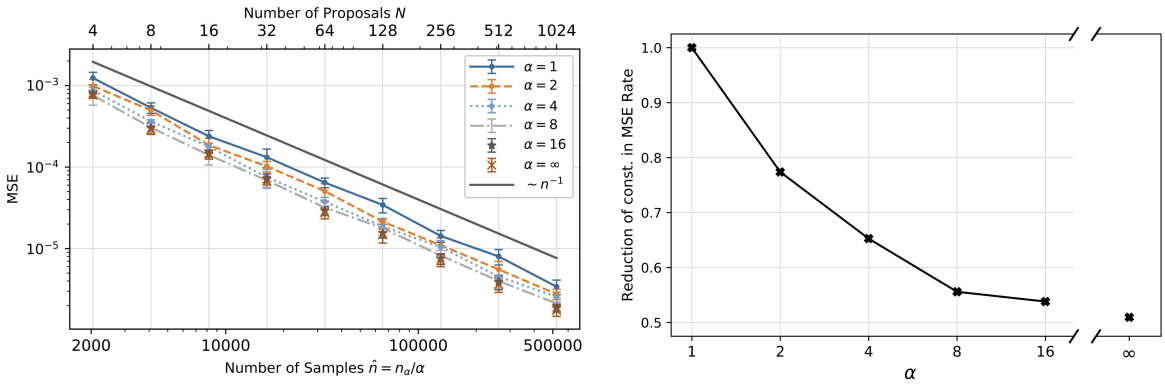


Figure 3: MSE of empirical means for a Gaussian in $d = 1$ for an increasing number of proposals N and sample size $n = LM$ with $L = 511$ in MP-MCMC; considered is an increase in number of subsamples $M = \alpha N$ per iteration for $\alpha \in \{1, 2, 4, 8, 16, \infty\}$ (*left*), where $\alpha = \infty$ represents RB-MP-MCMC. The error bars correspond to three standard deviations. Displayed is also the reduction factor in the MSE convergence rate compared to $\alpha = 1$ (*right*). The results are based on 500 simulations

Let $(\kappa_\Upsilon)_{\Upsilon \in \mathcal{Y}}$ denote a collection of proposal kernels indexed over some parameter space $\mathcal{Y}$. Given an initial Υ_1, the proposal kernel in iteration $\ell + 1$, $\kappa_{\Upsilon_{\ell+1}}$, is determined according to an update function $G_\ell : \mathcal{Y} \times \mathbb{R}^{(N+1)d} \to \mathcal{Y}$ via $\Upsilon_{\ell+1} = G_\ell(\Upsilon_\ell, \boldsymbol{y}^{(\ell)}_{1:N+1})$ based on Υ_ℓ and the current proposals $\boldsymbol{y}^{(\ell)}_{1:N+1}$. This results in an adaptive *Rao-Blackwellised multiple proposal MCMC* (ARB-MP-MCMC) method, see Algorithm 3.

4.1 Choice of proposal kernels

For adaptivity of proposal kernels we distinguish between two options. First, $\Upsilon = \Sigma$ with the parameter Σ denoting the covariance of the kernel. Second, $\Upsilon = (\boldsymbol{\mu}, \Sigma)$, where the parameter $\boldsymbol{\mu}$ denotes the mean of the kernel. We adapt Σ via G_ℓ by $\Sigma_{\ell+1} = \Sigma_\ell + \frac{1}{\ell+1}(\tilde{\Sigma}_{\ell+1} - \Sigma_\ell)$ and $\boldsymbol{\mu}$ by $\boldsymbol{\mu}_{\ell+1} = \boldsymbol{\mu}_\ell + \frac{1}{\ell+1}(\tilde{\boldsymbol{\mu}}_{\ell+1} - \boldsymbol{\mu}_\ell)$, where

$$\tilde{\boldsymbol{\mu}}_{\ell+1} = \sum_{i=1}^{N+1} p\left(I = i | \boldsymbol{y}^{(\ell)}_{1:N+1}, \Upsilon_\ell\right) \boldsymbol{y}^{(\ell)}_i, \tag{13}$$

$$\tilde{\Sigma}_{\ell+1} = \sum_{i=1}^{N+1} p\left(I = i | \boldsymbol{y}^{(\ell)}_{1:N+1}, \Upsilon_\ell\right) \left[\boldsymbol{y}^{(\ell)}_i - \boldsymbol{\mu}_{\ell+1}\right] \cdot \left[\boldsymbol{y}^{(\ell)}_i - \boldsymbol{\mu}_{\ell+1}\right]^T. \tag{14}$$

4.1.1 Gaussian kernel

Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ denote the *multivariate Gaussian* PDF with mean $\boldsymbol{\mu}$ and covariance Σ, evaluated at $\boldsymbol{x} \in \mathbb{R}^d$. If $\Upsilon = (\boldsymbol{\mu}, \Sigma)$, we set $\kappa_{(\boldsymbol{\mu},\Sigma)}(\boldsymbol{x}, \boldsymbol{y}) = \kappa_{(\boldsymbol{\mu},\Sigma)}(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma)$, i.e. an independence sampler. If $\Upsilon = \Sigma$, we set $\kappa_\Sigma(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{x}, \Sigma)$. For the latter, where only the covariance is updated, our algorithm generalises the adaptive MCMC algorithm introduced by Haario et al. (2001) allowing for multiple proposals. However, in this instance we choose to employ a slightly different covariance estimator than the standard empirical covariance used in (Haario et al., 2001): the proposal covariance in iteration $\ell + 1$ is itself based on a Rao-Blackwell estimate, since it has been shown that this exhibits lower asymptotic variance (Frenkel, 2006; Ceperley et al., 1977).

4.1.2 T-distribution kernel

The generalisation of the *Student t-distribution* to multiple dimensions introduced in (Kotz and Nadarajah, 2004) has the PDF,

$$t(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{(\pi(\nu-2))^{\frac{d}{2}}\Gamma(\frac{\nu}{2})\det(\Sigma)^{\frac{1}{2}}} \cdot \left[1 + \frac{1}{\nu-2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{\nu+d}{2}}, \tag{15}$$

for any $\boldsymbol{x} \in \mathbb{R}^d$, where $\boldsymbol{\mu}$ and Σ denote the mean and covariance, respectively. Here, Γ denotes the Gamma-function. Further, $\nu > 2$ is called the degree of freedom and determines how much heavier the tails of the t-distribution are relative to a Normal distribution. The latter is recovered for $\nu \to \infty$.

If $\Upsilon = (\boldsymbol{\mu}, \Sigma)$, then $\kappa_{(\boldsymbol{\mu},\Sigma)}(\boldsymbol{x}, \boldsymbol{y}) = \kappa_{(\boldsymbol{\mu},\Sigma)}(\boldsymbol{y}) = t(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma)$, i.e. an independence kernel. If $\Upsilon = \Sigma$, then $\kappa_\Sigma(\boldsymbol{x}, \boldsymbol{y}) = t(\boldsymbol{y}|\boldsymbol{x}, \Sigma)$.

4.2 Asymptotic unbiasedness

We derive conditions under which the adaptive RB-MP-MCMC, Algorithm 3, produces asymptotically unbiased estimates. Practically speaking, this means that after having discarded a sufficiently large burn-in period $L_B \ll L$ of weighted estimates $\{\tilde{\mu}^{(f)}_{\ell,N} : \ell = 1, ..., L_B\}$, we may assume the underlying chain is in stationarity, and the Rao-Blackwellised estimator defined by the remaining samples is unbiased.

Table 1: Summary of models and data in Bayesian logistic regression from (Michie et al., 1994; Ripley, 1996)

Name	# Covariates D	# Data points M	Dimension d
Ripley	2	250	3
Pima Indian	7	532	8
Heart	13	270	14
Australian	14	690	15
German	24	1000	25

Throughout this section we assume that for each fixed Υ, the chain underlying Algorithm 3 has the correct stationary distribution π, and that each individual update $P_\Upsilon^{(1)}(\boldsymbol{x}, \cdot)$ preserves π. Further, let the target π be absolutely continuous with respect to the Lebesgue measure λ and assume that any covariance matrix as part of the adaptation space is symmetric and positive definite. We say a set of covariance matrices $S \subset \mathbb{R}^{d\times d}$ is bounded if there are $0 < c_1 < c_2 < \infty$ such that $c_1 I_d \leq \Sigma \leq c_2 I_d$ for any $\Sigma \in S$, where the "$\leq$" is understood in the usual way for matrices: For two matrices $A, B \in \mathbb{R}^{d\times d}$, $A \leq B$ means that $B - A$ is positive semi-definite. We distinguish between two different assumptions.

Assumption 1 (Continuity)**.** *Suppose that the target π is continuous on a subset $S \in \mathbb{R}^d$ with $\lambda(S) > 0$ and strictly positive on S.*

The former assumption is met if the target is continuous. We now consider the case where proposals are generated independently of previous samples, except through their dependence via the adaptation parameters. In that case, the continuity assumption is not required.

Assumption 2 (Independence)**.** *Suppose that $\Upsilon = (\boldsymbol{\mu}, \Sigma)$ and that $\kappa_\Upsilon \in \{\mathcal{N}_\Upsilon, t_\Upsilon\}$ represents an independence kernel.*

Theorem 4.1. *Let $\Upsilon \in \mathcal{Y}$ be bounded. If either of the conditions in Assumption 1 or 2 are satisfied, then the sequence of estimators $(\hat{\mu}_{L,N}^{(f)})_{L\geq 1}$ from Algorithm 3 is asymptotically unbiased.*

Proof. For a proof we refer to Appendix A.5. □

5 Evaluation of ARB-MP-MCMC

5.1 Bayesian logistic regression

We consider the example of logistic regression for binary classification problems ((Robert and Casella, 1999; Gelman et al., 2013)), whereby categorical variables $y_m \in \{0, 1\}$ for $m = 1, ..., M$ depend on explanatory variables, which can be summarised by the design matrix $X = (\boldsymbol{X}_{1,\cdot}, ..., \boldsymbol{X}_{M,\cdot}) \in \mathbb{R}^{M\times D}$, where $\boldsymbol{X}_{m,\cdot} = (X_{m,1}, ..., X_{m,D})$. The conditional probability of y_m is defined by

$$P(y_m = 1|X, \boldsymbol{\theta}) = \sigma((1, X_{m,\cdot})^T \boldsymbol{\theta}), \tag{16}$$

and $P(y_m = 0|X, \boldsymbol{\theta}) = 1 - P(y_m = 1|X, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ with $d = D + 1$ and σ denotes the logistic function. We are interested in the statistical inference of the regression parameter $\boldsymbol{\theta}$, which by introducing the Gaussian prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \alpha I_d)$, with $\alpha = 100$, becomes a Bayesian inverse problem. The log likelihood of $\boldsymbol{y}$ can be derived from (16) as

$$\log(\boldsymbol{y}|X, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \tilde{X}^T \boldsymbol{y} - \sum_{m=1}^{M} \log\left(1 + \exp(\boldsymbol{\theta}^T \tilde{X}_{m,\cdot})\right),$$

where $\tilde{X}_{\cdot,j} = X_{\cdot,j-1}$ for any $j \geq 2$ and $\tilde{X}_{i,1} = 1$. The five data sets used are summarised in Table 1 (Michie et al., 1994; Ripley, 1996).

5.2 Algorithmic Setup

We compare the performance of RB-MP-MCMC and ARB-MP-MCMC in the context of the Bayesian logistic regression model introduced above employing a multivariate Gaussian independence sampler. In the non-adaptive case we utilise a fixed mean $\boldsymbol{\mu} = \arg\max_{\boldsymbol{x}} \pi(\boldsymbol{x})$, which is determined numerically. The covariance matrix is set equal to the Riemannian metric tensor (Girolami and Calderhead, 2011),

$$G_M(\boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\frac{\partial}{\partial \boldsymbol{x}^2} \log \pi(\boldsymbol{y}, \boldsymbol{x})\right], \tag{17}$$

evaluated at the posterior mode $\boldsymbol{\mu}$. Here, $\pi(\boldsymbol{y}, \boldsymbol{x})$ denotes the joint probability of observations and parameters.

In the adaptive version the proposal mean and covariance matrix are determined according to equations (13) and (14). As a reference we consider the standard, i.e. single proposal, random-walk Metropolis-Hastings algorithm. The associated proposals are multivariate Gaussian with covariance equal to the identity matrix[1],

[1]The seemingly better choice of the scaled Fisher-information evaluated at the posterior mode $\arg\max_{\boldsymbol{x}} \pi(\boldsymbol{x})$ only yields slight improvement. For the Australian Credit data set it actually leads to slightly larger variances

Table 2: Average reduction in empirical variance estimates from RB-MP-MCMC and adaptive RB-MP-MCMC, compared to SmMALA Metropolis-Hastings (M-H SmMALA) and standard Metropolis-Hastings (M-H) for the Bayesian logistic regression. The results are based on 25 MCMC simulations

M-H	Factor of variance reduction				
VS	Ripley	Pima Indian	Heart	Australian	German
M-H	1	1	1	1	1
M-H SmMALA	3.6	9.6	3.3	$28,8$	11.5
RB-MP-MCMC	13.1	27.9	26.8	148.0	18.7
ARB-MP-MCMC	10.3	21.1	22.3	202.7	57.6

and a step size chosen such that the asymptotically optimal acceptance rate of 20-25% is attained. As a further reference, we consider the simplified manifold Metropolis-adjusted Langevin (SmMALA) algorithm introduced in (Girolami and Calderhead, 2011), which makes use of the local metric tensor (17) at each step. The step size is tuned such that an approximately optimal acceptance rate of 50-60% is achieved.

We compare each multiple proposal algorithm with L iterations and N proposals, to N independent single proposal chains of each reference method, from which L samples per chain after burn-in are collected. Hence, we compare parallelisable multiple proposal methods to parallelisable single proposal methods. The number of likelihood evaluations $n = LN$ required for the simulation after burn-in is equal for both single and multiple proposal methods.

A burn-in of between 1024-8192 samples, increasing with dimensionality d, was discarded for all methods, the lengths of which were identified by an analysis of trace plots and histograms, to ensure that asymptotic regime had been reached.

5.3 Empirical Results

We analyse the empirical variance for posterior mean estimates of the proposed methods. For Metropolis-Hastings and SmMALA the arithmetic sample mean is used as an estimate. The outcomes of experiments for the two highest-dimensional cases among the data sets in Table 1 are displayed in Figure 4. For these two datasets, both multiple proposal Rao-Blackwellised methods outperform the reference algorithms, while the adaptive RB-MP-MCMC outperforms its non-adaptive counterpart. Numerical values for the reduction in MSE compared to Metropolis-Hastings are presented in Table 2.

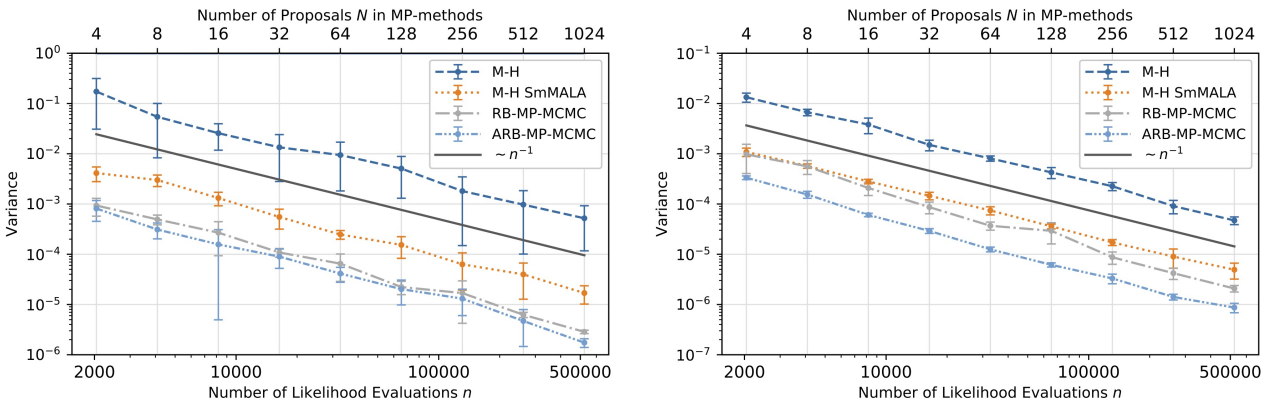
For the lower-dimensional examples similar outcomes were achieved, although the variance in adaptive RB-MP-MCMC compared to the non-adaptive version was slightly higher, which is due to the fact that the metric tensor evaluated at the mode is an already good approximation of the actual posterior covariance. Hence, the adaptive learning process does not achieve additional information about the posterior covariance.

Since we remove a burn-in for each of the N independent single proposal chains the majority of likelihood evaluations in their simulation happens during burn-in. Therefore, a less costly reference method in this example corresponds to a single chain with a single proposal and $n = LN$ collected samples after a single burn-in. Again, the number of likelihood evaluations between this single proposal method and the multiple proposal methods are equal.

We also performed this comparison, which lead to qualitative similar outcomes for the reference methods. Indeed, we typically observe slightly larger variances in the single chain case compared to multiple chains, which can be explained by an increased correlation among n collected samples from a single chain, compared to N sets of L collected samples from multiple independent chains, where $n = LN$. The reduction of variance due to using independent chains ranges from $0 - 15\%$ for SmMALA, and from $9 - 63\%$ for Metropolis-Hastings in the above example. As a result, the variance reduction by using the multiple proposal methods is even larger compared to the more economical single chain reference methods.

6 Conclusions

We have introduced an extension of the parallel MCMC method derived in (Calderhead, 2014) by Rao-Blackwellising the arithmetic mean of collected samples. We have proven that the resulting algorithm produces an asymptotically unbiased estimate for integrals with respect to the target, with an asymptotic variance that is bounded from above by the asymptotic variance of the original estimator. Furthermore, we have derived a generalisation of Haario's adaptive MCMC algorithm (Haario et al., 2001) allowing for multiple proposals within a Rao-Blackwellised scheme, considering two



(a) Australian Credit

(b) German Credit

Figure 4: Empirical variance estimates from (A)RB-MP-MCMC and (SmMALA) Metropolis-Hastings for the Bayesian logistic regression; results are displayed for increasing $n = LN$ of likelihood evaluations, where n is the total number of samples of the single-proposal methods, $L = 511$ and N the number of proposals in the respective multiple-proposal method. The results are based on 25 MCMC simulations, and the errors bands correspond to three standard deviations

common choices of adaptive proposal kernels. We formulated simple conditions that are easy to verify and that guarantee asymptotic unbiasedness of the resulting Rao-Blackwell estimates. In simulations the proposed adaptive method outperforms the original one in problems of higher dimensions, in which adaptivity leads to a more accurate approximation of posterior mean and covariance in the proposal kernel than for a fixed proposal kernel.

References

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11(5):815–828.

Au, S. and Beck, J. L. (1999). A new adaptive importance sampling scheme for reliability calculations. *Structural Safety*, 21(2):135–158.

Bai, Y., Roberts, G. O., and Rosenthal, J. S. (2011). On the containment condition for adaptive Markov chain Monte Carlo algorithms. *Advances and Applications in Statistics*, 21(1):1–54.

Barker, A. (1965). Monte Carlo calculations of the radial distribution functions for a proton electron plasma. *Australian Journal of Physics*, 18(2):119–134.

Calderhead, B. (2014). A general construction for parallelizing Metropolis-Hastings algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, 111(49):17408–17413.

Cappé, O., Douc, R., Guillin, A., Marin, J.-M., and Robert, C. P. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459.

Ceperley, D., Chester, G., and Kalos, M. (1977). Monte Carlo simulation of a many-fermion study. *Physical Review B*, 16(7):3081.

Cornuet, J.-M., Marin, J.-M., Mira, A., and Robert, C. P. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812.

Delmas, J.-F. and Jourdain, B. (2009). Does waste recycling really improve the multi-proposal Metropolis-Hastings algorithm? An analysis based on control variates. *Journal of Applied Probability*, 46(4):938–959.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. (2013). emcee: The MCMC hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306–312.

Frenkel, D. (2004). Speed-up of Monte Carlo simulations by sampling of rejected states. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17571–17575.

Frenkel, D. (2006). Waste-recycling Monte Carlo. In *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*, pages 127–137. Springer.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Giordani, P. and Kohn, R. (2008). Efficient Bayesian inference for multiple change-point and mixture in-

novation models. *Journal of Business and Economic Statistics*, 26(1):66–77.

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

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011). MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software*, 42(9):1–21.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Springer-Verlag London.

Michie, D., Spiegelhalter, D. J., and Taylor, C. C. (1994). *Machine learning, neural and statistical classification*. Ellis Horwood.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press.

Robert, C. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer-Verlag New York.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Tjelmeland, H. (2004). Using all Metropolis–Hastings proposals to estimate mean values. *Technical Report, Norwegian University of Science and Technology*.

Yang, S., Chen, Y., Bernton, E., and Liu, J. S. (2018). On parallelizable Markov chain Monte Carlo algorithms with waste-recycling. *Statistics and Computing*, 28(5):1073–1081.