Theoretical and methodological aspects of MCMC computations with noisy likelihoods

Christophe Andrieu, Anthony Lee and Matti Vihola

Approximate Bayesian computation (ABC) [11, 42] is a popular method for Bayesian inference involving an intractable, or expensive to evaluate, likelihood function but where simulation from the model is easy. The method consists of defining an alternative likelihood function which is also in general intractable but naturally lends itself to pseudo-marginal computations [5], hence making the approach of practical interest. The aim of this chapter is to show the connections of ABC Markov chain Monte Carlo with pseudo-marginal algorithms, review their existing theoretical results, and discuss how these can inform practice and hopefully lead to fruitful methodological developments.

0.1 The noisy likelihood perspective

Consider some data $y_{\mathrm{obs}} \in \mathcal{Y}$ assumed to arise from a family of probability distributions with densities $\{\ell(\cdot \mid \theta), \theta \in \Theta\}$, with respect to some appropriate reference measure $\lambda(\cdot)$, indexed by some unknown parameter $\theta \in \Theta \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$. In a Bayesian context θ is ascribed a prior distribution with density $\eta(\cdot)$ (with respect to some appropriate measure) and the posterior distribution has density

$$\pi(\theta \mid y_{\mathrm{obs}}) \propto \eta(\theta)\ell(y_{\mathrm{obs}} \mid \theta).$$

The intractability of the likelihood function may prevent the implementation of traditional sampling algorithms. To circumvent this problem it is natural to seek to approximate the desired likelihood function $\ell(y_{\mathrm{obs}} \mid \theta)$ and ABC methods do so by taking advantage of the fact that sampling from the family of distributions $\{\ell(\cdot \mid \theta)\lambda(\cdot), \theta \in \Theta\}$ may be simple. Standard practice consists first of defining a function of the data $s : \mathcal{Y} \to \mathbb{R}^q$ for some $q \in \mathbb{N}_+$, and thereby the "summary statistics" used to compare datasets. Then a distance $\|\cdot\|$ on $\mathbb{R}^q$ and a "kernel" $K : \mathbb{R}_+ \to [0,1]$ are chosen and combined to form $\psi(y^1, y^2) := K\big(\|s(y^1) - s(y^2)\|\big)$ for $y^1, y^2 \in \mathcal{Y}$ whose rôle is to measure the strength of the dissimilarity between datasets. One is naturally not constrained to this specific form of ψ and the only requirement is that $\psi : \mathcal{Y}^2 \to [0,1]$ and is statistically sensible. Note that there is no loss of generality in choosing the upper bound 1 for ψ since multiplicative constants do not affect Bayes' rule. A standard choice for ψ is

$$\psi(y^1, y^2) := \mathbb{I}\{\|s(y^1) - s(y^2)\| \leq \epsilon\}, \tag{0.1}$$

for some $\epsilon > 0$ although most of this chapter will deal with the general case, rather than this specific choice. Now, given $\psi : \mathcal{Y}^2 \to [0,1]$ one can the define the "ABC likelihood" function

$$\ell^{\psi}_{\mathrm{ABC}}(y_{\mathrm{obs}} \mid \theta) := \int \psi(y, y_{\mathrm{obs}})\ell(y \mid \theta)\lambda(\mathrm{d}y). \tag{0.2}$$

One can think of ABC likelihoods arising from the standard choice of ψ as being kernel density estimators of the probability density of the observed summary statistics under the assumed model for the data. The associated posterior distribution has density

$$\pi(\theta) := \pi_{\mathrm{ABC}}(\theta \mid y_{\mathrm{obs}}) \propto \eta(\theta)\ell^{\psi}_{\mathrm{ABC}}(y_{\mathrm{obs}} \mid \theta),$$

and we will use the simplified notation $\pi(\theta)$ in the remainder of the chapter. It seems at first sight that we have not made any progress since the new likelihood function is now an integral with respect to a distribution whose density is assumed to be intractable. This prevents direct implementation of the workhorse of Markov chain Monte Carlo (MCMC) methodology, the Metropolis–Hastings (MH) algorithm, described in Alg. 1 for a family of proposal distributions $\{q(\theta, \cdot), \theta \in \Theta\}$. For notational simplicity, we adopt the convention that for a random variable $X \sim \varpi(\cdot)$ where $\varpi(\cdot)$ is a probability distribution, $x \sim \varpi(\cdot)$ means that x is a realization of X, and do not use capital fonts for Greek letters representing random variables.

1 Given θ
2 Sample $\vartheta \sim q(\theta, \cdot)$
3 Return ϑ with probability

$$\min\left\{1, \frac{\eta(\vartheta)\ell^{\psi}_{\mathrm{ABC}}(y_{\mathrm{obs}} \mid \vartheta)q(\vartheta, \theta)}{\eta(\theta)\ell^{\psi}_{\mathrm{ABC}}(y_{\mathrm{obs}} \mid \theta)q(\theta, \vartheta)}\right\}$$

4 Otherwise return θ.

Algorithm 1: Exact ABC-MCMC update.

This is where the possibility to sample from $\ell(\cdot \mid \theta)\lambda(\cdot)$ comes into play, in combination with the standard "auxiliary variable trick". Define the probability distribution on $\Theta \times \mathcal{Y}$ with the following density

$$\pi(\theta, y) \propto \eta(\theta)\ell(y \mid \theta)\psi(y, y_{\mathrm{obs}}). \tag{0.3}$$

Evidently this distribution has $\pi(\mathrm{d}\theta)$ as a marginal and we aim to sample from this joint distribution. The rejection ABC algorithm proceeds, in its simplest form, by sampling $\theta \sim \eta$ and $y \sim \ell(\cdot \mid \theta)\lambda(\cdot)$, and accepting (θ, y) with probability $\psi(y, y_{\mathrm{obs}})$, therefore not requiring the evaluation of the likelihood function. This idea can also be used in the context of the MH algorithm by choosing a family of probability distributions $\{q(\theta, \cdot), \theta \in \Theta\}$ to update the parameter component and $\{\ell(\cdot \mid \theta), \theta \in \Theta\}$ to update the auxiliary dataset component. The resulting update is described in Alg. 2 (we add an index to the auxiliary datasets to indicate the distribution they are sampled from) and first appeared in [30].

We note that in some latent variable models, the ABC posterior π is not approximate. In particular, if $y_{\mathrm{obs}} \sim \psi(y, \cdot)$ and $y \sim \ell(\cdot \mid \theta)\lambda(\cdot)$ for some $\theta \in \Theta$ with s the identity function, then $\ell^{\psi}_{\mathrm{ABC}}(y_{\mathrm{obs}} \mid \theta)$ is the exact, albeit intractable, likelihood function. This has been stressed in [44] and is taken advantage of in, e.g., [21] and [15].

1 Given (θ, y_θ)
2 Sample $\vartheta \sim q(\theta, \cdot)$ and $y_\vartheta \sim \ell(\cdot \mid \vartheta)\lambda(\cdot)$
3 Return (ϑ, y_ϑ) with probability

$$\min\left\{1, \frac{\eta(\vartheta) \times \psi(y_\vartheta, y_{\text{obs}})q(\vartheta, \theta)}{\eta(\theta) \times \psi(y_\theta, y_{\text{obs}})q(\theta, \vartheta)}\right\} \tag{0.4}$$

4 otherwise return (θ, y_θ).

Algorithm 2: ABC-MCMC

0.2 Pseudo-marginal algorithms

We now develop another perspective on Alg. 2, which turns out to be fruitful in many respects and on which the remainder of the chapter is based. As we shall see this alternative point of view suggests many useful extensions, is both conceptually and notationally much simpler, and in fact covers scenarios of interest beyond ABC. Note however that despite the attractive generic nature of this perspective, one should in practice not forget about the initial problem at hand since it may possess additional specific structure one may exploit. The main starting point here is to notice that with $Y \sim \ell(\cdot \mid \theta)\lambda(\cdot)$ then $\psi(Y, y_{\text{obs}})$ is an unbiased estimator of (0.2) and that one can write the joint posterior density (0.3) as follows, in terms of the marginal $\pi(\theta)$,

$$\pi(\theta, y) \propto \eta(\theta)\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta)\frac{\psi(y, y_{\text{obs}})}{\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta)}\ell(y \mid \theta) \propto \pi(\theta)\frac{\psi(y, y_{\text{obs}})}{\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta)}\ell(y \mid \theta),$$

where we have assumed θ such that $\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta) \neq 0$. From this we conclude that $\pi(\theta)\psi(Y, y_{\text{obs}})/\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta)$ is an unbiased (and non-negative) estimator of $\pi(\theta)$ when $Y \sim \ell(\cdot \mid \theta)\lambda(\cdot)$. Clearly the acceptance probability (0.4) can be equally written as

$$\min\left\{1, \left[\pi(\vartheta)\frac{\psi(y_\vartheta, y_{\text{obs}})}{\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \vartheta)}q(\vartheta, \theta)\right] \Big/ \left[\pi(\theta)\frac{\psi(y_\theta, y_{\text{obs}})}{\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta)}q(\theta, \vartheta)\right]\right\},$$

that is Alg. 2 can be thought of as an approximate implementation of the exact update Alg. 1 where the expression for $\pi(\cdot)$ is replaced with an unbiased, "noisy" estimator. However, Alg. 2 targets the joint distribution with density $\pi(\theta, y)$, and is therefore exact in the sense that an ergodic Markov chain built on this type of update can produce samples of distribution arbitrarily close to $\pi(\mathrm{d}\theta)$. This remark leads in fact to a far more widely applicable idea [10, 5] and the resulting methods are referred to as pseudo-marginal algorithms (see [22, 28] for earlier, related, but different ideas). Indeed, assume that for any $\theta \in \Theta$ we can generate "unbiased measurements" of $\pi(\theta)$ of the form $\pi(\theta) \times W$ where $W \sim Q_\theta$, $Q_\theta\big(W \geq 0\big) = 1$ and such that $\mathbb{E}_{Q_\theta}[W] = C$ for some $C > 0$ independent of θ, and consider the probability distribution on $\Theta \times \mathsf{W}$ with density

$$\tilde{\pi}(\theta, w) = \pi(\theta) \times w \times Q_\theta(w). \tag{0.5}$$

For simplicity, we will hereafter assume that $C = 1$. Note that we do not assume here that $\pi(\theta)$ is tractable, but rather that $\pi(\theta) \times w$ is: W is purely conceptual and implicit in real scenarios. In the ABC scenario described above W is the positive real valued random variable such that for any $\theta \in \Theta$ and $A \in \mathcal{B}\big(\mathbb{R}\big)$, the Borel σ-algebra on $\mathbb{R}$,

$$Q_\theta(W_\theta \in A) = \int \mathbb{I}\{\psi(y, y_{\text{obs}})/\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta) \in A\}\ell(y \mid \theta)\lambda(\mathrm{d}y). \tag{0.6}$$

Now a MH update, with transition probability denoted $\tilde{P}$, targetting this distribution and with proposal distribution $q(\theta, \vartheta) \times Q_\vartheta(u)$ has acceptance probability

$$\tilde{\alpha}(\theta, w; \vartheta, u) = \min\left\{1, \frac{\pi(\vartheta) \times u \times Q_\vartheta(u)}{\pi(\theta) \times w \times Q_\theta(w)} \frac{q(\vartheta, \theta) Q_\theta(w)}{q(\theta, \vartheta) Q_\vartheta(u)}\right\} = \min\left\{1, \frac{\pi(\vartheta) \times u \; q(\vartheta, \theta)}{\pi(\theta) \times w \; q(\theta, \vartheta)}\right\}.$$

The Markov transition kernel, which we denote by $\tilde{P}$, is described algorithmically in Alg. 3

1 Given (θ, w)
2 Sample $\vartheta \sim q(\theta, \cdot)$ and $u \sim Q_\vartheta$
3 Return (ϑ, u) with probability

$$\min\left\{1, \frac{\pi(\vartheta) \times u \; q(\vartheta, \theta)}{\pi(\theta) \times w \; q(\theta, \vartheta)}\right\}$$

4 Otherwise return (θ, w)

Algorithm 3: Generic pseudo-marginal algorithm

Clearly a Markov chain Monte Carlo based on this update is exact in the sense outlined earlier and can be thought of as being an approximation of an exact MH update, which we denote by P, with acceptance probability $\min\{1, r(\theta, \vartheta)\}$ where

$$r(\theta, \vartheta) := \frac{\pi(\vartheta) \; q(\vartheta, \theta)}{\pi(\theta) \; q(\theta, \vartheta)},$$

which would use the exact values of the density $\pi(\theta)$.

What is the interest of these developments? First, $\tilde{\pi}(\theta, w)$ and Alg. 3 are notationally and conceptually simple, equivalent representations of $\pi(\theta, y)$ and Alg. 2, respectively, in that they lead to equivalent algorithms, provided we are only interested in the properties of the chain $\{\theta_i, i \geq 0\}$. Indeed, let $\{(\check{\theta}_i, Y_i), i \geq 0\}$ and $\{(\theta_i, W_i), i \geq 0\}$ be the Markov chains such that with μ (resp. ν_θ for any $\theta \in \Theta$) a probability distribution on Θ (resp. on $\mathcal{Y}$) and for any $\theta \in \Theta$ and $A \in \mathcal{B}(\mathbb{R})$

$$\tilde{\nu}_\theta(\tilde{W}_\theta \in A) := \int \mathbb{I}\{\psi(y, y_{\text{obs}})/\ell_{\text{ABC}}(y_{\text{obs}} \mid \theta) \in A\} \nu_\theta(\mathrm{d}y),$$

$(\check{\theta}_0, Y_0) \sim \mu \times \nu_\cdot$, $(\theta_0, W_0) \sim \mu \times \tilde{\nu}_\cdot$ and Markov transition probabilities as described in Alg. 2 and Alg. 3 respectively. With $\check{\mathbb{P}}(\cdot)$ and $\mathbb{P}(\cdot)$ the respective probabilities it can be checked easily that for any $m \in \mathbb{N}$, $i_1, i_2, \ldots, i_m \in \mathbb{N}$ and $A_1, A_2, \ldots, A_m \in \mathcal{B}(\Theta)^m$ we have

$$\check{\mathbb{P}}\left(\check{\theta}_{i_1} \in A_1, \ldots, \check{\theta}_{i_m} \in A_m\right) = \mathbb{P}\left(\theta_{i_1} \in A_1, \ldots, \theta_{i_m} \in A_m\right),$$

which indeed implies that the processes $\{\check{\theta}_i, i \geq 0\}$ and $\{\theta_i, i \geq 0\}$ are probabilistically indistinguishable. In particular, at a conceptual level this tells us on the one hand that the properties of the algorithm (i.e. $\{\check{\theta}_i, i \geq 0\}$) are fully characterized by the properties of the random variables induced by (0.6), but also that the algorithm can be thought of as random perturbation of the exact algorithm with exact acceptance probability $\min\{1, r(\theta, \vartheta)\}$, suggesting links between the exact algorithm and its perturbations.

A second feature of this representation is that it emphasises the fact that the central property exploited in ABC is the possibility to produce unbiased and non-negative estimators of (0.2) cheaply, and such estimators are not restricted to the standard choice

$\psi(Y, y_{\text{obs}})$ with $Y \sim \ell(\cdot \mid \theta)\lambda(\cdot)$. For example, one could average N multiple copies of the estimator as proposed in [12] and utilized in, e.g., [16]. That is, for any $\theta \in \Theta$ and $Y^1, Y^2, \ldots, Y^N \overset{\text{iid}}{\sim} \ell(\cdot \mid \theta)\lambda(\mathrm{d}\cdot)$ consider the unbiased estimator $T_\theta(N)$ of $\ell_{\text{ABC}}(y_{\text{obs}} \mid \theta)$

$$T_\theta(N) := \frac{1}{N}\sum_{i=1}^{N} \psi\big(Y^i, y_{\text{obs}}\big). \tag{0.7}$$

This leads to the unit expectation, non-negative random variable $W_\theta(N) = T_\theta(N)/\ell^{\psi}_{\text{ABC}}(y_{\text{obs}} \mid \theta)$, and in the light of the above there is no need to check the validity of a noisy MH algorithm that uses this estimator. One can check that the algorithm now targets

$$\pi(\theta) Q_\theta^N(w^1, w^2, \ldots, w^N) \frac{1}{N}\sum_{i=1}^{N} w^i,$$

but that it is also possible to aggregate, i.e. define $w(N) := N^{-1}\sum_{i=1}^{N} w^i$ such that the associated random variable has distribution $\mathcal{Q}_\theta^N\big(W_\theta(N) \in A\big) = \int \mathbb{I}\big\{w(N) \in A\big\} Q_\theta^N(\mathrm{d}(w^1, w^2, \ldots, w^N))$. We will return to such aggregation strategies in Section 0.4.1 and discuss alternatives to a product form of the joint distribution of density $Q_\theta^N(w^1, w^2, \ldots, w^N)$, since it is in fact sufficient that its N marginals all be $Q_\theta(\cdot)$ for the above to hold.

0.3 Performance measures

Before presenting various possible strategies to improve on standard ABC-MCMC algorithms in the next section we recall here standard performance measures for MCMC algorithms and a summary of some known theoretical results relating (essentially) the properties of $\big\{W_\theta, \theta \in \Theta\big\}$ to the performance of pseudo-marginal algorithms. As we shall see the variability and extreme behaviour of the estimators W_θ play a fundamental part in the (bad) behaviour of ABC-MCMC algorithms. The intuition goes as follows, upon recalling the expression for the acceptance probability of the noisy algorithm (Alg. 3) in terms of the acceptance ratio of the exact algorithm,

$$\min\left\{1, r(\theta, \vartheta)\frac{u}{w}\right\},$$

and that W_θ is a non-negative random variable of expectation one. Realizations of W_θ can take values larger than one and make leaving the state (θ, w) more difficult than for the exact algorithm (since u has to match w), resulting in the familiar "sticky behaviour" of the ABC-MCMC algorithm (see, e.g., [41]).

For μ a probability distribution defined on some measurable space $\big(\mathsf{E}, \mathcal{E}\big)$ and $\Pi : \mathsf{E} \times \mathcal{E} \to [0, 1]$ a Markov transition kernel with invariant distribution μ, i.e. $\mu\Pi = \mu$, we are interested in this section in two performance measures that address asymptotic variance and bias:

1. Letting $\Phi_0 \sim \mu$ and $\Phi_n \sim \Pi(\Phi_{n-1}, \cdot)$ for $n \geq 1$ one may be interested, for a function $f : \mathsf{E} \to \mathbb{R}$ in the behaviour of ergodic averages $S_T(\Phi) := T^{-1}\sum_{k=0}^{T-1} f(\Phi_k)$, their asymptotic variance is a natural performance measure and we will focus on the quantity
$$\text{var}\,(f, \Pi) := \lim_{T\to\infty} T\text{var}_\mu\big(S_T(\Phi)\big),$$

2. Letting for $x \in \mathsf{E}$, $\mathcal{L}_x(\Phi_n)$ be the law of the n-th state Φ_n of the Markov chain $\{\Phi_i, i \geq 0\}$ where $\Phi_0 = x$, we may be interested in the rate of convergence to equilibrium of the Markov chain. That is, for an appropriate norm $\|\cdot\|_*$ characterize the distance between $\mathcal{L}_x(\Phi_n)$ and μ for all $n \geq 0$ and, for example, establish the existence of $M > 0$, $V : \mathsf{E} \to \mathbb{R}_+$ and $\rho \in [0, 1]$ or $\alpha > 0$ or $\{r(n), n \geq 0\}$ such that either of the following inequalities hold for $n \geq 1$

$$\|\mathcal{L}_x(\Phi_n) - \mu\|_* \leq \begin{cases} M\rho^n & \text{(uniformly ergodic)} \\ MV(x)\rho^n & \text{(geometrically ergodic)} \\ MV(x)n^{-\alpha} & \text{(polynomially ergodic)} \\ V(x)r^{-1}(n), \quad r(n) \to \infty & \text{(ergodic).} \end{cases}$$

The role of V is to take into account the influence of the initialisation on convergence and we leave the norms unspecified, although it may be useful to know that they correspond to the supremum of $\left|\mathbb{E}\big(f(\Phi_n)\big) - \mu\big(f\big)\right|$ for f in certain classes of function. Such results will therefore provide us with a sense of the speed at which bias vanishes as n increases.

In the sequel, P is the noiseless algorithm, $\tilde{P}$ is the noisy algorithm using the family $\{Q_\theta, \theta \in \Theta\}$ and $\tilde{P}_N$ is the noisy algorithm which averages N samples marginally identically distributed according to $\{Q_\theta, \theta \in \Theta\}$ and joint distributions denoted $\{Q_\theta^N, \theta \in \Theta\}$. We will also consider more general families of noisy algorithms $\{\tilde{P}_\lambda, \lambda \in \mathbb{R}_+\}$ indexed by $\lambda > 0$, i.e. using weight distributions $\{Q_\theta^{(\lambda)}, \theta \in \Theta, \lambda \in \Lambda \subset \mathbb{R}_+\}$, with the convention that $\tilde{P}_{\lambda=0} = P$ (that is the latter is not defined on the same space as $\tilde{P}_\lambda$, $\lambda \neq 0$).

We start with simple results which confirm that $\tilde{P}$ is a suboptimal approximation of P but that concentration of W_θ on 1 allows $\tilde{P}$ to approach some performance measures of P arbitrarily closely.

0.3.1 Approximation of the noiseless algorithm

The pseudo-marginal algorithm, Alg. 3, never has a smaller asymptotic variance than its corresponding marginal algorithm [7, Theorem 7], therefore justifying attempts to approximate P in order to improve performance of $\tilde{P}$. A natural class of functions to consider are those with finite second moment under π, i.e. functions $\{f : \pi(f^2) < \infty\}$ where $\pi(f^2) := \int f(\theta)^2 \pi(\mathrm{d}\theta)$, or equivalently the functions f such that the random variable $f(X)$ has finite variance when $X \sim \pi$, i.e. $\mathrm{var}_\pi(f) < \infty$.

Theorem 1 (Noiseless is best)**.** *Assume $f : \Theta \to \mathbb{R}$ satisfies $\pi(f^2) < \infty$. The asymptotic variances of f with respect to the pseudo-marginal algorithm $\tilde{P}$ and the marginal algorithm P always satisfy*

$$\mathrm{var}(f, P) \leq \mathrm{var}(f, \tilde{P}).$$

Under general technical conditions, the asymptotic variance of the pseudo-marginal algorithm converges to the asymptotic variance of the marginal algorithm [7, Theorem 21]. Denote by $\{\tilde{\theta}_k^{(\lambda)}, k \geq 0\}$ the Markov chain with initial distribution $\tilde{\pi}_\lambda$ and kernel $\tilde{P}_\lambda$.

Theorem 2. *Assume that $\int |f(\theta)|^{2+\delta} \pi(\theta)\mathrm{d}\theta < \infty$ for some $\delta > 0$, that $\mathrm{var}(f, P) < \infty$ and there exists a constant $\lambda_0 \in \Lambda$ such that*

$$\lim_{n \to \infty} \sup_{0 \leq \lambda \leq \lambda_0} \left| \sum_{k=n}^{\infty} \mathbb{E}[\bar{f}(\tilde{\theta}_0^{(\lambda)}) \bar{f}(\tilde{\theta}_k^{(\lambda)})] \right| = 0 \quad \textit{where} \quad \bar{f} = f - \pi(f),$$

and that

$$\lim_{\lambda\to 0}\int |1-w|Q_\theta^{(\lambda)}(\mathrm{d}w) = 0 \qquad \textit{for all } \theta\in\Theta.$$

Then,

$$\lim_{\lambda\to 0}\mathrm{var}(f,\tilde{P}_\lambda) = \mathrm{var}(f,P).$$

The first condition simply says that the tails of the integrated autocovariances vanish uniformly for sufficiently good approximations of P, which should be the case for example if the approximation does not perturb the ergodicity properties of P significantly. This is further investigated in [7] and related to the results of Section 0.3.2. The second condition is very natural and formalizes the idea of concentration of $\tilde{\pi}_\lambda\big(\mathrm{d}(\theta,w)\big)$ on $\pi(\mathrm{d}\theta)\delta_1(\mathrm{d}w)$. The following result formalizes the fact that approximations involving unbounded noises are undesirable [5, Theorem 8].

Theorem 3. *If the weight distributions are such that*

$$\pi\big(\{\theta\in\Theta : \textstyle\int_M^\infty Q_\theta(\mathrm{d}w) > 0 \textit{ for all } M<\infty\}\big) > 0,$$

then the pseudo-marginal algorithm cannot be geometrically ergodic.

Corollary 1. *Even when P is geometrically ergodic, as soon as*

$$\left\{\theta\in\Theta : \int_M^\infty Q_\theta(\mathrm{d}w) > 0 \textit{ for all } M<\infty\right\}$$

has a positive π-probability, then for any $N\in\mathbb{N}_+$,

$$\left\{\theta\in\Theta : \int_M^\infty \mathcal{Q}_\theta^N(\mathrm{d}w) > 0 \textit{ for all } M<\infty\right\}$$

has a positive π-mass, and $\tilde{P}_N$ cannot be geometrically ergodic for any $N\in\mathbb{N}_+$.

Broadly speaking the result simply says that boundedness of the weights is required to ensure geometric ergodicity and the corollary that while averaging may ensure convergence of the integrated autocovariance, it will not always be the case that one can approach the rate of convergence of P: in fact the result says that one cannot even be geometric. In other words, despite the fact that "bad events" (e.g. the N weights we have drawn are all large simultaneously) have a vanishing probability of occurrence as N increases, their impact on the long term properties of the algorithm may still be felt. Such bad behaviour will however vanish, for example, for a fixed simulation length T as N increases, provided naturally that the algorithm is not initialized at points corresponding to such bad events.

0.3.2 Rates of convergence

The first result of this section holds under a condition (0.8) stronger than uniform ergodicity for P, but often used in practice to establish this property whenever the space Θ is compact, and the assumption that the noise is uniformly bounded (the condition in [5, Theorem 8] is slightly more general). In words the result says that uniform ergodicity of the noiseless algorithm is inherited by the noisy algorithm in this scenario.

Theorem 4. *Suppose there exist $\epsilon>0$, a probability measure ν on the measurable space $\big(\Theta,\mathcal{T}\big)$ such that for any $A\in\mathcal{T}$,*

$$\int_A q(\theta,\mathrm{d}\vartheta)\min\{1,r(\theta,\vartheta)\} \geq \epsilon\nu(A) \qquad \textit{for all } \theta\in\Theta, \tag{0.8}$$

and $M<\infty$ such that for all $\theta\in\Theta$, $Q_\theta\big(W_\theta\leq M\big)=1$, then $\tilde{P}$ is also uniformly ergodic.

It should be noted that even in this favourable scenario, $\tilde{P}_N$ may not achieve the rate of convergence of P for any $N \geq 1$ [5, Remark 1]. The interest of the next result is that it establishes in a simple, yet representative, scenario a direct link between the existence of general moments of W_θ and the rate of convergence to equilibrium one may expect from the algorithm.

Theorem 5. *Suppose there exist $\epsilon > 0$, a probability measure ν on the measurable space $(\Theta, \mathcal{T})$ such that for any $A \in \mathcal{T}$,*

$$\int_A q(\theta, \mathrm{d}\vartheta) \min\{1, r(\theta, \vartheta)\} \geq \epsilon \nu(A) \qquad \text{for all } \theta \in \Theta,$$

and a non-decreasing convex function $\phi : [0, \infty) \to [1, \infty)$ satisfying

$$\liminf_{t \to \infty} \frac{\phi(t)}{t} = \infty \qquad \text{and} \qquad M_W := \sup_{\theta \in \Theta} \int \phi(w) Q_\theta(\mathrm{d}w) < \infty. \tag{0.9}$$

Then $\tilde{P}$ is sub-geometrically ergodic with a rate of convergence characterized by ϕ.

Corollary 2. *Consider for example the case $\phi(w) = w^\beta + 1$ for some $\beta > 1$. Then there exists $C > 0$ such that for any function $f : \Theta \to \mathbb{R}$ and $n \in \mathbb{N}_+$,*

$$\left|\mathbb{E}\big(f(\theta_n)\big) - \pi\big(f\big)\right| \leq C \left|f\right|_\infty n^{-(\beta-1)},$$

where $\left|f\right|_\infty := \sup_{\theta \in \Theta} |f(\theta)|$. For readers familiar with the total variation distance, this implies that for any $\theta \in \Theta$,

$$\|\mathcal{L}_\theta(\theta_n) - \pi(\cdot)\|_{TV} \leq C n^{-(\beta-1)}.$$

Other rates of convergence may be obtained for other functions ϕ [19].

These last results will typically hold only in situations where Θ is bounded: extensions to more general scenarios have been considered in [7], but require one to be more specific about the type of MH updates considered and involve substantial additional technicalities. A rough summary of known results is presented in Table 0.1, we refer the reader to [7] for precise statements.

Marginal P	W_θ	Pseudo-marginal $\tilde{P}$
uniform	$W_\theta \leq c$ a.s.	uniform
geometric	$W_\theta \leq c$ a.s.	geometric if $\tilde{P}$ positive (conjecture in general [7, Section 3])
any	W_θ unbounded	not geometric
uniform	$\mathbb{E}_{Q_\theta}\left[W_\theta^{1+\epsilon}\right] \leq c$	polynomial
uniform	uniform integrability (0.9)	sub-geometric
IMH	–	IMH
geometric RWM	$\mathbb{E}_{Q_\theta}\left[W_\theta^{1+\epsilon} + W_\theta^{-\delta}\right] \leq c(\theta)$	polynomial [7, Theorem 38]

Table 0.1
Convergence inheritance. The constants are such that $\epsilon, \delta > 0$ and $c \in \mathbb{R}$, while $c(\cdot) : \Theta \to \mathbb{R}_+$ on the last line should satisfy some growth conditions [7, Theorem 38]. IMH and RWM stand for Independent MH and random walk Metropolis, respectively.

One clear distinction in Table 0.1 is the inheritance of geometric ergodicity of P by (at least) positive $\tilde{P}$ when W_θ is almost surely uniformly bounded and its failure to do so when W_θ is unbounded for all θ in some set of positive π-probability. In the case where W_θ is almost surely bounded but not uniformly so, characterization is not straightforward. Indeed, there are cases where $\tilde{P}$ does inherit geometric ergodicity and cases where it does not, see [7, Remark 14] and [27, Remark 2]. In [27] and its supplement, it is shown that failure to inherit geometric ergodicity in statistical applications is not uncommon when using "local" proposals such as a random walk, see Theorem 9 below.

One attractive property of uniformly and geometrically ergodic $\tilde{P}$ is that $\mathrm{var}(f, \tilde{P}) < \infty$ for all f with $\mathrm{var}_\pi(f) < \infty$. Conversely, when $\tilde{P}$ is not geometrically ergodic, and the chain is not almost periodic in a particular technical sense, then there do exist f with $\mathrm{var}_\pi(f) < \infty$ such that $\mathrm{var}(f, \tilde{P})$ is not finite (see [37] for more details). It is, however, not straightforward to identify which functions have finite asymptotic variance in the sub-geometric regime. It has recently been shown that uniformly bounded second moments of W_θ and geometric ergodicity of P is sufficient to ensure that all functions f of θ only with $\mathrm{var}_\pi(f) < \infty$ have finite asymptotic variance [18].

Theorem 6. *Let P be geometrically ergodic and $\sup_\theta \mathbb{E}_{Q_\theta}\left[W_\theta^2\right] < \infty$. Then for any $f : \Theta \to \mathbb{R}$ with $\mathrm{var}_\pi(f) < \infty$, $\mathrm{var}(f, \tilde{P}) < \infty$.*

We note that this result holds even in the case where W_θ is unbounded, in which case $\tilde{P}$ is not geometrically ergodic: the functions f with $\mathrm{var}_\pi(f) < \infty$ that have infinite asymptotic variance in this case are not functions of θ alone and must depend on w.

0.3.3 Comparison of algorithms

The results of the previous section are mostly concerned with comparisons of the noisy algorithm with its noiseless version. We consider now comparing different variations of the noisy algorithm. Intuitively one would, at comparable cost, prefer to use an algorithm which uses the estimators with the lowest variability. It turns out that the relevant notion of variability is the convex order [33].

Definition 1. The random variables $W_1 \sim F_1$ and $W_2 \sim F_2$ are *convex ordered*, denoted $W_1 \leq_{cx} W_2$ or $F_1 \leq_{cx} F_2$ hereafter, if for any convex function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\phi(W_1)] \leq \mathbb{E}[\phi(W_2)],$$

whenever the expectations are well-defined.

We note that the convex order $W^{(1)} \leq_{cx} W^{(2)}$ of square-integrable random variables automatically implies $\mathrm{var}(W^{(1)}) \leq \mathrm{var}(W^{(2)})$, but that the reverse is not true in general. As shown in [8, Example 13] while the convex order allows one to order performance of competing algorithms, the variance is not an appropriate measure of dispersion.

Let $\bar{P}_1$ and $\bar{P}_2$ be the corresponding competing pseudo-marginal implementations of the MH algorithm targeting $\pi(\cdot)$ marginally sharing the same family of proposal distributions $\{q(\theta, \cdot), \theta \in \Theta\}$ but using two families of weight distributions $\{Q_\theta^{(1)}, \theta \in \Theta\}$ and $\{Q_\theta^{(2)}, \theta \in \Theta\}$. Hereafter the property $Q_\theta^{(1)} \leq_{cx} Q_\theta^{(2)}$ for all $\theta \in \Theta$ is denoted $\{Q_\theta^{(1)}, \theta \in \Theta\} \leq_{cx} \{Q_\theta^{(2)}, \theta \in \Theta\}$. We introduce the notion of the right spectral gap, $\mathrm{Gap}_R(\Pi)$, of a $\mu-$reversible Markov chain evolving on E, noting that the MH update is reversible with respect to its invariant distribution. This can be intuitively understood as follows in the situation where E is finite, in which case the transition matrix Π can be diagonalized (in a certain sense) and its eigenvalues shown to be contained in $[-1, 1]$. In this scenario, $\mathrm{Gap}_R(\Pi) = 1 - \lambda_2$ where

λ_2 is the second largest eigenvalue of Π. These ideas can be generalized to more general spaces. The practical interest of the right spectral gap is that it is required to be positive for geometric ergodicity to hold, and provides information about the geometric rate of convergence when, for example, all the eigenvalues (in the finite scenario) are non-negative.

Theorem 7. *Let $\pi(\cdot)$ be a probability distribution on some measurable space $(\Theta, \mathcal{T})$ and let $\bar{P}_1$ and $\bar{P}_2$ be two pseudo-marginal approximations of P aiming to sample from $\pi(\cdot)$, sharing a common family of marginal proposal probability distributions $\{q(\theta, \cdot), \theta \in \Theta\}$ but with distinct weight distributions satisfying $\{Q_\theta^{(1)}, \theta \in \Theta\} \leq_{cx} \{Q_\theta^{(2)}, \theta \in \Theta\}$. Then,*

1. for any $\theta, \vartheta \in \Theta$, the conditional acceptance rates satisfy $\alpha_{\theta\vartheta}(\bar{P}_1) \geq \alpha_{\theta\vartheta}(\bar{P}_2)$,

2. for any $f : \Theta \to \mathbb{R}$ with $\mathrm{var}_\pi(f) < \infty$, the asymptotic variances satisfy $\mathrm{var}(f, \bar{P}_1) \leq \mathrm{var}(f, \bar{P}_2)$,

3. the spectral gaps satisfy $\mathrm{Gap}_R(\bar{P}_1) \geq \min\{\mathrm{Gap}_R(\bar{P}_2), 1 - \tilde{\rho}_2^*\}$, where $\tilde{\rho}_2^* := \tilde{\pi}_2\text{-ess}\sup_{(\theta,w)} \tilde{\rho}_2(\theta, w)$, the essential supremum of the rejection probability corresponding to $\bar{P}_2$,

4. if $\pi(\theta)$ is not concentrated on points, that is $\pi(\{\theta\}) = 0$ for all $\theta \in \Theta$, then $\mathrm{Gap}_R(\bar{P}_1) \geq \mathrm{Gap}_R(\bar{P}_2)$.

Various applications of this result are presented in [8], including the characterization of extremal bounds of performance measures for $\{Q_\theta, \theta \in \Theta\}$ belonging to classes of probability distributions (i.e. with given variance). As we shall see in the next section this result is also useful to establish that averaging estimators always improves the performance of algorithms, that introducing dependence between such copies may be useful or that stratification may be provably helpful in some situations.

0.4 Strategies to improve performance

0.4.1 Averaging estimators

Both intuition and theoretical results indicate that reducing variability of the estimates of $\pi(\cdot)$ in terms of the convex order ensures improved performance. We briefly discuss here some natural strategies. The first one consists of averaging estimators of the density, that is consider for any $\theta \in \Theta$ estimators of $\pi(\theta)$ of the form

$$\frac{1}{N}\sum_{i=1}^{N} \pi(\theta) W^i = \pi(\theta) \frac{1}{N}\sum_{i=1}^{N} W^i$$

for $(w^1, w^2, \ldots, w^N) \sim Q_\theta^N(\,\cdot\,)$ for a probability distribution $Q_\theta^N(\,\cdot\,)$ on $\mathbb{R}_+^N$ such that for $i = 1, \ldots, N$ [5]

$$\int w^i Q_\theta^N(w^1, w^2, \ldots, w^N) \mathrm{d}(w^1, w^2, \ldots, w^N) = 1,$$

and chosen in such a way that it reduces the variability of the estimator. A possibly useful application of this idea in an ABC framework could consist of using a stationary Markov chain with a Q_θ-invariant transition kernel to sample $W^1, W^2, \ldots$, such as a Gibbs sampler which may not require evaluation of the probability density of the observations.

Increasing N is, broadly speaking, always a good idea, at least provided one can perform computations in parallel at no extra cost.

Proposition 1 (see e.g. [33, Corollary 1.5.24])**.** *For exchangeable random variables* $(W^1, W^2, \dots)$*, for any* $N \geq 1$

$$\frac{1}{N+1}\sum_{i=1}^{N+1} W^i \leq_{cx} \frac{1}{N}\sum_{i=1}^{N} W^i.$$

Letting $\tilde{P}_N$ denote the noisy transition kernel using N exchangeable random variables $(W^1, W^2, \dots)$, this leads to the following result [8] by a straightforward application of Theorem 7.

Theorem 8. *Let* $(W^1, W^2, \dots)$ *be exchangeable, and* f *satisfy* $\pi(f^2) < \infty$*. Then for* $N \geq 2$*,* $N \mapsto \mathrm{var}(f, \tilde{P}_N)$ *is non-increasing.*

It can also be shown, with an additional technical condition, that $N \mapsto \mathrm{Gap}_R(\tilde{P}_N)$ is non-decreasing, suggesting improved convergence to equilibrium for positive algorithms.

A simple question arising from Theorem 8, is whether $\mathrm{var}(f, \tilde{P}_N)$ approaches $\mathrm{var}(f, P)$ as $N \to \infty$ for all f such that $\pi(f^2) < \infty$ under weaker conditions than in Theorem 2. In the ABC setting, however, it can be shown under fairly weak assumptions when Θ is not compact that this is not the case [27, Theorem 2].

Theorem 9. *Assume that* ψ *satisfies (0.1),* $\pi(\theta) > 0$ *for all* $\theta \in \Theta$*,* $\ell_{\mathrm{ABC}}(y_{\mathrm{obs}} \mid \theta) \to 0$ *as* $|\theta| \to \infty$ *and*

$$\lim_{r\to\infty} \sup_{\theta\in\Theta} q(\theta, B_r^{\complement}(\theta)) = 0,$$

where $B_r^{\complement}(\theta)$ *is the complement of the* $|\cdot|$ *ball of radius* r *around* θ*. Then* $\tilde{P}_N$ *cannot be geometrically ergodic for any* N *and there exist functions* f *with* $\pi(f^2) < \infty$ *such that* $\mathrm{var}(f, \tilde{P}_N)$ *is not finite.*

The assumptions do not preclude the possibility that P is geometrically ergodic and hence $\mathrm{var}(f, P)$ being finite for all f with $\pi(f^2) < \infty$, and so this result represents a failure to inherit geometric ergodicity in such cases. This result can be generalized to other choices of ψ under additional assumptions, see the supplement to [27].

The inability of $\tilde{P}_N$ more generally to escape the fate of $\tilde{P}_1$ is perhaps not surprising: from Table 0.1 we can see that apart from the case where P is a geometric RWM, the conditions on W_θ are unaffected by simple averaging. Further results in this direction are provided by quantitative bounds on asymptotic variances established by [13, Proposition 4] when ψ satisfies (0.1) and by [39] for general pseudo-marginal algorithms; more detailed results than below can be found in the latter.

Theorem 10. *Assume that* $\tilde{P}_N$ *is positive. Then for* $f : \Theta \to \mathbb{R}$ *with* $\mathrm{var}_\pi(f) < \infty$*,*

$$\mathrm{var}(f, \tilde{P}_1) \leq (2N-1)\mathrm{var}(f, \tilde{P}_N).$$

This shows that if the computational cost simulating the Markov chain with transition kernel $\tilde{P}_N$ is proportional to N then there is little to no gain in using $N \geq 1$. This is, however, not always the case: there may be some significant overhead associated with generating the first sample, or parallel implementation may make using some $N > 1$ beneficial [39]. It also implies that that $\mathrm{var}(f, \tilde{P}_N) < \infty$ implies $\mathrm{var}(f, \tilde{P}_1) < \infty$ so the class of π-finite variance functions with finite $\mathrm{var}(f, \tilde{P}_N)$ does not depend on N. The selection of parameters governing the concentration of W_θ around 1 in order to maximize computational efficiency has been considered more generally in [40] and [20], although the assumptions in these analyses are less specific to the ABC setting.

We now consider dependent random variables $W^1, \ldots, W^N$. It is natural to ask whether for a fixed $N \geq 1$ introducing dependence can be either beneficial or detrimental. We naturally expect that introducing some form of negative dependence between estimates could be helpful. For probability distributions $F_1, F_2, \ldots, F_N$ defined on some measurable space $\big(\mathsf{E}, \mathcal{E}\big)$ the associated Fréchet class $\mathcal{F}\big(F_1, F_2, \ldots, F_N\big)$ is the set of probability distributions on E^N with $F_1, F_2, \ldots, F_N$ as marginals. There are various ways one can compare the dependence structure of elements of $\mathcal{F}\big(F_1, F_2, \ldots, F_N\big)$ and one of them is the supermodular order, denoted

$$\big(W^1, W^2, \ldots, W^N\big) \leq_{sm} \Big(\tilde{W}^1, \tilde{W}^2, \ldots, \tilde{W}^N\Big) \tag{0.10}$$

hereafter (see [33] for a definition). In the case $N = 2$ this can be shown to be equivalent to

$$\mathbb{E}\big(f\big(W^1\big)g\big(W^2\big)\big) \leq \mathbb{E}\big(f\big(\tilde{W}^1\big)g\big(\tilde{W}^2\big)\big)$$

for any non-decreasing functions $f, g : \mathsf{E} \to \mathbb{R}$ for which the expectations exist. Interestingly (0.10) implies the following convex order

$$\sum_{i=1}^{N} W^i \leq_{cx} \sum_{i=1}^{N} \tilde{W}^i,$$

see, for example, the results in [38, Section 9.A]. An immediate application of Theorem 7 allows us then to order corresponding noisy algorithms in terms of the dependence structure of $\big(W^1, W^2, \ldots, W^N\big)$ and $\big(\tilde{W}^1, \tilde{W}^2, \ldots, \tilde{W}^N\big)$. We note however that it may be difficult in practical situations to check that the supermodular order holds between two sampling schemes.

0.4.2 Rejuvenation

We have seen that introducing multiple copies and averaging improves performance of the algorithm, at the expense of additional computation which may offset the benefits if parallel architectures cannot be used. In this section we show that the introduction of $N \geq 2$ copies also allows for the development of other algorithms which may address the sticky behaviour of some ABC algorithms. We observe that averaging also induces a discrete mixture structure of the distribution targetted,

$$\begin{aligned}\pi(\theta)Q_\theta^N(w^1, w^2, \ldots, w^N)\frac{1}{N}\sum_{i=1}^{N} w_i &= \sum_{i=1}^{N}\frac{1}{N}\pi(\theta)Q_\theta^N(w^1, w^2, \ldots, w^N)w^i \\ &= \sum_{k=1}^{N}\tilde{\pi}(k, \theta, w^1, w^2, \ldots, w^N)\end{aligned}$$

with

$$\tilde{\pi}(k, \theta, w^1, w^2, \ldots, w^N) := \frac{1}{N}\pi(\theta)Q_\theta^N(w^1, w^2, \ldots, w^N)w^k.$$

This is one of the other (hidden) ideas of [2] which can also be implicitly found in [43, 9], where such a distribution is identified as target distribution of the algorithm. This means that the mechanism described in Alg. 3 is not the sole possibility in order to define MCMC updates targetting $\pi(\theta)$ marginally. In particular, notice the form of the following two conditional distributions

$$\tilde{\pi}(k \mid \theta, w^1, w^2, \ldots, w^N) = \frac{w^k}{\sum_{i=1}^{N} w^i}, \tag{0.11}$$

and with $w^{-k} := \big(w^1, w^2, \ldots, w^k, w^{k+1}, \ldots, w^N\big)$

$$\tilde{\pi}(w^{-k} \mid k, \theta, w^k) = Q_\theta^N(w^{-k} \mid w^k), \tag{0.12}$$

which can be used as Gibbs type MCMC updates leaving $\tilde{\pi}$ invariant. In addition, we have the standard decomposition

$$\begin{aligned}\tilde{\pi}(k, \theta, w^1, w^2, \ldots, w^N) &= \tilde{\pi}(k \mid \theta, w^1, w^2, \ldots, w^N) \times \tilde{\pi}(\theta, w^1, w^2, \ldots, w^N) \\ &= \tilde{\pi}(k \mid \theta, w^1, w^2, \ldots, w^N) \times \pi(\theta) Q_\theta^N(w^1, w^2, \ldots, w^N) \frac{1}{N} \sum_{i=1}^N w^i,\end{aligned}$$

which tells us that at equilibrium k can always be recovered by sampling from (0.11). Then (0.12) suggests that one can rejuvenate $N-1$ of the aggregated pseudo-marginal estimators $w^1, w^2, \ldots, w^N$, provided sampling from $Q_\theta(w_{-k} \mid w_k)$ is simple: this is always the case when independence is assumed. From above, the following MCMC update leaves $\tilde{\pi}(\theta, w^1, w^2, \ldots, w^N)$ invariant

$$R(\theta, w; \mathrm{d}(\vartheta, u)) := \sum_{k=1}^N \tilde{\pi}(k \mid \theta, w) Q_\theta^N(u^{-k} \mid w^k) \delta_{\theta, w_k}(\mathrm{d}\vartheta \times \mathrm{d}u^k)$$

and is described algorithmically in Alg. 4 (where $\mathcal{P}\big(w^1, w^2, \ldots, w^N\big)$ is the probability distribution of a discrete valued random variable such that $\mathbb{P}\big(K = k\big) \propto w_k$).

1 Given $\theta, w^1, w^2, \ldots, w^N$
2 Sample $k \sim \mathcal{P}\big(w^1, w^2, \ldots, w^N\big)$
3 Sample $u^{-k} \sim Q_\theta^N(\cdot \mid w^k)$ and set $u^k = w^k$
4 Return $\theta, u^1, \ldots, u^N$

Algorithm 4: iSIR algorithm

Such algorithms are of general interest, and are analyzed in [4]. This update can, however, also be intertwined with the standard ABC update. In the context of ABC this gives Alg. 5.

1 Given $\theta, y^1, \ldots, y^N$
2 Sample $k \sim \mathcal{P}\big(\psi(y^1, y_{\mathrm{obs}}), \psi(y^2, y_{\mathrm{obs}}), \ldots, \psi(y^N, y_{\mathrm{obs}})\big)$
3 Sample $\tilde{y}^i \sim \ell\big(\cdot \mid \theta\big)\lambda(\mathrm{d}\cdot)$, $i \in \{1, \ldots, N\} \setminus \{k\}$ and set $\tilde{y}^k = y^k$
4 Return $\theta, \tilde{y}^1, \ldots, \tilde{y}^N$

Algorithm 5: iSIR with ABC

and intertwining Alg. 5 with Alg. 3 one obtains Alg. 6.

In fact a recent result [29, Theorem 17] ensures that the resulting algorithm has a better asymptotic variance, the key observation being that one can compare two different inhomogeneous Markov chains with alternating transition kernels $\tilde{P}$ and R. For the standard pseudo-marginal algorithm R is the identity whereas for ABC with rejuvenation, R is the iSIR.

0.4.3 Playing with the Us

In practice, simulation of the random variables Y on a computer often involves using d (pseudo-)random numbers uniformly distributed on the unit interval $[0, 1]$, which are then

Given $\theta, y^1, \ldots, y^N$
Sample $k \sim \mathcal{P}\big(\psi(y^1, y_{\text{obs}}), \psi(y^2, y_{\text{obs}}), \ldots, \psi(y^N, y_{\text{obs}})\big)$
Sample $\tilde{y}^i \sim \ell\big(\cdot \mid \theta\big)\lambda(\mathrm{d}\cdot)$, $i \in \{1, \ldots, N\} \setminus \{k\}$ and set $\tilde{y}^k = y^k$
Sample $\vartheta \sim q(\theta, \cdot)$
Sample $\check{y}^i \sim \ell\big(\cdot \mid \vartheta\big)\lambda(\mathrm{d}\cdot)$, $i \in \{1, \ldots, N\}$
Return $(\vartheta, \check{y}^1, \ldots, \check{y}^N)$ with probability

$$\min\left\{1, \frac{\eta(\vartheta) \times \frac{1}{N}\sum_{i=1}^N \psi(\tilde{y}^i, y_{\text{obs}}) q(\vartheta, \theta)}{\eta(\theta) \times \frac{1}{N}\sum_{i=1}^N \psi(\tilde{y}^i, y_{\text{obs}}) q(\theta, \vartheta)}\right\},$$

otherwise return $(\theta, \tilde{y}^1, \ldots, \tilde{y}^N)$

Algorithm 6: ABC-MCMC with rejuvenation

mapped to form one Y^i. That is, there is a mapping from the unit cube $[0,1]^d$ to $\mathcal{Y}$, and with an inconsequential abuse of notation, if $U^i \sim \mathcal{U}(\big[0,1\big]^d)$ then $Y(U^i) \sim \ell(y \mid \theta)\lambda(\mathrm{d}y)$ and

$$T_\theta(N) := \frac{1}{N}\sum_{i=1}^N \psi\big(Y\big(U^i\big), y_{\text{obs}}\big)$$

is an unbiased estimator of (0.2), equivalent in fact to (0.7). This representation is discussed in [3] as a general reparametrization strategy to circumvent intractability, and we show here how this can be also exploited in order to improve the performance of ABC-MCMC. An extremely simple illustration of this is the situation where $d = 1$ and an inverse cdf method is used, that is $Y(U) = F^{-1}(U)$ where F is the cumulative distribution function (cdf) of Y. This is the case, for example, for the g-and-k model, whose inverse cdf is given by [21].

0.4.3.1 Stratification

Stratification is a classical variance reduction strategy with applications to Monte Carlo methods. It proceeds as follows in the present context. Let $\mathcal{A} := \big\{A_1, \ldots, A_N\big\}$ be a partition of the unit cube $[0,1]^d$ such that $\mathbb{P}(U^1 \in A_i) = 1/N$, and such that it is possible to sample uniformly from each A_i. Perhaps the simplest example of this is when $\mathcal{A}$ corresponds to the dyadic sub-cubes of $[0,1]^d$. Let $V^i \sim \mathcal{U}(A_i)$ for $i = 1, \ldots, N$ be independent. We may now replace the estimator in (0.7) with

$$T_\theta^{\text{strat}}(N) := \frac{1}{N}\sum_{i=1}^N \psi\big(Y\big(V^i\big), y_{\text{obs}}\big).$$

It is straightforward to check that this is a non-negative unbiased estimator of $\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta)$ which means that $W_\theta^{\text{strat}}(N) := T_\theta^{\text{strat}}(N)/\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta)$ has unit expectation as required. It has been shown in [8, Section 6.2] that when ψ satisfies (0.1), Theorem 7 applies directly in this scenario and that $\tilde{P}^{\text{strat}}$, the approximation of P corresponding to using $\{W_\theta^{\text{strat}}(N), \theta \in \Theta\}$ instead of $\{W_\theta(N), \theta \in \Theta\}$ always dominates $\tilde{P}$. These results extend to more general choices of $\psi(\cdot, \cdot)$ but require a much better understanding of this function.

0.4.3.2 Introducing dependence between estimators

Ideally for the acceptance probability of $\tilde{P}$,

$$\min\left\{1, r(\theta, \vartheta)\frac{w_\vartheta}{w_\theta}\right\},$$

to be reasonably large we would like w_ϑ large when w_θ is large. That is, we would like large and nefarious realisations of w_θ to be compensated by larger values of w_ϑ. A natural idea is therefore to attempt to introduce "positive dependence" between the estimators. [26] discuss the introduction of dependence in the discussion of [2], and a more sophisticated methodology that seeks to correlate w_θ and w_ϑ has recently been proposed in [17].

Viewing the introduction of positive dependence very generally, for $\theta, \vartheta \in \Theta$ let $Q_{\theta\vartheta}(\mathrm{d}w \times \mathrm{d}u)$ have marginals $Q_\theta(\mathrm{d}w)$ and $Q_\vartheta(\mathrm{d}w)$. It is shown in [8] that if in addition

$$Q_{\theta\vartheta}(A \times B) = Q_{\vartheta\theta}(B \times A), \quad \theta, \vartheta \in \Theta, \quad A, B \in \mathcal{B}(\mathbb{R}_+),$$

then the algorithm with the acceptance ratio above and proposal $Q_{\theta\vartheta}(A \times B)/Q_\theta(A)$ remains exact, i.e. reversible with respect to (0.5). A natural question is how one may implement the abstract condition above. Let $\mathcal{U}(S)$ denote the uniform distribution over the set S. In the context of ABC applications where the observations are functions of $U \sim \mathcal{U}(\left[0,1\right]^d)$, that is $Y = Y(U)$ it is possible to introduce dependence on the uniforms involved. Assume for now that $d = 1$ and let $C(\cdot, \cdot)$ be a copula, that is a probability distribution on $\big([0,1]^2, \mathcal{B}\big([0,1]^2\big)\big)$ with the uniform distribution on $[0,1]$ as marginals [34]. Some copulas induce positive or negative dependence between the pair of uniforms involved. It is possible to define a partial order among copulas (and in fact more generally for distributions with fixed marginals) which ranks copulas in terms of the "strength" of the dependence. The concordance order $(W_\theta^{(2)}, W_\vartheta^{(2)}) \leq_c (W_\theta^{(1)}, W_\vartheta^{(1)})$ holds if and only if for any non-decreasing functions for which the expectations exist,

$$\mathbb{E}\big(f(W_\theta^{(2)})g(W_\vartheta^{(2)})\big) \leq \mathbb{E}\big(f(W_\theta^{(1)})g(W_\vartheta^{(1)})\big).$$

Now let for any $\theta \in \Theta$, $W_\theta = w_\theta\big(U_1\big) = \psi \circ y_\theta\big(U_1\big)$ and $\big(U_1, U_2\big) \sim C(\cdot, \cdot)$. If $C(\cdot, \cdot)$ is symmetric, then for any $\theta, \vartheta \in \Theta$, with $A_\theta^{-1} := \big\{u \in [0,1] : w_\theta(u) \in A\big\}$

$$\begin{aligned} Q_{\theta\vartheta}\big(W_\theta \in A_\theta, W_\vartheta \in A_\vartheta\big) &= \mathbb{P}\big(U_1 \in A_\theta^{-1}, U_2 \in A_\vartheta^{-1}\big) \\ &= C\big(A_\theta^{-1}, A_\vartheta^{-1}\big) \\ &= C\big(A_\vartheta^{-1}, A_\theta^{-1}\big) \\ &= Q_{\vartheta\theta}\big(W_\vartheta \in A_\vartheta, W_\theta \in A_\theta\big), \end{aligned}$$

that is the required condition is satisfied for the algorithm to remain exact. Now, for example, if for any $\theta \in \Theta$ $w_\theta : [0,1] \to \mathbb{R}_+$ is monotone one should choose a copula which induces positive dependence. Indeed, as stated in [8] $(W_\theta^{(2)}, W_\vartheta^{(2)}) \leq_c (W_\theta^{(1)}, W_\vartheta^{(1)})$ implies that the expected acceptance ratio is larger for the choice $(W_\theta^{(1)}, W_\vartheta^{(1)})$ than the $(W_\theta^{(2)}, W_\vartheta^{(2)})$. However increasing the expected acceptance probability does not guarantee improved performance of the algorithm: for example, in the limiting case, the Fréchet–Hoeffding bounds will lead to reducible Markov chains in many scenarios. More specifically consider the copula defined as a mixture of the independent copula and the "copy" copula with weights $(\lambda, 1-\lambda)$

$$C_\lambda(A, B) = \lambda\mathcal{U}(A \cap B) + (1 - \lambda)\mathcal{U}(A)\mathcal{U}(B).$$

This copula is symmetric for all $\lambda \in [0,1]$ and is such that sampling from its conditionals is simple: copy with probability λ or draw afresh from a uniform. Note that $\lambda = 0$ corresponds to the standard pseudo marginal. It is not difficult to show that if $0 \leq \lambda \leq \lambda' \leq 1$ then $(W_\theta^{(\lambda)}, W_\vartheta^{(\lambda)}) \leq_c (W_\theta^{(\lambda')}, W_\vartheta^{(\lambda')})$, meaning that the conditional acceptance ratio is a non-decreasing function of λ in terms of the concordance order. However the choice $\lambda = 1$, which corresponds to the upper Fréchet–Hoeffding bound will obviously lead to a reducible

Markov chain. This is therefore a scenario where λ needs to be optimized and where adaptive MCMC may be used [6]. Such schemes can naturally be extended to the multivariate scenario but require the tuning of more parameters and an understanding of the variations of $w_\theta(\cdot)$ along each of its coordinates.

0.4.4 Locally adaptive ABC-MCMC

Given the inability of $\tilde{P}_N$ to inherit geometric ergodicity from P in fairly simple ABC settings, one may wonder if an alternative Markov kernel that uses a locally adaptive number of pseudo-samples can. A variety of such kernels are presented in [23], but we restrict our interest here to the "1-hit" ABC-MCMC method proposed in [25] for the specific setting where ψ satisfies (0.1).

1 Given θ
2 Sample $\vartheta \sim q(\theta, \cdot)$
3 With probability

$$1 - \min\left\{1, \frac{\eta(\vartheta)q(\vartheta,\theta)}{\eta(\theta)q(\theta,\vartheta)}\right\},$$

stop and output θ.
4 Sample $Y_\theta \sim \ell(\cdot \mid \theta)\lambda(\mathrm{d}\cdot)$ and $Y_\vartheta \sim \ell(\cdot \mid \vartheta)\lambda(\mathrm{d}\cdot)$ independently until

$$\mathbb{I}\{\|s(Y_\theta) - s(y_{\text{obs}})\| \leq \epsilon\} + \mathbb{I}\{\|s(Y_\vartheta) - s(y_{\text{obs}})\| \leq \epsilon\} \geq 1,$$

and output ϑ if $\mathbb{I}\{\|s(Y_\vartheta) - s(y_{\text{obs}})\| \leq \epsilon\} = 1$. Otherwise, output θ

Algorithm 7: 1-hit ABC-MCMC

This algorithm defines a Markov chain evolving on Θ that is not a Metropolis–Hastings Markov chain. Indeed, the algorithm defines a transition kernel $\check{P}$ in which the proposal is q as in Algorithm 1 but the acceptance probability is

$$\min\left\{1, \frac{\eta(\vartheta)q(\vartheta,\theta)}{\eta(\theta)q(\theta,\vartheta)}\right\} \times \frac{\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \vartheta)}{\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta) + \ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \vartheta) - \ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta)\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \vartheta)}.$$

From this, it can be verified directly that the Markov chain is reversible with respect to π.

An interesting feature of this algorithm is that a random number of pseudo-observations from the distributions associated with both θ and ϑ are sampled in what can intuitively be viewed as a race between the two parameter values. In fact, the number of paired samples required for the race to terminate, N, is a geometric random variable depending on both $\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \theta)$ and $\ell^\psi_{\text{ABC}}(y_{\text{obs}} \mid \vartheta)$ in such a way that N is typically larger when these quantities are both small and smaller when either of these are large. This can be interpreted broadly as a local adaptation of the computational effort expended in simulating the Markov chain, in contrast to the fixed N strategy outlined in Section 0.4.1. In [27, Proposition 3] it is shown that the expected computational cost of simulating each iteration of the resulting Markov chain is bounded whenever η defines a proper prior. In [27, Theorem 4] it is shown that $\check{P}$ can inherit (under additional assumptions) geometric ergodicity from P even when Theorem 9 holds, i.e. $\tilde{P}_N$ is not geometrically ergodic for any N. Consequently, $\check{P}$ can provide superior estimates of expectations, in comparison to $\tilde{P}_N$, for some functions f such that $\pi(f^2) < \infty$.

0.4.5 Inexact algorithms

A natural question in the context of ABC-MCMC methods is how pseudo-marginal methods compare to inexact variants. Of course, there are a large number of inexact methods and we consider here only the simple case arising from a small modification of the pseudo-marginal algorithm known as Monte Carlo within Metropolis (MCWM) [35]. In this algorithm, a Markov chain is defined by the transition kernel $\hat{P}_\lambda$ whose algorithmic description is given in Algorithm 8. Here, as in the pseudo-marginal approach, one has for any $\lambda > 0$

$$\mathbb{E}_{Q_\theta^{(\lambda)}}[W_\theta] = 1, \quad \theta \in \Theta$$

and one defines a Markov chain evolving on Θ as follows.

1 Given θ
2 Sample $\vartheta \sim q(\theta, \cdot)$
3 Sample $W_\theta \sim Q_\theta^{(\lambda)}$
4 Sample $W_\vartheta \sim Q_\vartheta^{(\lambda)}$
5 Return the realisation ϑ with probability

$$\min\left\{1, \frac{\pi(\vartheta) \times W_\vartheta q(\vartheta, \theta)}{\pi(\theta) \times W_\theta q(\theta, \vartheta)}\right\},$$

otherwise return θ

Algorithm 8: MCWM ABC

This Markov chain has been studied in [5] in the case where P is uniformly ergodic, see also [1]. Extensions to the case where P is geometrically ergodic are considered in [32]. One such result is [32, Theorem 4.1]:

Theorem 11. *Assume P is geometrically ergodic, that*

$$\lim_{\lambda \to 0} \sup_{\theta \in \Theta} Q_\theta^{(\lambda)}(|W_\theta - 1| > \delta) = 0, \quad \forall \delta > 0,$$

and

$$\lim_{\lambda \to 0} \sup_{\theta \in \Theta} \mathbb{E}_{Q_\theta^{(\lambda)}}\left[W_\theta^{-1}\right] = 1.$$

Then $\hat{P}_\lambda$ is geometrically ergodic for all sufficiently small λ. In addition, under a very mild technical assumption on P,

$$\lim_{\lambda \to 0} \|\pi - \hat{\pi}_\lambda\|_{\mathrm{TV}} = 0,$$

where $\hat{\pi}_\lambda$ is the unique invariant distribution associated with $\hat{P}_\lambda$ for sufficiently small λ.

Results of this kind invite comparison with corresponding results for $\tilde{P}$, the pseudo-marginal Markov transition kernel. While Theorem 11 is reassuring, the assumptions above correspond to "uniform in θ" assumptions on $Q_\theta^{(\lambda)}$, similar to assumptions for analyzing $\tilde{P}$, that may not hold in practical applications. Nevertheless there are differences, and $\check{P}$ can be geometrically ergodic when $\tilde{P}$ is not when the same distributions $\{Q_\theta, \theta \in \Theta\}$ are employed for both. On the other hand, examples in which $\tilde{P}$ is geometrically ergodic but $\hat{P}$ is transient can also be constructed, so a degree of caution is warranted.

This result can be interpreted in the specific case where N is a number of i.i.d. pseudo-samples. When $W_\theta(N)$ is a simple average of N i.i.d. random variables W_θ, one can see that the assumptions of Theorem 11 are weaker than corresponding assumptions for $\tilde{P}_N$.

In fact, the conditions are satisfied when the weights W_θ are uniformly integrable (rather than being uniformly bounded) and there exist constants $M > 0$, $\beta > 0$ and $\gamma \in (0, 1)$ such that

$$\sup_{\theta \in \Theta} Q_\theta(W_\theta \leq w) \leq M w^\beta, \qquad w \in (0, \gamma).$$

The latter condition imposes, e.g., the requirement that $Q_\theta(W_\theta = 0) = 0$ for all $\theta \in \Theta$ and indeed the MCWM algorithm is not defined without this assumption. When the weights W_θ additionally have a uniformly bounded $1 + k$ moment then one can obtain bounds on the rate of convergence of $\hat{\pi}_N$ to π in total variation [32, Proposition 4.1].

0.5 Remarks

In this chapter, we have presented a relevant subset of theory and directions in methodological research pertaining to ABC-MCMC algorithms. Given the recent prevalence of applications involving ABC, this survey has not been exhaustive and has focused on specific aspects in which the theory is particularly clear. For example, we have not treated the question of which kernels are most appropriate in various applications, and indeed methodological innovations such as the incorporation of a tolerance parameter ϵ (such as that found in (0.1) but which may parameterize alternative kernels as well) as a state variable of the ABC Markov chain (see, e.g., [14, 36]).

Theoretical and methodological advances continue to be made in this area. For example, the spectral gap of the Markov kernel $\tilde{P}_N$ discussed in Section 0.4.1 does not increase as a function of N in general, and it is of interest to consider whether alternative Markov kernels can overcome this issue. In [24] it is shown that in many cases the spectral gap of the ABC-MCMC kernel with rejuvenation in Section 0.4.2 does improve as N grows, and alternative methodology similar to that in Section 0.4.4 is both introduced and analyzed. Finally, we note that when the model admits specific structure, alternatives to the simple ABC method presented here may be more computationally efficient. Indeed, in the case where the observations are temporally ordered then it is natural and typically much more efficient to consider Sequential Monte Carlo methods for generating W_θ (see, e.g., [31]).

Index

Bibliography

[1] P. Alquier, N. Friel, R. Everitt, and A. Boland. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Stat. Comput.*, 16(1):29–47, 2016.

[2] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):269–342, 2010. (with discussion).

[3] C. Andrieu, A. Doucet, and A. Lee. Discussion of "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation" by Fearnhead and Prangle. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74(3):451–452, 2012.

[4] C. Andrieu, A. Lee, and M. Vihola. Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 2017. To appear.

[5] C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725, 2009.

[6] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Stat. Comput.*, 18(4):343–373, 2008.

[7] C. Andrieu and M. Vihola. Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.*, 25(2):1030–1077, 2015.

[8] C. Andrieu and M. Vihola. Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.*, 26(5):2661–2696, 2016.

[9] H. M. Austad. Parallel multiple proposal MCMC algorithms. Master's thesis, Norwegian University of Science and Technology, 2007.

[10] M. A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160, 2003.

[11] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[12] C. Becquet and M. Przeworski. A new approach to estimate parameters of speciation models with application to apes. *Genome research*, 17(10):1505–1519, 2007.

[13] L. Bornn, N. S. Pillai, A. Smith, and D. Woodard. The use of a single pseudo-sample in approximate Bayesian computation. *Stat. Comput.*, 2017. To appear.

[14] P. Bortot, S. G. Coles, and S. A. Sisson. Inference for stereological extremes. *J. Am. Stat. Assoc.*, 102(477):84–92, 2007.

[15] T. A. Dean, S. S. Singh, A. Jasra, and G. W. Peters. Parameter estimation for hidden Markov models with intractable likelihoods. *Scand. J. Statist.*, 41(4):970–987, 2014.

[16] P. Del Moral, A. Doucet, and A. Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.*, 22(5):1009–1020, 2012.

[17] G. Deligiannidis, A. Doucet, and M. K. Pitt. The correlated pseudo-marginal method. *arXiv preprint arXiv:1511.04992*, 2015.

[18] G. Deligiannidis and A. Lee. Which ergodic averages have finite asymptotic variance? *arXiv preprint arXiv:1606.08373*, 2016.

[19] R. Douc, G. Fort, E. Moulines, and P. Soulier. Practical drift conditions for subgeometric rates of convergence. *Ann. Appl. Probab.*, 14(3):1353–1377, 2004.

[20] A. Doucet, M. K. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.

[21] P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74(3):419–474, 2012.

[22] A. Kennedy and J. Kuti. Noise without noise: a new Monte Carlo method. *Physical Review Letters*, 54(23):2473, 1985.

[23] A. Lee. On the choice of MCMC kernels for approximate Bayesian computation with SMC samplers. In *Proceedings of the Winter Simulation Conference*, 2012.

[24] A. Lee, C. Andrieu, and A. Doucet. An active particle perspective of MCMC and its application to locally adaptive MCMC algorithms. In preparation.

[25] A. Lee, C. Andrieu, and A. Doucet. Discussion of "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation" by Fearnhead and Prangle. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74(3):449–450, 2012.

[26] A. Lee and C. Holmes. Discussion of "particle Markov chain Monte Carlo" by Andrieu, Doucet & Holenstein. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):327–329, 2010.

[27] A. Lee and K. Łatuszyński. Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika*, 101(3):655–671, 2014.

[28] L. Lin and J. Sloan. A stochastic Monte Carlo algorithm. *Phys. Rev. D*, 61(hep-lat/9905033):074505, 1999.

[29] F. Maire, R. Douc, J. Olsson, et al. Comparison of asymptotic variances of inhomogeneous Markov chains with application to Markov chain Monte Carlo methods. *Ann. Statist.*, 42(4):1483–1510, 2014.

[30] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 100(26):15324–15328, 2003.

[31] J. S. Martin, A. Jasra, S. S. Singh, N. Whiteley, P. Del Moral, and E. McCoy. Approximate Bayesian computation for smoothing. *Stoch. Anal. Appl.*, 32(3):397–420, 2014.

[32] F. J. Medina-Aguayo, A. Lee, and G. O. Roberts. Stability of noisy Metropolis–Hastings. *Stat. Comput.*, 26(6):1187–1211, 2016.

[33] A. Müller and D. Stoyan. *Comparison methods for stochastic models and risks*. Wiley, 2002.

[34] R. B. Nelsen. *An introduction to copulas*, volume 139. Springer, 1999.

[35] P. D. O'Neill, D. J. Balding, N. G. Becker, M. Eerola, and D. Mollison. Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. C Applied Statistics*, 49(4):517–542, 2000.

[36] O. Ratmann, C. Andrieu, C. Wiuf, and S. Richardson. Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Natl. Acad. Sci. USA*, 106(26):10576–10581, 2009.

[37] G. O. Roberts and J. S. Rosenthal. Variance bounding Markov chains. *Ann. Appl. Probab.*, 18(3):1201–1214, 2008.

[38] M. Shaked and J. G. Shanthikumar. *Stochastic orders*. Springer, 2007.

[39] C. Sherlock, A. Thiery, and A. Lee. Pseudo-marginal Metropolis–Hastings using averages of unbiased estimators. *Biometrika*, 2017. To appear.

[40] C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.*, 43(1):238–275, 02 2015.

[41] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 104(6):1760–1765, 2007.

[42] S. Tavare, D. J. Balding, R. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

[43] H. Tjelmeland. Using all Metropolis–Hastings proposals to estimate mean values. Technical Report 4, Norwegian University of Science and Technology, 2004.

[44] R. D. Wilkinson. Approximate Bayesian computation gives exact results under the assumption of model error. *Stat. Appl. Genet. Mol. Biol.*, 12(2):129–141, 2013.