Biometrika (2018), **xx**, x, *pp*. 1–8
Printed in Great Britain

Unbiased approximations of products of expectations

BY A. LEE
School of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, U.K.
anthony.lee@bristol.ac.uk

S. TIBERI
Institute of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
simone.tiberi@uzh.ch

AND G. ZANELLA
Department of Decision Sciences, BIDSA and IGIER, Bocconi University, Via Roentgen 1, 20136, Milan, Italy
giacomo.zanella@unibocconi.it

SUMMARY

We consider the problem of approximating the product of n expectations with respect to a common probability distribution μ. Such products routinely arise in statistics as values of the likelihood in latent variable models. Motivated by pseudo-marginal Markov chain Monte Carlo schemes, we focus on unbiased estimators of such products. The standard approach is to sample N particles from μ and assign each particle to one of the expectations. This is wasteful and typically requires the number of particles to grow quadratically with the number of expectations. We propose an alternative estimator that approximates each expectation using most of the particles while preserving unbiasedness, which is computationally more efficient when the cost of simulations greatly exceeds the cost of likelihood evaluations. We carefully study its properties, showing that in latent variable contexts the proposed estimator needs only $\mathcal{O}(n)$ particles to match the performance of the standard approach with $\mathcal{O}(n^2)$ particles. We demonstrate the procedure on two latent variable examples from approximate Bayesian computation and single-cell gene expression analysis, observing computational gains by factors of about 25 and 450 respectively.

Some key words: latent variables; Markov chain Monte Carlo; pseudo-marginal; approximate Bayesian computation.

1. INTRODUCTION

Let X be a random variable with probability measure μ on a measurable space $(\mathsf{X}, \mathcal{X})$, and let $L^1(\mu)$ be the class of integrable, real-valued functions, i.e. $L^1(\mu) = \{f : \int_{\mathsf{X}} |f(x)|\, \mu(\mathrm{d}x) < \infty\}$. For a sequence of non-negative potential functions $G_1, \ldots, G_n \in L^1(\mu)$, we consider approximations of products of n expectations

$$\gamma = \prod_{p=1}^{n} E\{G_p(X)\} = \prod_{p=1}^{n} \mu(G_p), \tag{1}$$

where we denote $\mu(f) = \int_{\mathsf{X}} f(x)\mu(\mathrm{d}x)$ for $f \in L^1(\mu)$. These arise, e.g., as values of the likelihood function in latent variable models. We concentrate on unbiased approximations of γ; these can be used, e.g., within pseudo-marginal Markov chain methods for approximating posterior expectations. Pseudo-marginal methods (Beaumont, 2003; Andrieu & Roberts, 2009) are a variation of classical Metropolis–

Hastings algorithms, where the target density function is replaced by an unbiased estimator while still preserving the correct invariant distribution.

To motivate this general problem, and because our main result in the sequel relates to latent variable models, we provide the following generic example of such a model.

Example 1 *(Latent variable model).* Let g be a Markov transition density and $Y_1, \ldots, Y_n$ be independent and identically distributed Y-valued random variables distributed according to the probability density function ν where $\nu(y) = E\{g(X, y)\} = \int_E g(x, y)\mu(\mathrm{d}x)$. That is, the Y_p are independent and distributed according to $g(X_p, \cdot)$ where $X_p \sim \mu$. For observations $y_1, \ldots, y_n$, respectively, of $Y_1, \ldots, Y_n$, we can write $\prod_{p=1}^n \nu(y_p) = \prod_{p=1}^n E\{g(X, y_p)\} = \prod_{p=1}^n E\{G_p(X)\} = \gamma$, where the potential functions are defined via $G_p(x) = g(x, y_p)$, for $p \in \{1, \ldots, n\}$.

Remark 1. To see that γ can be viewed as a value of the likelihood function, let $\theta \in \Theta$ be a statistical parameter and let $\{(\mu_\theta, g_\theta) : \theta \in \Theta\}$ be parameterized families of distributions and Markov transition densities. The likelihood function L is then $L(\theta) = \prod_{p=1}^n \nu_\theta(y_p)$ where $\nu_\theta(y) = E_\theta\{g_\theta(X, y)\} = \int_E g_\theta(x, y)\mu_\theta(\mathrm{d}x)$, and clearly $L(\theta)$ is of the form (1) for any $\theta \in \Theta$.

The focus of this paper is approximations of γ using N independent and μ-distributed random variables $\zeta = (\zeta_1, \ldots, \zeta_N)$, which we will refer to throughout as particles. A straightforward approach to constructing an unbiased approximation of γ is to approximate each expectation $E\{G_p(X)\} = \mu(G_p)$ independently using M particles, where $N = Mn$. That is, we define

$$\gamma^N_{\mathrm{simple}} = \prod_{p=1}^n \frac{1}{M} \sum_{i=1}^M G_p(\zeta_{(p-1)M+i}). \tag{2}$$

The following lack-of-bias, consistency, second moment and variance properties are easily established.

PROPOSITION 1. *We have* $E(\gamma^N_{\mathrm{simple}}) = \gamma$, $\gamma^N_{\mathrm{simple}} \to \gamma$ *in probability as* $N \to \infty$ *and*

$$E\{(\gamma^N_{\mathrm{simple}}/\gamma)^2\} = \prod_{p=1}^n [1 + \{\mu(\bar{G}_p^2) - 1\}/M], \tag{3}$$

where $\bar{G}_p = G_p/\mu(G_p)$ *for each* $p \in \{1, \ldots, n\}$. *Also,* $\mathrm{var}(\gamma^N_{\mathrm{simple}})$ *is finite and converges to* 0 *as* $M \to \infty$ *if and only if*

$$\max_{p \in \{1,\ldots,n\}} \mu(G_p^2) < \infty. \tag{4}$$

The approximation $\gamma^N_{\mathrm{simple}}$ is straightforward to compute and analyze since it is a product of averages of independent random variables. However, each particle is only used to approximate one of the expectations in the product, and in situations where these particles are expensive to obtain this may be wasteful. An alternative approach is to use

$$\gamma^N_{\mathrm{biased}} = \prod_{p=1}^n \frac{1}{N} \sum_{i=1}^N G_p(\zeta_i), \tag{5}$$

which is consistent and not wasteful, but also not unbiased in general.

PROPOSITION 2. *We have* $\gamma^N_{\mathrm{biased}} \to \gamma$ *in probability as* $N \to \infty$ *but* $E(\gamma^N_{\mathrm{biased}}) \neq \gamma$ *in general.*

We propose in the sequel an approximation $\gamma^N_{\mathrm{recycle}}$ that is unbiased like $\gamma^N_{\mathrm{simple}}$ but similar to $\gamma^N_{\mathrm{biased}}$ in that it uses most of the particles to approximate each expectation in the product while remaining computationally tractable. The approximation $\gamma^N_{\mathrm{recycle}}$ can be viewed as an unbiased approximation of γ^N_{perm}, the rescaled permanent of a particular rectangular matrix of random variables, which is never worse in terms of variance than $\gamma^N_{\mathrm{simple}}$ but is very computationally costly to compute in general. The approximation $\gamma^N_{\mathrm{recycle}}$ is an extension of the importance sampling approximation of the permanent of a square matrix proposed by Kuznetsov (1996) to the case of rectangular matrices. While it is possible for $\gamma^N_{\mathrm{recycle}}$ to have a higher

variance than γ^N_{simple}, we show that in many statistical scenarios it requires far fewer particles to obtain a given variance, e.g. in the latent variable setting described above. In particular, under weak assumptions, one needs to take $N = \mathcal{O}(n)$ to control the relative variance of $\gamma^N_{\text{recycle}}$ but one requires $N = \mathcal{O}(n^2)$ to control the relative variance of γ^N_{simple}. Ultimately, this provides large computational savings when the cost of simulating from μ is much greater than the cost of evaluating each G_p.

2. The associated permanent and its approximation

An alternative approximation of γ which uses the particles $\zeta = (\zeta_1, \ldots, \zeta_N)$ is obtained by first rewriting γ in (1) as $\gamma = \prod_{p=1}^n E\{G_p(X)\} = E\{\prod_{p=1}^n G_p(X_p)\}$, with $X_1, \ldots, X_p$ independent μ-distributed random variables. Indeed, γ^N_{biased} is a V-statistic of order n for γ, and the corresponding U-statistic for γ is

$$\gamma^N_{\text{perm}} = \sum_{k \in P(N,n)} |P(N,n)|^{-1} \prod_{p=1}^n G_p(\zeta_{k_p}), \tag{6}$$

where $P(N,n) = \{k \in \{1,\ldots,N\}^n : k_i = k_j \iff i = j\}$ is the set of n-permutations of N, whose cardinality is $|P(N,n)| = N!/(N-n)!$. We observe that γ^N_{perm} is exactly $|P(N,n)|^{-1}$ times the permanent of the rectangular matrix A (see, e.g., Ryser, 1963, p. 25) with entries $A_{ij} = G_i(\zeta_j)$ since then $\text{perm}(A) = \sum_{k \in P(N,n)} \prod_{p=1}^n A_{p,k_p} = \sum_{k \in P(N,n)} \prod_{p=1}^n G_p(\zeta_{k_p})$. The approximation γ^N_{perm} is unbiased and consistent since it is a U-statistic and moreover it is less variable than γ^N_{simple} in terms of the convex order (see, e.g., Shaked & Shanthikumar, 2007, Section 3.A), defined by $X \preceq_{\text{cx}} Y$ if $E\{\phi(X)\} \leq E\{\phi(Y)\}$ for all convex functions $\phi : \mathbb{R} \to \mathbb{R}$ such that the expectations are well-defined. Since $x \mapsto x^2$ is convex, $X \preceq_{\text{cx}} Y$ implies $\text{var}(X) \leq \text{var}(Y)$. Convex-ordered families of random variables also allow one to order the asymptotic variances of associated pseudo-marginal Markov chains (Andrieu & Vihola, 2016, Theorem 10). We now state basic properties of γ^N_{perm}, which can be compared with Proposition 1.

THEOREM 1. *We have* $E(\gamma^N_{\text{perm}}) = \gamma$, $\gamma^N_{\text{perm}} \preceq_{\text{cx}} \gamma^N_{\text{simple}}$ *and* $\gamma^N_{\text{perm}} \to \gamma$ *in probability as* $N \to \infty$. *Given* $K \sim \text{Uniform}\{P(N,n)\}$, *it holds that* $E\{(\gamma^N_{\text{perm}}/\gamma)^2\} = E\{\prod_{p=1}^n \bar{G}_p(\zeta_p)\bar{G}_p(\zeta_{K_p})\}$. *Also,* $\text{var}(\gamma^N_{\text{perm}})$ *is finite and* $\text{var}(\gamma^N_{\text{perm}}) \to 0$ *as* $N \to \infty$ *if and only if* (4) *holds.*

Theorem 1 suggests that γ^N_{perm} is a superior approximation of γ in comparison to γ^N_{simple}. However, computing γ^N_{perm} is equivalent to computing the permanent of a rectangular matrix, which has no known polynomial-time algorithm. In fact, computing the permanent of a square matrix is #P-hard (Valiant, 1979). Using an extension of the importance sampling estimator of the permanent of a square matrix due to Kuznetsov (1996), we define the following unbiased approximation of γ^N_{perm} and hence γ,

$$\gamma^N_{\text{recycle}} = \prod_{p=1}^n \frac{1}{N-p+1} \sum_{j=1}^N G_p(\zeta_j) \mathbb{I}\left(j \notin \{K_1, \ldots, K_{p-1}\}\right), \tag{7}$$

where $K = (K_1, \ldots, K_n)$ is a random variable with values in $\{1,\ldots,N\}^n$ whose distribution given ζ is defined by the sequence of conditional probabilities

$$\text{pr}\left(K_p = i \mid \zeta, K_1, \ldots, K_{p-1}\right) \propto G_p(\zeta_i) \mathbb{I}\left(i \notin \{K_1, \ldots, K_{p-1}\}\right). \tag{8}$$

In (8) we take $G_p(\zeta_i) / \sum_{j=1}^N G_p(\zeta_j) \mathbb{I}\left(j \notin \{K_1, \ldots, K_{p-1}\}\right)$ to be 1 when the denominator is equal to 0, in which case $K_p \mid (\zeta, K_1, \ldots, K_{p-1}) \sim \text{Uniform}(\{1,\ldots,N\} \setminus \{K_1, \ldots, K_{p-1}\})$. The choice of the conditional distribution of K_p when $\sum_{j=1}^N G_p(\zeta_j) \mathbb{I}\left(j \notin \{K_1, \ldots, K_{p-1}\}\right) = 0$ is in some sense arbitrary, as in any case $\gamma^N_{\text{recycle}} = 0$ whenever this happens. We now state basic properties of $\gamma^N_{\text{recycle}}$, which can be compared with Theorem 1.

THEOREM 2. *We have* $E(\gamma^N_{\mathrm{recycle}} \mid \zeta) = \gamma^N_{\mathrm{perm}}$, $E(\gamma^N_{\mathrm{recycle}}) = \gamma$, $\gamma^N_{\mathrm{perm}} \preceq_{\mathrm{cx}} \gamma^N_{\mathrm{recycle}}$ *and* $\gamma^N_{\mathrm{recycle}} \to \gamma$ *in probability as* $N \to \infty$. *Given a vector of independent random variables* $S = (S_1, \ldots, S_n)$ *with* $S_p \sim \mathrm{Uniform}(\{p, \ldots, N\})$ *for* $p \in \{1, \ldots, n\}$, *it holds* $E\{(\gamma^N_{\mathrm{recycle}}/\gamma)^2\} = E\{\prod_{p=1}^n \bar{G}_p(\zeta_p)\bar{G}_p(\zeta_{S_p})\}$. *Also,* $\mathrm{var}(\gamma^N_{\mathrm{recycle}})$ *is finite and* $\mathrm{var}(\gamma^N_{\mathrm{recycle}}) \to 0$ *as* $N \to \infty$ *if and only if*

$$\max_{p \in \{1,\ldots,n\},\, B \subseteq \{1,\ldots,p\}} \mu\Big(G_p \prod_{j \in B} G_j\Big) < \infty. \tag{9}$$

COROLLARY 1. *If* $\max_{p \in \{1,\ldots,n\}} \mu(G_p^{n+1}) < \infty$ *then* $\mathrm{var}(\gamma^N_{\mathrm{recycle}}) \to 0$ *as* $N \to \infty$.

Remark 2. While (4) is sufficient for γ^N_{perm} and $\gamma^N_{\mathrm{simple}}$ to have finite variance converging to 0 as $N \to \infty$, this is not sufficient in general for $\gamma^N_{\mathrm{recycle}}$, which requires (9) instead.

The estimator $\gamma^N_{\mathrm{simple}}$ uses only N/n out of N particles to estimate each expectation in the product; in contrast $\gamma^N_{\mathrm{recycle}}$ uses $N - p$ particles for the pth expectation $\mu(G_p)$. In this sense, the latter recycles most of the particles for each term, and we therefore refer to $\gamma^N_{\mathrm{recycle}}$ as the recycled estimator. While Remark 2 implies that it is not possible for $\mathrm{var}(\gamma^N_{\mathrm{recycle}}) \leq \mathrm{var}(\gamma^N_{\mathrm{simple}})$ in general, we show in the coming section that $\mathrm{var}(\gamma^N_{\mathrm{recycle}})$ can be orders of magnitude smaller than $\mathrm{var}(\gamma^N_{\mathrm{simple}})$ in many statistical settings. Let us first provide a result motivated by approximate Bayesian computation applications, in which it is often the case that the potential functions are indicator functions. In this case, it is always true that $\gamma^N_{\mathrm{recycle}}$ has a smaller variance than $\gamma^N_{\mathrm{simple}}$.

PROPOSITION 3. *Let* $A_1, \ldots, A_n \in \mathcal{X}$ *satisfy* $\mu(A_p) > 0$ *for* $p \in \{1, \ldots, n\}$ *and let* $G_p = \mathbb{I}_{A_p}$ *for* $p \in \{1, \ldots, n\}$. *Then* $E\{(\gamma^N_{\mathrm{recycle}}/\gamma)^2\} \leq E\{(\gamma^N_{\mathrm{simple}}/\gamma)^2\}$.

We observe also that an algorithm computing $\gamma^N_{\mathrm{recycle}}$ by accumulating averages and drawing K_p for $p = 1, \ldots, n$ does not need to construct the matrix of potential function evaluations A, but can construct each row one at a time without any recomputation. This makes the algorithm feasible to implement even for values of n and N such that storing A would exhaust memory on a typical computer.

In the supplementary material, we compare $\gamma^N_{\mathrm{recycle}}$ and $\gamma^N_{\mathrm{biased}}$ in terms of mean squared error, and the results suggest that they are of the same order so that there is no appreciable bias-variance tradeoff.

3. SCALING OF THE NUMBER OF PARTICLES WITH n IN LATENT VARIABLE MODELS

We investigate the variance of $\gamma^N_{\mathrm{recycle}}$ in comparison to $\gamma^N_{\mathrm{simple}}$ in the large n regime. In particular, we show that only $N = \mathcal{O}(n)$ particles are required to control the relative variance of $\gamma^N_{\mathrm{recycle}}$ in some scenarios in which $N = \mathcal{O}(n^2)$ particles are required to control the relative variance of $\gamma^N_{\mathrm{simple}}$. In the supplement, we also show that this cannot always be true, in some situations $N = \mathcal{O}(n^2)$ is a lower bound on the number of particles required to control the relative variance of γ^N_{perm}, and therefore $\gamma^N_{\mathrm{recycle}}$. To simplify the presentation, we define $c_p = \mu(\bar{G}_p^2) - 1$, for $p \in \{1, \ldots, n\}$. We will occasionally make reference to the following assumption when considering the large n regime

$$0 < \inf_{p \geq 1} c_p \leq \sup_{p \geq 1} c_p < \infty. \tag{10}$$

We begin by observing that from Proposition 1, if (10) holds and $M = \lceil \alpha n^\beta \rceil$, where $\lceil x \rceil$ denotes the least integers greater or equal than x, then the second moment of $\gamma^N_{\mathrm{simple}}/\gamma$ is bounded above as $n \to \infty$ if and only if $\alpha > 0$ and $\beta \geq 1$. Since $N = Mn$, this implies that to stabilize the relative variance of $\gamma^N_{\mathrm{simple}}$ in the large n regime one must take $N = \mathcal{O}(n^2)$.

The second moment of $\gamma^N_{\mathrm{recycle}}/\gamma$ is more complex to analyze because it involves interactions between different potential functions. However, if a mutual independence condition is satisfied, the following proposition implies that $N = \lceil \alpha n \rceil$ with $\alpha > 1$ is sufficient for $E\{(\gamma^N_{\mathrm{recycle}}/\gamma)^2\}$ to be uniformly bounded over n, for example by $\exp\{c/(\alpha - 1)\}$ with $c = \sup_{p \geq 1} c_p$.

PROPOSITION 4. *Assume $G_1(X), \ldots, G_n(X)$ are mutually independent when $X \sim \mu$. Then*

$$E\left\{\left(\gamma^N_{\text{recycle}}/\gamma\right)^2\right\} = \prod_{p=1}^{n} \{1 + c_p/(N-p+1)\}. \tag{11}$$

The assumption of mutual independence in Proposition 4 is very strong in statistical settings. However, we show now that in latent variable models the expected second moment of $\gamma^N_{\text{recycle}}/\gamma$ is very similar to (11), where the expectation is with respect to the law of the independent and identically distributed random variables $Y_1, \ldots, Y_n \sim \nu$ and we denote it by E_Y. For the remainder of this section, we denote by $\bar{G}_1$ the random function $x \mapsto g(x, Y_1)/\nu(Y_1)$ for $Y_1 \sim \nu$. We begin by verifying that, for latent variable models, a finite expected second moment for $\bar{G}_1(X)$ when $X \sim \mu$ is sufficient for $\text{var}(\gamma^N_{\text{recycle}})$ to be finite.

PROPOSITION 5. *In the setting of Example* 1*, assume that $E_Y\{\mu(\bar{G}_1^2)\} < \infty$. Then* (9) *holds almost surely. Also, if $Y_1, \ldots, Y_n$ are independent and identically distributed random variables with common distribution ν_0 that is absolutely continuous with respect to ν, then* (9) *holds almost surely.*

Remark 3. The condition $E_Y\{\mu(\bar{G}_1^2)\} < \infty$ is not very strong, but is not always satisfied. For example, if μ is $\text{Uniform}(0,1)$ and $g(x,\cdot)$ is $\text{Uniform}(0,x)$ for each $x \in (0,1)$ then simple calculations show that $E_Y\{\mu(\bar{G}_1^2)\} = \infty$. See Remark A4 in the supplementary material for further details and references.

The following Theorem is our main result in terms of applicability to statistical scenarios. It suggests that when considering the expected second moment of $\gamma^N_{\text{recycle}}/\gamma$, it is as if the random variables $G_1(X), \ldots, G_n(X)$ are mutually independent on average, and allows easy comparison with the corresponding expected second moment of $\gamma^N_{\text{simple}}/\gamma$.

THEOREM 3. *In the setting of Example* 1*, and letting E_Y denoting expectation with respect to $Y_1, \ldots, Y_n$,*

$$E_Y\left[E\left\{\left(\gamma^N_{\text{recycle}}/\gamma\right)^2\right\}\right] = \prod_{p=1}^{n} \{1 + C/(N-p+1)\}, \qquad C = E_Y\left\{\mu(\bar{G}_1^2)\right\} - 1. \tag{12}$$

In the setting of Example 1, it is straightforward to obtain from Proposition 1 that $E_Y[E\{(\gamma^N_{\text{simple}}/\gamma)^2\}] = (1 + C/M)^n$, where C is as in Theorem 3. Hence, one requires $N = \lceil \alpha n \rceil$ for $\alpha > 1$ to control the expected relative variance of $\gamma^N_{\text{recycle}}$ but one requires $M = \mathcal{O}(n)$ and hence $N = \mathcal{O}(n^2)$ to control the expected relative variance of γ^N_{simple} when $0 < C < \infty$. In addition, it is clear that $E_Y[E\{(\gamma^N_{\text{recycle}}/\gamma)^2\}] < E_Y[E\{(\gamma^N_{\text{simple}}/\gamma)^2\}]$ for any N that is an integer multiple of $n > 1$. Theorem 3 can be combined with Markov's inequality to bound the probability of $\text{var}(\gamma^N_{\text{recycle}}/\gamma)$ being large. In particular, since $E\{(\gamma^N_{\text{recycle}}/\gamma)^2\} \geq 0$, we obtain that if $N = \lceil \alpha n \rceil$ with $\alpha > 1$ then $\text{pr}_Y[E\{(\gamma^N_{\text{recycle}}/\gamma)^2\} \geq \lambda \exp\{C/(\alpha-1)\}] \leq \lambda^{-1}$.

In Section 2 of the supplementary material we provide additional scaling analysis for $\text{var}(\gamma^N_{\text{recycle}})$ and $\text{var}(\gamma^N_{\text{simple}})$ in two alternative scenarios, where the potential functions exhibit, respectively, negative and positive correlation (see Propositions A1-A3).

4. EXAMPLE APPLICATIONS

4·1. *Adaptive pseudo-marginal random-walk Metropolis*

We consider Bayesian inference in two latent variable model applications, employing $\gamma^N_{\text{recycle}}$ or γ^N_{simple} to approximate $L(\theta)$ in a pseudo-marginal adaptive random-walk Metropolis Markov chain. The likelihood estimators are not simple averages and have a variance that decreases initially more rapidly than $\mathcal{O}(1/N)$, so the results of Sherlock et al. (2017) are not relevant and we follow Doucet et al. (2015) and Sherlock et al. (2015) and choose N such that the relative variance of the estimator is close to 2. While the relative variance typically varies with θ, if the posterior distribution for θ is reasonably concentrated near the true parameter θ_0, in practice one can often choose N so that the estimator has a

relative variance of around 2 at some point close to θ_0. Following Haario et al. (2001), the adaptation of the Markov chain involves tuning the covariance matrix of the proposal so that it is approximately $q(\theta, \theta') = \mathcal{N}(\theta'; \theta, d^{-1/2}2.38\Sigma)$, where Σ is the posterior covariance matrix.

Using γ^N_{perm} instead of $\gamma^N_{\mathrm{simple}}$ to approximate each $L(\theta)$ in a pseudo-marginal Markov chain can only decrease the asymptotic variance of ergodic averages of functions φ with $\mathrm{var}_\pi(\varphi) < \infty$. This is a consequence of Andrieu & Vihola (2016, Theorem 10) and Theorem 1. Using $\gamma^N_{\mathrm{recycle}}$ does not have the same guarantee in general, but Theorem 3 suggests that if the estimators perform similarly for a set of θ with large posterior mass, then this should result in greatly improved performance over $\gamma^N_{\mathrm{simple}}$ for large n. The results of the following simulation studies are in agreement with such theoretical considerations.

4·2. *Approximate Bayesian computation: g-and-k model*

Approximate Bayesian computation is a branch of simulation-based inference used when the likelihood function cannot be evaluated pointwise but one can simulate from the model for any value of the statistical parameter. While there are a number of variants, in general the methodology involves comparing a summary statistic associated with the observed data with summary statistics associated with pseudo-data simulated using different parameter values (see Marin et al., 2012, for a review). When the data are modelled as n observations of independent and identically distributed random variables with distribution μ, it is commonplace to summarize the data using some fixed-dimensional summary statistic independent of n, for computational rather than statistical reasons. This summarization, or dimension reduction, can in principle involve little loss of information about the parameters—in exponential families sufficient statistics of fixed dimension exist and could be computed or approximated—but in practice this is not always easy to achieve. An alternative approach that we adopt here is to eschew dimension reduction altogether and treat the model as a standard latent variable model using noisy approximate Bayesian computation (Fearnhead & Prangle, 2012). This may be viewed as an alternative to the construction of summaries using the Wasserstein distance recently proposed by Bernton et al. (2017). A possible outcome is that less data may be required to achieve a given degree of posterior concentration; a theoretical treatment of this is beyond the scope of this paper.

We consider the g-and-k distribution, which is a common example application for approximate Bayesian computation methods. The supplementary material details the simulation set-up, which follows Allingham et al. (2009). Here $n = 100$ and the likelihood $L(\theta)$ is equivalent to $\gamma(\theta) = \prod_{p=1}^n \mu_\theta(G_p)$, where G_p are potential functions and μ_θ follows a g-and-k model with parameter θ. In order to have a relative variance of $\gamma^N_{\mathrm{recycle}}(\theta_0)$ of roughly 2, it was sufficient to take $N = 80n = 8000$ whereas for $\gamma^N_{\mathrm{simple}}(\theta_0)$ we required $N = 80n^2 = 800000$. Using both estimators resulted in very similar Markov chains, but the computational cost of using the simple estimator was over 24 times greater; it took 18.1 hours to simulate a simple chain and 44 minutes to simulate a recycled chain of length 10^6. In the supplement, we plot posterior density estimates associated with the recycled chain and provide effective sample sizes for each component. Finally, we observe that the posterior distribution for θ places most of its mass near θ_0 despite using $n = 100$; in contrast Allingham et al. (2009) used $n = 10^5$ and our estimated posteriors show more concentration overall and better identification of the g parameter than their Figure 3. This suggests that this type of latent variable approach may be preferable to dimension-reducing summaries in some independent and identically distributed models.

4·3. *Poisson-Beta model for gene expression*

We now consider a model for single-cell gene expression levels originally proposed in Peccoud & Ycart (1995). Section 3·3 of the supplement contains a description of the model together with the simulation set-up. Collaborators working on an extension of this model and facing computational difficulties were a major motivation for this paper, whose methodology is now used in Tiberi et al. (2018). Here $n = 1000$ and the likelihood function values $L(\theta)$ are exactly of the form described in Remark 1, with Gaussian potential functions G_p and Poisson-Beta distribution μ_θ. In order to have a relative variance of $\gamma^N_{\mathrm{recycle}}(\theta_0)$ of roughly 2, it was sufficient to take $N = 40n = 4 \times 10^4$ whereas for $\gamma^N_{\mathrm{simple}}(\theta_0)$ we required $N = 40n^2 = 4 \times 10^7$. Using both estimators resulted in very similar Markov chains, but the computational

cost of using the simple estimator was approximately 440 times greater; it took 2.8 hours to simulated a recycled chain of length 10^6 and a simple chain of equal length was only feasible to simulate by using a parallel algorithm on a much more powerful 18-core processor, which still took 1.93 days. Using the same processor as the rest of the simulations would have taken approximately 52.5 days. The recycled estimator is particularly suitable here because the latent variables are discrete so that by avoiding recomputations, the number of potential function evaluations grows sub-linearly with N.

5. Discussion

We have demonstrated that the use of the recycled estimator proposed here successfully reduces computational time for Bayesian inference using pseudo-marginal Markov chain Monte Carlo from days or months to hours in some cases. Relating the results on numbers of samples required to common notions of asymptotic time complexity, however, requires some care. For a given relative variance in the setting of Theorem 3, one can choose α such that the following approximately holds. The number of samples required for the recycled estimator is αn and the number of function evaluations is slightly less than αn^2 while for the simple estimator we require αn^2 samples and αn^2 function evaluations. The computational time for the recycled estimator can be expressed as $\alpha n(c_s + nc_g + nc_r)$ where c_s is the cost of sampling from μ, c_g the cost of evaluating a potential function, and c_r is the problem-independent time per particle associated with sampling from (8). For the simple estimator, the computational time is $\alpha n^2(c_s + c_g)$ and so the recycled estimator is $(c_g + c_s)/(c_g + c_r)$ times faster than the simple estimator as $n \to \infty$, so that the improvement depends almost entirely on the relative differences between c_s, c_g and c_r. For sophisticated latent variable models, it is common for c_s to be orders of magnitude larger than c_g.

There are alternative unbiased approximations of the permanent of a rectangular matrix that could be used in place of the approach due to Kuznetsov (1996). In particular, it is straightforward to extend the algorithm of Kou & McCullagh (2009) to the rectangular case, or to use the Godsil–Gutman estimator (Godsil & Gutman, 1981; Friedland et al., 2004). We provide empirical comparisons of their variance in the supplementary material, which indicate very little improvement and potentially a much larger variance in the case of the Kou–McCullagh estimator; the most important issue in their use is that they require the $n \times N$ matrix of potential function evaluations to be constructed, which can exhaust memory for large applications. Another alternative unbiased approximation of γ is to average multiple versions of γ^N_{simple} by randomly reassigning the particles to different potential functions, but we show in Remark A2 of the supplement that this would scale poorly with n. It would be of interest to obtain accurate, lower bounds for the second moment of γ^N_{perm}, particularly in the setting of Example 1 to complement Theorem 3. This would determine whether more computationally expensive approximations of the permanent, such as Wang & Jasra (2016), are worth pursuing in this context. We have been able to show that in the setting of Proposition 4, $E\{(\gamma^N_{\text{perm}}/\gamma)^2\} \geq \prod_{p=1}^n (1 + c_p/N)$, so that significant improvement over $\gamma^N_{\text{recycle}}$ is impossible, but the argument did not extend naturally to the setting of Theorem 3.

One can define $\gamma^N_{\text{recycle}}$ alternatively by choosing a permutation σ of $\{1, \ldots, n\}$ according to any distribution and re-ordering the $G_1, \ldots, G_n$ as $G_{\sigma(1)}, \ldots, G_{\sigma(n)}$. The corresponding condition to (9), if the distribution for σ places mass on every possible permutation of $\{1, \ldots, n\}$, is then $\max_{p \in \{1,\ldots,n\},\, B \subseteq \{1,\ldots,n\}} \mu(G_p \prod_{j \in B} G_j) < \infty$.

One can define a recycled estimator of a product of n expectations, each with respect to a different distribution. Let $\mu_1, \ldots, \mu_n$ denote the distributions, take a common dominating probability distribution $\tilde{\mu}$ and define, for each $p \in \{1, \ldots, n\}$, $\tilde{G}_p = G_p \cdot \mathrm{d}\mu_p/\mathrm{d}\tilde{\mu}$ so that $\tilde{\mu}(\tilde{G}_p) = \mu_p(G_p)$. That is, one can re-express the product of expectations as a product of expectations all with respect to $\tilde{\mu}$. The recycled estimator could be useful when $\tilde{\mu}(\tilde{G}_p^2)/\tilde{\mu}(\tilde{G}_p)^2$ is not too large for any $p \in \{1, \ldots, n\}$.

Finally, it would be interesting to see if the use of the recycled estimator in the context of pseudo-marginal Markov chain Monte Carlo could be combined with the methodology of Deligiannidis et al. (2018) to bring further improvements.

Acknowledgments

Some of this research was undertaken while all three authors were at the University of Warwick. The authors acknowledge helpful comments from Christophe Andrieu, Christopher Jennison, Matti Vihola, a referee and an associate editor. AL was supported by The Alan Turing Institute and EPSRC. GZ was supported by the "New Directions in Bayesian NonParametrics" ERC grant.

Supplementary materials

Supplementary material available at *Biometrika* includes further analysis of the variances of $\gamma^N_{\mathrm{recycle}}$, $\gamma^N_{\mathrm{biased}}$ and γ^N_{perm}, additional information for the example applications, links to software to reproduce the simulations and proofs of the theoretical results.

BIBLIOGRAPHY

ALLINGHAM, D., KING, R. A. R. & MENGERSEN, K. L. (2009). Bayesian estimation of quantile distributions. *Stat. Comput.* **19**, 189–201.

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

ANDRIEU, C. & VIHOLA, M. (2016). Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.* **26**, 2661–2696.

BEAUMONT, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160.

BERNTON, E., JACOB, P. E., GERBER, M. & ROBERT, C. P. (2017). Inference in generative models using the Wasserstein distance. *arXiv preprint arXiv:1701.05146* .

DELIGIANNIDIS, G., DOUCET, A. & PITT, M. K. (2018). The correlated pseudo-marginal method. *J. R. Stat. Soc. B* To appear.

DOUCET, A., PITT, M. K., DELIGIANNIDIS, G. & KOHN, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102**, 295–313.

FEARNHEAD, P. & PRANGLE, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. B* **74**, 419–474.

FRIEDLAND, S., RIDER, B., ZEITOUNI, O. et al. (2004). Concentration of permanent estimators for certain large matrices. *Ann. Appl. Probab.* **14**, 1559–1576.

GODSIL, C. D. & GUTMAN, I. (1981). On the matching polynomial of a graph. In *Algebraic methods in graph theory*, L. Lovász & V. T. Sós, eds., vol. I. Amsterdam: North-Holland, pp. 241–249.

HAARIO, H., SAKSMAN, E., TAMMINEN, J. et al. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242.

KOU, S. C. & MCCULLAGH, P. (2009). Approximating the α-permanent. *Biometrika* **96**, 635–644.

KUZNETSOV, N. Y. (1996). Computing the permanent by importance sampling method. *Cybernet. Systems Anal.* **32**, 749–755.

MARIN, J.-M., PUDLO, P., ROBERT, C. P. & RYDER, R. J. (2012). Approximate Bayesian computational methods. *Stat. Comput.* **22**, 1167–1180.

PECCOUD, J. & YCART, B. (1995). Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234.

RYSER, H. J. (1963). *Combinatorial Mathematics*, vol. 14 of *Carus Mathematical Monographs*. Math. Assoc. America.

SHAKED, M. & SHANTHIKUMAR, J. G. (2007). *Stochastic Orders*. New York: Springer.

SHERLOCK, C., THIERY, A. H. & LEE, A. (2017). Pseudo-marginal Metropolis–Hastings sampling using averages of unbiased estimators. *Biometrika* **104**, 727–734.

SHERLOCK, C., THIERY, A. H., ROBERTS, G. O. & ROSENTHAL, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43**, 238–275.

TIBERI, S., WALSH, M., CAVALLARO, M., HEBENSTREIT, D. & FINKENSTÄDT, B. (2018). Bayesian inference on stochastic gene transcription from flow cytometry data. *Bioinformatics (In Press)* .

VALIANT, L. G. (1979). The complexity of computing the permanent. *Theoret. Comput. Sci.* **8**, 189–201.

WANG, J. & JASRA, A. (2016). Monte Carlo algorithms for computing α-permanents. *Stat. Comput.* **26**, 231–248.