Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation

By ANTHONY LEE AND KRZYSZTOF ŁATUSZYŃSKI

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.
anthony.lee@warwick.ac.uk   k.g.latuszynski@warwick.ac.uk

SUMMARY

Approximate Bayesian computation has emerged as a standard computational tool when dealing with intractable likelihood functions in Bayesian inference. We show that many common Markov chain Monte Carlo kernels used to facilitate inference in this setting can fail to be variance bounding and hence geometrically ergodic, which can have consequences for the reliability of estimates in practice. This phenomenon is typically independent of the choice of tolerance in the approximation. We prove that a recently introduced Markov kernel can inherit the properties of variance bounding and geometric ergodicity from its intractable Metropolis–Hastings counterpart, under reasonably weak conditions. We show that the computational cost of this alternative kernel is bounded whenever the prior is proper, and present indicative results for an example where spectral gaps and asymptotic variances can be computed, as well as an example involving inference for a partially and discretely observed, time-homogeneous, pure jump Markov process. We also supply two general theorems, one providing a simple sufficient condition for lack of variance bounding for reversible kernels and the other providing a positive result concerning inheritance of variance bounding and geometric ergodicity for mixtures of reversible kernels.

Some key words: Approximate Bayesian computation; Geometric ergodicity; Local adaptation; Markov chain Monte Carlo; Variance bounding.

1. INTRODUCTION

Approximate Bayesian computation refers to a branch of Monte Carlo methodology that uses the ability to simulate data according to a parameterized likelihood function in lieu of computation of that likelihood to perform approximate, parametric Bayesian inference. These methods have been used in an increasingly diverse range of applications since their inception in the context of population genetics (Tavaré et al., 1997; Pritchard et al., 1999), particularly in cases where the likelihood function is either impossible or computationally prohibitive to evaluate.

We consider a standard Bayesian setting with data $y \in \mathsf{Y}$, a parameter space Θ, a prior $p : \Theta \to \mathbb{R}_+$ and, for each $\theta \in \Theta$, a likelihood $f_\theta : \mathsf{Y} \to \mathbb{R}_+$. We assume that Y is a metric space and consider the artificial likelihood

$$f_\theta^\epsilon(y) = V(\epsilon)^{-1} \int_{\mathsf{Y}} I(y \in B_{\epsilon,x}) f_\theta(x)\, \mathrm{d}x = V(\epsilon)^{-1} f_\theta(B_{\epsilon,y}), \qquad (1)$$

which is commonly employed in approximate Bayesian computation. The value of ϵ can be interpreted as the tolerance of the approximation. Here, $B_{r,z}$ denotes a metric ball of radius r around z, $V(r) = \int_{\mathsf{Y}} I(x \in B_{r,0})\, \mathrm{d}x$ is the volume of a ball of radius r in Y, and I denotes the

indicator function. With a slight abuse of language, we refer to densities as distributions and, where convenient, employ the measure-theoretic notation $\mu(A) = \int_A \mu(d\lambda)$. We consider situations in which both ϵ and y are fixed, and so define functions $h : \Theta \to [0, 1]$ and $w : Y \to [0, 1]$ by

$$h(\theta) = f_\theta(B_{\epsilon,y}) \qquad (2)$$

and $w(x) = I(y \in B_{\epsilon,x})$ to simplify the presentation. The value $h(\theta)$ can be interpreted as the probability of hitting $B_{\epsilon,y}$ with a sample drawn from f_θ.

While the artificial likelihood (1) is also intractable in general, the approximate posterior it induces, $\pi(\theta) = h(\theta)p(\theta)/\int_\Theta h(\vartheta)p(\vartheta)\,d\vartheta$, can be dealt with by using constrained versions of standard methods when sampling from f_θ is possible for any $\theta \in \Theta$ (see, e.g., Marin et al., 2012). In particular, one typically uses f_θ as a proposal in such a way that its explicit computation is avoided. We are often interested in computing $\pi(\varphi) = \int_\Theta \varphi(\theta)\pi(\theta)\,d\theta$, the posterior expectation of some function φ, and it is this type of quantity that can be approximated using Monte Carlo methodology. We focus on one such method, Markov chain Monte Carlo, whereby a Markov chain is constructed by sampling iteratively from an irreducible Markov kernel P with unique stationary distribution π. We can use such a chain to estimate $\pi(\varphi)$ directly with appropriately normalized partial sums, i.e., given the realization $\theta_1, \theta_2, \ldots$ of a chain started at θ_0, where $\theta_i \sim P(\theta_{i-1}, \cdot)$, for $i \in \mathbb{N}$ we compute the estimate

$$\frac{1}{m}\sum_{i=1}^{m}\varphi(\theta_i) \qquad (3)$$

for some m. Alternatively, the Markov kernels can be used within other methods such as sequential Monte Carlo (Del Moral et al., 2006). In the former case, it is desirable that a central limit theorem hold for (3) and that the asymptotic variance $\mathrm{var}(P, \varphi)$ of (3) be reasonably small, while in the latter case it is desirable that the kernel be geometrically ergodic, i.e., the m-fold iterate of P, $P^m(\theta_0, \cdot)$, should converge at a geometric rate in m to π in total variation (see, e.g., Roberts & Rosenthal, 2004; Meyn & Tweedie, 2009), at least because this property is often assumed in analyses (see, e.g., Jasra & Doucet, 2008; Whiteley, 2012). In addition, consistent estimation of $\mathrm{var}(P, \varphi)$ is well established (Hobert et al., 2002; Jones et al., 2006; Bednorz & Łatuszyński, 2007; Flegal & Jones, 2010) for geometrically ergodic chains.

Motivated by these considerations, we study both the variance bounding (Roberts & Rosenthal, 2008) and geometric ergodicity properties of a number of reversible kernels used for approximate Bayesian computation. For reversible P, a central limit theorem holds for all $\varphi \in L^2(\pi)$ if and only if P is variance bounding (Roberts & Rosenthal, 2008, Theorem 7), where $L^2(\pi)$ is the space of square-integrable functions with respect to π. Reversible kernels that are not variance bounding can still produce Markov chains for which (3) satisfies a central limit theorem for some, but not all, functions in $L^2(\pi)$.

Much of the literature is concerned with controlling the trade-off associated with the quality of approximation (1), controlled by ϵ and manipulation of y, and counteracting computational difficulties (see, e.g., Fearnhead & Prangle, 2012). We address here a separate issue, namely that many Markov kernels used in this context are neither variance bounding nor geometrically ergodic, for any finite ϵ under widely met assumptions when using local proposal distributions. As a partial remedy, we show that under reasonably mild conditions, a kernel proposed in Lee et al. (2012) can inherit the properties of variance bounding and geometric ergodicity from its intractable Metropolis–Hastings (Metropolis et al., 1953; Hastings, 1970) counterpart. This allows the specification of a broad class of models for which we can be assured that this particular kernel will

be geometrically ergodic. In addition, conditions ensuring inheritance of either property can be met without knowledge of f_θ, e.g., by using a symmetric proposal and a prior that is continuous and everywhere positive and has exponential or heavier tails.

To assist in the interpretation of results and the quantitative example discussed later, we provide some background on the spectral properties of variance bounding and geometrically ergodic Markov kernels. Both variance bounding and geometric ergodicity of a reversible Markov kernel P are related to $\sigma_0(P)$, the spectrum of P considered as an operator on $L_0^2(\pi)$, the restriction of $L^2(\pi)$ to zero-mean functions (see, e.g., Geyer & Mira, 2000; Mira, 2001). Variance bounding is equivalent to $\sup \sigma_0(P) < 1$ (Roberts & Rosenthal, 2008, Theorem 14), and geometric ergodicity is equivalent to $\sup |\sigma_0(P)| < 1$ (Roberts & Rosenthal, 1997, Theorem 2.1; Kontoyiannis & Meyn, 2012). The spectral gap, $\mathrm{Gap}(P) = 1 - \sup |\sigma_0(P)|$, of a geometrically ergodic Markov kernel is closely related to its aforementioned geometric rate of convergence to π, with faster rates associated with larger spectral gaps. In particular, its convergence in total variation satisfies

$$\|\pi(\cdot) - P^m(\theta_0, \cdot)\|_{\mathrm{TV}} \leqslant C_\rho(\theta_0)\rho^m \tag{4}$$

for some $1 > \rho \geqslant \sup |\sigma_0(P)|$ and some function $C_\rho : \Theta \to \mathbb{R}_+$ (cf. Baxendale, 2005).

2. The Markov kernels

In this section we describe the algorithmic specification of the π-invariant, reversible Markov kernels under study. The algorithms specify how to sample from each kernel; in each, a candidate ϑ is proposed according to a common proposal $q(\theta, \cdot)$ and is accepted or rejected, possibly along with other auxiliary variables, using simulations from the likelihoods f_ϑ and f_θ. We assume that for all $\theta \in \Theta$, $q(\theta, \cdot)$ and p are densities with respect to a common dominating measure, e.g., the Lebesgue or counting measure. In the following, we define $a \wedge b = \min(a, b)$.

The first and simplest Markov kernel in this setting was proposed by Marjoram et al. (2003) and is a special case of a pseudo-marginal kernel (Beaumont, 2003; Andrieu & Roberts, 2009). Such kernels have been used in the context of approximate Bayesian computation for the estimation of parameters in speciation models (Becquet & Przeworski, 2007; Chen et al., 2009; Li et al., 2010; Kim et al., 2011) and as a methodological component within a sequential Monte Carlo sampler (Drovandi & Pettitt, 2011; Del Moral et al., 2012). They evolve on $\Theta \times \mathsf{Y}^N$ and involve sampling auxiliary variables $z_{1:N} \sim f_\vartheta^{\otimes N}$ for a fixed $N \in \mathbb{N}$. We denote kernels of this type for any N by $P_{1,N}$, and describe their simulation in Algorithm 1. It is readily verified (Beaumont, 2003; Andrieu & Roberts, 2009) that $P_{1,N}$ is reversible with respect to

$$\bar\pi(\theta, x_{1:N}) \propto p(\theta) \prod_{j=1}^{N} f_\theta(x_j) \frac{1}{N} \sum_{j=1}^{N} w(x_j),$$

and we have $\bar\pi(\theta) = \int \bar\pi(\theta, x_{1:N}) \, dx_{1:N} = \pi(\theta)$, i.e., the θ-marginal of $\bar\pi$ is $\pi(\theta)$.

Algorithm 1. To sample from $P_{1,N}(\theta, x_{1:N}; \cdot)$:

Step 1. Sample $\vartheta \sim q(\theta, \cdot)$ and $z_{1:N} \sim f_\vartheta^{\otimes N}$.

Step 2. With probability

$$1 \wedge \frac{p(\vartheta)q(\vartheta, \theta) \sum_{j=1}^{N} w(z_j)}{p(\theta)q(\theta, \vartheta) \sum_{j=1}^{N} w(x_j)},$$

output $(\vartheta, z_{1:N})$; otherwise, output $(\theta, x_{1:N})$.

In Lee et al. (2012), two alternative kernels were proposed in this context, both of which evolve on Θ. One of the kernels, denoted by $P_{2,N}$ and described in Algorithm 2, is an alternative pseudo-marginal kernel which, in addition to sampling $z_{1:N} \sim f_\vartheta^{\otimes N}$, also samples auxiliary variables $x_{1:N-1} \sim f_\theta^{\otimes N-1}$. Detailed balance can be verified directly upon interpreting $\sum_{j=1}^{N} w(z_j)$ and $\sum_{j=1}^{N-1} w(x_j)$ as $\mathrm{Bi}\{N, h(\vartheta)\}$ and $\mathrm{Bi}\{N-1, h(\theta)\}$ random variables, respectively. The other kernel, denoted by P_3 and described in Algorithm 3, also involves sampling according to f_θ and f_ϑ but does not sample a fixed number of auxiliary variables. This kernel also satisfies detailed balance (Lee, 2012, Proposition 1).

Algorithm 2. To sample from $P_{2,N}(\theta, \cdot)$:

Step 1. Sample $\vartheta \sim q(\theta, \cdot)$, $x_{1:N-1} \sim f_\theta^{\otimes N-1}$ and $z_{1:N} \sim f_\vartheta^{\otimes N}$.

Step 2. With probability

$$1 \wedge \frac{p(\vartheta)q(\vartheta, \theta)\sum_{j=1}^{N} w(z_j)}{p(\theta)q(\theta, \vartheta)\{1 + \sum_{j=1}^{N-1} w(x_j)\}},$$

output ϑ; otherwise, output θ.

Algorithm 3. To sample from $P_3(\theta, \cdot)$:

Step 1. Sample $\vartheta \sim q(\theta, \cdot)$.

Step 2. With probability

$$1 - \left\{1 \wedge \frac{p(\vartheta)q(\vartheta, \theta)}{p(\theta)q(\theta, \vartheta)}\right\},$$

stop and output θ.

Step 3. For $i = 1, 2, \ldots$, until $\sum_{j=1}^{i} w(z_j) + w(x_j) \geqslant 1$, sample $x_i \sim f_\theta$ and $z_i \sim f_\vartheta$. Set $N \leftarrow i$.

Step 4. If $w(z_N) = 1$, output ϑ; otherwise, output θ.

Our first results in § 3 concern $P_{1,N}$ and $P_{2,N}$. One would typically expect better performance from these kernels for larger values of N (Andrieu & Vihola, 2014), and such behaviour can often be demonstrated empirically. However, we establish that both of these kernels can nevertheless fail to be variance bounding regardless of the value of N when q proposes moves locally. This suggests that increasing N may only bring improvement up to a certain point. On the other hand, subsequent results for P_3 show that by expending more computational effort in particular places, one can successfully inherit variance bounding and/or geometric ergodicity from P_{MH}, the Metropolis–Hastings kernel with proposal q.

Because many of our positive results for P_3 are in relation to P_{MH}, we provide in Algorithm 4 the algorithmic specification for sampling from P_{MH}. In the approximate Bayesian computation setting, use of P_{MH} is ruled out by assumption since h cannot be computed. However, the preceding kernels are all, in some sense, exact approximations of P_{MH}.

Algorithm 4. To sample from $P_{\mathrm{MH}}(\theta, \cdot)$:

Step 1. Sample $\vartheta \sim q(\theta, \cdot)$.

Step 2. With probability

$$1 \wedge \frac{p(\vartheta)h(\vartheta)q(\vartheta, \theta)}{p(\theta)h(\theta)q(\theta, \vartheta)},$$

output ϑ; otherwise, output θ.

The kernels share a similar structure, and $P_{2,N}$, P_3 and P_{MH} can each be written as

$$P(\theta, \mathrm{d}\vartheta) = q(\theta, \mathrm{d}\vartheta)\alpha(\theta, \vartheta) + \left\{ 1 - \int_{\Theta} q(\theta, \mathrm{d}\theta')\alpha(\theta, \theta') \right\} \delta_{\theta}(\mathrm{d}\vartheta), \tag{5}$$

where δ_x denotes the Dirac measure centred at x. Evidently, only the acceptance probability $\alpha(\theta, \vartheta)$ differs for the three kernels. The kernel $P_{1,N}$ can be represented similarly, with modifications to account for its evolution on the extended space $\Theta \times \mathsf{Y}^N$. The representation (5) is used extensively in our analysis, and for $P_{2,N}$, P_3 and P_{MH} we have, respectively,

$$\alpha_{2,N}(\theta, \vartheta) = \int_{\mathsf{Y}^N} \int_{\mathsf{Y}^{N-1}} \left[1 \wedge \frac{c(\vartheta, \theta) \sum_{j=1}^{N} w(z_j)}{c(\theta, \vartheta)\{1 + \sum_{j=1}^{N-1} w(x_j)\}} \right] f_{\theta}^{\otimes N-1}(\mathrm{d}x_{1:N-1}) f_{\vartheta}^{\otimes N}(\mathrm{d}z_{1:N}), \tag{6}$$

$$\alpha_3(\theta, \vartheta) = \left\{ 1 \wedge \frac{c(\vartheta, \theta)}{c(\theta, \vartheta)} \right\} \frac{h(\vartheta)}{h(\theta) + h(\vartheta) - h(\theta)h(\vartheta)}, \tag{7}$$

$$\alpha_{\mathrm{MH}}(\theta, \vartheta) = 1 \wedge \frac{c(\vartheta, \theta)h(\vartheta)}{c(\theta, \vartheta)h(\theta)}, \tag{8}$$

where $c(\theta, \vartheta) = p(\theta)q(\theta, \vartheta)$ and (7) is obtained as in Lee (2012), for example.

3. THEORETICAL PROPERTIES

We assume that Θ is a metric space and

$$H = \int_{\Theta} p(\theta)h(\theta)\,\mathrm{d}\theta$$

satisfies $H \in (0, \infty)$ so that π is well-defined. We allow p to be improper, i.e., $\int_{\Theta} p(\theta)\,\mathrm{d}\theta$ to be infinite, but when p is proper we assume it is normalized so that $\int_{\Theta} p(\theta)\,\mathrm{d}\theta = 1$. Letting A^{C} denote the complement of a set A, we define the set of local proposals as

$$\mathcal{Q} = \left\{ q : \text{for all } \delta > 0 \text{ there exists } r \in (0, \infty) \text{ such that for all } \theta \in \Theta, \ q(\theta, B_{r,\theta}^{\mathrm{C}}) < \delta \right\}.$$

Membership of $\mathcal{Q}$ corresponds to tightness of $\{q(\theta, \cdot)\}_{\theta \in \Theta}$, when suitably centred, and this definition encompasses a broad range of common choices in practice, e.g., random walk proposals.

Let $\mathcal{V}$ and $\mathcal{G}$ denote the collections of reversible kernels that are, respectively, variance bounding (Roberts & Rosenthal, 2008) and geometrically ergodic (see, e.g., Roberts & Rosenthal, 2004; Meyn & Tweedie, 2009); so $\mathcal{G} \subset \mathcal{V}$. In our analysis, we make use of the following conditions.

Condition 1. The proposal q is a member of $\mathcal{Q}$. In addition, $\pi(B_{r,0}^{\mathrm{C}}) > 0$ for all $r > 0$ but $\lim_{v \to \infty} \sup_{\theta \in B_{v,0}^{\mathrm{C}}} h(\theta) = 0$.

Condition 2. The proposal q is a member of $\mathcal{Q}$. In addition, for all $K > 0$, there exists an $M_K \in [1, \infty)$ such that for all (θ, ϑ) in the set

$$\big\{ (\theta, \vartheta) \in \Theta^2 : \vartheta \in B_{K,\theta}, \ \pi(\theta)q(\theta, \vartheta) \wedge \pi(\vartheta)q(\vartheta, \theta) > 0 \big\},$$

either $h(\vartheta)/h(\theta) \in [M_K^{-1}, M_K]$ or $c(\vartheta, \theta)/c(\theta, \vartheta) \in [M_K^{-1}, M_K]$.

Condition 1 ensures that the posterior has mass arbitrarily far from zero but that $h(\theta)$ gets arbitrarily small as we move away from some compact set in Θ, while Condition 2 constrains the interplay between the likelihood and the prior-proposal pair. For example, Condition 2 is satisfied for symmetric q when p is continuous and everywhere positive with exponential or heavier tails; alternatively, it is satisfied if the likelihood is continuous and everywhere positive and decays at most exponentially fast. Conditions 1 and 2 are not mutually exclusive.

Remark 1. A global variant of Condition 2 can be defined where q need not be a member of $\mathcal{Q}$ but there exists an $M \in [1, \infty)$ such that for all (θ, ϑ) in the set $\{(\theta, \vartheta) \in \Theta^2 : \pi(\theta)q(\theta, \vartheta) \wedge \pi(\vartheta)q(\vartheta, \theta) > 0\}$, either $h(\vartheta)/h(\theta) \in [M^{-1}, M]$ or $c(\vartheta, \theta)/c(\theta, \vartheta) \in [M^{-1}, M]$. Theorems 3 and 4, which hold under Condition 2, also hold under this variant, with simplified proofs that are omitted here.

We first present a general theorem that supplements Theorem 5.1 of Roberts & Tweedie (1996) for reversible kernels, indicating that lack of geometric ergodicity due to arbitrarily sticky states coincides with lack of variance bounding. The proofs are given in the Appendix and the Supplementary Material.

THEOREM 1. *For any ν not concentrated at a single point and any reversible, irreducible, ν-invariant Markov kernel P such that $P(\theta, \{\theta\})$ is a measurable function, if $\nu -$ ess $\sup_{\theta} P(\theta, \{\theta\}) = 1$, then P is not variance bounding.*

Our first result concerning the kernels under study is negative, and indicates that performance of $P_{1,N}$ and $P_{2,N}$ under Condition 1 can be poor, irrespective of the value of N.

THEOREM 2. *Under Condition 1, $P_{1,N} \notin \mathcal{V}$ and $P_{2,N} \notin \mathcal{V}$ for all $N \in \mathbb{N}$.*

Remark 2. Theorem 2 immediately implies that under Condition 1, $P_{1,N} \notin \mathcal{G}$ and $P_{2,N} \notin \mathcal{G}$ by Theorem 1 of Roberts & Rosenthal (2008). The former implication is not covered by Theorem 8 of Andrieu & Roberts (2009) or Propositions 9 and 12 of Andrieu & Vihola (2014), because what they call weights in this context, namely $w(x)/h(\theta)$, are bounded above by $h(\theta)^{-1}$ for π-almost every $\theta \in \Theta$ and f_θ-almost every $x \in \mathsf{Y}$ but are not uniformly bounded in θ.

We emphasize that the choice of q is crucial in establishing Theorem 2. Since $H > 0$, if $q(\theta, \vartheta) = g(\vartheta)$, for instance, and $\sup_\theta p(\theta)/g(\theta) < \infty$, then by Theorem 2.1 of Mengersen & Tweedie (1996) one has that $P_{1,N}$ is uniformly ergodic and hence belongs to $\mathcal{G}$. Uniform ergodicity, however, does little to motivate the use of an independent proposal in challenging scenarios, particularly when Θ is high-dimensional.

Remark 3. We observe from (2) that when $\lim_{v \to \infty} \sup_{\theta \in B_{v,0}^C} h(\theta) = 0$ holds for a given $\epsilon = \epsilon_0$, it must hold for all $\epsilon \in (0, \epsilon_0]$. Furthermore, often this condition holds because $\lim_{v \to \infty} \sup_{\theta \in B_{v,0}^C} f_\theta(C) = 0$ for any compact subset C of Y. In such cases, $\lim_{v \to \infty}$

$\sup_{\theta \in B^C_{v,0}} h(\theta) = 0$ for any finite $\epsilon > 0$, and Theorem 2 will correspondingly hold for any finite $\epsilon > 0$ such that $\pi(B^C_{r,0}) > 0$ for all $r > 0$.

Our negative result above is not exclusive to the particular approximate Bayesian computation set-up considered here. In the Appendix we provide supplementary results which indicate that the theorem can be extended to the use of autoregressive proposals not covered by $\mathcal{Q}$, approximations of the likelihood of a more general form than (1), and Markov kernels with an invariant distribution in which ϵ is a nondegenerate auxiliary variable, cases which do arise in practice (see, e.g., Bortot et al., 2007; Sisson & Fan, 2011). However, the following results do not apply to these alternative settings, since P_3 lacks an obvious analogue when the artificial likelihood is not given by (1).

Our next three results concern P_3, and demonstrate first that variance bounding of P_{MH} is a necessary condition for variance bounding of P_3, and further that P_{MH} is at least as good as P_3 in terms of the asymptotic variance of estimates such as (3). More importantly, and in contrast to $P_{1,N}$ and $P_{2,N}$, P_3 can systematically inherit variance bounding and geometric ergodicity properties from P_{MH} under Condition 2.

PROPOSITION 1. *The Markov kernels P_3 and P_{MH} are ordered in the sense of Peskun (1973) and Tierney (1998), so $P_3 \in \mathcal{V} \Rightarrow P_{\mathrm{MH}} \in \mathcal{V}$ and $\mathrm{var}(P_{\mathrm{MH}}, \varphi) \leqslant \mathrm{var}(P_3, \varphi)$.*

THEOREM 3. *Under Condition 2, $P_{\mathrm{MH}} \in \mathcal{V} \Rightarrow P_3 \in \mathcal{V}$.*

THEOREM 4. *Under Condition 2, $P_{\mathrm{MH}} \in \mathcal{G} \Rightarrow P_3 \in \mathcal{G}$.*

Remark 4. Proposition 1 and Theorems 3 and 4 are precise in the following sense: there exist models for which $P_3 \in \mathcal{V} \setminus \mathcal{G}$ and $P_{\mathrm{MH}} \in \mathcal{V} \setminus \mathcal{G}$ and there exist models for which $P_3 \in \mathcal{G}$ and $P_{\mathrm{MH}} \in \mathcal{V} \setminus \mathcal{G}$; that is, under Condition 2, $P_{\mathrm{MH}} \in \mathcal{V} \nRightarrow P_3 \in \mathcal{G}$ and $P_3 \in \mathcal{G} \nRightarrow P_{\mathrm{MH}} \in \mathcal{G}$. These possibilities are illustrated in § 4·1.

Remark 5. While Condition 2 is only a sufficient condition, counterexamples can be constructed to show that some assumptions are necessary for Theorems 3 and 4 to hold. Condition 2 allows us to ensure that $\alpha_{\mathrm{MH}}(\theta, \vartheta)$ and $\alpha_3(\theta, \vartheta)$ differ only in a controlled manner, for all θ and enough ϑ, and hence that P_{MH} and P_3 are not too different. As an example of the possible differences between P_{MH} and P_3 more generally, consider the case where $p(\theta) = \tilde{p}(\theta)/\psi(\theta)$ and $h(\theta) = \tilde{h}(\theta)\psi(\theta)$ for some $\psi : \Theta \to (0, 1]$. Then properties of P_{MH} depend only on $\tilde{p}$ and $\tilde{h}$, while those of P_3 can additionally be dramatically altered by the choice of ψ.

Theorem 4 can be used to provide sufficient conditions for $P_3 \in \mathcal{G}$ through $P_{\mathrm{MH}} \in \mathcal{G}$ and Condition 2. The regular contour condition obtained in Theorem 4.3 of Jarner & Hansen (2000), for example, implies the following corollary.

COROLLARY 1. *Assume that* (a) *h decays super-exponentially and p has exponential or heavier tails, or* (b) *p has super-exponential tails and h decays exponentially or slower. Suppose, moreover, that π is continuous and everywhere positive, q is symmetric and satisfies $q(\theta, \vartheta) \geqslant \varepsilon_q$ whenever $|\theta - \vartheta| \leqslant \delta_q$ for some $\varepsilon_q, \delta_q > 0$, and*

$$\limsup_{|\theta| \to \infty} \frac{\theta}{|\theta|} \cdot \frac{\nabla \pi(\theta)}{|\nabla \pi(\theta)|} < 0,$$

where $\cdot$ denotes the Euclidean scalar product. Then $P_3 \in \mathcal{G}$.

Following Remark 1, an alternative condition, independent of the choice of q, which ensures inheritance of variance bounding and geometric ergodicity of P_3 from P_{MH} is that $\inf_{\theta \in \Theta} h(\theta) > 0$, i.e., that h is bounded below away from zero. This condition will usually only hold when Θ is compact. Under this condition, both $P_{1,N}$ and $P_{2,N}$ will also successfully inherit these properties; the former has already been shown in Andrieu & Vihola (2014, Proposition 9), and for $P_{2,N}$ the same type of argument can be used. This allows us to state the following corollary, which can be verified by the arguments in Roberts & Rosenthal (2004, § 3.3).

COROLLARY 2. *Let Θ be compact with q, p and h all continuous, such that $\inf_{\theta, \vartheta \in \Theta} q(\theta, \vartheta) > 0$ and $\inf_{\theta \in \Theta} h(\theta) > 0$. Then $P_{1,N}$, $P_{2,N}$ and P_3 are all geometrically ergodic.*

Remark 6. Under the conditions of Corollary 2, $P_{1,N}$, $P_{2,N}$ and P_3 are all uniformly ergodic since the ratio of the acceptance probabilities, $\alpha_{\mathrm{MH}}(\theta, \vartheta)/\alpha_i(\theta, \vartheta)$, is bounded above by a constant for $i \in \{1, 2, 3\}$. This suggests that in approximate Bayesian computation, a conservative choice is to restrict inference to a compact set Θ in which h is bounded below.

The proofs of Theorems 3 and 4 can be extended to cover the case where $\tilde{P}_{\mathrm{MH}}$ is a finite, countable or continuous mixture of P_{MH} kernels associated with a collection of proposals $\{q_s\}_{s \in S}$ and $\tilde{P}_3$ is the corresponding mixture of P_3 kernels. With a modification of Condition 2, the following proposition is stated without proof, and could be used in conjunction with Theorem 3 of Fort et al. (2003), for instance.

Condition 3. Each proposal q is a member of $\mathcal{Q}$. In addition, for all $K > 0$, there exists an $M_K \in [1, \infty)$ such that for all $q_t \in \{q_s\}_{s \in S}$ and (θ, ϑ) in the set

$$\left\{ (\theta, \vartheta) \in \Theta^2 : \vartheta \in B_{K,\theta}, \ \pi(\theta)q_t(\theta, \vartheta) \wedge \pi(\vartheta)q_t(\vartheta, \theta) > 0 \right\},$$

either $\quad h(\vartheta)/h(\theta) \in [M_K^{-1}, M_K] \quad$ or $\quad c_t(\vartheta, \theta)/c_t(\theta, \vartheta) \in [M_K^{-1}, M_K], \quad$ where $\quad c_t(\theta, \vartheta) = p(\theta)q_t(\theta, \vartheta)$.

PROPOSITION 2. *Let $\tilde{P}_{\mathrm{MH}}(\theta, \mathrm{d}\vartheta) = \int_S \mu(\mathrm{d}s) P_{\mathrm{MH}}^{(s)}(\theta, \mathrm{d}\vartheta)$, where μ is a mixing distribution on S and each $P_{\mathrm{MH}}^{(s)}$ is a π-invariant Metropolis–Hastings kernel with proposal q_s. Let $\tilde{P}_3(\theta, \mathrm{d}\vartheta) = \int_S \mu(\mathrm{d}s) P_3^{(s)}(\theta, \mathrm{d}\vartheta)$ be defined analogously. Then $\tilde{P}_3 \in \mathcal{V} \Rightarrow \tilde{P}_{\mathrm{MH}} \in \mathcal{V}$ and $\mathrm{var}(\tilde{P}_{\mathrm{MH}}, \varphi) \leqslant \mathrm{var}(\tilde{P}_3, \varphi)$, and under Condition 3, both $\tilde{P}_{\mathrm{MH}} \in \mathcal{V} \Rightarrow \tilde{P}_3 \in \mathcal{V}$ and $\tilde{P}_{\mathrm{MH}} \in \mathcal{G} \Rightarrow \tilde{P}_3 \in \mathcal{G}$.*

We also provide a general result that can justify using, for example, P_3 as one component of a mixture of reversible kernels, some of which may not be variance bounding or geometrically ergodic.

THEOREM 5. *Let $\tilde{K} = \sum_{i=1}^{\infty} a_i K_i$ be a mixture of reversible Markov kernels with invariant distribution π, where $\sum_{i=1}^{\infty} a_i = 1$ and $a_i \geqslant 0$ for $i \in \mathbb{N}$. Suppose that K_1 has the unique invariant distribution π and $a_1 > 0$. Then $K_1 \in \mathcal{V} \Rightarrow \tilde{K} \in \mathcal{V}$ and $K_1 \in \mathcal{G} \Rightarrow \tilde{K} \in \mathcal{G}$.*

While the sampling of a random number of auxiliary variables in the implementation of P_3 appears to be helpful for the inheritance of qualitative properties from P_{MH}, one concern is that the computational effort associated with the kernel may be unbounded. Our final result indicates that this is not the case whenever p is proper.

PROPOSITION 3. *Let (N_i) be the sequence of random variables associated with step 3 of Algorithm 3 when one iterates P_3, with $N_j = 0$ if at iteration j the kernel outputs at step 2.*

Then, if $\int p(\theta) \, d\theta = 1$, $H > 0$ and P_3 is irreducible,

$$n = \lim_{m \to \infty} m^{-1} \sum_{i=1}^{m} N_i \leqslant H^{-1} < \infty.$$

When p is proper, H is a natural quantity. If n_R is the expected number of proposals to obtain a sample from π using the rejection sampler of Pritchard et al. (1999), we have $n_R = H^{-1}$; and if we construct $P_{1,N}$ with the proposal $q(\theta, \vartheta) = p(\vartheta)$, then H places a lower bound on its spectral gap. In fact, n can be arbitrarily smaller than n_R, as we illustrate in § 4·1, and for a realistic example in § 4·3 the average number of samples required per iteration was much smaller than H^{-1}.

One potential issue with all three of the kernels $P_{1,N}$, $P_{2,N}$ and P_3, when implemented using local proposals, is that their performance for a fixed computational budget will be poor if the Markov chain is initialized in a region of the state space with little posterior mass. This can be circumvented by trying to identify regions of high posterior mass and initializing the chain at a point in such a region. Finally, Remark 6 suggests that a conservative choice is to take Θ to be a compact set in which h is bounded below; this would contain most of the interesting values of θ.

4. Examples

4·1. *A posterior density with compact support*

We begin with a simple example that clarifies the comments in Remark 4 and some of those following Proposition 3. In particular, let $\theta \in \Theta = \mathbb{R}_+$, $p(\theta) = I(0 \leqslant \theta \leqslant a)/a$ and $h(\theta) = bI(0 \leqslant \theta \leqslant 1)$ for $(a, b) \in [1, \infty) \times (0, 1]$, with π supported on $[0, 1]$.

We have $H^{-1} = a/b$ and $n \leqslant b^{-1}$ for any q, so $n_R/n \geqslant a$. Furthermore, even if p is improper, n is finite. Regarding Remark 4, for any $a \geqslant 1$, consider the proposal $q(\theta, \vartheta) = 2I(0 \leqslant \theta \leqslant 1/2) I(1/2 < \vartheta \leqslant 1) + 2I(1/2 < \theta \leqslant 1) I(0 \leqslant \vartheta \leqslant 1/2)$. If $b = 1$, then $P_3 \in \mathcal{V} \setminus \mathcal{G}$ and $P_{\mathrm{MH}} \in \mathcal{V} \setminus \mathcal{G}$. However, if $b \in (0, 1)$, then $P_3 \in \mathcal{G}$ and $P_{\mathrm{MH}} \in \mathcal{V} \setminus \mathcal{G}$.

4·2. *Geometric distribution*

We consider the situation where $\theta \in \Theta = \mathbb{Z}_+$, $p(\theta) = I(\theta \in \mathbb{N})(1 - a)a^{\theta-1}$ and $h(\theta) = b^{\theta}$ for $(a, b) \in (0, 1)^2$. The posterior π is a geometric distribution with success parameter $1 - ab$, and geometric series manipulations shown in the Supplementary Material give the expected number of proposals needed in the rejection sampler, $n_R = (1 - ab)/\{b(1 - a)\}$. If $q(\theta, \vartheta) = \{I(\vartheta = \theta - 1) + I(\vartheta = \theta + 1)\}/2$, then

$$\frac{(1 - ab)}{2} \left\{ \frac{(a + b)}{b(1 - a)(1 + b)} - 1 \right\} \leqslant n \leqslant \frac{(1 - ab)}{2} \left\{ \frac{a + b}{b(1 - a)} - 1 \right\}, \qquad (9)$$

where n is as in Proposition 3, and so $n_R/n \geqslant 2/\{a(1 + b)\}$, which grows without bound as $a \to 0$. Regarding the propriety condition on p, we observe that $n_R \to \infty$ and $n \to \infty$ as $a \to 1$ with b fixed.

To supplement the qualitative results on variance bounding and geometric ergodicity of the kernels, we investigated a modification of this example with a finite number of states. More specifically, we considered the case where the prior is truncated to the set $\{1, \ldots, D\}$ for some $D \in \mathbb{N}$. In this context, we can calculate explicit transition probabilities and hence spectral gaps $1 - |\sigma_0(P)|$ and asymptotic variances $\mathrm{var}(P, \varphi)$ of (3) for $P_{2,N}$, P_3 and P_{MH}. Figure 1(a) shows the log spectral gaps for a range of values of D for each kernel and $a = b = 0.5$. The spectral gaps
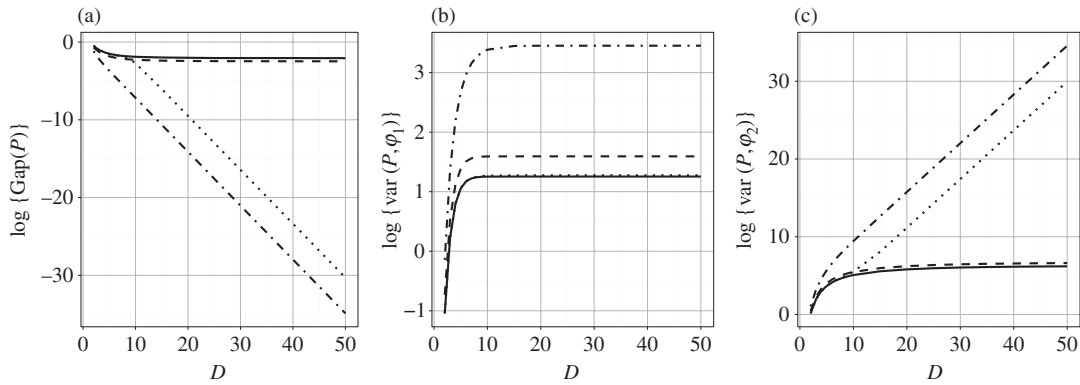
Fig. 1. Logarithmic plots of (a) the spectral gap, (b) var(P, φ_1), and (c) var(P, φ_2) against D for $P_{2,1}$ (dot-dash), $P_{2,100}$ (dotted), P_3 (dashed) and P_{MH} (solid), with $a = b = 0\cdot5$.

of P_3 and P_{MH} stabilize, whereas those of $P_{2,N}$ decrease exponentially fast in D, albeit with some improvement for larger N. The spectral gaps obtained with (4) suggest that the convergence of $P_{2,N}$ to π can be extremely slow for some θ_0, even when D is relatively small. Indeed, in this finite, discrete setting with reversible P, the bounds

$$\frac{1}{2}\left\{\max|\sigma_0(P)|\right\}^m \leqslant \max_{\theta_0} \|\pi(\cdot) - P^m(\theta_0, \cdot)\|_{TV} \leqslant \frac{1}{2}\left\{\max|\sigma_0(P)|\right\}^m \left\{\frac{1 - \min_\theta \pi(\theta)}{\min_\theta \pi(\theta)}\right\}^{1/2}$$

hold (Montenegro & Tetali, 2006, §2 and Theorem 5.9), which clearly indicate that $P_{2,N}$ can converge exceedingly slowly when P_3 and P_{MH} converge reasonably quickly. The value of n in this case stabilized at $0\cdot847$, within the bounds of (9) and considerably smaller than 100.

In Fig. 1(b) and (c), log$\{$var$(P, \varphi)\}$ is plotted against D for $\varphi_1(\theta) = \theta$ and $\varphi_2(\theta) = (ab)^{-\theta/2\cdot1}$, respectively, computed using the expression in Kemeny & Snell (1969, p. 84). The choice of φ_2 is motivated by the fact that when p is not truncated, $\varphi(\theta) = (ab)^{-\theta/(2+\delta)}$ is in $L^2(\pi)$ if and only if $\delta > 0$. While var(P, φ_1) is stable for all the kernels, var(P, φ_2) increases rapidly with D for $P_{2,1}$ and $P_{2,100}$. Although var$(P_{2,N}, \varphi_1)$ can be lower than var(P_3, φ_1), the former requires many more simulations from the likelihood. Indeed, while the results we have obtained pertain to qualitative properties of the Markov kernels, this example illustrates that P_3 can significantly outperform $P_{2,100}$ for estimating even the more well-behaved $\pi(\varphi_1)$, when cost per iteration of each kernel is taken into account. Additional figures in the Supplementary Material show similar graphs for the case where $a = 0\cdot5$ and $b \in \{0\cdot1, 0\cdot9\}$, and for these values of b the value of n stabilized at $4\cdot77$ and $0\cdot502$, respectively.

Figure 2 plots log$\{$var$(P, \varphi_{3,t})/\pi(\varphi_{3,t})\}$ against t for $\varphi_{3,t}(\theta) = I(\theta \geqslant t)$ so that $\pi(\varphi_{3,t})$ is a tail probability. The division by $\pi(\varphi_{3,t})$ makes this an appropriately scaled relative asymptotic variance, since one needs $1/\pi(\varphi_{3,t})$ perfect samples from π in expectation to get a single sample in the set $\{\theta : \theta \geqslant t\}$. While P_{MH} and P_3 have constant log$\{$var$(P, \varphi_{3,t})/\pi(\varphi_{3,t})\}$ as t increases, $P_{2,1}$ and $P_{2,100}$ do not, a result of their inability to estimate tail probabilities accurately. In various applications, approximate Bayesian computation could be used to infer such quantities, but our results here indicate that $P_{1,N}$ and $P_{2,N}$ may not be appropriate for this purpose.

The bounds in (9) imply that n will grow as $a \to 1$ with b fixed, and one might be concerned that P_3 could consequently be less computationally advantageous in comparison with $P_{2,N}$. We calculated asymptotic variances associated with the kernels for $b = 0\cdot5$ and $a \in \{0\cdot9, 0\cdot99, 0\cdot999\}$, and the corresponding values of n for P_3 were approximately 5, 50 and
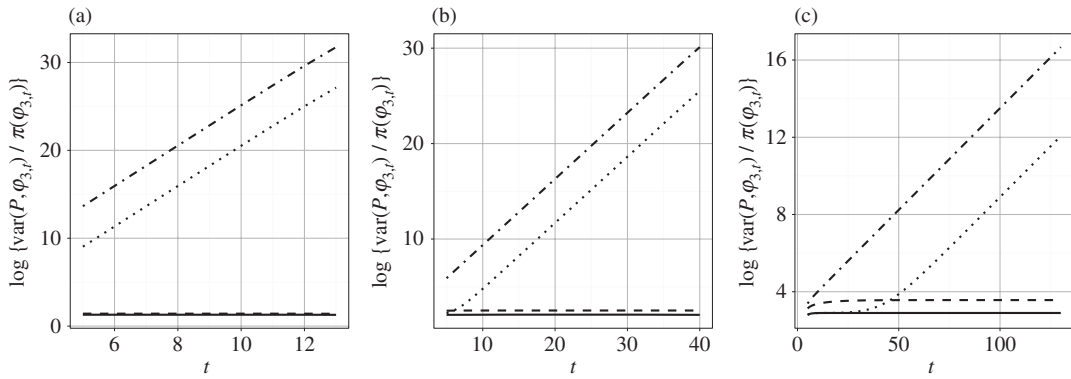
Fig. 2. Plots of $\log\{\mathrm{var}(P, \varphi_{3,t})/\pi(\varphi_{3,t})\}$ against t for $P = P_{2,1}$ (dot-dash), $P = P_{2,100}$ (dotted), $P = P_3$ (dashed) and $P = P_{\mathrm{MH}}$ (solid), with $a = 0.5$ and (a) $b = 0.1$, (b) $b = 0.5$, (c) $b = 0.9$.

500. Graphs are shown in the Supplementary Material. To take into account the cost of the kernels, we can compare $N\mathrm{var}(P_{2,N}, \varphi_1)$ with $n\,\mathrm{var}(P_3, \varphi_1)$, and for these values of a we have $\mathrm{var}(P_{2,1}, \varphi_1) \approx 100\,\mathrm{var}(P_{2,100}, \varphi_1)$, although $P_{2,100}$ can more feasibly be implemented in parallel on many-core devices such as graphics processing units (see, e.g., Lee et al., 2010). On the other hand, $\mathrm{var}(P_{2,1}, \varphi_1)/\{n\,\mathrm{var}(P_3, \varphi_1)\}$ is about 75, 5000 and well over 60 000 for $a = 0.9, 0.99$ and 0.999, respectively, indicating that the relative performance of P_3 increases rapidly as $a \to 1$.

4·3. *Stochastic Lotka–Volterra model*

We now turn to stochastic kinetic models for which the posterior does not take a simple form and exhibits strong correlations between components of θ. Such models are used, for example, in systems biology, where Bayesian inference has been investigated by Wilkinson (2006) and Boys et al. (2008). We consider a simple member of this class of models, the Lotka–Volterra predator-prey model (Lotka, 1925; Volterra, 1926), which was also considered as an example for approximate Bayesian computation in Toni et al. (2009) and Fearnhead & Prangle (2012).

In this setting, $X_{1:2}(t)$ is a bivariate, integer-valued pure jump Markov process with $X_{1:2}(0) = (50, 100)$. For small Δt, we have

$$\mathrm{pr}\{X_{1:2}(t + \Delta t) = z_{1:2} \mid X_{1:2}(t) = x_{1:2}\}$$

$$= \begin{cases} \theta_1 x_1 \Delta t + o(\Delta t), & z_{1:2} = (x_1 + 1, x_2), \\ \theta_2 x_1 x_2 \Delta t + o(\Delta t), & z_{1:2} = (x_1 - 1, x_2 + 1), \\ \theta_3 x_2 \Delta t + o(\Delta t), & z_{1:2} = (x_1, x_2 - 1), \\ 1 - \Delta t\,(\theta_1 x_1 + \theta_2 x_1 x_2 + \theta_3 x_2) + o(\Delta t), & z_{1:2} = x_{1:2}, \\ o(\Delta t) & \text{otherwise}, \end{cases}$$

where the first three cases correspond to prey birth, prey consumption and predator death. Theory and methods relating to simulation of this type of time-homogeneous, pure jump Markov process and its historical uses in statistics can be traced through Feller (1940), Doob (1945) and Kendall (1949, 1950), and the method was rediscovered by Gillespie (1977) in the context of stochastic kinetic models. These articles develop a straightforward way to simulate the full process $X_{1:2}(t)$, $t \in [0, 10]$, as the interjump times are exponential random variables, although more sophisticated approaches are possible (see, e.g., Wilkinson, 2006, Ch. 8).

The data were simulated with $\theta = (1, 0.005, 0.6)$ (Wilkinson, 2006, p. 152). Our observations are both partial and discrete, with $y = \{88, 165, 274, 268, 114, 46, 32, 36, 53, 92\}$ being the
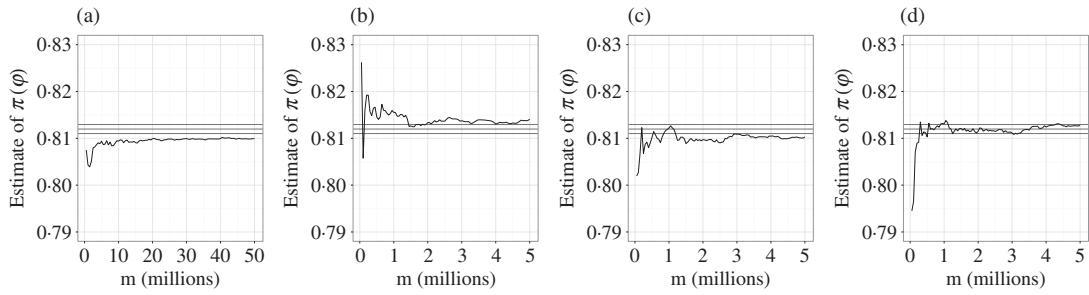
Fig. 3. Estimates of the posterior mean of θ_3 by iteration using the kernels (a) $P_{1,1}$, (b) $P_{1,15}$, (c) $P_{2,15}$, and (d) P_3. In each panel, the three horizontal lines represent the estimate obtained using the rejection sampler, with two estimated standard deviations added and subtracted.
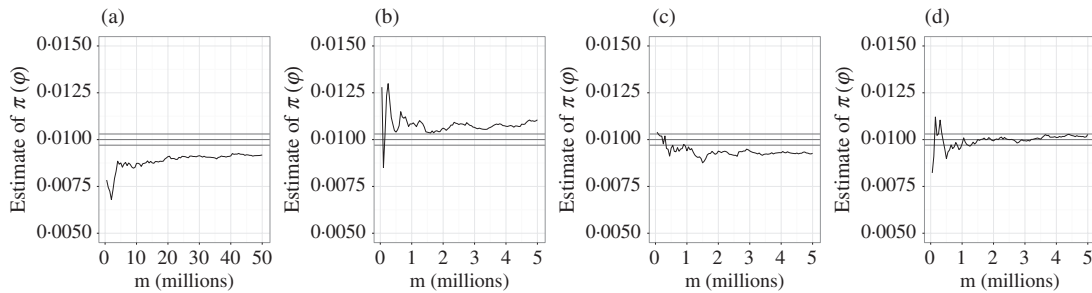

Fig. 4. Estimates of $\pi(\theta_3 \geqslant 1\cdot79)$ by iteration using the kernels (a) $P_{1,1}$, (b) $P_{1,15}$, (c) $P_{2,15}$, and (d) P_3. In each panel, the three horizontal lines represent the estimate obtained using the rejection sampler, with two estimated standard deviations added and subtracted.

simulated values of X_1 at times $\{1, 2, \ldots, 10\}$; for approximate Bayesian computation we use a log transformation of $X_1(t)$ and $y(t)$ with $\epsilon = 1$, i.e.,

$$B_{\epsilon,y} = \big\{ X_1(t) : \log\{X_1(i)\} - \log\{y(i)\} \leqslant \epsilon, \quad i \in \{1, \ldots, 10\} \big\}.$$

We first model $\theta \in \Theta = [0, \infty)^3$ with $p(\theta) = 100 \exp(-\theta_1 - 100\,\theta_2 - \theta_3)$ and take $q(\theta, \vartheta) = \mathcal{N}(\vartheta; \theta, \Sigma)$ where $\Sigma = \mathrm{diag}(0\cdot25, 0\cdot0025, 0\cdot25)$. The choice of independent exponential priors on θ is motivated by Condition 2. Estimated marginal posterior densities, obtained using 10^6 samples from π with a rejection sampler, can be found in the Supplementary Material. The posterior of θ_1 is tighter than that of θ_3, and the samples indicate strong positive correlation between θ_2 and θ_3. In this setting, P_3 for 5×10^6 iterations gave an average value of 15 for n, and we also ran kernels $P_{1,1} = P_{2,1}$ for 5×10^7 iterations as well as $P_{1,15}$ and $P_{2,15}$ both for 5×10^6 iterations. All the kernels gave density estimates visibly indistinguishable from those in Fig. S5 of the Supplementary Material, but inspection of their partial sums by iteration reveals important differences. For each chain, we show in Fig. 3 estimates of the posterior mean of θ_3 and in Fig. 4 estimates of the probability that $\theta_3 \geqslant 1\cdot79$, accompanied by lines corresponding to the estimate obtained using the samples from the rejection sampler. The choice of $1\cdot79$ corresponds to an estimate of the 90th percentile using these latter samples. It seems that P_3 accurately estimates the correct value, with the uncertainty of the estimate being correlated with perturbations of the partial sum. However, the other kernels miss the value of interest by some amount, and particularly in the case of $P_{1,1}$, the perturbations of the partial sum over time are small, which may lead practitioners to believe mistakenly that the estimate has converged.

We performed a second analysis using a slightly different prior, with $p(\theta) = 0{\cdot}01 \exp(-\theta_1 - 0{\cdot}01\,\theta_2 - \theta_3)$, such that differences in the kernels are accentuated. Here, only the independent prior for θ_2 has changed to become less informative. In this case, a rejection sampler cannot practically be used to verify results, as the expected number of proposals required to obtain one sample by rejection is around $4{\cdot}5 \times 10^5$. The average value of n for P_3, however, was 13. Although not shown here, marginal posterior density estimates for the parameters using each kernel are close to those in Fig. S5 of the Supplementary Material, but the estimates corresponding to $P_{1,1}$ exhibit characteristic bumps in their tails. As above, we can inspect each chain's corresponding partial sums by iteration to uncover important differences. Figures S6 and S7 in the Supplementary Material show, respectively, estimates of the posterior mean of θ_2 and the posterior probability that $\theta_3 \geqslant 2$ for each chain, and the latter is illustrative of the inability of P_1 and P_2 to produce chains without long tail excursions.

In practical applications such as this, it may not be possible to determine easily whether or not P_{MH} is variance bounding or geometrically ergodic. However, Theorems 3 and 4 do establish that if P_{MH} has either of these properties, then P_3 will inherit them from P_{MH}. In practice, it is not unusual for the conditions of Corollary 1 to hold, and that might be the case here. Similarly, it is also quite common for Condition 1 to hold, and so one might expect that P_1 and P_2 are not variance bounding here.

5. Discussion

Our analysis shows that P_3 may be geometrically ergodic and/or variance bounding in a wide variety of situations where the kernels $P_{1,N}$ and $P_{2,N}$ are not. In practice, Condition 2 can be verified and used in prior and proposal choice to ensure that P_3 systematically inherits these properties from P_{MH}. Condition 2 is not necessary, but weaker conditions are likely to be complicated.

Theorems 2 and 3 together with Proposition 3, whose assumptions are not mutually exclusive, allow us to conclude that the behaviour of P_3 is characteristically different from that of $P_{1,N}$ and $P_{2,N}$ in some settings. A large expected number of simulations from f_θ and f_ϑ using P_3 could be viewed as analogous to being stuck for many iterations using $P_{1,N}$ or $P_{2,N}$. However, while the expected number of simulations and the asymptotic variance of (3) for any $\varphi \in L^2(\pi)$ are finite when using P_3 under the conditions of Theorem 3, there exist $\varphi \in L^2(\pi)$ for which a central limit theorem does not hold for (3) when using $P_{1,N}$ or $P_{2,N}$ under the conditions of Theorem 2.

Variance bounding and geometric ergodicity are likely to coincide in most applications of interest, as Metropolis–Hastings kernels that are variance bounding but not geometrically ergodic exhibit periodic behaviour rarely encountered in statistical inference. Bounds on the second largest eigenvalue and/or the spectral gap of P_3 in relation to properties of P_{MH} could be obtained through Cheeger-like inequalities using conductance arguments as in the proofs of Theorems 3 and 4, although these may be quite loose in some situations (see, e.g., Diaconis & Stroock, 1991) and we have not pursued them here. Finally, Roberts & Rosenthal (2011) have demonstrated that some simple Markov chains that are not geometrically ergodic can converge extremely slowly and that properties of such algorithms can be very sensitive to even slight parameter changes.

The theoretical results obtained in § 3 and the examples in § 4 provide some understanding of the relative qualitative merits of P_3 over $P_{1,N}$ and $P_{2,N}$. However, the results do not prove that P_3 should necessarily be uniformly preferred over $P_{2,N}$, although the examples do suggest that P_3 may have better asymptotic variance properties when taking the cost of simulations into account in a variety of scenarios. In addition, Theorem 5 can be used to justify the mixture of P_3 with alternative reversible kernels such as $P_{2,N}$, if desired.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs that are not given in the Appendix, extensions of Theorem 2, calculations for §4·2, and additional figures.

Appendix

Many of our proofs make use of the relationship between conductance, the spectrum of a Markov kernel, and variance bounding for reversible Markov kernels P. In particular, the conductance being positive, $\kappa > 0$, is equivalent to $\sup S(P) < 1$ (Lawler & Sokal, 1988, Theorem 2.1), which, as stated earlier, is equivalent to variance bounding. The conductance κ for a π-invariant transition kernel P on Θ is defined as

$$\kappa = \inf_{A:0<\pi(A)\leqslant 1/2} \kappa(A), \quad \kappa(A) = \pi(A)^{-1} \int_A P(\theta, A^C)\pi(\mathrm{d}\theta) = \int_\Theta P(\theta, A^C)\pi_A(\mathrm{d}\theta),$$

where $\pi_A(\mathrm{d}\theta) = \pi(\mathrm{d}\theta)I(\theta \in A)/\pi(A)$.

Finally, we make use of the fact that if $q \in \mathcal{Q}$, we can define the function

$$r_q(\delta) = \inf\left\{r : \text{for all } \theta \in \Theta,\ q(\theta, B_{r,\theta}^C) < \delta\right\}.$$

Proof of Theorem 1. If $\nu - \mathrm{ess\,sup}_\theta P(\theta, \{\theta\}) = 1$ and $P(\theta, \{\theta\})$ is measurable, then the set $A_\tau = \{\theta \in \Theta : P(\theta, \{\theta\}) \geqslant 1 - \tau\}$ is measurable and $\nu(A_\tau) > 0$ for every $\tau > 0$. Moreover, $a_0 = \lim_{\tau \searrow 0} \nu(A_\tau)$ exists, since $A_{\tau_2} \subset A_{\tau_1}$ for $\tau_2 < \tau_1$. Now, assume $a_0 > 0$ and define $A_0 = \{\theta \in \Theta : P(\theta, \{\theta\}) = 1\} = \bigcap_n A_{\tau_n}$, where $\tau_n \searrow 0$. By continuity from above, $\nu(A_0) = a_0 > 0$, and since ν is not concentrated at a single point, P is reducible, which is a contradiction. Hence $a_0 = 0$. Consequently, by taking $\tau_n \searrow 0$ with τ_1 small enough, we have $\nu(A_{\tau_n}) < 1/2$ for every n; and since

$$\kappa \leqslant \lim_n \kappa(A_{\tau_n}) = \lim_n \int_{A_{\tau_n}} P(\theta, A_{\tau_n}^C)\nu_{A_{\tau_n}}(\mathrm{d}\theta) \leqslant \lim_n \int_{A_{\tau_n}} P(\theta, \{\theta\}^C)\nu_{A_{\tau_n}}(\mathrm{d}\theta) = \lim_n \tau_n = 0,$$

we have that $P \notin \mathcal{V}$, which proves the theorem. $\square$

Proof of Theorem 2. We prove the result for $P_{2,N}$. The proof for $P_{1,N}$ is essentially identical, with minor adjustments for the extended state space, and so we omit it. By Theorem 1, it suffices to show that $\pi - \mathrm{ess\,sup}_\theta P_{2,N}(\theta, \{\theta\}) = 1$, i.e., for all $\tau > 0$, there exists $A \subseteq \Theta$ with $\pi(A) > 0$ such that for all $\theta \in A$, $P_{2,N}(\theta, \{\theta\}^C) \leqslant \tau$.

From Condition 1, $q \in \mathcal{Q}$. Given $\tau > 0$, let $r = r_q(\tau/2)$, $v = \inf\{v : \sup_{\theta \in B_v^c(0)} h(\theta) < 1 - (1 - \tau/2)^{1/N}\}$ and $A = B_{v+r,0}^C$. From Condition 1, $\pi(A) > 0$, and by (5) and (6) we have that for all $\theta \in A$,

$$P_{2,N}(\theta, \{\theta\}^C) = \int_{\{\theta\}^C} \int_{\mathsf{Y}^N} \int_{\mathsf{Y}^{N-1}} \left[1 \wedge \frac{c(\vartheta, \theta)\sum_{j=1}^N w(z_j)}{c(\theta, \vartheta)\{1 + \sum_{j=1}^{N-1} w(x_j)\}}\right] f_\theta^{\otimes N-1}(\mathrm{d}x_{1:N-1}) f_\vartheta^{\otimes N}(\mathrm{d}z_{1:N})q(\theta, \mathrm{d}\vartheta)$$

$$\leqslant \sup_{\theta \in \Theta} q\left(\theta, B_{r,\theta}^C\right) + \int_{B_{r,\theta}} \int_{\mathsf{Y}^N} I\left\{\sum_{i=1}^N w(z_i) \geqslant 1\right\} f_\vartheta^{\otimes N}(\mathrm{d}z_{1:N})q(\theta, \mathrm{d}\vartheta)$$

$$\leqslant \frac{\tau}{2} + \int_{B_{r,\theta}} \left[1 - \left\{1 - \sup_{\vartheta \in B_{r,\theta}} h(\vartheta)\right\}^N\right] q(\theta, \mathrm{d}\vartheta) \leqslant \tau,$$

as desired. $\square$

The following two lemmas are pivotal in the proofs of Proposition 1 and Theorems 3 and 4, and make extensive use of (5), (7) and (8). Their proofs can be found in the Supplementary Material.

LEMMA A1. *For any $\theta \in \Theta$, $P_3(\theta, \{\theta\}) \geqslant P_{MH}(\theta, \{\theta\})$.*

LEMMA A2. *Assume Condition 2. For π-almost all θ and any $A \subseteq \Theta$ such that $\theta \in A$ and $r > 0$,*

$$P_{MH}(\theta, A^C) \leqslant \sup_\theta q(\theta, B_{r,\theta}^C) + (1 + M_r)P_3(\theta, A^C),$$

where M_r is as defined in Condition 2.

Proof of Theorem 3. We prove the result under Condition 2. Let κ_{MH} and κ_3 be the conductances of P_{MH} and P_3, respectively, and let A be a measurable set with $\pi(A) > 0$. Since $q \in \mathcal{Q}$, we let $R = r_q(\kappa_{MH}/2)$ and take M_R as in Condition 2. Then, by Lemma A2,

$$\kappa_{MH}(A) = \int_\Theta P_{MH}(\theta, A^C)\pi_A(d\theta) \leqslant \frac{\kappa_{MH}}{2} + (1 + M_R)\int_\Theta P_3(\theta, A^C)\pi_A(d\theta)$$
$$= \frac{\kappa_{MH}}{2} + (1 + M_R)\kappa_3(A).$$

Since A is arbitrary, we conclude that $\kappa_{MH} \leqslant 2(1 + M_R)\kappa_3$, and so $\kappa_{MH} > 0 \Rightarrow \kappa_3 > 0$. □

References

ANDRIEU, C. & ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37**, 697–725.

ANDRIEU, C. & VIHOLA, M. (2014). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Prob.* to appear.

BAXENDALE, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Prob.* **15**, 700–38.

BEAUMONT, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–60.

BECQUET, C. & PRZEWORSKI, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* **17**, 1505–19.

BEDNORZ, W. & ŁATUSZYŃSKI, K. (2007). A few remarks on "Fixed-width output analysis for Markov chain Monte Carlo" by Jones et al. *J. Am. Statist. Assoc.* **102**, 1485–6.

BORTOT, P., COLES, S. G. & SISSON, S. A. (2007). Inference for stereological extremes. *J. Am. Statist. Assoc.* **102**, 84–92.

BOYS, R. J., WILKINSON, D. J. & KIRKWOOD, T. B. L. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statist. Comp.* **18**, 125–35.

CHEN, J., KÄLLMAN, T., GYLLENSTRAND, N. & LASCOUX, M. (2009). New insights on the speciation history and nucleotide diversity of three boreal spruce species and a Tertiary relict. *Heredity* **104**, 3–14.

DEL MORAL, P., DOUCET, A. & JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Statist. Soc.* B **68**, 411–36.

DEL MORAL, P., DOUCET, A. & JASRA, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statist. Comp.* **22**, 1009–20.

DIACONIS, P. & STROOCK, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Prob.* **1**, 36–61.

DOOB, J. L. (1945). Markoff chains—denumerable case. *Trans. Am. Math. Soc.* **58**, 455–73.

DROVANDI, C. C. & PETTITT, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics* **67**, 225–33.

FEARNHEAD, P. & PRANGLE, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Statist. Soc.* B **74**, 419–74.

FELLER, W. (1940). On the integro-differential equations of purely discontinuous Markoff processes. *Trans. Am. Math. Soc.* **48**, 488–515.

FLEGAL, J. M. & JONES, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.* **38**, 1034–70.

FORT, G., MOULINES, E., ROBERTS, G. O. & ROSENTHAL, J. S. (2003). On the geometric ergodicity of hybrid samplers. *J. Appl. Prob.* **40**, 123–46.

GEYER, C. J. & MIRA, A. (2000). On non-reversible Markov chains. In *Monte Carlo Methods*, vol. 26 of *Fields Institute Communications*. Providence, Rhode Island: American Mathematical Society, pp. 93–108.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–61.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hobert, J. P., Jones, G. L., Presnell, B. & Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika* **89**, 731–43.

Jarner, S. F. & Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stoch. Proces. Appl.* **85**, 341–61.

Jasra, A. & Doucet, A. (2008). Stability of sequential Monte Carlo samplers via the Foster–Lyapunov condition. *Statist. Prob. Lett.* **78**, 3062–9.

Jones, G. L., Haran, M., Caffo, B. S. & Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Am. Statist. Assoc.* **101**, 1537–47.

Kemeny, J. G. & Snell, J. L. (1969). *Finite Markov Chains*. Princeton, New Jersey: Van Nostrand.

Kendall, D. G. (1949). Stochastic processes and population growth. *J. R. Statist. Soc.* B **11**, 230–82.

Kendall, D. G. (1950). An artificial realization of a simple "birth-and-death" process. *J. R. Statist. Soc.* B **12**, 116–9.

Kim, S. K., Carbone, L., Becquet, C., Mootnick, A. R., Li, D. J., de Jong, P. J. & Wall, J. D. (2011). Patterns of genetic variation within and between gibbon species. *Molec. Biol. Evol.* **28**, 2211–8.

Kontoyiannis, I. & Meyn, S. P. (2012). Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Prob. Theory Rel. Fields* **154**, 327–39.

Lawler, G. F. & Sokal, A. D. (1988). Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Trans. Am. Math. Soc.* **309**, 557–80.

Lee, A. (2012). On the choice of MCMC kernels for approximate Bayesian computation with SMC samplers. In *Proc. 2012 Winter Simul. Conf.*, C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose & A. M. Uhrmacher, eds. New York: Association for Computing Machinery. Article no. 27.

Lee, A., Andrieu, C. & Doucet, A. (2012). Discussion of paper by P. Fearnhead and D. Prangle. *J. R. Statist. Soc.* B **74**, 419–74.

Lee, A., Yau, C., Giles, M. B., Doucet, A. & Holmes, C. C. (2010). On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *J. Comp. Graph. Statist.* **19**, 769–89.

Li, Y., Stocks, M., Hemmilä, S., Källman, T., Zhu, H., Zhou, Y., Chen, J., Liu, J. & Lascoux, M. (2010). Demographic histories of four spruce (Picea) species of the Qinghai-Tibetan Plateau and neighboring areas inferred from multiple nuclear loci. *Molec. Biol. Evol.* **27**, 1001–14.

Lotka, A. J. (1925). *Elements of Physical Biology*. Baltimore: Williams and Wilkins Co.

Marin, J.-M., Pudlo, P., Robert, C. P. & Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statist. Comp.* **22**, 1167–80.

Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Nat. Acad. Sci.* **100**, 15324–8.

Mengersen, K. L. & Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101–21.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–92.

Meyn, S. P. & Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge: Cambridge University Press.

Mira, A. (2001). Ordering and improving the performance of Monte Carlo Markov chains. *Statist. Sci.* **16**, 340–50.

Montenegro, R. & Tetali, P. (2006). Mathematical aspects of mixing times in Markov chains. *Foundat. Trends Theor. Comp. Sci.* **1**, 237–354.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60**, 607–12.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molec. Biol. Evol.* **16**, 1791–8.

Roberts, G. O. & Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Prob.* **2**, 13–25.

Roberts, G. O. & Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Prob. Surveys* **1**, 20–71.

Roberts, G. O. & Rosenthal, J. S. (2008). Variance bounding Markov chains. *Ann. Appl. Prob.* **18**, 1201–14.

Roberts, G. O. & Rosenthal, J. S. (2011). Quantitative non-geometric convergence bounds for independence samplers. *Methodol. Comp. Appl. Prob.* **13**, 391–403.

Roberts, G. O. & Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83**, 95–110.

Sisson, S. A. & Fan, Y. (2011). Likelihood-free Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones & X.-L. Meng, eds. Boca Raton, Florida: Chapman & Hall/CRC, pp. 313–33.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–18.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Ann. Appl. Prob.* **8**, 1–9.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**, 187–202.

Volterra, V. (1926). Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. R. Acad. Naz. dei Lincei* **2**, 31–113.

Whiteley, N. (2012). Sequential Monte Carlo samplers: Error bounds and insensitivity to initial conditions. *Stoch. Anal. Appl.* **30**, 774–98.

Wilkinson, D. J. (2006). *Stochastic Modelling for Systems Biology*. Mathematical and Computational Biology Series. Boca Raton: Chapman & Hall/CRC.

[*Received October* 2012. *Revised April* 2014]