



Title	Exploiting Multi-Core Architectures for Reduced-Variance Estimation with Intractable Likelihoods
Authors(s)	Friel, Nial, Mira, Antonietta, Oates, Chris J.
Publication date	2015
Publication information	Friel, Nial, Antonietta Mira, and Chris J. Oates. "Exploiting Multi-Core Architectures for Reduced-Variance Estimation with Intractable Likelihoods" 11, no. 1 (2015).
Publisher	International Society for Bayesian Analysis (ISBA)
Item record/more information	http://hdl.handle.net/10197/6729
Publisher's version (DOI)	10.1214/15-BA948

Downloaded 2023-10-05T14:16:07Z

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)



© Some rights reserved. For more information

Exploiting Multi-Core Architectures for Reduced-Variance Estimation with Intractable Likelihoods

Nial Friel^{*}, Antonietta Mira[†], and Chris. J. Oates[‡]

Abstract. Many popular statistical models for complex phenomena are intractable, in the sense that the likelihood function cannot easily be evaluated. Bayesian estimation in this setting remains challenging, with a lack of computational methodology to fully exploit modern processing capabilities. In this paper we introduce novel control variates for intractable likelihoods that can dramatically reduce the Monte Carlo variance of Bayesian estimators. We prove that our control variates are well-defined and provide a positive variance reduction. Furthermore, we show how to optimise these control variates for variance reduction. The methodology is highly parallel and offers a route to exploit multi-core processing architectures that complements recent research in this direction. Indeed, our work shows that it may not be necessary to parallelise the sampling process itself in order to harness the potential of massively multi-core architectures. Simulation results presented on the Ising model, exponential random graph models and non-linear stochastic differential equation models support our theoretical findings.

Keywords: control variates, MCMC, parallel computing, zero variance.

1 Introduction

Many models of interest are intractable, by which it is understood that the likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$ that describes how data \mathbf{y} arise from a model parametrised by $\boldsymbol{\theta} \in \Theta$ is unavailable in closed form. The predominant sources of intractability that are encountered in statistical modelling can be classified as follows:

- Type I The need to compute a normalising constant $\mathfrak{Z}(\boldsymbol{\theta}) = \int f(\mathbf{y}'; \boldsymbol{\theta}) d\mathbf{y}'$ that depends on parameters $\boldsymbol{\theta}$ such that $p(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}; \boldsymbol{\theta})/\mathfrak{Z}(\boldsymbol{\theta})$.
- Type II The need to marginalise over a set of latent variables \mathbf{x} such that $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x}$.

Bayesian estimation in both of these settings can be extremely challenging as many established computational techniques (e.g. Gibbs sampling and Metropolis–Hastings) are incompatible with intractable likelihoods. This has motivated researchers to propose several approximations to the likelihood function that are tractable (e.g. Marjoram et

^{*}School of Mathematical Sciences and Insight: The National Centre for Data Analytics, University College Dublin, Ireland, nial.friel@udc.ie

[†]Interdisciplinary Institute of Data Science and Institute of Finance, University of Lugano, Switzerland, antonietta.mira@usi.ch

[‡]Department of Statistics, University of Warwick, UK, c.oates@warwick.ac.uk

al., 2003; Møller et al., 2006; Murray et al., 2006; Rue et al., 2009). In the other direction, several (exact) Markov chain Monte Carlo (MCMC) algorithms have been proposed that facilitate inference in intractable models (e.g. Beskos et al., 2006; Andrieu and Roberts, 2009; Andrieu et al., 2010; Lyne et al., 2013). However, for MCMC methodology, it remains the case that estimator variance can be heavily inflated relative to the tractable case, due to the need to perform auxiliary calculations on extended state spaces in order to address the intractability (Sherlock et al., 2014). Below we elaborate on the two types of intractability and on the related references in the literature that have addressed them.

Type I Intractability arises from the need to compute a parameter-dependent normalising constant (sometimes called a partition function). This paper focuses on the sub-class of Type I intractable models known as Gibbs random fields (GRFs) where data \mathbf{y} arises from a model of the form

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{s}(\mathbf{y}) - \log \mathfrak{P}(\boldsymbol{\theta}) \quad (1)$$

such that the partition function

$$\mathfrak{P}(\boldsymbol{\theta}) = \int \exp(\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})) d\mathbf{y} \quad (2)$$

is intractable. In a Bayesian context, this leads to a “doubly intractable” distribution and is the subject of current research in the statistical community. Below we survey applications of, and methodology for, models exhibiting this form of intractability:

Example 1 (Spatial statistics). *A GRF-type intractability arises in classical spatial statistics where we seek to model the joint distribution of variables Y_j that are subject to local interactions. The autologistic distribution (Besag, 1996) is a well-studied model for the analysis of binary spatial data defined on a lattice. This model has been applied in diverse contexts including ecology (Augustin et al., 1996), the spatial analysis of plant species (Huffer and Wu, 1998; He et al., 2009) and dentistry (Bandyopadhyay et al., 2009). The canonical Ising model is a special case of the autologistic distribution and is defined on a regular lattice of size $n \times n$, where j is used to index each of the $n \times n$ different lattice locations. Here the random variable $\mathbf{Y} \in \{-1, 1\}^{n \times n}$ has a probability distribution defined in terms of a single sufficient statistic*

$$s(\mathbf{y}) = \sum_{j=1}^{n \times n} \sum_{i \sim j} y_i y_j,$$

where the notation $i \sim j$ means that the lattice point i is a neighbour of lattice point j . Interactions are modelled between neighbouring lattice points $i \sim j$, being captured by the energy term $y_i y_j$. The likelihood for this model takes the form of a GRF where the partition function

$$\mathfrak{P}(\boldsymbol{\theta}) = \sum_{\mathbf{y}' \in \{-1, 1\}^{n \times n}} \exp(\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y}')) \quad (3)$$

involves the summation over $2^{n \times n}$ different possible state vectors \mathbf{y}' . Typically, this summation is infeasible and leads to Type I intractability for all but small values of the lattice size n .

Example 2 (Social network analysis). *Exponential random graph (ERG) models are widely used in social network analysis (see Robins et al., 2014, and the references therein). The ERG model is defined on a random adjacency matrix $\mathbf{Y} = \{Y_{ij} : i = 1, \dots, n; j = 1, \dots, n\}$ of a graph with n nodes where $Y_{ij} = 1$ if nodes i and j are connected by an edge, and $Y_{ij} = 0$ otherwise. An edge connecting a node to itself is not permitted so $Y_{ii} = 0$. The edges in an ERG may be undirected, whereby $Y_{ij} = Y_{ji}$, or directed, whereby a directed edge from node i to node j is not necessarily reciprocated. Write $\mathcal{G}(n)$ for the set of all permitted graphs on n vertices. The likelihood of an observed graph \mathbf{y} is modelled in terms of a collection of sufficient statistics $\mathbf{s}(\mathbf{y}) = (s_1(\mathbf{y}), \dots, s_k(\mathbf{y}))$ and corresponding parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. For example, typical statistics include $s_1(\mathbf{y}) = \sum_{i < j} y_{ij}$ and $s_2(\mathbf{y}) = \sum_{i < j < k} y_{ik}y_{jk}$ that encode, respectively, the observed number of edges and two-stars, that is, the number of configurations of pairs of edges that share a common node. It is also possible to consider statistics that count the number of configuration of k edges that share a node in common, for $k > 2$. The likelihood takes the form of a GRF where the partition function*

$$\mathfrak{P}(\boldsymbol{\theta}) = \sum_{\mathbf{y}' \in \mathcal{G}(n)} \exp(\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y}')) \quad (4)$$

involves the summation over $|\mathcal{G}(n)| = O(2^{n \times n})$ possible different graphs and leads to Type I intractability for all but small values of the number n of vertices.

The dependence of the partition function $\mathfrak{P}(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ leads to difficulties in inferring this parameter. An early attempt to circumvent this difficulty is the pseudo-likelihood approach of Besag (1972), which in turn has been generalised to composite likelihood approximations, see for example Davison et al. (2012). An alternative class of inferential approaches results from realising that, although one cannot evaluate the likelihood function, it is possible to sample pseudo-data from the generative model, so-called “forward simulation”. The Monte Carlo MLE approach of Geyer and Thompson (1992) exploits forward simulation to allow maximum likelihood estimation. From a Bayesian perspective, simulating from the likelihood has also played an influential role in several approaches, for example, the auxiliary variable method of Møller et al. (2006), that was subsequently extended by Murray et al. (2006) to the exchange algorithm. The exchange algorithm avoids the need to directly evaluate the partition function by considering an augmented target distribution $p(\boldsymbol{\theta}, \boldsymbol{\theta}', \mathbf{y}' | \mathbf{y})$ that includes a second copy $\boldsymbol{\theta}'$ of the parameter vector and forward-simulated pseudo-data \mathbf{y}' drawn from the likelihood function $p(\mathbf{y}' | \boldsymbol{\theta}')$, defined in such a way that the Markov chain transition kernel for the parameter vector $\boldsymbol{\theta}$ of interest involves partition functions for the current and proposed values of $\boldsymbol{\theta}$ that cancel in the numerator and denominator of the Metropolis–Hastings ratio, thus circumventing the Type I intractability issue (see Algorithm 1) at the expense of increased Monte Carlo variance.

An emerging research direction is the construction of approximate Monte Carlo algorithms, providing convergence guarantees, in situations where it is expensive or impossible to calculate the likelihood. This is particularly pressing in cases where the exchange algorithm is applied, since forward simulating from Gibbs random fields is challenging. Perfect sampling is often prohibitively expensive or impossible to carry

Algorithm 1 Exchange algorithm for Type I intractability (Murray et al., 2006)

- 1: Initialise $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(0)'}$, $\mathbf{y}^{(0)'}$.
- 2: **for** $i = 1, \dots, I$ **do**
- 3: Obtain $\boldsymbol{\theta}' \sim h(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i-1)})$.
- 4: Obtain $\mathbf{y}' \sim p(\mathbf{y}'|\boldsymbol{\theta}')$.
- 5: Exchange $\boldsymbol{\theta}' \mapsto \boldsymbol{\theta}^{(i)}$, $\boldsymbol{\theta}^{(i-1)} \mapsto \boldsymbol{\theta}^{(i)'}$, $\mathbf{y}' \mapsto \mathbf{y}^{(i)'}$ with probability

$$\alpha = \min \left\{ 1, \frac{p(\mathbf{y}|\boldsymbol{\theta}') p(\mathbf{y}'|\boldsymbol{\theta}^{(i-1)}) p(\boldsymbol{\theta}') h(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}')}{p(\mathbf{y}'|\boldsymbol{\theta}') p(\mathbf{y}|\boldsymbol{\theta}^{(i-1)}) p(\boldsymbol{\theta}) h(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(i-1)})} \right\}, \quad (5)$$

otherwise set $\boldsymbol{\theta}^{(i-1)} \mapsto \boldsymbol{\theta}^{(i)}$, $\boldsymbol{\theta}^{(i-1)'}$ $\mapsto \boldsymbol{\theta}^{(i)'}$, $\mathbf{y}^{(i-1)'}$ $\mapsto \mathbf{y}^{(i)'}$.

6: **end for**

out, and in this case Everitt (2012) has provided convergence results for the case where one uses the final draw from a Gibbs sampler targeting the likelihood as an approximate realisation. In a similar vein, Alquier et al. (2014) and Pillai and Smith (2014) develop convergence results for approximate MCMC algorithms resulting from approximating the transition kernel due to the intractability of the likelihood function. It is worth noting that several authors have used this type of approach to develop approximate algorithms for large datasets by using subsets of the data to approximate the likelihood, (Welling and Teh, 2011; Anh et al., 2012; Korattikara et al., 2014).

Type II Intractability arises from the need to marginalise over latent variables \mathbf{x} such that the marginal likelihood

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x} \quad (6)$$

is unavailable in closed form. Such problems arise frequently in applied statistics and examples include inference for the parameters of spatio-temporal models (Rue et al., 2009; Lyne et al., 2013), regression models with random effects (Fahrmeir and Lang, 2001), time-series models (West and Harrison, 1997), and selection between competing models based on Bayes factors (e.g. Caimo and Friel, 2013; Armond et al., 2014). Below we provide examples of, and survey methodology for, models exhibiting this form of intractability:

Example 3 (Hidden Markov model). *Applications of hidden Markov models abound in many areas, including finance, economics and biology. See Cappé et al. (2005) for a detailed analysis of this general area. In a hidden Markov model, the parameters $\boldsymbol{\theta}$ that specify a Markov chain*

$$\mathbf{x}_{n+1} \sim p(\mathbf{x}_{n+1}|\mathbf{x}_n, \boldsymbol{\theta}) \quad (7)$$

may be of interest, whilst the latent sample path $\{\mathbf{x}_n\}_{n=0}^N$ of the Markov chain that gives rise to observations $\mathbf{y}_n \sim p(\mathbf{y}_n|\mathbf{x}_n)$ may not be of interest and must be marginalised.

Even in discrete cases where $\mathbf{x}_n \in \mathcal{X}$ for a finite state space \mathcal{X} , the number of possible samples paths $\{\mathbf{x}_n\}_{n=0}^N$ grows exponentially in N and this renders the marginalisation

$$p(\{\mathbf{y}_n\}_{n=0}^N | \boldsymbol{\theta}) = \sum_{\mathbf{x}_0, \dots, \mathbf{x}_N \in \mathcal{X}} p(\{\mathbf{y}_n\}_{n=0}^N | \{\mathbf{x}_n\}_{n=0}^N, \boldsymbol{\theta}) p(\{\mathbf{x}_n\}_{n=0}^N | \boldsymbol{\theta}) \quad (8)$$

corresponding to (6) computationally intractable.

Example 4 (Stochastic differential equations). *Stochastic differential equations (SDEs) are widely used in several fields including biology (Wilkinson, 2011) and finance (Lamberton and Lapeyre, 2007). See Øksendal (2003) for an excellent introduction to SDEs, including a focus on several application areas. A general stochastic diffusion is defined as*

$$d\mathbf{X}(t) = \boldsymbol{\alpha}(\mathbf{X}(t); \boldsymbol{\theta})dt + \boldsymbol{\beta}^{1/2}(\mathbf{X}(t); \boldsymbol{\theta})d\mathbf{W}(t), \quad \mathbf{X}(0) = \mathbf{X}_0, \quad (9)$$

where $\mathbf{X}(t)$ is a stochastic process taking values in \mathbb{R}^d , $\boldsymbol{\alpha} : \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^d$ is a drift function, $\boldsymbol{\beta} : \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^d \times \mathbb{R}^d$ is a diffusion function, $\mathbf{W}(t)$ is a d -dimensional Wiener process, $\boldsymbol{\theta} \in \Theta$ are unknown model parameters, and $\mathbf{X}_0 \in \mathbb{R}^d$ is an initial state (assumed known here). For general SDEs, an analytic form for the distribution of sample paths is unavailable. An excellent review of approximate likelihood methods for SDEs is provided in Fuchs (2013). To facilitate inference here, a popular approach is to introduce a fine discretisation t_1, \dots, t_T of time with mesh size δt . Write $\mathbf{X}_i = \mathbf{X}(t_i)$. The Euler–Maruyama approximation to the SDE likelihood is then given by

$$p(\mathbf{X} | \boldsymbol{\theta}) \propto \prod_{i=2}^T \psi(\mathbf{X}_i | \mathbf{X}_{i-1} + \boldsymbol{\alpha}_i \delta t, \boldsymbol{\beta}_i \delta t) \quad (10)$$

where $\psi(\cdot | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the probability density function for a Gaussian random variable with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ and where we have used the shorthand $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}(\mathbf{X}_{i-1}; \boldsymbol{\theta})$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}(\mathbf{X}_{i-1}; \boldsymbol{\theta})$. We partition $\mathbf{X} = [\mathbf{X}^o \ \mathbf{X}^u]$ such that $\mathbf{y} = \mathbf{X}^o$ are observed (for simplicity here without noise) and $\mathbf{x} = \mathbf{X}^u$ are unobserved. This is essentially a hidden Markov model with a continuous latent state and therefore exhibits Type II intractability, since to draw inferences on $\boldsymbol{\theta}$ it is required to marginalise the unobserved variables \mathbf{X}^u .

A popular contemporary approach to inference under Type II intractability is the pseudo-marginal MCMC of Andrieu and Roberts (2009) that replaces the marginal likelihood $p(\mathbf{y} | \boldsymbol{\theta})$ in the Metropolis–Hastings acceptance ratio with an unbiased estimate that can either be obtained by forward-simulation from $p(\mathbf{x} | \boldsymbol{\theta})$, or using importance sampling techniques. The pseudo-marginal MCMC typically leads to reduced efficiency relative to the (unavailable) marginal algorithm, but improved efficiency relative to a Markov chain constructed on the extended space $(\boldsymbol{\theta}, \mathbf{x})$ (Sherlock et al., 2014). When combined with particle MCMC (Andrieu et al., 2010), the pseudo-marginal algorithm represents a popular technique to deal with general forms of Type II intractability. Within specific model classes it may be possible to design specialised approaches to estimation; for example, Kou et al. (2012) and Beskos et al. (2013) both present sophisticated schemes for parameter inference in discretely observed stochastic differential

equation models. Other attempts to address Type II intractability include the popular approximation scheme of Rue et al. (2009) and the references therein. Such schemes trade exactness of computation for substantial reduction in computational effort, but many questions surround the extent of approximation error (e.g. Lindgren et al., 2011).

In summary, applications involving statistical models with both types of intractability are widespread in the literature. Moreover, as detailed above, statistical methodology to overcome both types of intractability is at the frontier of research in computational statistics. Indeed, one might anticipate that even wider applicability will result as these methods disseminate in the scientific community, whereby hitherto intractable statistical models will be amenable to statistical inference.

Outline of the paper The present contribution addresses the problem of estimating posterior expectations via MCMC when data arise from an intractable likelihood:

Problem. *Estimate the posterior expectation $\mu = \mathbb{E}_{\theta|\mathbf{y}}[g(\theta)]$ for some known function $g : \Theta \rightarrow \mathbb{R}$, where data \mathbf{y} arise from an intractable likelihood of either Type I or II.*

Our focus is on the use of control variates for the reduction of Monte Carlo variance (Glasserman, 2004). The basic idea behind control variate schemes in Bayesian computation is that a modified function $\tilde{g}(\theta) = g(\theta) + \phi_1 h_1(\theta) + \dots + \phi_m h_m(\theta)$ is constructed such that $\tilde{g}(\theta)$ has the same posterior expectation but a reduced posterior variance compared to $g(\theta)$. This can occur when (i) each of the $h_i(\theta)$ have zero posterior expectation, (ii) the collection $[h_1(\theta), \dots, h_m(\theta)]$ has strong posterior canonical correlation with the target $g(\theta)$, and (iii) the coefficients ϕ_1, \dots, ϕ_m are chosen appropriately. Recently Mira et al. (2013) proposed to use the score vector $\mathbf{u}(\theta|\mathbf{y}) := \nabla_{\theta} \log p(\theta|\mathbf{y})$ as the basis for a set of control variates, since this can be guaranteed to have zero expectation under mild boundary conditions (described below). There it was shown that these score-based control variates can significantly reduce Monte Carlo variance, sometimes dramatically. Indeed, the methodology was named “zero variance” (ZV) by Mira et al. (2013), following Assaraf and Caffarel (1999), since in several special cases the score has perfect canonical correlation with the target, generating an estimate that has zero sampling variance. Further support for the use of the score as a control variate was provided in Papamarkou et al. (2014); Oates et al. (2015), who demonstrated that the approach fits naturally within Hamiltonian-type and Langevin-type MCMC schemes that themselves make use of the score, requiring essentially no additional computational effort. It would therefore be extremely desirable to design control variates for intractable likelihoods, where sampling variance is acutely problematic. However, for Bayesian inference with intractable likelihoods, the score is unavailable as it requires the derivative of unknown quantities. Our work is motivated by overcoming this impasse.

The main contribution of this paper is to introduce a stochastic approximation to ZV control variates, called “reduced-variance” (RV) control variates, that can be computed for intractable likelihoods of both Type I and II. Specifically, we study the effect of replacing the true score function $\mathbf{u}(\theta|\mathbf{y})$ for the intractable models in the ZV methodology with an unbiased estimate $\hat{\mathbf{u}}(\theta|\mathbf{y})$ that can be obtained via repeated forward-simulation. Importantly, these forward-simulations can be performed in parallel,

offering the opportunity to exploit modern multi-core processing architectures (Suchard et al., 2010; Lee et al., 2010) in a straightforward manner that directly complements (and is compatible with) related research efforts for parallelisation of MCMC methodology (Alquier et al., 2014; Angelino et al., 2014; Bardenet et al., 2014; Calderhead, 2014; Korattikara et al., 2014; Maclaurin and Adams, 2014).

From a theoretical perspective, we prove that RV control variates are well-defined and provide a positive variance reduction. Furthermore, we propose default tuning parameters that are proven to maximise variance reduction and prove that the optimal estimator for serial computation requires essentially the same computational effort as the state-of-the-art estimate obtained under either the exchange algorithm or the pseudo-marginal algorithm. These results are orthogonal to recent work by Doucet et al. (2012) and Sherlock et al. (2014) that deals with implementation of MCMC samplers themselves. Empirical results presented on the Ising model, exponential random graphs and nonlinear stochastic differential equations support our theoretical findings.

2 Methods

2.1 Control variates and intractable likelihoods

Our presentation of control variate methodology below focuses on the problem of evaluating posterior expectations, but the methodology itself applies more broadly. In this restricted setting, control variates can be employed when the aim is to estimate, with high precision, the posterior expectation $\mu = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[g(\boldsymbol{\theta})]$ of a (real-valued) function $g(\boldsymbol{\theta})$ of an unknown parameter $\boldsymbol{\theta}$. In this paper, we focus on a real-valued random parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$. The generic control variate principle relies on constructing an auxiliary function $\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + h(\boldsymbol{\theta})$ where $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[h(\boldsymbol{\theta})] = 0$ and so $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[\tilde{g}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[g(\boldsymbol{\theta})]$. In many cases it is possible to choose $h(\boldsymbol{\theta})$ such that the variance $\mathbb{V}_{\boldsymbol{\theta}|\mathbf{y}}[\tilde{g}(\boldsymbol{\theta})] < \mathbb{V}_{\boldsymbol{\theta}|\mathbf{y}}[g(\boldsymbol{\theta})]$, leading to a Monte Carlo estimator with strictly smaller variance:

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^n \tilde{g}(\boldsymbol{\theta}^{(i)}), \quad (11)$$

where $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(n)}$ are independent samples from $p(\boldsymbol{\theta}|\mathbf{y})$. Intuitively, greater variance reduction can occur when $h(\boldsymbol{\theta})$ is negatively correlated with $g(\boldsymbol{\theta})$ in the posterior, since much of the randomness ‘‘cancels out’’ in the auxiliary function $\tilde{g}(\boldsymbol{\theta})$.

In classical literature, the function $h(\boldsymbol{\theta})$ is often formed as a sum $\phi_1 h_1(\boldsymbol{\theta}) + \dots + \phi_m h_m(\boldsymbol{\theta})$ where the $h_i(\boldsymbol{\theta})$ each have zero posterior expectation (under the target) and are known as control variates, whilst ϕ_i are coefficients that must be specified (Glasserman, 2004). Alternative constructions also exist (e.g. ratio control variates; Evans and Swartz, 2000) but here we focus only on control variates with an additive structure. For estimation based on Markov chains, Andradóttir et al. (1993) proposed control variates for discrete state spaces. Later Mira et al. (2003) extended the approach of Assaraf and Caffarel (1999) observing that the optimal choice of $h(\boldsymbol{\theta})$ is intimately associated

with the solution of the Poisson equation $h(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[g(\boldsymbol{\theta})] - g(\boldsymbol{\theta})$ and proposing to solve this equation numerically. Further work to construct control variates for Markov chains includes Hammer and Tjelmeland (2008) for Metropolis–Hastings samplers and Dellaportas and Kontoyiannis (2012) for Gibbs samplers.

In this paper, we consider the particularly elegant class of control variates that are expressed as functions of the score vector $\mathbf{u}(\boldsymbol{\theta}|\mathbf{y})$ of the log-posterior density. Mira et al. (2013) proposed the ZV control variates

$$h(\boldsymbol{\theta}|\mathbf{y}) = \Delta_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] + \nabla_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] \cdot \mathbf{u}(\boldsymbol{\theta}|\mathbf{y}) \quad (12)$$

where $\nabla_{\boldsymbol{\theta}} = [\partial/\partial\theta_1, \dots, \partial/\partial\theta_d]^T$ is the gradient operator, $\Delta_{\boldsymbol{\theta}} = (\partial^2/\partial\theta_1^2 + \dots + \partial^2/\partial\theta_d^2)$ is the Laplacian operator, and the “trial function” $P(\boldsymbol{\theta})$ belongs to the family \mathcal{P} of polynomials in $\boldsymbol{\theta}$. In this paper, we adopt the convention that both $\boldsymbol{\theta}$ and $\mathbf{u}(\boldsymbol{\theta}|\mathbf{y})$ are $d \times 1$ vectors. Mira et al. (2013) showed, in particular, that any posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ approximating a Gaussian forms a suitable candidate for implementing the ZV scheme. The ZV approach has recently been extended to encompass non-parametric trial functions $P(\boldsymbol{\theta})$. Oates et al. (2014) proves that the associated estimators possess superior convergence rates relative to estimation that does not use control variates. A consequence of this latter approach is that large variance reductions can be achieved outside of the Gaussian setting. For a comprehensive review of the ZV methodology see Papamarkou et al. (2015). Unfortunately, ZV methods are not directly compatible with intractable likelihoods:

Type I A naive application of ZV methods to GRFs with Type I intractability would require the score function that is obtained by differentiating

$$\log p(\boldsymbol{\theta}|\mathbf{y}) = \boldsymbol{\theta}^T \mathbf{s}(\mathbf{y}) - \log \mathfrak{P}(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + C, \quad (13)$$

where C is a constant in $\boldsymbol{\theta}$, to obtain

$$\mathbf{u}(\boldsymbol{\theta}|\mathbf{y}) = \mathbf{s}(\mathbf{y}) - \nabla_{\boldsymbol{\theta}} \log \mathfrak{P}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}). \quad (14)$$

It is clear that (14) will not have a closed form when the partition function $\mathfrak{P}(\boldsymbol{\theta})$ is intractable. In the sections below, we demonstrate how forward-simulation can be used to approximate $\nabla_{\boldsymbol{\theta}} \log \mathfrak{P}(\boldsymbol{\theta})$ and then leverage this fact to reduce Monte Carlo variance.

Type II Similarly, a naive application of ZV within Type II intractable likelihood problems would require the evaluation of the score function

$$\mathbf{u}(\boldsymbol{\theta}|\mathbf{y}) = \nabla_{\boldsymbol{\theta}} \log \int p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x} + \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}). \quad (15)$$

It is clear that (15) will not have a closed form when the integral over the latent variable \mathbf{x} is intractable. In the sections below, we demonstrate how forward-simulation can be used to approximate $\nabla_{\boldsymbol{\theta}} \log \int p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x}$, before again leveraging this fact to reduce Monte Carlo variance.

2.2 Unbiased estimation of the score

Our approach relies on the ability to construct an unbiased estimator for the score function in both Type I and Type II intractable models.

Type I An unbiased estimator for $\mathbf{u}(\boldsymbol{\theta}|\mathbf{y})$ that can be computed for Type I models of GRF form is constructed by noting that

$$\nabla_{\boldsymbol{\theta}} \log \mathfrak{P}(\boldsymbol{\theta}) = \frac{1}{\mathfrak{P}(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \mathfrak{P}(\boldsymbol{\theta}) \quad (16)$$

$$= \frac{1}{\mathfrak{P}(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \int \exp(\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})) d\mathbf{y} \quad (17)$$

$$= \frac{1}{\mathfrak{P}(\boldsymbol{\theta})} \int \mathbf{s}(\mathbf{y}) \exp(\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})) d\mathbf{y} \quad (18)$$

$$= \mathbb{E}_{\mathbf{Y}|\boldsymbol{\theta}}[\mathbf{s}(\mathbf{Y})], \quad (19)$$

where we have assumed regularity conditions that permit the interchange of derivative and integral operators (including that the domain of \mathbf{Y} does not depend on $\boldsymbol{\theta}$). Specifically, combining (14) and (19) we estimate the score function by exploiting multiple forward-simulations

$$\hat{\mathbf{u}}(\boldsymbol{\theta}|\mathbf{y}) := \mathbf{s}(\mathbf{y}) - \left[\frac{1}{K} \sum_{k=1}^K \mathbf{s}(\mathbf{Y}_k) \right] + \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) \quad (20)$$

where the $\mathbf{Y}_1, \dots, \mathbf{Y}_K$ are independent simulations from the GRF with density $p(\mathbf{y}|\boldsymbol{\theta})$. Forward-simulation for GRF has previously been leveraged to facilitate estimation (e.g. Potamianos and Goutsias, 1997) and can be achieved using, for example, perfect sampling (Propp and Wilson, 1996; Mira et al., 2001). We make two important observations: Firstly, one realisation \mathbf{Y}_1 must be drawn in any case to perform the exchange algorithm, so that this requires no additional computation. Secondly, these K simulations can be performed in parallel, enabling the exploitation of multi-core processing architectures.

Type II For intractable models of Type II an alternative approach to construct an unbiased estimate for the score is required. Specifically, we notice that the score $\mathbf{u}(\boldsymbol{\theta}, \mathbf{x}) := \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})$ of the extended posterior is typically available in closed form and this can be leveraged as follows:

$$\mathbf{u}(\boldsymbol{\theta}|\mathbf{y}) = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{y}) = \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y})}{p(\boldsymbol{\theta}|\mathbf{y})} \quad (21)$$

$$= \frac{1}{p(\boldsymbol{\theta}|\mathbf{y})} \nabla_{\boldsymbol{\theta}} \int p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) d\mathbf{x} \quad (22)$$

$$= \int \frac{[\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})] p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})}{p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y})} d\mathbf{x} \quad (23)$$

$$= \int [\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})] p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) d\mathbf{x} = \mathbb{E}_{\mathbf{X}|\boldsymbol{\theta}, \mathbf{y}}[\mathbf{u}(\boldsymbol{\theta}, \mathbf{X})] \quad (24)$$

where again we have assumed regularity conditions that allow us to interchange the integral and the derivative operators. We therefore have a simulation-based estimator

$$\hat{\mathbf{u}}(\boldsymbol{\theta}|\mathbf{y}) := \frac{1}{K} \sum_{k=1}^K \mathbf{u}(\boldsymbol{\theta}, \mathbf{X}_k) \quad (25)$$

where the $\mathbf{X}_1, \dots, \mathbf{X}_K$ are independent simulations from the posterior conditional $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$. ((24) is sometimes called ‘‘Fisher’s identity’’; Nemeth et al., 2014). We note that it is straightforward to implement pseudo-marginal MCMC in such a way that samples \mathbf{X}_i are obtained as a by-product, so that estimation of the score requires no additional computation.

2.3 Reduced-variance control variates

This paper advocates constructing control variates using an unbiased estimator for the score as follows:

$$\hat{h}(\boldsymbol{\theta}|\mathbf{y}) := \Delta_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] + \nabla_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] \cdot \hat{\mathbf{u}}(\boldsymbol{\theta}|\mathbf{y}), \quad (26)$$

where again $P \in \mathcal{P}$ is a polynomial trial function. The coefficients $\boldsymbol{\phi}$ of this polynomial $P(\boldsymbol{\theta})$ must be specified, and we will also write $P(\boldsymbol{\theta}|\boldsymbol{\phi})$ to emphasise this point. These will be referred to as ‘‘reduced-variance’’ control variates from the fact that (26) is a stochastic approximation to the ZV control variates in (12) and can therefore be expected to have similar properties. Pseudocode is provided in Algorithm 2.

For this idea to work, it must be the case that the RV control variates $\hat{h}(\boldsymbol{\theta}|\mathbf{y})$ have zero expectation. This is guaranteed under mild assumptions stated below:

Lemma 1. *Assume that Θ is possibly unbounded, B_r are bounded sets increasing to Θ and $\lim_{r \rightarrow \infty} \int_{\partial B_r} p(\boldsymbol{\theta}|\mathbf{y}) \nabla P(\boldsymbol{\theta}) \cdot \mathbf{n}(\boldsymbol{\theta}) d\boldsymbol{\theta} = 0$, where $\mathbf{n}(\boldsymbol{\theta})$ is the outward pointing unit normal field of the boundary ∂B_r . Then, for Type I models, $\mathbb{E}_{\boldsymbol{\theta}, \mathbf{Y}_1, \dots, \mathbf{Y}_K | \mathbf{y}}[\hat{h}(\boldsymbol{\theta}|\mathbf{y})] = 0$, whilst, for Type II models, $\mathbb{E}_{\boldsymbol{\theta}, \mathbf{X}_1, \dots, \mathbf{X}_K | \mathbf{y}}[\hat{h}(\boldsymbol{\theta}|\mathbf{y})] = 0$, so that in both cases $\hat{h}(\boldsymbol{\theta}|\mathbf{y})$ is a well-defined control variate.*

Proof. From unbiasedness of $\hat{\mathbf{u}}(\boldsymbol{\theta}|\mathbf{y})$ we have, for Type I models,

$$\mathbb{E}_{\boldsymbol{\theta}, \mathbf{Y}_1, \dots, \mathbf{Y}_K | \mathbf{y}}[\hat{h}(\boldsymbol{\theta}|\mathbf{y})] = \mathbb{E}_{\boldsymbol{\theta} | \mathbf{y}} [\mathbb{E}_{\mathbf{Y}_1, \dots, \mathbf{Y}_K | \boldsymbol{\theta}} [\Delta_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) \cdot \hat{\mathbf{u}}(\boldsymbol{\theta}|\mathbf{y})]] \quad (31)$$

$$= \mathbb{E}_{\boldsymbol{\theta} | \mathbf{y}} [\Delta_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) \cdot \mathbf{u}(\boldsymbol{\theta}|\mathbf{y})], \quad (32)$$

with the analogous result holding for Type II models. The remainder follows from Mira et al. (2013): Using the definition of the score $\mathbf{u}(\boldsymbol{\theta}|\mathbf{y})$, we have

$$= \int_{\Theta} [\Delta_{\boldsymbol{\theta}} P(\boldsymbol{\theta})] p(\boldsymbol{\theta}|\mathbf{y}) + [\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})] \cdot [\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y})] d\boldsymbol{\theta}. \quad (33)$$

Then applying the divergence theorem (see, e.g. Kendall and Bourne, 1992) we obtain

Algorithm 2 Reduced-variance estimation for intractable likelihoods

-
- 1: Obtain $\boldsymbol{\theta}^{(i)} \sim \boldsymbol{\theta}|\mathbf{y}$, $i = 1, \dots, I$ ▷ using MCMC
 - 2: **for** $i = 1, \dots, I$ **do**
 - 3: **if** Type I **then**
 - 4: Obtain $\mathbf{y}^{(i,k)} \sim \mathbf{Y}|\boldsymbol{\theta}^{(i)}$, $k = 1, \dots, K$ ▷ simulate from the model
 - 5: Construct an approximation to the score at $\boldsymbol{\theta}^{(i)}$:

$$\hat{\mathbf{u}}^{(i)} = \mathbf{s}(\mathbf{y}) - \left[\frac{1}{K} \sum_{k=1}^K \mathbf{s}(\mathbf{y}^{(i,k)}) \right] + \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}^{(i)}) \quad (27)$$

- 6: **else if** Type II **then**
- 7: Obtain $\mathbf{x}^{(i,k)} \sim \mathbf{x}|\boldsymbol{\theta}^{(i)}, \mathbf{y}$, $k = 1, \dots, K$ ▷ simulate from the posterior
- 8: Construct an approximation to the score at $\boldsymbol{\theta}^{(i)}$:

$$\hat{\mathbf{u}}^{(i)} = \frac{1}{K} \sum_{k=1}^K \mathbf{u}(\boldsymbol{\theta}, \mathbf{x}^{(i,k)}) \quad (28)$$

- 9: **end if**
- 10: **end for**
- 11: Estimate optimal polynomial coefficients $\boldsymbol{\phi}^*$ by $\hat{\boldsymbol{\phi}}$ ▷ see Section 2.4
- 12: **for** $i = 1, \dots, I$ **do**
- 13: Construct the reduced-variance control variates

$$\hat{h}^{(i)} = \Delta_{\boldsymbol{\theta}}[P(\boldsymbol{\theta}^{(i)}|\hat{\boldsymbol{\phi}})] + \nabla_{\boldsymbol{\theta}}[P(\boldsymbol{\theta}^{(i)}|\hat{\boldsymbol{\phi}})] \cdot \hat{\mathbf{u}}^{(i)}. \quad (29)$$

- 14: **end for**
- 15: Estimate the expectation μ using

$$\hat{\mu} := \frac{1}{I} \sum_{i=1}^I g(\boldsymbol{\theta}^{(i)}) + \hat{h}^{(i)}. \quad (30)$$

$$= \int_{\Theta} \nabla_{\boldsymbol{\theta}} \cdot [[\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})] p(\boldsymbol{\theta}|\mathbf{y})] d\boldsymbol{\theta} \quad (34)$$

$$= \oint_{\partial\Theta} [[\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})] p(\boldsymbol{\theta}|\mathbf{y})] \cdot \mathbf{n}(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (35)$$

The assumption of the Lemma forces this integral to equal zero, as required. \square

To illustrate the mildness of these conditions, observe that in the case of a scalar parameter $\theta \in \mathbb{R}$ and a degree-one polynomial P , the boundary condition is satisfied whenever $\lim_{\theta \rightarrow \pm\infty} p(\theta|\mathbf{y}) = 0$. More generally, it follows from the work of Oates et al. (2015) that, for unbounded state spaces $\Theta \subseteq \mathbb{R}^d$, a sufficient condition for unbiasedness is that the tails of $p(\boldsymbol{\theta}|\mathbf{y})$ vanish faster than $\|\boldsymbol{\theta}\|^{d+k-2}$ where k is the degree of the

polynomial P . (Here $\|\cdot\|$ can be taken to be any norm on \mathbb{R}^d , due to the equivalence of norms in finite dimensions.)

2.4 Optimising the tuning parameters

Our proposed estimator has two tuning parameters: (i) the polynomial coefficients ϕ , and (ii) the number K of forward-simulations from $\mathbf{Y}|\boldsymbol{\theta}$, in the case of Type I intractability, or from $\mathbf{X}|\boldsymbol{\theta}, \mathbf{y}$ in the case of Type II intractability. In this section, we derive optimal choices for both of these tuning parameters. Here optimality is defined as maximising the variance reduction factor, that in the case of Type I models is defined as

$$R := \frac{\mathbb{V}_K[g(\boldsymbol{\theta})]}{\mathbb{V}_K[g(\boldsymbol{\theta}) + \hat{h}(\boldsymbol{\theta}|\mathbf{y})]} \quad (36)$$

where the subscript K indicates that randomness arises from the augmented posterior $p(\boldsymbol{\theta}, \mathbf{Y}_1, \dots, \mathbf{Y}_K|\mathbf{y})$. The case of Type II models simply replaces $\mathbf{Y}_1, \dots, \mathbf{Y}_K$ with $\mathbf{X}_1, \dots, \mathbf{X}_K$. Below we proceed by firstly deriving the optimal coefficients ϕ^* for fixed number K of simulations and subsequently deriving the optimal value of K assuming the use of optimal coefficients.

Polynomial coefficients ϕ

First we consider the optimal choice of polynomial coefficients ϕ ; this follows fairly straightforwardly from classical results. For general degree polynomials $P(\boldsymbol{\theta}|\phi)$ with coefficients ϕ , we can write $\hat{h}(\boldsymbol{\theta}|\mathbf{y}) = \boldsymbol{\phi}^T \mathbf{m}(\boldsymbol{\theta}, \hat{\mathbf{u}})$, where in the case of degree-one polynomials $\mathbf{m}(\boldsymbol{\theta}, \hat{\mathbf{u}}) = \hat{\mathbf{u}}$ and for higher polynomials the map \mathbf{m} is more complicated: Suppose that we employ a polynomial

$$P(\boldsymbol{\theta}) = \sum_{i=1}^d a_i \theta_i + \sum_{i,j=1}^d b_{i,j} \theta_i \theta_j + \sum_{i,j,k=1}^d c_{i,j,k} \theta_i \theta_j \theta_k + \dots \quad (37)$$

with coefficients $\phi = \{a_i, b_{i,j}, c_{i,j,k}, \dots\}$. For convenience, we assume symmetries $b_{\tau(i,j)} = b_{i,j}$, $c_{\tau(i,j,k)} = c_{i,j,k}$, etc., for all permutations τ . Then from (26)

$$\begin{aligned} \hat{h}(\boldsymbol{\theta}|\mathbf{y}) &= \left[2 \sum_{i=1}^d b_{i,i} + 6 \sum_{i,j=1}^d c_{i,i,j} \theta_j + \dots \right] \\ &+ \sum_{i=1}^d \left[a_i + 2 \sum_{j=1}^d b_{i,j} \theta_j + 3 \sum_{j,k=1}^d c_{i,j,k} \theta_j \theta_k + \dots \right] \hat{u}_i(\boldsymbol{\theta}|\mathbf{y}). \end{aligned} \quad (38)$$

This can in turn be re-written as $\hat{h}(\boldsymbol{\theta}|\mathbf{y}) = \boldsymbol{\phi}^T \mathbf{m}(\boldsymbol{\theta}, \hat{\mathbf{u}})$ where the components $\{a_i, b_{i,j}, c_{i,j,k}, \dots\}$ of ϕ and $\mathbf{m}(\boldsymbol{\theta}, \hat{\mathbf{u}})$ are identified in the inner product as

$$a_i \leftrightarrow \hat{u}_i, \quad (39)$$

$$b_{i,i} \leftrightarrow 2 + 2\theta_i \hat{u}_i, \quad (40)$$

$$b_{i,j} \leftrightarrow 2\theta_j\hat{u}_i + 2\theta_i\hat{u}_j \quad (i < j), \quad (41)$$

$$c_{i,i,i} \leftrightarrow 6\theta_i + 3\theta_i^2\hat{u}_i, \quad (42)$$

$$c_{i,i,j} \leftrightarrow 12\theta_j + 12\theta_i\theta_j\hat{u}_i + 6\theta_i^2\hat{u}_k \quad (i < j), \quad (43)$$

$$c_{i,j,k} \leftrightarrow 6\theta_j\theta_k\hat{u}_i + 6\theta_i\theta_k\hat{u}_j + 6\theta_i\theta_j\hat{u}_k \quad (i < j < k) \quad \dots \quad (44)$$

An optimal choice of coefficients for general degree polynomials is given by the following:

Lemma 2. *For Type I models, the variance reduction factor R is maximised over all possible coefficients ϕ by the choice*

$$\phi^*(\mathbf{y}) := -\nabla_K^{-1}[\mathbf{m}(\boldsymbol{\theta}, \hat{\mathbf{u}})]\mathbb{E}_K[g(\boldsymbol{\theta})\mathbf{m}(\boldsymbol{\theta}, \hat{\mathbf{u}})] \quad (45)$$

and, at the optimal value $\phi = \phi^*$, we have

$$R^{-1} = 1 - \rho(K)^2 \quad (46)$$

where $\rho(K) = \text{Corr}_K[g(\boldsymbol{\theta}), \hat{h}(\boldsymbol{\theta}|\mathbf{y})]$. An analogous result holds for Type II models, replacing $\mathbf{Y}_1, \dots, \mathbf{Y}_K$ with $\mathbf{X}_1, \dots, \mathbf{X}_K$.

Proof. This is a standard result in control variate theory for a linear combination of (well-defined) control variates (e.g. p. 664, Rubinstein and Marcus, 1985). \square

Following the recommendations of Mira et al. (2013); Papamarkou et al. (2014); Oates et al. (2015), we mainly restrict attention to polynomials of degree at most two. Indeed, degree-two polynomials are sufficient for exactness in the special cases discussed in Papamarkou et al. (2014). Similarly, following Mira et al. (2013), we estimate ϕ^* by plugging in the empirical variance and covariance matrices into (45) to obtain an estimate $\hat{\phi}$. This introduces estimator bias since the same samples are “used twice”; however, Glasserman (2004) argues that this bias vanishes more quickly than the Monte Carlo error and hence the error due to this plug-in procedure is typically ignored. (Any bias could alternatively be removed via a sample-splitting step, but this does not seem necessary for the examples that we consider below.)

Number of forward-simulations K

Now we derive an optimal number K of forward-simulations to generate at each state $\boldsymbol{\theta}^{(i)}$ visited in the MCMC sample path, assuming the use of optimal coefficients as derived above. Assuming that parallel computations incur no additional cost, this optimum will depend on the number K_0 of cores that are available for parallel processing in the computing architecture and we consider the general case below. We present the following Lemma for Type I models, but the analogous result holds for Type II models by simply replacing $\mathbf{Y}_1, \dots, \mathbf{Y}_K$ with $\mathbf{X}_1, \dots, \mathbf{X}_K$.

Lemma 3. *Assume that (i) the condition of Lemma 1 is satisfied, (ii) perfect transitions of the Markov chain (i.e. perfect mixing) is achieved, (iii) $\mathbb{E}_K[(g(\boldsymbol{\theta}) + \hat{h}(\boldsymbol{\theta}|\mathbf{y}))^2] < \infty$, and (iv) $\phi = \phi^*$. Then*

$$\sqrt{I}(\hat{\mu} - \mu) \xrightarrow{d} N(0, (1 - \rho(K)^2)\nabla_{\boldsymbol{\theta}|\mathbf{y}}[g(\boldsymbol{\theta})]). \quad (47)$$

Proof. From (i) we have that $\mathbb{E}_K[g(\boldsymbol{\theta}) + \hat{h}(\boldsymbol{\theta}|\mathbf{y})] = \mu$. From (ii), (iii) and the central limit theorem, we have that

$$\sqrt{I}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \mathbb{V}_K[g(\boldsymbol{\theta}) + \hat{h}(\boldsymbol{\theta}|\mathbf{y})]). \quad (48)$$

Then from (iv) and (36) we have that

$$\mathbb{V}_K[g(\boldsymbol{\theta}) + \hat{h}(\boldsymbol{\theta}|\mathbf{y})] = (1 - \rho(K)^2)\mathbb{V}_{\boldsymbol{\theta}|\mathbf{y}}[g(\boldsymbol{\theta})], \quad (49)$$

as required. \square

Write I for the number of MCMC iterations. Then, under the hypotheses of Lemma 3, the key quantity that we aim to minimise is the cost-normalised variance ratio

$$r(K, I) := \frac{1 - \rho(K)^2}{I}, \quad (50)$$

where the optimisation is constrained by fixed computational cost $c = I[K/K_0]$ on a K_0 -core architecture. In other words, for fixed computational cost c , should we focus on obtaining more MCMC samples (large I) or better estimating the RV control variates (large K)? (Note that we assume the calculation of the score vector incurs negligible computational cost – this is certainly true whenever the score is itself a prerequisite for MCMC sampling.) This resource-allocation problem can be solved analytically:

Lemma 4. *The optimum variance for fixed computational cost (i.e. $c = I[K/K_0]$) is always achieved by setting $K = K_0$, the available number of cores.*

Proof. See Appendix A. \square

Our findings may be concisely summarised as follows: For serial computation, choose $K = 1$ and I as large as possible. This typically requires no additional computation relative to standard estimation since one forward-simulation \mathbf{Y} is generated as part of the exchange algorithm and at least one forward-simulation \mathbf{X} is used as the basis for the pseudo-marginal algorithm. For parallel computation, choose $K = K_0$ equal to the number of available cores (but no more) and then let I be as large as possible.

Finally, we note that RV control variates extend easily to the case where multiple expectations $\mu_j = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[g_j(\boldsymbol{\theta})]$ are of interest. Indeed, the same MCMC output can be used to construct control variates $h^{(j)}(\boldsymbol{\theta}|\mathbf{y})$ specific to problem j simply by re-estimating the optimal coefficients

$$\boldsymbol{\phi}^{*,(j)}(\mathbf{y}) = -\mathbb{V}_K^{-1}[\mathbf{m}(\boldsymbol{\theta}, \hat{\boldsymbol{u}})]\mathbb{E}_K[g_j(\boldsymbol{\theta})\mathbf{m}(\boldsymbol{\theta}, \hat{\boldsymbol{u}})] \quad (51)$$

based on the target function g_j and proceeding as above. In this way multiple expectations can be estimated without requiring any additional sampling or simulation.

3 Applications

Here we provide empirical results for an analytically tractable example, along with a version of the Ising model (Type I intractability), an exponential random graph model (Type I), and a nonlinear stochastic differential equation model (Type II).

3.1 Example 1. Tractable exponential

As a simple and analytically tractable example, consider inference for the posterior mean $\mu = \mathbb{E}_{\theta|y}[\theta]$, so that $g(\theta) = \theta$, where data y arise from the exponential distribution $p(y|\theta) = \theta \exp(-\theta y)$ and inference is performed using an improper prior $p(\theta) \propto 1$. The exponential likelihood can be formally viewed as a GRF with sufficient statistic $s(y) = -y$ and partition function $\mathfrak{P}(\theta) = \frac{1}{\theta}$; however, the model is sufficiently simple that all quantities of interest are available in closed form. Indeed, it can easily be verified that $p(\theta|y) = y^2 \theta \exp(-\theta y)$, so that the posterior is directly seen to satisfy the boundary condition of Lemma 1 for any polynomial. The true posterior expected value is $\mu = \frac{2}{y}$ and similarly the score function can be computed exactly as $u(\theta|y) = -y + \frac{1}{\theta}$.

All of the estimators that we consider are (essentially) unbiased (as noted before, the negligible bias resulting from estimation of $\hat{\phi}$ can trivially be removed by data-splitting); in this section, we therefore restrict attention to examining the estimator variances. The maximum variance reduction that we achieve with access to the exact score can be obtained from $\rho(\infty) = \text{Corr}_{\theta|y}(\theta, \phi^*(y)^T \mathbf{m}(\theta, u))$. For degree-one polynomials $P(\theta) = a\theta$ the ZV method corresponds to $m(\theta, u) = u = -y + \frac{1}{\theta}$ and, since θ is not strongly linearly correlated with $\frac{1}{\theta}$, the maximum variance reduction that can be achieved by degree-one polynomials is not substantial. However, the use of degree-two polynomials $P(\theta) = a\theta + b\theta^2$ leads to $\mathbf{m}(\theta, u) = [u, 2 + 2\theta u] = [-y + \frac{1}{\theta}, 4 - 2y\theta]$ and taking $\phi = [0, \frac{1}{2y}]$ leads to a control variate $\phi^T \mathbf{m}(\theta, u) = \frac{2}{y} - \theta$. Thus the ZV estimator $g(\theta) + \phi^T \mathbf{m}(\theta, u)$ is equal to $\frac{2}{y}$, which is independent of θ , i.e. exact zero variance is achieved.

In general, the score $u(\theta|y)$ will be unavailable for GRF but may be estimated by $\hat{u}(\theta|y)$ as described above, with the estimate becoming exact as $K \rightarrow \infty$. We investigate through simulation the effect of employing finite values of K . Intuitively, the proposed approach will be more effective when the target function $g(\theta)$ of interest is strongly correlated (under the posterior) with a linear combination $\phi^T \mathbf{m}(\theta, \hat{u})$. Figure S1¹ demonstrates that when K is large, the RV control variates (for degree-two polynomials) are closely correlated with the ZV control variates (left column) and, hence, with the target function $g(\theta)$ (right column). We would therefore expect to see a large reduction in Monte Carlo variance using RV estimation in this regime.

The main conclusions to be drawn from this tractable example are summarised in Figure 1 where we display estimates for the estimator standard deviation $\text{std}[\hat{\mu}]$, computed as the standard error of the mean over all I Monte Carlo samples. In total the estimation procedure was repeated 100 times, and we report the mean value of $\text{std}[\hat{\mu}]$ along with the standard error of this mean computed over the 100 realisations. We considered

¹All supplementary tables and figures can be found in Friel et al. (2015).

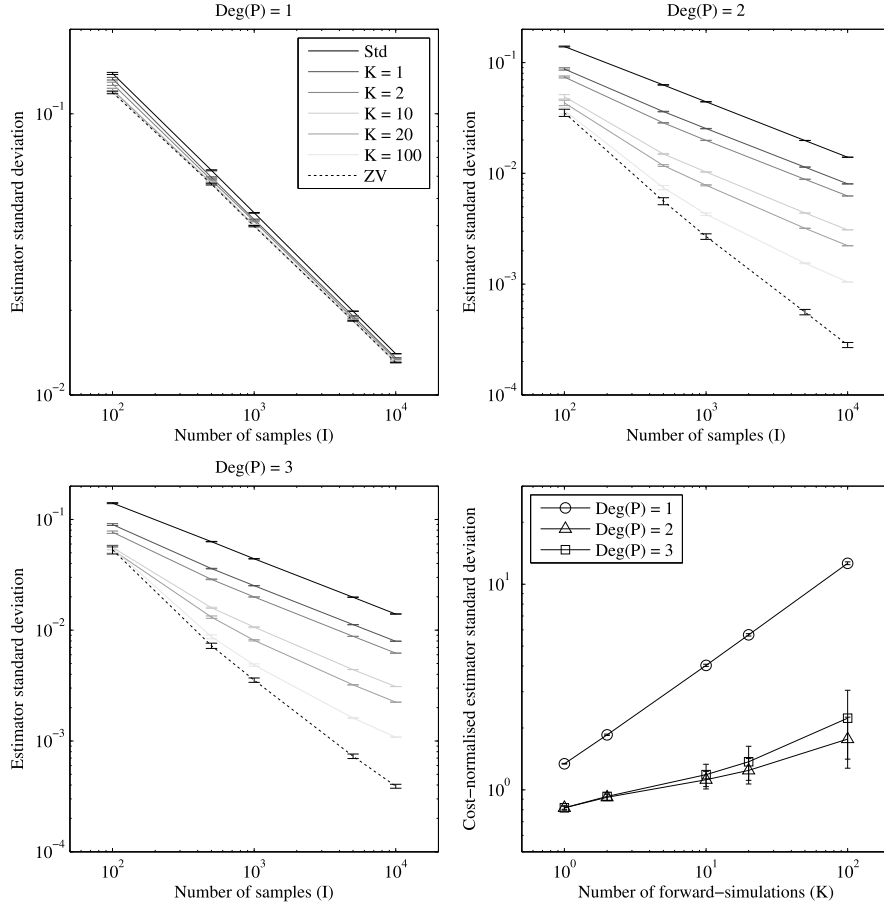


Figure 1: Tractable exponential example. Comparing the standard deviations of Monte Carlo estimators, including the default estimator (“Std”; the regular MCMC estimator with no variance reduction), the reduced-variance estimator with $K = 1, 2, 10, 20$ and 100 , and the ZV estimator. [The ZV estimate has some non-zero standard deviation here because, in practice, the optimal coefficients $\phi(y)^*$ must be estimated using Monte Carlo.] The final panel (bottom right) displays estimator standard deviation normalised by computational cost.

varying the number of Monte Carlo samples $I = 100, 500, 1000, 5000, 10000$, the number of forward-simulations $K = 1, 2, 10, 20, 100$ and the degree of the polynomial trial function $\text{deg}(P) = 1, 2, 3$. Results demonstrate that estimator variance reduces as either I or K is increased, as expected. A comparison *between* the plots (full data provided in Table S1) shows that degree-two polynomials considerably out-perform the degree-one polynomials, whereas the degree-three polynomials tend to slightly under-perform the degree-two polynomials. (The theoretical best ZV control variates are degree-two

polynomials and therefore degree-three polynomials require that additional coefficients associated to higher order control variates – that we know, theoretically, should be equal to zero – are estimated from data, thus adding extra noise.)

To assess computational efficiency, we also report the quantity $\sqrt{TK}\text{std}[\hat{\mu}]$ that has units “standard deviation per unit serial computational cost” and can be used to evaluate the computational efficiency of competing strategies (bottom right panel of Figure 1 and Table S1). Here we see that $K = 1$ minimises $\sqrt{TK}\text{std}[\hat{\mu}]$ and is consistent with the theoretical result that $K = 1$ is optimal for serial computation. Figure S2 plots the canonical correlation coefficient between $\mathbf{m}(\theta, \hat{u})$ and $g(\theta)$ for values of $K = 1, 2, \dots, 10$. Here we notice that over 80% of the correlation is captured by just one forward-simulation from the likelihood ($K = 1$), further supporting our theoretical result that $K = 1$ is optimal for serial computation. Indeed, a theoretical prediction $\rho(K) = (K/(K + C))^{1/2}$ resulting from Lemma 5 in Appendix A, shown as a solid line in Figure S2, closely matches these simulation results.

3.2 Example 2. Ising model

In the experiments below, we consider an Ising model of size $n = 16$, about the limit for exact solution, as defined in Section 1, Example 1, above. Assuming that the lattice points have been indexed from top to bottom in each column and that columns are ordered from left to right, then an interior point y_i in a first order neighbourhood model has neighbours $\{y_{i-n}, y_{i-1}, y_{i+1}, y_{i+n}\}$. Each point along the edges of the lattice has either two or three neighbours. We focus on estimating the posterior mean $\mu = \mathbb{E}_{\theta|\mathbf{y}}[\theta]$ under a prior $\theta \sim N(0, 5^2)$. Since the tails of the prior vanish exponentially and the likelihood is bounded, the posterior automatically satisfies the boundary conditions of Lemma 1. Here data \mathbf{y} were simulated exactly from the likelihood using $\theta = 0.4$, via the recursive scheme of Friel and Rue (2007). This recursive algorithm also allows exact calculation of the partition function. In turn this allows a very precise estimate of the posterior mean; for the data that we consider below this posterior mean is $\mu = 0.43455$, calculated numerically over a very fine grid of θ values.

Figure 2 displays the MCMC trace plots, obtained using the exchange algorithm, for $g(\theta) = \theta$ (blue) and the RV version $g(\theta) + \hat{\phi}^T \mathbf{m}(\theta, \hat{u})$. Trace plots are presented for increasing values of $K \in \{1, 20, 100, 500\}$ and using degree-one (red) and degree-two (green) polynomials.² For $K = 1$ we observe little difference between controlled (i.e. using RV control variates) and uncontrolled trajectories, suggesting that RV control variates do not justify the additional coding effort in the case of serial computation. However, it is evident that as K increases, the Monte Carlo variance of the controlled trajectory decreases; indeed when $K = 500$ the variance is dramatically reduced compared to the (uncontrolled) MCMC samples of θ . These findings are summarised in Table 1. Additionally, we find that degree-two polynomials offer a substantial improvement over degree-one polynomials in terms of variance reduction, but that this is mainly realised

²For convenience, forward-simulation was performed on a single core using a Gibbs sampler with 1,000 burn-in iterations. A sample of size K were collected from this chain at a lag of 500 iterations in order to ensure that dependence between samples was negligible. This accurately mimics the setting of independent samples that corresponds to performing multiple forward-simulations in parallel.

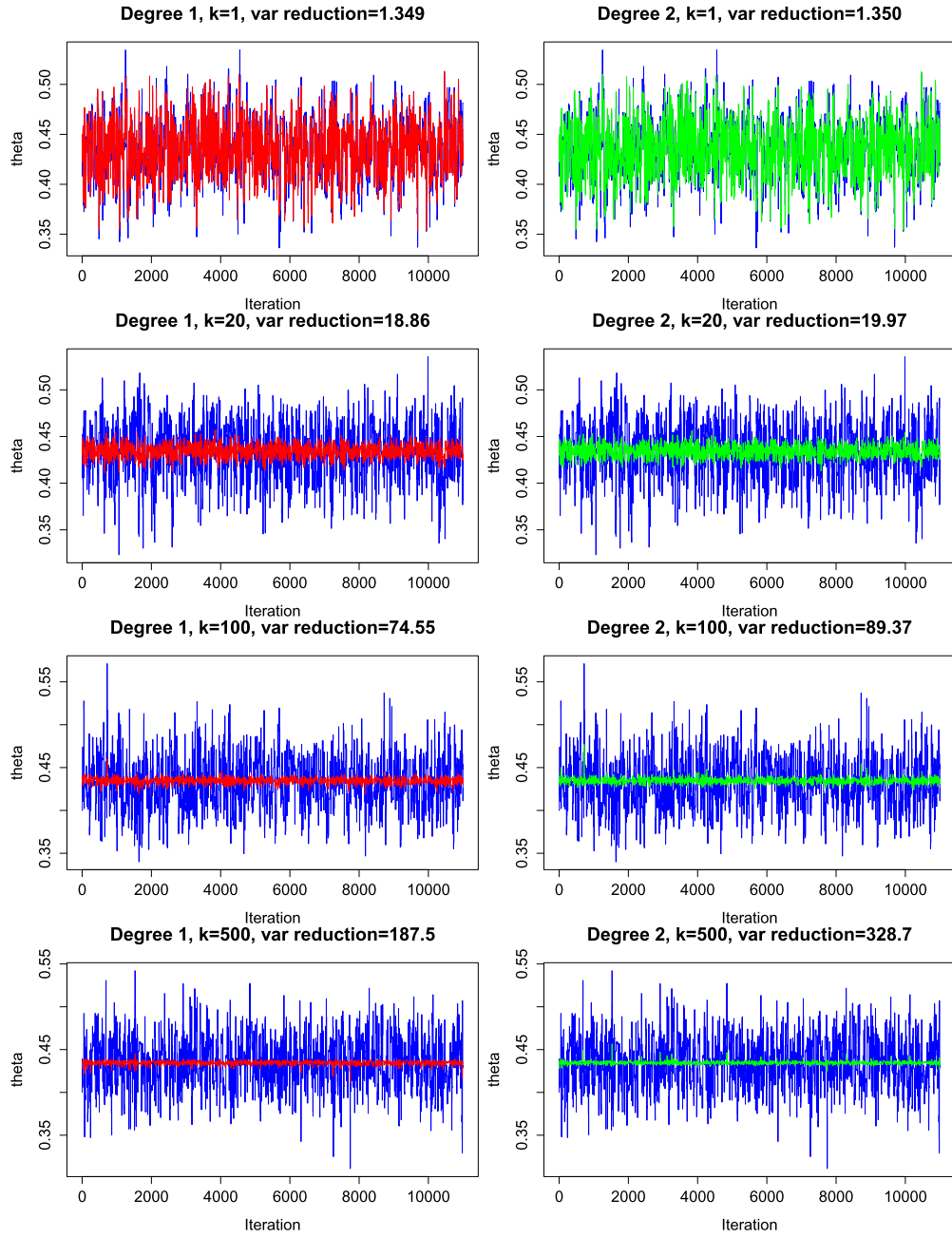


Figure 2: Ising model. As the number of forward-simulations, K , increases, the precision of the controlled estimate of the posterior mean for θ increases. The degree-two polynomial yields greater precision compared to the degree-one polynomial, particularly for larger values of K .

	$K = 1$	$K = 20$	$K = 100$	$K = 500$
$\hat{\mu}$	0.4340	0.4345	0.4322	0.4340
$\hat{\mu}_1$	0.4351	0.4345	0.4347	0.4346
$\hat{\mu}_2$	0.4351	0.4344	0.4346	0.4346
$R = \mathbb{V}[\hat{\mu}]/\mathbb{V}[\hat{\mu}_1]$	1.349	18.86	74.55	187.5
$R = \mathbb{V}[\hat{\mu}]/\mathbb{V}[\hat{\mu}_2]$	1.350	19.97	89.37	328.7

Table 1: Ising model. As the number of forward-simulations, K , used to estimate the score function increases, the reduction in variance becomes more substantial. Here $\hat{\mu}$ is the standard Monte Carlo estimate, $\hat{\mu}_1$ is the reduced-variance estimate using degree-one polynomials and $\hat{\mu}_2$ is the reduced-variance estimate using degree-two polynomials. Brute-force calculation produces a value $\mu = 0.43455$ for this example. [Variances were estimated with respect to empirical means.]

for larger values of K . These results present a powerful approach to exploit multi-core processing to deliver a real-time acceleration in the convergence of MCMC estimators.

3.3 Example 3. Exponential random graph models

In the experiment below, we consider the *Gamaneg* network (Read, 1954), displayed in Figure 3, that consists of $n = 16$ sub-tribes of the Eastern central highlands of New Guinea. In this graph, an edge represents an antagonistic relationship between two sub-tribes. Here we consider an ERG model as defined in Section 1, Example 2, with $k = 2$, i.e. two sufficient statistics, where $s_1(y)$ counts the total number of observed edges and the two-star statistic $s_2(y)$ is also as defined in Section 1 above. Here the parameters

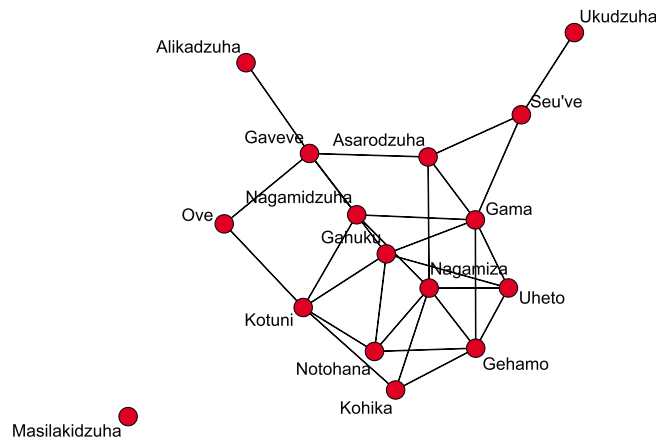


Figure 3: Gamaneg graph. The vertices represent 16 sub-tribes of the Eastern central highlands of New Guinea and edges represent an antagonistic relationship between two sub-tribes.

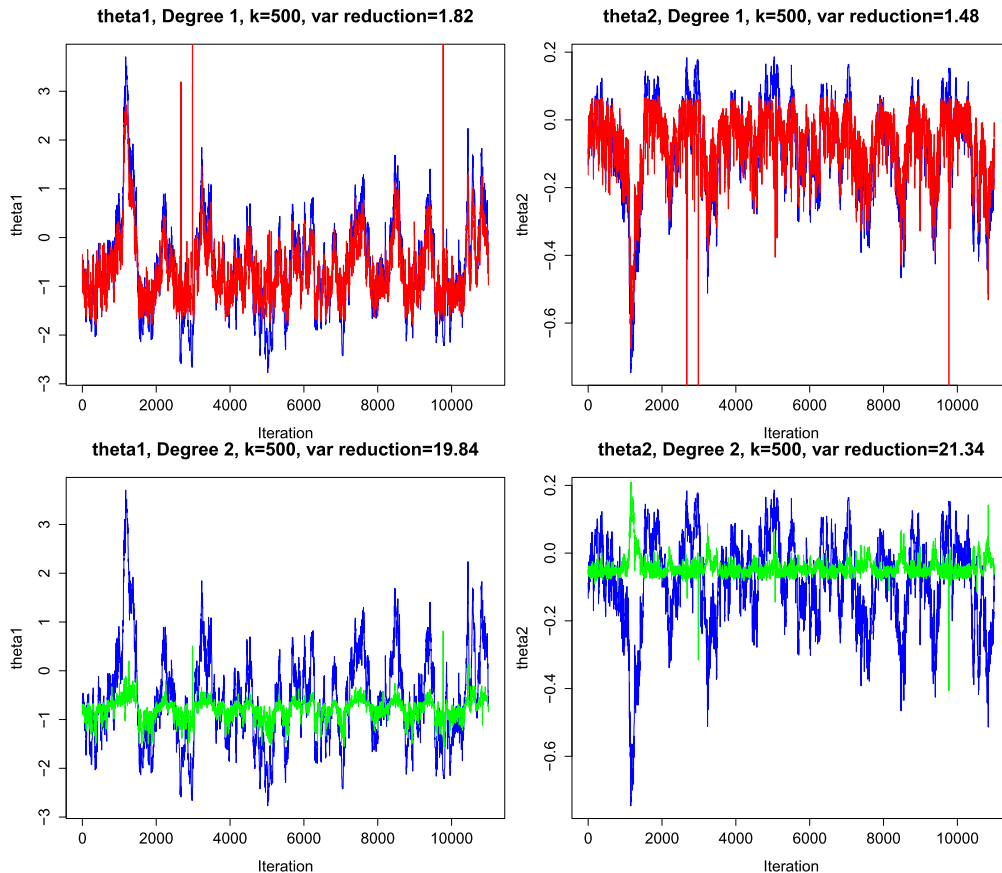


Figure 4: Exponential random graph model. The top row displays the trace plot for θ_1 and θ_2 in uncontrolled (blue) and controlled versions for a degree-one (red) polynomial, while the bottom row is similar but for a degree-two (green) polynomial.

θ_1, θ_2 control the propensity of edges and two-star configurations, respectively, in the network. Positive values of θ_1 and θ_2 tend to lead to, respectively, over-representation of edges and two-star configurations in networks realised from the likelihood. The prior distributions for θ_1 and θ_2 were both set to be independent $N(0, 5^2)$, from which it follows that the boundary condition of Lemma 1 is satisfied. This is a benchmark dataset that has previously been used to assess Monte Carlo methodology (Friel, 2013), making it well-suited to our purposes.

Again we focus on the challenge of estimating the posterior mean $\mu = \mathbb{E}_{\theta|y}[\theta]$, in this case performing independent estimation with $g(\theta) = \theta_j$ for $j = 1, 2$. Recently Caimo and Friel (2011), Caimo and Friel (2014) developed Bayesian methodology for this model, based on the exchange algorithm, that can be directly utilised for the RV

framework developed in this paper. The exchange algorithm was run for $I = 11,000$ iterations, where at each iteration $K = 500$ forward-simulations were used to estimate the score.³ Figure 4 illustrates that a variance reduction of about 20 times is possible using a degree-two polynomial for each of the two components of the parameter vector. From the uncontrolled trajectories it is difficult to comment on the relative posterior means μ_1 and μ_2 of θ_1 and θ_2 , respectively, but from the controlled trajectories it is visually clear that we have $\mu_1 < \mu_2 < 0$. This suggests that posterior predictions of network structure typically contain more two-stars than edges.

We note that Caimo and Mira (2014) recently proposed the use of delayed rejection to reduce autocorrelation in the exchange algorithm for ERG models, demonstrating an approximate two-fold variance reduction; the delayed rejection exchange algorithm is fully compatible with our methodology and, if combined, should yield a further reduction in variance.

3.4 Example 4. Nonlinear stochastic differential equations

For our final example we consider performing Bayesian inference for a system of nonlinear stochastic differential equations (SDEs) as defined in Section 1, Example 4, (9). This problem is well-known to pose challenges for Bayesian computation, and recent work in this direction includes (Beskos et al., 2006; Golightly and Wilkinson, 2008). We estimate the score using

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathbf{y}) \approx \frac{1}{K} \sum_{k=1}^K \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \mathbf{x}^{(k)}|\mathbf{y}) \quad (52)$$

where $\mathbf{x}^{(k)}$ are independent samples from $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$. Such samples can be generated using MCMC techniques, and in this paper we make use of a Metropolis–Hastings sampler with “diffusion bridge” proposals (Fuchs, 2013). Note that since $p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, we have that

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}|\boldsymbol{\theta}). \quad (53)$$

Direct calculation shows that, assuming $\boldsymbol{\beta}$ is invertible,

$$\begin{aligned} & \nabla_{\theta_j} \log p(\mathbf{X}|\boldsymbol{\theta}) \\ &= \sum_{i=2}^T \begin{aligned} & -\frac{1}{2} \text{tr}(\boldsymbol{\beta}_i^{-1} \nabla_{\theta_j} \boldsymbol{\beta}_i) + (\nabla_{\theta_j} \boldsymbol{\alpha}_i)^T \boldsymbol{\beta}_i^{-1} (\mathbf{X}_i - \mathbf{X}_{i-1} - \boldsymbol{\alpha}_i \delta t) \\ & + \frac{1}{2\delta t} (\mathbf{X}_i - \mathbf{X}_{i-1} - \boldsymbol{\alpha}_i \delta t)^T \boldsymbol{\beta}_i^{-1} (\nabla_{\theta_j} \boldsymbol{\beta}_i) \boldsymbol{\beta}_i^{-1} (\mathbf{X}_i - \mathbf{X}_{i-1} - \boldsymbol{\alpha}_i \delta t) \end{aligned} \end{aligned} \quad (54)$$

Consider the specific example of the Susceptible-Infected-Recovered (SIR) model from epidemiology. Letting X_1, X_2 denote respectively the proportions of susceptible

³For convenience, the forward-simulation step was achieved using a Gibbs sampler where a burn-in phase of 1,000 iterations. We drew K samples from this chain at a lag of 1,000 iterations. This accurately mimics the setting of independent samples that corresponds to performing multiple forward-simulations in parallel.

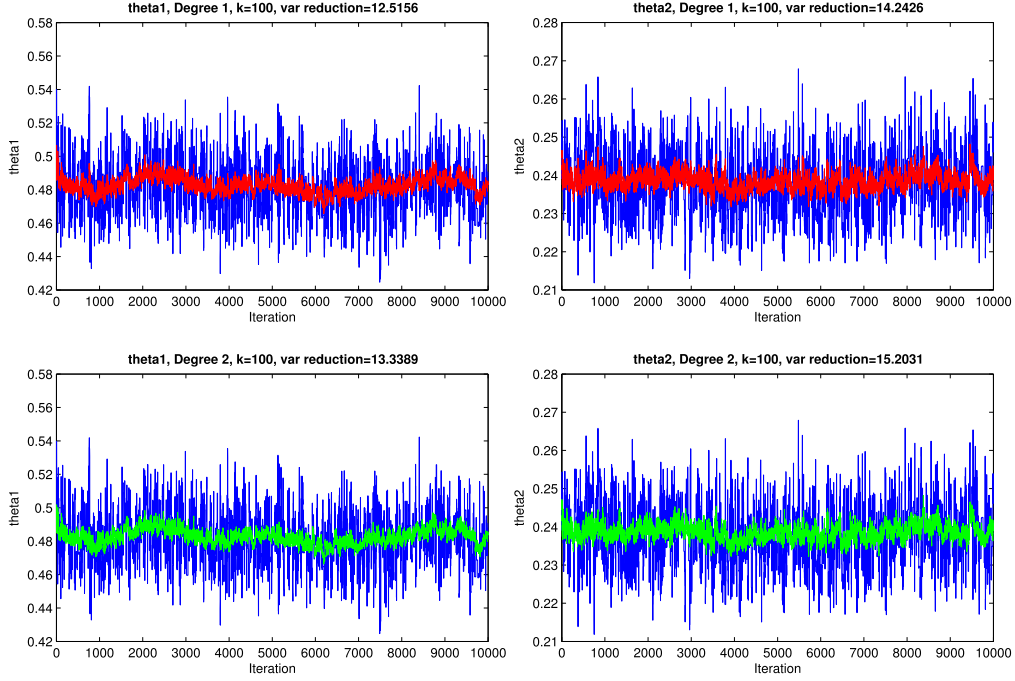


Figure 5: SIR model. The top row displays the trace plots for θ_1 and θ_2 in uncontrolled (blue) and controlled (red) versions for a degree-one polynomial. The bottom row is similar, but for degree-two polynomials.

and infected individuals in a population, modelled as continuous random variables, the SIR model has a stochastic representation given by

$$\alpha(\mathbf{X}; \boldsymbol{\theta}) = \begin{bmatrix} -\theta_1 X_1 X_2 \\ \theta_1 X_1 X_2 - \theta_2 X_2 \end{bmatrix}, \quad \beta(\mathbf{X}; \boldsymbol{\theta}) = \frac{1}{N} \begin{bmatrix} \theta_1 X_1 X_2 & -\theta_1 X_1 X_2 \\ -\theta_1 X_1 X_2 & \theta_1 X_1 X_2 + \theta_2 X_2 \end{bmatrix} \quad (55)$$

where N is a fixed population size and the rate parameters $\boldsymbol{\theta} \in [0, \infty)^2$ are unknown. We assess our methodology by attempting to estimate the posterior mean of $\boldsymbol{\theta}$, taking $g(\boldsymbol{\theta}) = \theta_j$ for $j = 1, 2$ in turn. Here each θ_j was assigned an independent Gamma prior with shape and scale hyperparameters both equal to 2. This prior vanishes at the origin and has exponentially decaying tails, so that the boundary condition of Lemma 1 is satisfied by all polynomials. Data were generated using the initial condition $\mathbf{X}_0 = [0.99, 0.01]$, population size $N = 1,000$ and parameters $\boldsymbol{\theta} = [0.5, 0.25]$. Observations were made at 20 evenly spaced intervals in the period from $t = 0$ to $t = 35$. Five latent data points were introduced between each observed data point, so that the latent process has dimension $2 \times (20 - 1) \times 5 = 190$. At each Monte Carlo iteration we sampled $K = 100$ realisations of the latent data process \mathbf{X}^u .

Figure 5 demonstrates that a variance reduction of about 12–14 times is possible using degree-one polynomials and 13–15 times using degree-two polynomials. Again, these

results highlight the potential to exploit multi-core processing for variance reduction in Monte Carlo methodology.

4 Conclusions

In this paper, we have shown how repeated forward-simulation enables reduced-variance estimation in models that have intractable likelihoods. The examples that we have considered illustrate the value of the proposed methodology in spatial statistics, social network analysis and inference for latent data models such as SDEs. The RV methodology provides a straightforward means to leverage multi-core architectures for Bayesian estimation that compliments recent work for MCMC in this direction by Alquier et al. (2014); Angelino et al. (2014); Bardenet et al. (2014); Calderhead (2014); Korattikara et al. (2014); Maclaurin and Adams (2014). Our theoretical analysis revealed that the number K of forward-simulations should be taken equal to the number of cores in order to provide the optimal variance reduction per unit (serial) computation. Furthermore, it was shown that the proposed RV estimator converges to the (intractable) ZV estimator of Mira et al. (2013) as the number K of cores becomes infinite. Our theoretical findings are supported by empirical results on standard benchmark datasets that demonstrate a substantial variance reduction can be realised in practice. In particular, results for the Ising model demonstrate that a 200–300 times variance reduction can be achieved by exploiting a $K = 500$ core architecture. More generally, our work shows that *it may not be necessary to parallelise the sampling process itself*; the potential of massively multi-core architectures can be harnessed in post-processing MCMC samples using control variates.

To conclude, we suggest interesting directions for further research:

- The approach that we pursued was a post-processing procedure that does not require modification to the MCMC sampling mechanism itself. However, an interesting possibility would be to also use the output of forward-sampling to construct gradient-based proposal mechanisms for the underlying MCMC sampler, following recent work in this direction by Alquier et al. (2014); Nemeth et al. (2014). This would retain the inherently parallel nature of the simulation procedure whilst yielding useful approximate Monte Carlo schemes that converge to an “idealised” (i.e. marginal) sampler as the number of forward-simulations, K , increases. In particular, our procedure can be implemented within the various schemes developed in Alquier et al. (2014); Nemeth et al. (2014) without any additional computational cost. Stochastic approximation of the score function was also recently considered by Atchadé et al. (2014) in the context of designing proximal gradient algorithms. Our work therefore combines to illustrate the wide range of statistical models for which such an approach could prove practically useful. Alternative approaches to handling Type I intractability, such as Approximate Bayesian Computation (Marjoram et al., 2003) typically also require a forward-simulation step and thus could also be embedded within our framework.
- In terms of statistical efficiency it would be interesting to extend the reduced-variance methodology to the non-parametric setting recently considered by Oates

et al. (2014) that provides a mechanism to learn a suitable trial function $P(\boldsymbol{\theta})$ that need not be polynomial, leading (in some cases) to improved convergence rates. A second interesting possibility would be to allow the number of forward-simulations, K to depend upon the current state $\boldsymbol{\theta}$; in this way fewer simulations could be performed when it is expected that the score estimate $\hat{\mathbf{u}}(\boldsymbol{\theta}|\mathbf{y})$ is likely to have a low variance. A third direction would be to move beyond independent estimation of the score $\mathbf{u}(\boldsymbol{\theta}|\mathbf{y})$ for each value of $\boldsymbol{\theta}$; here non-parametric regression techniques could play a role and this should yield further reductions in estimator variance, which again has a close analogy with Oates et al. (2014).

- Finally, a referee suggested the intriguing possibility to exploit control variates for reduced-variance density estimation. Specifically, we would take $g_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}) = K_h(\boldsymbol{\theta}^* - \boldsymbol{\theta})$ for a bandwidth- h kernel smoother estimate for the posterior density at a point $\boldsymbol{\theta}^* \in \Theta$, such that $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[g_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})] \rightarrow p(\boldsymbol{\theta}^*|\mathbf{y})$ as $h \rightarrow 0$. This is a direction that we are keen to explore further.

Of course, in the era of big data, as statisticians are increasingly interested in analysing larger datasets an immediate challenge is the issue of dealing with intractable likelihoods, due to the volume of data. Moreover, one would anticipate that statistical methodology will focus on the development of inferential algorithms that exploit modern multi-core computer architectures. Both of these will inevitably lead to further development and extensions of the methodology described in this paper.

Appendix A

Below we prove Lemma 4 from the Main Text. First, we require a technical result:

Lemma 5. Write $\rho(\infty) = \text{Corr}_{\boldsymbol{\theta}|\mathbf{y}}[g(\boldsymbol{\theta}), h(\boldsymbol{\theta}|\mathbf{y})]$. There exists $C \in (0, \infty)$ such that

$$\rho(K)^2 = \left(\frac{1}{\rho(\infty)^2} + \frac{C}{K} \right)^{-1}. \quad (56)$$

Proof. For Type I models we write $\hat{h}(\boldsymbol{\theta}, \mathbf{Y}) = h(\boldsymbol{\theta}) + \epsilon(\boldsymbol{\theta}, \mathbf{Y})$ where $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_K)$ and suppress dependence on the data \mathbf{y} in this notation. It follows that the discrepancy between the reduced-variance and ZV control variates is given by

$$\epsilon(\boldsymbol{\theta}, \mathbf{Y}) = \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) \cdot \left[\frac{1}{K} \sum_{k=1}^K s(\mathbf{Y}_k) - \mathbb{E}_{\mathbf{Y}'|\boldsymbol{\theta}}[s(\mathbf{Y}')] \right]. \quad (57)$$

Taking an analogous approach to Type II models, we obtain

$$\epsilon(\boldsymbol{\theta}, \mathbf{X}) = \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) \cdot \left[\frac{1}{K} \sum_{k=1}^K \mathbf{u}(\boldsymbol{\theta}, \mathbf{X}_k) - \mathbb{E}_{\mathbf{X}'|\boldsymbol{\theta}, \mathbf{y}}[\mathbf{u}(\boldsymbol{\theta}, \mathbf{X}')] \right]. \quad (58)$$

Note that $\mathbb{E}_{\mathbf{Y}|\boldsymbol{\theta}}[\epsilon(\boldsymbol{\theta}, \mathbf{Y})] = 0$ and hence

$$\mathbb{E}_{\mathbf{Y}, \boldsymbol{\theta}|\mathbf{y}}[\epsilon(\boldsymbol{\theta}, \mathbf{Y})] = \mathbb{E}_{\mathbf{Y}|\boldsymbol{\theta}, \mathbf{y}} \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}[\epsilon(\boldsymbol{\theta}, \mathbf{Y})] = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}} \mathbb{E}_{\mathbf{Y}|\boldsymbol{\theta}}[\epsilon(\boldsymbol{\theta}, \mathbf{Y})] = 0, \quad (59)$$

with an analogous result holding for Type II models. Using these results, we have that, for Type I models,

$$\mathbb{V}_{\mathbf{Y}, \boldsymbol{\theta} | \mathbf{y}}[\epsilon(\boldsymbol{\theta}, \mathbf{Y})] = \mathbb{E}_{\mathbf{Y}, \boldsymbol{\theta} | \mathbf{y}}[\epsilon(\boldsymbol{\theta}, \mathbf{Y})^2] \quad (60)$$

$$= \mathbb{E}_{\boldsymbol{\theta} | \mathbf{y}} \mathbb{E}_{\mathbf{Y} | \boldsymbol{\theta}}[\epsilon(\boldsymbol{\theta}, \mathbf{Y})^2] = \mathbb{E}_{\boldsymbol{\theta} | \mathbf{y}} \mathbb{V}_{\mathbf{Y} | \boldsymbol{\theta}}[\epsilon(\boldsymbol{\theta}, \mathbf{Y})] \quad (61)$$

$$= \mathbb{E}_{\boldsymbol{\theta} | \mathbf{y}}[K^{-1} \mathbb{V}_{\mathbf{Y} | \boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) \cdot \mathbf{s}(\mathbf{Y})]] = C_1 K^{-1} \quad (62)$$

where $C_1 = \mathbb{E}_{\boldsymbol{\theta} | \mathbf{y}} \mathbb{V}_{\mathbf{Y} | \boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) \cdot \mathbf{s}(\mathbf{Y})]$. Also observe that, for any function $g(\boldsymbol{\theta})$,

$$\text{Cov}_{\mathbf{Y}, \boldsymbol{\theta} | \mathbf{y}}[g(\boldsymbol{\theta}), \epsilon(\boldsymbol{\theta}, \mathbf{Y})] = \mathbb{E}_{\mathbf{Y}, \boldsymbol{\theta} | \mathbf{y}}[g(\boldsymbol{\theta})\epsilon(\boldsymbol{\theta}, \mathbf{Y})] - \mathbb{E}_{\mathbf{Y}, \boldsymbol{\theta} | \mathbf{y}}[g(\boldsymbol{\theta})] \mathbb{E}_{\mathbf{Y}, \boldsymbol{\theta} | \mathbf{y}}[\epsilon(\boldsymbol{\theta}, \mathbf{Y})] \quad (63)$$

$$= \mathbb{E}_{\boldsymbol{\theta} | \mathbf{y}} \mathbb{E}_{\mathbf{Y} | \boldsymbol{\theta}}[g(\boldsymbol{\theta})\epsilon(\boldsymbol{\theta}, \mathbf{Y})] = 0. \quad (64)$$

Putting these results together, we obtain

$$\rho(K) = \frac{\text{Cov}_{\mathbf{Y}, \boldsymbol{\theta} | \mathbf{y}}[g(\boldsymbol{\theta}), \hat{h}(\boldsymbol{\theta}, \mathbf{Y})]}{\sqrt{\mathbb{V}_{\mathbf{Y}, \boldsymbol{\theta} | \mathbf{y}}[g(\boldsymbol{\theta})]} \sqrt{\mathbb{V}_{\mathbf{Y}, \boldsymbol{\theta} | \mathbf{y}}[\hat{h}(\boldsymbol{\theta}, \mathbf{Y})]}} = \frac{\text{Cov}_{\boldsymbol{\theta} | \mathbf{y}}[g(\boldsymbol{\theta}), h(\boldsymbol{\theta})]}{\sqrt{\mathbb{V}_{\boldsymbol{\theta} | \mathbf{y}}[g(\boldsymbol{\theta})]} \sqrt{\mathbb{V}_{\boldsymbol{\theta} | \mathbf{y}}[h(\boldsymbol{\theta})] + C_1 K^{-1}}} \quad (65)$$

from which it follows that

$$\frac{1}{\rho(K)^2} = \frac{\mathbb{V}_{\boldsymbol{\theta} | \mathbf{y}}[g(\boldsymbol{\theta})] (\mathbb{V}_{\boldsymbol{\theta} | \mathbf{y}}[h(\boldsymbol{\theta})] + C_1 K^{-1})}{\text{Cov}_{\boldsymbol{\theta} | \mathbf{y}}[g(\boldsymbol{\theta}), h(\boldsymbol{\theta})]^2} = \frac{1}{\rho(\infty)^2} + \frac{C}{K} \quad (66)$$

where $C = C_1 / \text{Cov}_{\boldsymbol{\theta} | \mathbf{y}}[g(\boldsymbol{\theta}), h(\boldsymbol{\theta})]^2$. The analogous derivation for Type II models completes the proof. \square

A simple corollary of Lemma 5 is that the reduced-variance estimator converges to the (unavailable) ZV estimator as $K \rightarrow \infty$. Moreover, we can derive an optimal choice for K subject to fixed computational cost:

Proof of Lemma 4. Starting from (50), we substitute $I = c / \lceil K / K_0 \rceil$ and use the identity in Lemma 5 to obtain

$$r(K) = \frac{1}{c} \left\lceil \frac{K}{K_0} \right\rceil \left(1 - \frac{K \rho(\infty)^2}{K + C \rho(\infty)^2} \right). \quad (67)$$

Figure S3 displays typical cost-normalised variance ratios $r(K)$, for both non-Gaussian (i.e. $\rho(\infty) < 1$; left) and Gaussian (i.e. $\rho(\infty) = 1$; right) cases. In each case, the minimum is attained at $K = K_0$. In general, we see from first principles that (i) $r > 0$, (ii) $r(k)$ is increasing for $k = K_0, 2K_0, 3K_0, \dots$, and (iii) $r(k)$ is decreasing on the interval $((n-1)K_0, nK_0]$ and bounded below by $r((n-1)K_0)$, for any $n \in \mathbb{N}$. Thus we have $\arg \min_{K=1,2,3,\dots} r(K) = K_0$, as required. \square

Supplementary Material

Supplementary Figures and Tables (DOI: [10.1214/15-BA948SUPP](https://doi.org/10.1214/15-BA948SUPP); .pdf).

References

- Augustin, N., Muggleston, M., and Buckland, S. (1996). “An autologistic model for spatial distribution of wildlife.” *Journal of Applied Ecology*, 33(2):339–347. doi: <http://dx.doi.org/10.2307/2404755>. 2
- Alquier, P., Friel, N., Everitt, R., and Boland, A. (2014). “Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels.” arXiv:1403.5496. 4, 7, 23
- Andradóttir, S., Heyman, D. P., and Teunis, J. O. (1993). “Variance reduction through smoothing and control variates for Markov Chain simulations.” *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 3(3):167–189. doi: <http://dx.doi.org/10.1145/174153.174154>. 7
- Andrieu, C., and Roberts, G. O. (2009). “The pseudo-marginal approach for efficient Monte Carlo computations.” *The Annals of Statistics*, 37(2):697–725. MR2502648. doi: <http://dx.doi.org/10.1214/07-AOS574>. 2, 5
- Andrieu, C., Doucet, A., and Holenstein, R. (2010). “Particle Markov chain Monte Carlo (with Discussion).” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 72(3):269–342. MR2758115. doi: <http://dx.doi.org/10.1111/j.1467-9868.2009.00736.x>. 2, 5
- Angelino, E., Kohler, E., Waterland, A., Seltzer, M., and Adams, R. P. (2014). “Accelerating MCMC via Parallel Predictive Prefetching.” arXiv:1403.7265. 7, 23
- Ahn, S., Korattikara, A., and Welling, M. (2012). “Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring.” In: *Proceedings of the 29th International Conference on Machine Learning*, 1591–1598. 4
- Armond, J., Saha, K., Rana, A. A., Oates, C. J., Jaenisch, R., Nicodemi, M., Mukherjee, S. (2014). “A stochastic model dissects cellular states and heterogeneity in transition processes”. *Nature Scientific Reports*, 4:3692. 4
- Assaraf, R., and Caffarel, M. (1999), Zero-Variance Principle for Monte Carlo Algorithms. *Physical Review Letters*, 83(23):4682–4685. doi: <http://dx.doi.org/10.1103/PhysRevLett.83.4682>. 6, 7
- Atchadé, Y, Fort, G., and Moulines, E. (2014). “On stochastic proximal gradient algorithms.” arXiv:1402.2365. 23
- Bandyopadhyay, D., Reich, B. J., and Slate, E. (2009). “Bayesian Modeling of Multivariate Spatial Binary Data with applications to Dental Caries.” *Statistics in Medicine*, 28(28):3492–3508. MR2744378. doi: <http://dx.doi.org/10.1002/sim.3647>. 2
- Bardenet, R., Doucet, A., and Holmes, C. (2014). “Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach.” In: *Proceedings of the 31st International Conference on Machine Learning*, 405–413. 7, 23
- Besag, J. E. (1972). “Nearest-neighbour systems and the auto-logistic model for binary data.” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 34(1):697–725. MR0323276. 3

- Besag, J. E. (1974) “Spatial interaction and the statistical analysis of lattice systems (with discussion).” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 36(2):192–236. [MR0373208](#). 2
- Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006). “Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion).” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 68(3):333–382. [MR2278331](#). doi: <http://dx.doi.org/10.1111/j.1467-9868.2006.00552.x>. 2, 21
- Beskos, A., Kalogeropoulos, K., and Pazos, E. (2013). “Advanced MCMC methods for sampling on diffusion pathspace.” *Stochastic Processes and Their Applications*, 123(4):1415–1453. [MR3016228](#). doi: <http://dx.doi.org/10.1016/j.spa.2012.12.001>. 5
- Caimo, A., and Friel, N. (2011). “Bayesian inference for exponential random graph models.” *Social Networks*, 33:41–55. doi: <http://dx.doi.org/10.1016/j.socnet.2010.09.004>. 20
- Caimo, A., and Friel, N. (2013). “Bayesian model selection for exponential random graph models.” *Social Networks*, 35:11–24. doi: <http://dx.doi.org/10.1016/j.socnet.2012.10.003>. 4
- Caimo, A., and Friel, N. (2014). “Bergm: Bayesian inference for exponential random graphs using R.” *Journal of Statistical Software*, 61(2). 20
- Caimo, A., and Mira, A. (2014). “Efficient computational strategies for Bayesian social networks.” *Statistics and Computing*, 25(1):113–125. doi: <http://dx.doi.org/10.1007/s11222-014-9516-7>. 21
- Calderhead, B. (2014). “A general construction for parallelizing Metropolis–Hastings algorithms.” *Proceedings of the National Academy of Sciences, USA*, 111(49):17408–17413. doi: <http://dx.doi.org/10.1073/pnas.1408184111>. 7, 23
- Cappé, O., Moulines, E., and Ryden, T. (2005). “Inference in hidden Markov models.” Springer, New York. [MR2159833](#). 4
- Davison, A. C., Padoan, S. A., and Ribatet, M. (2009). “Statistical modelling of spatial extremes.” *Statistical Science*, 27:161–186. [MR2963980](#). doi: <http://dx.doi.org/10.1214/11-STS376>. 3
- Dellaportas, P., and Kontoyiannis, I. (2012). “Control variates for estimation based on reversible Markov chain Monte Carlo samplers.” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 74(1):133–161. [MR2885843](#). doi: <http://dx.doi.org/10.1111/j.1467-9868.2011.01000.x>. 8
- Doucet, A., Pitt, M., Deligiannidis, G., and Kohn, R. (2012). “Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator.” arXiv:1210.1871. 7
- Evans, M., and Swartz, T. (2000). “Approximating integrals via Monte Carlo and deterministic methods.” Oxford University Press. [MR1859163](#). 7

- Everitt, R. (2012). “Bayesian parameter estimation for latent Markov random fields and social networks.” *Journal of Computational and graphical Statistics*, 21(4):940–960. MR3005805. doi: <http://dx.doi.org/10.1080/10618600.2012.687493>. 4
- Fahrmeir, L., and Lang, S. (2001). “Bayesian inference for generalized additive mixed models based on Markov random field priors.” *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 50(2):201–220. MR1833273. doi: <http://dx.doi.org/10.1111/1467-9876.00229>. 4
- Friel, N., and Rue, H. (2007). “Recursive computing and simulation-free inference for general factorizable models.” *Biometrika*, 94:661–672. MR2410015. doi: <http://dx.doi.org/10.1093/biomet/asm052>. 17
- Friel, N. (2013). “Estimating the evidence for Gibbs random fields.” *Journal of Computational and Graphical Statistics*, 22:518–532. MR3173729. doi: <http://dx.doi.org/10.1080/10618600.2013.778780>. 20
- Friel, N., Mira, A., and Oates, Ch. J. (2015). “Supplementary Figures and Tables.” *Bayesian Analysis*. doi: <http://dx.doi.org/10.1214/15-BA948SUPP>. 15
- Fuchs, C. (2013). *Inference for Diffusion Processes with Applications in Life Sciences*. Springer, Heidelberg. MR3015023. doi: <http://dx.doi.org/10.1007/978-3-642-25969-2>. 5, 21
- Glasserman, P. (2004). *Monte Carlo methods in financial engineering*. Springer, New York. MR1999614. 6, 7, 13
- Golightly, A., and Wilkinson, D. J. (2008). “Bayesian inference for nonlinear multivariate diffusion models observed with error.” *Computational Statistics and Data Analysis*, 52(3):1674–1693. MR2422763. doi: <http://dx.doi.org/10.1016/j.csda.2007.05.019>. 21
- Geyer, C. J., and Thompson, E. A. (1992). “Constrained Monte Carlo maximum likelihood for dependent data (with discussion).” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 54(3):657–699. MR1185217. 3
- Hammer, H., and Tjelmeland, H. (2008). “Control variates for the Metropolis–Hastings algorithm.” *Scandinavian Journal of Statistics*, 35(3):400–414. MR2446727. doi: <http://dx.doi.org/10.1111/j.1467-9469.2008.00601.x>. 8
- He, F., Zhou, J., and Zhu, H. (2003). “Autologistic regression model for the distribution of vegetation.” *Journal of Agricultural, Biological, and Environmental Statistics*, 8(2):205–222. doi: <http://dx.doi.org/10.1198/1085711031508>. 2
- Huffer, F. W., and Wu, H. (1998). “Markov Chain Monte Carlo for Autologistic Regression Models with Application to the Distribution of Plant Species.” *Biometrics*, 54:509–524. doi: <http://dx.doi.org/10.2307/3109759>. 2
- Kendall, P. C., and Bourne, D. E. (1992). “Vector analysis and Cartesian tensors (3rd ed.).” CRC Press, Florida. doi: <http://dx.doi.org/10.1007/978-1-4899-4427-6>. 10

- Korattikara, A., Chen, Y., and Welling, M. (2014). “Austerity in MCMC Land: Cutting the Metropolis–Hastings Budget.” In: *Proceedings of the 31st International Conference on Machine Learning*, 181–189. 4, 7, 23
- Kou, S. C., Olding, B. P., Lysy, M., and Liu, J. S. (2012). “A multiresolution method for parameter estimation of diffusion processes.” *Journal of the American Statistical Association*, 107(500):1558–1574. MR3036416. doi: <http://dx.doi.org/10.1080/01621459.2012.720899>. 5
- Lamberton, D., and Lapeyre, B. (2007). *Introduction to stochastic calculus applied to finance*. CRC Press. MR2362458. 5
- Lee, A., Yau, C., Giles, M., Doucet, A., and Holmes, C. (2010). “On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods.” *Journal of Computational and Graphical Statistics* 19(4):769–789. doi: <http://dx.doi.org/10.1198/jcgs.2010.10039>. 7
- Lindgren, F., Rue, H., and Lindström, J. (2011). “An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach.” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 73(4):423–498. MR2853727. doi: <http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x>. 6
- Lyne, A. M., Girolami, M., Atchade, Y., Strathmann, H., and Simpson, D. (2013). “Playing Russian Roulette with Intractable Likelihoods.” arXiv:1306.4032. 2, 4
- Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). “Markov chain Monte Carlo without likelihoods.” *Proceedings of the National Academy of Sciences, U.S.A.*, 100:15324–15328. doi: <http://dx.doi.org/10.1073/pnas.0306899100>. 1, 23
- Maclaurin, D., and Adams, R. P. (2014). “Firefly Monte Carlo: Exact MCMC with Subsets of Data.” In: *Proceedings of the 30th Annual Conference on Uncertainty in Artificial Intelligence*, 543–552. 7, 23
- Mira, A., Möller, J., and Roberts, G. O. (2001). “Perfect Slice Samplers.” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 63(3):593–606. MR1858405. doi: <http://dx.doi.org/10.1111/1467-9868.00301>. 9
- Mira, A., Tenconi, P., and Bressanini, D. (2003). “Variance reduction for MCMC.” Technical Report 2003/29, Università degli Studi dell’ Insubria, Italy. 7
- Mira, A., Solgi, R., and Imparato, D. (2013). “Zero Variance Markov Chain Monte Carlo for Bayesian Estimators.” *Statistics and Computing* 23(5):653–662. MR3094805. doi: <http://dx.doi.org/10.1007/s11222-012-9344-6>. 6, 8, 10, 13, 23
- Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). “An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants.” *Biometrika*, 93:451–458. MR2278096. doi: <http://dx.doi.org/10.1093/biomet/93.2.451>. 1, 3
- Murray, I., Ghahramani, Z., and MacKay, D. (2006). “MCMC for doubly-intractable distributions.” In: *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, 359–366. 1, 3, 4

- Nemeth, C., Sherlock, C., and Fearnhead, P. (2014). “Particle Metropolis adjusted Langevin algorithms.” arXiv:1412.7299. 10, 23
- Oates, C. J., Girolami, M., and Chopin, N. (2014). “Control functionals for Monte Carlo integration.” *CRiSM Working Paper, The University of Warwick*, 14:22. 8, 23, 24
- Oates, C. J., Papamarkou, T., and Girolami, M. (2015). “The Controlled Thermodynamic Integral for Bayesian Model Comparison.” *Journal of the American Statistical Association*, to appear. doi: <http://dx.doi.org/10.1080/01621459.2015.1021006>. 6, 11, 13
- Øksendal, B. (2003). *Stochastic differential equations*. Springer-Verlag, Berlin. MR2001996. doi: <http://dx.doi.org/10.1007/978-3-642-14394-6>. 5
- Papamarkou, T., Mira, A., and Girolami, M. (2014). “Zero Variance Differential Geometric Markov Chain Monte Carlo Algorithms.” *Bayesian Analysis*, 9(1):97–128. MR3188301. doi: <http://dx.doi.org/10.1214/13-BA848>. 6, 13
- Papamarkou, T., Mira, A., and Girolami, M. (2015). “Hamiltonian Methods and Zero-Variance Principle.” In: *Current Trends in Bayesian Methodology with Applications* (eds. Dipak K. Dey, Umesh Singh and A. Loganathan), Chapman and Hall/CRC Press. 8
- Pillai, N. S., and Smith, A. (2014) “Ergodicity of Approximate MCMC Chains with Applications to Large Data Sets.” arXiv:1405.0182. 4
- Potamianos, G., and Goutsias, J. (1997). “Stochastic approximation algorithms for partition function estimation of Gibbs random fields.” *IEEE Transactions on Information Theory*, 43(6):1948–1965. MR1481051. doi: <http://dx.doi.org/10.1109/18.641558>. 9
- Propp, J. G., and Wilson, D. B. (1996). “Exact sampling with coupled Markov chains and applications to statistical mechanics.” *Random Structures and Algorithms*, 9(1):223–252. MR1611693. doi: [http://dx.doi.org/10.1002/\(SICI\)1098-2418\(199608/09\)9:1/2<223::AID-RSA14>3.3.CO;2-R](http://dx.doi.org/10.1002/(SICI)1098-2418(199608/09)9:1/2<223::AID-RSA14>3.3.CO;2-R). 9
- Read, K. E. (1954). “Cultures of the Central Highlands, New Guinea.” *Southwestern Journal of Anthropology* 10(1):1–43. 19
- Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). “An introduction to exponential random graph models for social networks.” *Social Networks*, 29:173–191. doi: <http://dx.doi.org/10.1016/j.socnet.2006.08.002>. 3
- Rubinstein, R. Y., and Marcus, R. (1985). “Efficiency of Multivariate Control Variates in Monte Carlo Simulation.” *Operations Research*, 33(3):661–677. MR0791714. doi: <http://dx.doi.org/10.1287/opre.33.3.661>. 13
- Rue, H., Martino, S., and Chopin, N. (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion).” *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 71(2):319–392. MR2649602. doi: <http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x>. 1, 4, 6

- Sherlock, C., Thiery, A., Roberts, G. O., and Rosenthal, J. S. (2014). “On the efficiency of pseudo-marginal random walk Metropolis algorithm.” *The Annals of Statistics*, 43(1), 238–275. MR3285606. doi: <http://dx.doi.org/10.1214/14-AOS1278>. 2, 5, 7
- Suchard, M., Wang, Q., Chan, C., Frelinger, J., Cron, A., and West, M. (2010). “Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures.” *Journal of Computational and Graphical Statistics* 19(2):419–438. MR2758309. doi: <http://dx.doi.org/10.1198/jcgs.2010.10016>. 7
- Welling, M., and Teh, Y. W. (2011). “Bayesian Learning via Stochastic Gradient Langevin Dynamics.” In: *Proceedings of the 28th International Conference on Machine Learning*, 681–688. 4
- West, M., and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models (2nd ed.)*. Springer-Verlag, New York. MR1482232. 4
- Wilkinson, D. J. (2011). *Stochastic Modelling for Systems Biology*. CRC Press. MR2222876. 5

Acknowledgments

The authors are grateful for the constructive feedback they received from the Editors and Reviewers at Bayesian Analysis. NF was supported by the Science Foundation Ireland [12/IP/1424]. The Insight Centre for Data Analytics is supported by Science Foundation Ireland [SFI/12/RC/2289]. AM was supported by the Swiss National Science Foundation [CR12I1-156229]. CJO was supported by the EPSRC Centre for Research in Statistical Methodology [EP/D002060/1]. The authors thank Mark Girolami for helpful discussions.