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

Nial Friel[⋆], Antonietta Mira[†], Chris J. Oates[‡]

August 29, 2014

[⋆]*School of Mathematical Sciences and Insight: The National Centre for Data Analytics, University College Dublin, Ireland.*
[†]*Swiss Finance Institute, University of Lugano, Switzerland.*
[‡]*Department of Statistics, University of Warwick, UK.*

Abstract

Many popular statistical models for complex phenomena are intractable, in the sense that the likelihood function cannot easily be evaluated, even up to proportionality. 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 reduce the Monte Carlo variance of Bayesian estimators, in some cases dramatically. We prove that these control variates are well-defined, provide a positive variance reduction and derive optimal tuning parameters that are targeted at optimising this variance reduction. Moreover, the methodology is highly parallelisable and offers a route to exploit multi-core processing architectures for Bayesian computation. Results presented on the Ising model, exponential random graphs and nonlinear stochastic differential equations support our theoretical findings.

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

1 Introduction

Many models of interest are intractable, by which it is understood that the likelihood function $p(\boldsymbol{y}|\boldsymbol{\theta})$, that describes how data $\boldsymbol{y}$ arise from a model parametrised by $\boldsymbol{\theta}$, is unavailable in closed form, even up to proportionality. 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 $\mathfrak{P}(\boldsymbol{\theta}) = \int f(\boldsymbol{y}';\boldsymbol{\theta})d\boldsymbol{y}'$ that depends on parameters $\boldsymbol{\theta}$, such that $p(\boldsymbol{y}|\boldsymbol{\theta}) = f(\boldsymbol{y};\boldsymbol{\theta})/\mathfrak{P}(\boldsymbol{\theta})$.

Type II: The need to marginalise over a set of latent variables $\boldsymbol{x}$, such that $p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x}$.

Bayesian estimation in both of these settings is extremely challenging as many established computational techniques are incompatible with intractable likelihoods. This has motivated researchers to propose several approximations to the likelihood function that are tractable (e.g. Marjoram *et 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).

In the next subsections we elaborate on the two types of intractability and on the related references in the literature that have addressed them.

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

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

such that the partition function

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

is intractable. The dependence of the partition function on $\boldsymbol{\theta}$ leads to difficulties in inferring this parameter, which cannot proceed based on the sufficient statistic $\boldsymbol{s}(\boldsymbol{y})$ alone. An early attempt to circumvent this difficulty is the pseudolikelihood 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 simulate realisations from this model. The Monte Carlo MLE approach of Geyer and Thompson (1992) exploits this fact to allow maximum likelihood estimation. From a Bayesian perspective, simulating from the likelihood has also played an influential role in several approach, 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 cleverly avoids the need to directly evaluate the partition function by considering an augmented target distribution, 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. Other recent attempts to address this problem include Liang (2010); Everitt (2012); Lyne *et al.* (2013); Atchadé *et al.* (2013); Alquier *et al.* (2014); Rao *et al.* (2014).

Type II: Type II intractability arises from the need to marginalise over latent variables $\boldsymbol{x}$ such that the marginal likelihood

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \tag{3}$$

is unavailable in closed form. For example, in a hidden Markov model the parameters $\boldsymbol{\theta}$ that specify the Markov chain may be of interest, whilst the latent sample path $\boldsymbol{x}$ of the Markov chain that gives rise to observations $\boldsymbol{y}$ may not be of interest and must be marginalised. Typically in this case the number of possible samples paths $\boldsymbol{x}$ is very large and this renders the integral in Eqn. 3 intractable. The main approach to inference under Type II intractability is the pseudo-marginal MCMC of Andrieu and Roberts (2009), that replaces the marginal likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ in the Metropolis-Hastings acceptance ratio with an unbiased estimate that can either be obtained by forward-simulation from $p(\boldsymbol{x}|\boldsymbol{\theta})$, 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 $(\boldsymbol{\theta}, \boldsymbol{x})$ (Sherlock *et al.*, 2014). When combined with particle MCMC (Andrieu *et al.*, 2010), the pseudo-marginal algorithm represents the current state-of-the-art for Type II intractability. Other attempts to address this problem include the popular approximation scheme of Rue *et al.* (2009) and the references therein.

Outline of the paper: We address the problem of estimating posterior expectations via MCMC when data arise from an intractable likelihood:

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

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), network models (Caimo and Friel, 2011), 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; Oates *et al.*, 2014).

In this paper we focus on the use of control variates for the reduction of Monte Carlo variance (Glasserman, 2004). Here a modified function $\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \phi_1 h_1(\boldsymbol{\theta}) + \cdots + \phi_m h_m(\boldsymbol{\theta})$ is constructed such that $\tilde{g}(\boldsymbol{\theta})$ has the same posterior expectation but a reduced variance compared to $g(\boldsymbol{\theta})$. This can be achieved when each of the $h_i(\boldsymbol{\theta})$ have zero posterior expectation and $\boldsymbol{h}(\boldsymbol{\theta})$ has strong canonical correlation with $g(\boldsymbol{\theta})$. Recently Mira *et al.* (2013) proposed to use the score vector $\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) := \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{y})$ as a set of control variates, since this can be guaranteed to have zero mean under mild boundary conditions (described below). There it was shown that these "zero variance" (ZV) estimators can significantly reduced Monte Carlo variance, sometimes dramatically. Further support for the use of the score as a control variate was provided in Papamarkou *et al.* (2014); Oates *et al.* (2014), who demonstrated that the approach fits naturally within Hamiltonian-type MCMC schemes

that themselves make use of the score, requiring essentially no additional computational effort. 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, which we call "reduced-variance" control variates, that can be computed for intractable likelihoods of both Type I and Type II. Specifically we study the effect of replacing the true score function $\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y})$ for the intractable models in the ZV methodology with an unbiased estimate $\hat{\boldsymbol{u}}(\boldsymbol{\theta}|\boldsymbol{y})$ 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 straight-forward manner that complements related research efforts for parallelisation of MCMC methodology (Alquier *et al.*, 2014; Maclaurin and Adams, 2014; Angelino *et al.*, 2014; Korattikara *et al.*, 2014; Bardenet *et al.*, 2014). We prove that these reduced-variance control variates are well-defined and provide a positive variance reduction. Furthermore we derive optimal tuning parameters that are targeted at optimising this 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. The proposed methodology should therefore be considered a default choice in both serial and parallel computational approaches to Bayesian estimation with intractable likelihoods. Empirical results presented on the Ising model, exponential random graphs and nonlinear stochastic differential equations support our theoretical findings.

2 Methods

2.1 ZV control variates and intractable likelihoods

Control variates are often employed when the aim is to estimate, with high precision, the expectation $\mu = \mathbb{E}_\pi[g(\boldsymbol{X})]$ of a function $g(\boldsymbol{X})$ of a random variable $\boldsymbol{X}$ that is distributed according to a (possibly unnormalised) density $\pi(\boldsymbol{X})$. In this paper we focus on a real-valued random variable $\boldsymbol{X} = \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ and take the density π to be the posterior distribution of the parameter $\boldsymbol{\theta}$ given data $\boldsymbol{y}$. The generic control variate principle relies on constructing an auxiliary function $\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + h(\boldsymbol{\theta})$ where $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{\theta})] = 0$ and so $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[\tilde{g}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta})]$.

In many cases it is possible to choose $h(\boldsymbol{\theta})$ such that the variance $\mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{y}}[\tilde{g}(\boldsymbol{\theta})] < \mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta})]$, leading to a Monte Carlo estimator with strictly smaller variance:

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} \tilde{g}(\boldsymbol{\theta}^{(i)}), \tag{4}$$

where $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ are independent samples from $p(\boldsymbol{\theta}|\boldsymbol{y})$. 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 $h(\boldsymbol{\theta})$ is formed as a sum $\phi_1 h_1(\boldsymbol{\theta}) + \ldots \phi_m h_m(\boldsymbol{\theta})$ where the $h_i(\boldsymbol{\theta})$ each have zero mean under $\pi(\boldsymbol{\theta})$ and are known as control variates, whilst ϕ_i are coefficients that must be specified (Glasserman, 2004). For estimation based on Markov chains, Andradóttir *et al.* (1993) proposed control variates for discrete state spaces. Later Mira *et al.* (2003) extended this approach to continuous state spaces, observing that the optimal choice of $h(\boldsymbol{\theta})$ is intimately associated with the solution of the Poisson equation $h(\boldsymbol{\theta}) = \mathbb{E}_\pi[g(\boldsymbol{\theta})] - g(\boldsymbol{\theta})$ and proposing to solve this equation numerically. Further work on constructing control variates for Markov chains includes Hammer and Tjelmeland (2008) for Metropolis-Hasings chains and Dellaportas and Kontoyiannis (2012) for Gibbs samplers.

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

$$h(\boldsymbol{\theta}|\boldsymbol{y}) = \Delta_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] + \nabla_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] \cdot \boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) \tag{5}$$

where $\nabla_{\boldsymbol{\theta}} = [\partial/\partial_{\theta_1}, \ldots, \partial/\partial_{\theta_d}]^T$ is the gradient operator, $\Delta_{\boldsymbol{\theta}} = (\partial^2/\partial^2_{\theta_1} + \cdots + \partial^2/\partial^2_{\theta_d})$ is the Laplacian operator and "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 $\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y})$ are $d \times 1$ vectors. From the work of Mira *et al.* (2013) we know that any posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$ that approximates a Gaussian forms a suitable candidate for implementing the ZV scheme.

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

$$\log p(\boldsymbol{\theta}|\boldsymbol{y}) = \boldsymbol{\theta}^T \boldsymbol{s}(\boldsymbol{y}) - \log \mathfrak{P}(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + C, \tag{6}$$

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

$$\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) = \boldsymbol{s}(\boldsymbol{y}) - \nabla_{\boldsymbol{\theta}} \log \mathfrak{P}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}). \tag{7}$$

It is clear that Eqn. 7 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 that we can evaluate the score function

$$\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) = \nabla_{\boldsymbol{\theta}} \log \int p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} + \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}). \tag{8}$$

It is clear that Eqn. 8 will not have a closed-form when the integral over the latent variable $\boldsymbol{x}$ is intractable. In the sections below we demonstrate how forward-simulation can be used to approximate $\nabla_{\boldsymbol{\theta}} \log \int p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}$, 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 $\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y})$, that can be computed for Type I models of GRF form, is constructed by noting that

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log \mathfrak{P}(\boldsymbol{\theta}) &= \frac{1}{\mathfrak{P}(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \mathfrak{P}(\boldsymbol{\theta}) && (9)\\
&= \frac{1}{\mathfrak{P}(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \int \exp(\boldsymbol{\theta}^T \boldsymbol{s}(\boldsymbol{y})) d\boldsymbol{y} && (10)\\
&= \frac{1}{\mathfrak{P}(\boldsymbol{\theta})} \int \boldsymbol{s}(\boldsymbol{y}) \exp(\boldsymbol{\theta}^T \boldsymbol{s}(\boldsymbol{y})) d\boldsymbol{y} && (11)\\
&= \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}[\boldsymbol{s}(\boldsymbol{Y})], && (12)
\end{aligned}$$

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

$$\hat{\boldsymbol{u}}(\boldsymbol{\theta}|\boldsymbol{y}) := \boldsymbol{s}(\boldsymbol{y}) - \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{s}(\boldsymbol{Y}_k) + \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) \tag{13}$$

where the $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$ are independent simulations from the GRF with density $p(\boldsymbol{y}|\boldsymbol{\theta})$. Forward-simulation for GRF can be achieved using, for example, perfect sampling (Propp and Wilson, 1996; Mira *et al.*, 2001). We make two important observations: Firstly, one realisation $\boldsymbol{Y}_1$ 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 (almost trivial) 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 $\boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{x}) := \nabla_{\boldsymbol{\theta},\boldsymbol{x}} \log p(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})$ of the extended posterior is typically available in closed form and this can be leveraged as follows:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{\theta}|\boldsymbol{y})} && (14)\\
&= \frac{1}{p(\boldsymbol{\theta}|\boldsymbol{y})} \nabla_{\boldsymbol{\theta}} \int p(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} && (15)\\
&= \int \frac{[\nabla_{\boldsymbol{\theta}} p(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})]}{p(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})} \frac{p(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})}{p(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{x} && (16)\\
&= \int [\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})] p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) d\boldsymbol{x} = \mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{y}}[\boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{X})] && (17)
\end{aligned}$$

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

$$\hat{\boldsymbol{u}}(\boldsymbol{\theta}|\boldsymbol{y}) := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{X}_k) \tag{18}$$

where the $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ are independent simulations from the posterior conditional $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. Observe that it is straight-forward to implement pseudo-marginal MCMC in such a way that samples $\boldsymbol{X}_i$ are obtained as a by-product, so that estimation of the score requires no additional computation.

2.3 Reduced-variance control variates

Reduced-variance control variates are constructed using an unbiased estimator for the score as follows:

$$\hat{h}(\boldsymbol{\theta}|\boldsymbol{y}) := \Delta_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] + \nabla_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] \cdot \hat{\boldsymbol{u}}(\boldsymbol{\theta}|\boldsymbol{y}), \tag{19}$$

where again $P \in \mathcal{P}$ is a polynomial. 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. The nomenclature "reduced-variance" derives from the fact that Eqn. 19 is a stochastic approximation to the ZV control variates and can therefore be expected to have similar properties. Pseudocode for our methodology is provided in Alg. 1.

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

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

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

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K|\boldsymbol{y}}[\hat{h}(\boldsymbol{\theta}|\boldsymbol{y})] &= \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\mathbb{E}_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K|\boldsymbol{\theta}}\left[\Delta_{\boldsymbol{\theta}}P(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}P(\boldsymbol{\theta}) \cdot \hat{\boldsymbol{u}}(\boldsymbol{\theta}|\boldsymbol{y})\right]\right] && (24)\\ &= \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\Delta_{\boldsymbol{\theta}}P(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}P(\boldsymbol{\theta}) \cdot \boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y})\right], && (25) \end{aligned}$$

with the analogous result holding for Type II models. Using the definition of the score $\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y})$ we have

$$= \int_{\Theta} [\Delta_{\boldsymbol{\theta}}P(\boldsymbol{\theta})]p(\boldsymbol{\theta}|\boldsymbol{y}) + [\nabla_{\boldsymbol{\theta}}P(\boldsymbol{\theta})] \cdot [\nabla_{\boldsymbol{\theta}}p(\boldsymbol{\theta}|\boldsymbol{y})] d\boldsymbol{\theta}. \tag{26}$$

Algorithm 1 Reduced-variance estimation for intractable likelihoods

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

$$\hat{\boldsymbol{u}}^{(i)} = \boldsymbol{s}(\boldsymbol{y}) - \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{s}(\boldsymbol{y}^{(i,k)}) + \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}^{(i)}) \tag{20}$$

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

$$\hat{\boldsymbol{u}}^{(i)} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{x}^{(i,k)}) \tag{21}$$

  end if
end for
Estimate optimal polynomial coefficients $\boldsymbol{\phi}^*$ by $\hat{\boldsymbol{\phi}}$ ▷ see section 2.4.1
for $i = 1, \ldots, I$ **do**
  Construct the reduced-variance control variates

$$\hat{h}^{(i)} = \Delta_{\boldsymbol{\theta}}[P(\boldsymbol{\theta}^{(i)}|\hat{\boldsymbol{\phi}})] + \nabla_{\boldsymbol{\theta}}[P(\boldsymbol{\theta}^{(i)}|\hat{\boldsymbol{\phi}})] \cdot \hat{\boldsymbol{u}}^{(i)}. \tag{22}$$

end for
Estimate the expectation μ using

$$\hat{\mu} := \frac{1}{I}\sum_{i=1}^{I} g(\boldsymbol{\theta}^{(i)}) + \hat{h}^{(i)}. \tag{23}$$

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

$$= \int_{\Theta} \nabla_{\boldsymbol{\theta}} \cdot [[\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})] p(\boldsymbol{\theta}|\boldsymbol{y})] d\boldsymbol{\theta} \tag{27}$$

$$= \oint_{\partial\Theta} [[\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})] p(\boldsymbol{\theta}|\boldsymbol{y})] \cdot \boldsymbol{n}(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{28}$$

It is now apparent that the assumption of the Lemma forces this integral to equal zero. □

To illustrate the mildness of these conditions, observe that in the case of a scalar parameter $\theta \in [0, \infty)$ and a degree-one polynomial P, the boundary condition requires that both

$p(\boldsymbol{\theta} = 0|\boldsymbol{y}) = 0$ and $\lim_{\theta\to\infty} p(\theta|\boldsymbol{y}) = 0$. More generally, it follows from the work of Oates *et al.* (2014) that, for unbounded state spaces $\Theta \subseteq \mathbb{R}^d$, a sufficient condition for unbiasedness is that the tails of $p(\boldsymbol{\theta}|\boldsymbol{y})$ 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 $\boldsymbol{\phi}$, and (ii) the number K of forward-simulations from $\boldsymbol{Y}|\boldsymbol{\theta}$, in the case of Type I intractability, or from $\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{y}$ 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

$$R := \frac{\mathbb{V}_{\boldsymbol{\theta},\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K|\boldsymbol{y}}[g(\boldsymbol{\theta})]}{\mathbb{V}_{\boldsymbol{\theta},\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K|\boldsymbol{y}}[g(\boldsymbol{\theta}) + \hat{h}(\boldsymbol{\theta}|\boldsymbol{y})]} \tag{29}$$

and in the case of Type II models replaces $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$ with $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$. Below we proceed by firstly deriving the optimal coefficients $\boldsymbol{\phi}^*$ for fixed number K of simulations and subsequently deriving the optimal value of K assuming the use of optimal coefficients.

2.4.1 Polynomial coefficients $\boldsymbol{\phi}$

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

$$P(\boldsymbol{\theta}) = \sum_{i=1}^{d} a_i \theta_i + \sum_{i,j=1}^{d} b_{i,j} \theta_i \theta_j + \sum_{i,j,k=1}^{d} c_{i,j,k} \theta_i \theta_j \theta_k + \ldots \tag{30}$$

with coefficients $\boldsymbol{\phi} = \{a_i, b_{i,j}, c_{i,j,k}, \ldots\}$. 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 Eqn. 19 we have that

$$\begin{aligned} \hat{h}(\boldsymbol{\theta}|\boldsymbol{y}) \quad = \quad & \left[2 \sum_{i=1}^{d} b_{i,i} + 6 \sum_{i,j=1}^{d} c_{i,i,j} \theta_j + \ldots \right] \\ & + \sum_{i=1}^{d} \left[a_i + 2 \sum_{j=1}^{d} b_{i,j} \theta_j + 3 \sum_{j,k=1}^{d} c_{i,j,k} \theta_j \theta_k + \ldots \right] \hat{u}_i(\boldsymbol{\theta}|\boldsymbol{y}). \end{aligned} \tag{31}$$

This can in turn be re-written as $\hat{h}(\boldsymbol{\theta}|\boldsymbol{y}) = \boldsymbol{\phi}^T \boldsymbol{m}(\boldsymbol{\theta}, \hat{\boldsymbol{u}})$ where the components of $\boldsymbol{\phi}$ and $\boldsymbol{m}(\boldsymbol{\theta}, \hat{\boldsymbol{u}})$ are identified as

$$\begin{aligned}
a_i &\leftrightarrow \hat{u}_i && (32)\\
b_{i,i} &\leftrightarrow 2 + 2\theta_i \hat{u}_i && (33)\\
b_{i,j} &\leftrightarrow 2\theta_j \hat{u}_i + 2\theta_i \hat{u}_j \quad (i < j) && (34)\\
c_{i,i,i} &\leftrightarrow 6\theta_i + 3\theta_i^2 \hat{u}_i && (35)\\
c_{i,i,j} &\leftrightarrow 12\theta_j + 12\theta_i\theta_j \hat{u}_i + 6\theta_i^2 \hat{u}_k \quad (i < j) && (36)\\
c_{i,j,k} &\leftrightarrow 6\theta_j\theta_k \hat{u}_i + 6\theta_i\theta_k \hat{u}_j + 6\theta_i\theta_j \hat{u}_k \quad (i < j < k) && (37)\\
&\vdots
\end{aligned}$$

Following the recommendations of Mira *et al.* (2013); Papamarkou *et al.* (2014); Oates *et al.* (2014) we mainly restrict attention to polynomials of degree at most two. 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 $\boldsymbol{\phi}$ by the choice*

$$\boldsymbol{\phi}^*(\boldsymbol{y}) := -\mathbb{V}^{-1}_{\boldsymbol{\theta},\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_K|\boldsymbol{y}}[\boldsymbol{m}(\boldsymbol{\theta}, \hat{\boldsymbol{u}})]\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_K|\boldsymbol{y}}[g(\boldsymbol{\theta})\boldsymbol{m}(\boldsymbol{\theta}, \hat{\boldsymbol{u}})] \tag{38}$$

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

$$R^{-1} = 1 - \rho(K)^2 \tag{39}$$

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

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). □

In practice, following Mira *et al.* (2013), we estimate $\boldsymbol{\phi}^*$ by plugging in the empirical variance and covariance matrices into Eqn. 38 to obtain an estimate $\hat{\boldsymbol{\phi}}$. This introduces estimator bias since the data 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 "data-splitting" step, but this does not seem necessary for the examples that we consider below.)

2.4.2 Number of forward-simulations K

Now we consider the optimal number K of forward-simulations, assuming the use of optimal coefficients as derived above. This 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. Write I for the number of MCMC iterations. We present the following Lemma for Type I models, but the analogous result holds for Type II models by simply replacing $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_K$ with $\boldsymbol{X}_1, \dots, \boldsymbol{X}_K$.

Lemma 3. *Assume that (i) the condition of Lemma 1 is satisfied, (ii) perfect transitions of the Markov chain (i.e. perfect mixing) (iii)* $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K|\boldsymbol{y}}[(g(\boldsymbol{\theta})+\hat{h}(\boldsymbol{\theta}|\boldsymbol{y}))^2]<\infty$*, and (iv)* $\boldsymbol{\phi}=\boldsymbol{\phi}^*$*. Then*

$$\sqrt{I}(\hat{\mu}-\mu)\xrightarrow{d}N\left(0,(1-\rho(K)^2)\mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta})]\right). \tag{40}$$

Proof. From (i) we have that $\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K|\boldsymbol{y}}[g(\boldsymbol{\theta})+\hat{h}(\boldsymbol{\theta}|\boldsymbol{y})]=\mu$. From (ii), (iii) and the central limit theorem we have that

$$\sqrt{I}(\hat{\mu}-\mu)\xrightarrow{d}N(0,\mathbb{V}_{\boldsymbol{\theta},\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K|\boldsymbol{y}}[g(\boldsymbol{\theta})+\hat{h}(\boldsymbol{\theta}|\boldsymbol{y})]). \tag{41}$$

Then from (iv) and Eqn. 29 we have that $\mathbb{V}_{\boldsymbol{\theta},\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_K|\boldsymbol{y}}[g(\boldsymbol{\theta})+\hat{h}(\boldsymbol{\theta}|\boldsymbol{y})]=(1-\rho(K)^2)\mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta})]$. □

So, under the hypotheses of Lemma 3, the key quantity that we aim to minimise is the cost-normalised variance ratio

$$r(K,I):=\frac{1-\rho(K)^2}{I}, \tag{42}$$

where the optimisation is constrained by fixed computational cost $c=I\lceil K/K_0\rceil$ on a K_0-core architecture. This can be achieved analytically with help from the following technical lemma:

Lemma 4. *Write* $\rho(\infty)$ *for* $\mathrm{Corr}_{\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta}),h(\boldsymbol{\theta}|\boldsymbol{y})]$*. There exists* $C\in(0,\infty)$ *such that*

$$\rho(K)^2=\left(\frac{1}{\rho(\infty)^2}+\frac{C}{K}\right)^{-1}. \tag{43}$$

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

$$\epsilon(\boldsymbol{\theta},\boldsymbol{Y})=\nabla_{\boldsymbol{\theta}}P(\boldsymbol{\theta})\cdot\left[\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{s}(\boldsymbol{Y}_k)-\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}[\boldsymbol{s}(\boldsymbol{Y})]\right]. \tag{44}$$

Taking an analogous approach to Type II models we obtain

$$\epsilon(\boldsymbol{\theta},\boldsymbol{X})=\nabla_{\boldsymbol{\theta}}P(\boldsymbol{\theta})\cdot\left[\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{X}_k)-\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{y}}[\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{X})]\right]. \tag{45}$$

Note that $\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}[\epsilon(\boldsymbol{\theta},\boldsymbol{Y})]=0$ and hence $\mathbb{E}_{\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{y}}[\epsilon(\boldsymbol{\theta},\boldsymbol{Y})]=\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta},\boldsymbol{y}}\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[\epsilon(\boldsymbol{\theta},\boldsymbol{Y})]=\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}[\epsilon(\boldsymbol{\theta},\boldsymbol{Y})]=0$, with an analogous result holding for Type II models. Using these results we have that, for Type I models

$$\begin{aligned}\mathbb{V}_{\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{y}}[\epsilon(\boldsymbol{\theta},\boldsymbol{Y})] &= \mathbb{E}_{\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{y}}[\epsilon(\boldsymbol{\theta},\boldsymbol{Y})^2]=\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}[\epsilon(\boldsymbol{\theta},\boldsymbol{Y})^2] &(46)\\ &= \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\mathbb{V}_{\boldsymbol{Y}|\boldsymbol{\theta}}[\epsilon(\boldsymbol{\theta},\boldsymbol{Y})]=\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[K^{-1}\mathbb{V}_{\boldsymbol{Y}|\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}}P(\boldsymbol{\theta})\cdot\boldsymbol{s}(\boldsymbol{Y})]]=C_1K^{-1} &(47)\end{aligned}$$

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

$$\begin{aligned} \mathrm{Cov}_{\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta}), \epsilon(\boldsymbol{\theta}, \boldsymbol{Y})] &= \mathbb{E}_{\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta})\epsilon(\boldsymbol{\theta}, \boldsymbol{Y})] - \mathbb{E}_{\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta})]\mathbb{E}_{\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{y}}[\epsilon(\boldsymbol{\theta}, \boldsymbol{Y})] \quad (48) \\ &= \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}[g(\boldsymbol{\theta})\epsilon(\boldsymbol{\theta}, \boldsymbol{Y})] = 0. \quad (49) \end{aligned}$$

Putting these results together we obtain

$$\rho(K) = \frac{\mathrm{Cov}_{\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta}), \hat{h}(\boldsymbol{\theta}, \boldsymbol{Y})]}{\sqrt{\mathbb{V}_{\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta})]}\sqrt{\mathbb{V}_{\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{y}}[\hat{h}(\boldsymbol{\theta}, \boldsymbol{Y})]}} = \frac{\mathrm{Cov}_{\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta}), h(\boldsymbol{\theta})]}{\sqrt{\mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta})]}\sqrt{\mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{y}}[h(\boldsymbol{\theta})] + C_1 K^{-1}}} \quad (50)$$

from which it follows that

$$\frac{1}{\rho(K)^2} = \frac{\mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta})](\mathbb{V}_{\boldsymbol{\theta}|\boldsymbol{Y}}[h(\boldsymbol{\theta})] + C_1 K^{-1})}{\mathrm{Cov}_{\boldsymbol{\theta}|\boldsymbol{y}}[g(\boldsymbol{\theta}), h(\boldsymbol{\theta})]^2} = \frac{1}{\rho(\infty)^2} + \frac{C}{K} \quad (51)$$

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

A simple corollary of Lemma 4 is that the reduced-variance estimator converges to the (unavailable) ZV estimator as $K \to \infty$. Moreover we can derive an optimal choice for K subject to fixed computational cost:

Lemma 5. *The optimum variance for fixed computational cost (i.e. $c = I\lceil K/K_0 \rceil$) is always achieved by setting $K = K_0$, the available number of cores.*

Proof. Starting from Eqn. 42, we substitute $I = c/\lceil K/K_0 \rceil$ and use the identity in Lemma 4 to obtain

$$r(K) = \frac{1}{c}\left\lceil \frac{K}{K_0} \right\rceil \left(1 - \frac{K\rho(\infty)^2}{K + C\rho(\infty)^2}\right). \quad (52)$$

We see from first principles that $\arg\min_{k=1,2,3,\ldots} r(K) = K_0$, as required. □

Our findings may be concisely summarised as follows: For serial computation, choose $K = 1$. (This typically requires no additional computation since one forward-simulation $\boldsymbol{Y}$ is generated as part of the exchange algorithm and forward-simulations can be 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).

3 Applications

Here we provide empirical results for an analytically tractable example, along with a version of the Ising model (Type I), 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 $\mathcal{P}(\theta) = \frac{1}{\theta}$, 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 (the negligible bias resulting from estimation of $\hat{\boldsymbol{\phi}}$ 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 if we had access to the exact score $u(\theta)$ can be obtained from $\rho(\infty) = \mathrm{Corr}_{\theta|y}(\theta, \boldsymbol{\phi}^*(y)^T \boldsymbol{m}(\theta, u))$. 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 $\boldsymbol{m}(\theta, u) = [u, 2 + 2\theta u] = [-y + \frac{1}{\theta}, 4 - 2y\theta]$ and taking $\boldsymbol{\phi} = [0, \frac{1}{2y}]$ leads to a control variate $\boldsymbol{\phi}^T \boldsymbol{m}(\theta, u) = \frac{2}{y} - \theta$. Thus the ZV estimator $g(\theta) + \boldsymbol{\phi}^T \boldsymbol{m}(\theta, u)$ 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 $K \to \infty$. 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 $\boldsymbol{\phi}^T \boldsymbol{m}(\theta, \hat{u})$. Fig. 1 demonstrates that (for degree-two polynomials) when K is large, the reduced-variance control variates 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 when employing reduced-variance estimation in this regime.

Fig. 2 plots the canonical correlation coefficient between $\boldsymbol{m}(\theta, \hat{u})$ and $g(\theta)$ for values of $K = 1, 2, \ldots, 10$. Here we notice that over 80% of the correlation is captured by just one forward-simulation from the likelihood ($K = 1$), supporting our theoretical result that $K = 1$ is optimal for serial computation. Indeed, our theoretical prediction $\rho(K) = \left(\frac{K}{K+C}\right)^{1/2}$ resulting from Lemma 4, shown as a solid line in Fig. 2, closely matches this data.

Table 1 displays estimates of the estimator standard deviation $\mathrm{std}[\hat{\mu}]$, 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 $\mathrm{std}[\hat{\mu}]$ along with the standard error of this mean computed over the 100 realisations. We considered varying the number of Monte Carlo samples $I = 100, 500, 1000, 5000, 10000$, the number of forward-simulations $K = 1, 10, 100$ and the degree of the polynomial trial function $\deg(P) = 1, 2, 3$. Results

demonstrate that estimator variance reduces as either I or K is increased, as expected. The 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 the coefficients associated to higher order control variates - that we know, theoretically, should be equal to zero - are estimated from data, thus adding extra noise.) We also report the quantity $\sqrt{IK}\text{std}[\hat{\mu}]$ that has units "standard deviation per unit serial computational cost" and can be used to evaluate the computational efficiency of competing strategies. Indeed, we see that $K = 1$ minimises $\sqrt{IK}\text{std}[\hat{\mu}]$ and is consistent with the theoretical result that $K = 1$ is optimal for serial computation.

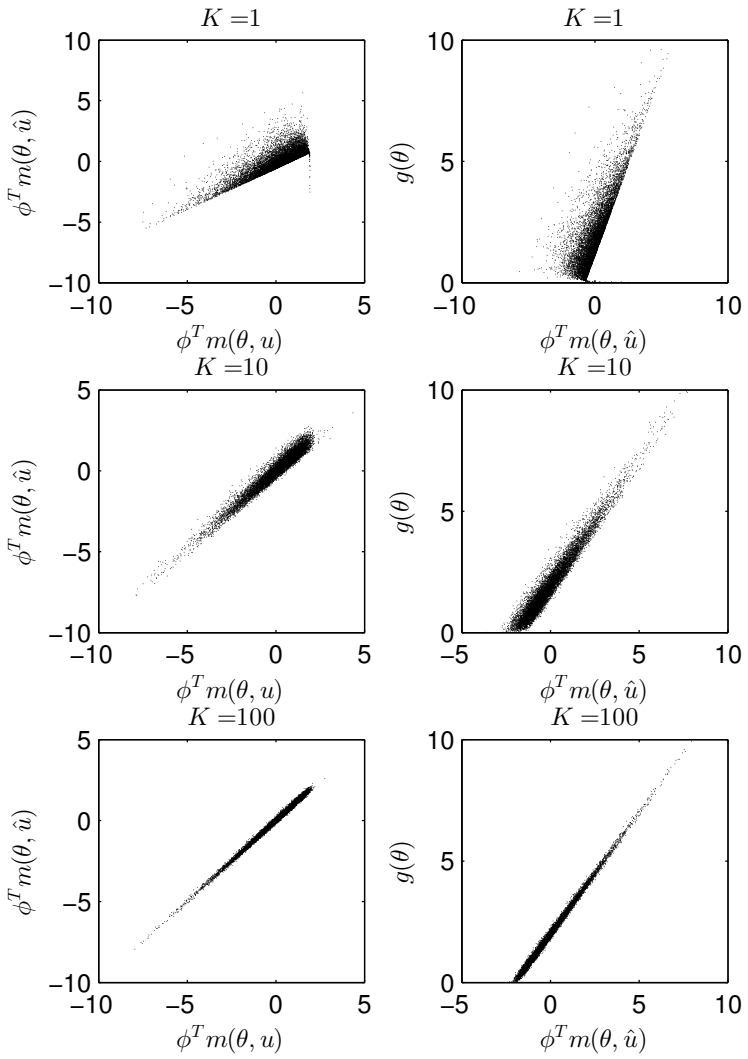

Figure 1: Tractable exponential example. Here we plot the best linear combination $\boldsymbol{\phi}^T \boldsymbol{m}(\theta, \hat{u})$ against (i) the best linear combination $\boldsymbol{\phi}^T \boldsymbol{m}(\theta, u)$ (left column) and (ii) the target function $g(\theta)$ (right column), for values of $K = 1, 10, 100$. [For the simulation we used $I = 10,000$ Monte Carlo samples, data $y = 1$ and polynomial trial functions of degree $\deg(P) = 2$.]

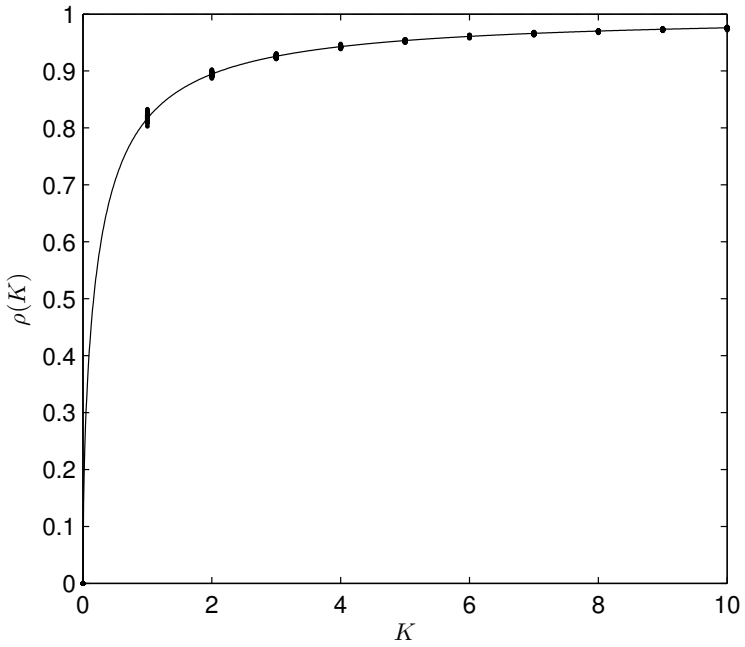

Figure 2: Tractable exponential example. We ran 100 simulations and in each case computed the correlation $\rho(K)$ between the target function $g(\theta) = \theta$ and the fitted control variate $\boldsymbol{\phi}^T \boldsymbol{m}(\theta, \hat{u})$ based on K forward-simulations. Results, shown here as dots, show that correlation is quite high already when $K = 1$, supporting the theoretical result that $K = 1$ is optimal for serial computation. The solid curve shows a fit of $\rho(K) = \left(\frac{K}{K+C}\right)^{1/2}$ with $C = 1/2$, in line with the theory. As $K \to \infty$ we have $\rho(K) \to \rho(\infty)$ and in this case $\rho(\infty) = 1$ since the zero-variance estimator is exact. [For the simulation we used $I = 10,000$ Monte Carlo samples, data $y = 1$ and polynomial trial functions of degree $\deg(P) = 2$.]

$\text{Deg}(P) = 1$	Std	$K = 1$	$K = 10$	$K = 100$	ZV
$I = 100$	0.14 ± 0.0015	0.13 ± 0.0014	0.12 ± 0.0017	0.12 ± 0.0019	0.12 ± 0.0019
$I = 500$	0.063 ± 0.00028	0.059 ± 0.00026	0.056 ± 0.00037	0.056 ± 0.00041	0.056 ± 0.00041
$I = 1000$	0.044 ± 0.00015	0.042 ± 0.00019	0.041 ± 0.00022	0.04 ± 0.00021	0.04 ± 0.00021
$I = 5000$	0.02 ± 3.1e-05	0.019 ± 3.8e-05	0.019 ± 5.4e-05	0.018 ± 6.6e-05	0.018 ± 6.6e-05
$I = 10000$	0.014 ± 1.4e-05	0.014 ± 2.2e-05	0.013 ± 3.4e-05	0.013 ± 4e-05	0.013 ± 4.1e-05
$\times\sqrt{IK}$		1.3 ± 0.0093	4 ± 0.052	13 ± 0.16	
$\text{Deg}(P) = 2$	Std	$K = 1$	$K = 10$	$K = 100$	ZV
$I = 100$	0.14 ± 0.0014	0.086 ± 0.0014	0.048 ± 0.0021	0.04 ± 0.0026	0.038 ± 0.0028
$I = 500$	0.062 ± 0.00027	0.036 ± 0.00019	0.015 ± 0.00018	0.0075 ± 0.00031	0.0055 ± 0.00038
$I = 1000$	0.044 ± 0.00015	0.025 ± 9.9e-05	0.01 ± 5.7e-05	0.0045 ± 0.00013	0.0029 ± 0.00018
$I = 5000$	0.02 ± 3e-05	0.011 ± 1.9e-05	0.0044 ± 7.5e-06	0.0015 ± 1.5e-05	0.00056 ± 3.3e-05
$I = 10000$	0.014 ± 1.5e-05	0.008 ± 1e-05	0.0031 ± 3.6e-06	0.0011 ± 6.2e-06	0.0003 ± 1.7e-05
$\times\sqrt{IK}$		0.81 ± 0.011	1.1 ± 0.1	1.8 ± 0.55	
$\text{Deg}(P) = 3$	Std	$K = 1$	$K = 10$	$K = 100$	ZV
$I = 100$	0.14 ± 0.0014	0.091 ± 0.0022	0.059 ± 0.0029	0.045 ± 0.0036	0.043 ± 0.0037
$I = 500$	0.062 ± 0.00028	0.036 ± 0.0002	0.016 ± 0.00025	0.0093 ± 0.0003	0.008 ± 0.00034
$I = 1000$	0.044 ± 0.00015	0.025 ± 9e-05	0.01 ± 5.9e-05	0.0051 ± 0.00019	0.0037 ± 0.00023
$I = 5000$	0.02 ± 2.9e-05	0.011 ± 1.7e-05	0.0044 ± 8.4e-06	0.0016 ± 1.5e-05	0.00067 ± 2.9e-05
$I = 10000$	0.014 ± 1.8e-05	0.0079 ± 1.1e-05	0.0031 ± 3.5e-06	0.0011 ± 7.4e-06	0.00042 ± 1.8e-05
$\times\sqrt{IK}$		0.82 ± 0.022	1.2 ± 0.17	2.1 ± 0.63	

Table 1: Tractable exponential example. Comparing the standard errors of Monte Carlo estimators, including the default estimator ("Std"; the regular MCMC estimator with no variance reduction), the reduced-variance estimator with $K = 1$, 10 and 100, and the ZV estimator. [The ZV estimate has some non-zero variance here because, in practice, the optimal coefficients $\boldsymbol{\phi}(y)^*$ must be estimated using Monte Carlo.]

3.2 Example 2: Ising model

In this section we model the spatial distribution of binary variables, Y_j, 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. We will focus on the classical Ising model where the random variable $\boldsymbol{Y} \in \{-1, 1\}^{n \times n}$ has likelihood defined in terms of a single sufficient statistic

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

where the notation $i \sim j$ means that the lattice point i is a neighbour of lattice point j. 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 $\{y_{i-m}, y_{i-1}, y_{i+1}, y_{i+m}\}$. Each point along the edges of the lattice has either 2 or 3 neighbours. The likelihood for this model takes the form of a GRF where the partition function

$$\mathfrak{P}(\theta) = \sum_{\boldsymbol{y}' \in \{-1,1\}^{n \times n}} \exp(\theta s(\boldsymbol{y}')) \tag{53}$$

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

In the experiments below we consider lattices of size $n = 16$, about the limit for exact solution, and focus on estimating the posterior mean $\mu = \mathbb{E}_{\theta|\boldsymbol{y}}[\theta]$ 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 $\boldsymbol{y}$ 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 allow 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.

Fig. 3 below displays the MCMC trace plots, obtained using the exchange algorithm, for $g(\theta) = \theta$ (blue) and the reduced-variance version $g(\theta) + \hat{\boldsymbol{\phi}}^T \boldsymbol{m}(\theta, \hat{u})$. Trace plots are presented for increasing values of $K \in \{1, 20, 100, 500\}$ and using degree-one (red) and degree-two (green) polynomials.[1] It is evident that as K increases, the Monte Carlo variance of the controlled random variable decreases; indeed when $K = 500$ the variance is dramatically reduced compared to the (uncontrolled) MCMC samples of θ. These findings are summarised in Table 2. 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 for larger values of K. These results present a powerful approach to exploit multi-core processing architectures to deliver a real-time acceleration in the convergence of MCMC estimators.

[1] For 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.

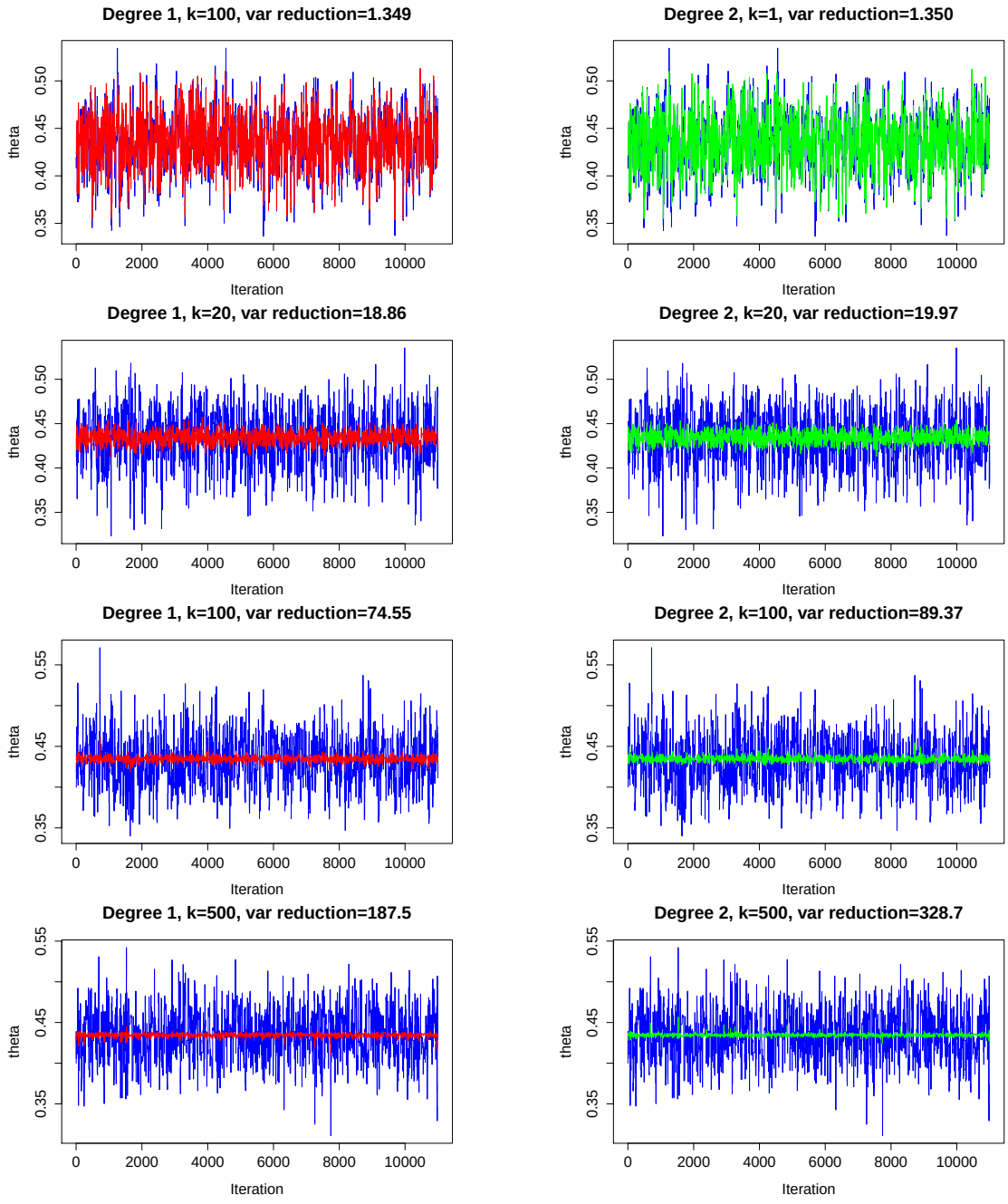

Figure 3: Ising model: As the number of forward-simulations, K, increases, the precision of the controlled estimate of the posterior mean for θ improves. The degree-two polynomial yields greater precision compared to the degree-one polynomial, particularly for larger values of K.

	$K=1$	$K=20$	$K=100$	$K=500$
$\hat{\mu}$	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 2: 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.

3.3 Example 3: Exponential random graph models

The exponential random graph (ERG) model is defined on a random adjacency matrix $\boldsymbol{Y}$ of a graph on n nodes (or actors) and a set of edges (dyadic relationships) $\{Y_{ij} : i = 1, \ldots, n; j = 1, \ldots, n\}$ where $Y_{ij} = 1$ if the pair (i, j) is connected by an edge, and $Y_{ij} = 0$ otherwise. An edge connecting a node to itself is not permitted so $Y_{ii} = 0$. The dyadic variables in an ERG may be undirected, whereby $Y_{ij} = Y_{ji}$ for each pair (i, j), 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.

ERG models are now widely used in social network analysis; see Robins *et al.* (2014) for an introduction. The likelihood of an observed network $\boldsymbol{y}$ is modelled in terms of a collection of sufficient statistics $\boldsymbol{s}(\boldsymbol{y}) = (s_1(\boldsymbol{y}), \ldots, s_k(\boldsymbol{y}))$ and corresponding parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. For example, typical statistics include $s_1(y) = \sum_{i<j} y_{ij}$ and $s_2(y) = \sum_{i<j<k} y_{ik}y_{jk}$ which are, 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 $k > 2$. Specifically, the likelihood takes the form of a GRF where the partition function

$$\mathfrak{P}(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}' \in \mathcal{G}(n)} \exp(\boldsymbol{\theta}^T \boldsymbol{s}(\boldsymbol{y}')) \tag{54}$$

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.

In experiment below we consider the *Gamaneg* network (Read, 1954), displayed in Fig. 4, 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 with two sufficient statistics, where $s_1(y)$ counts the total number

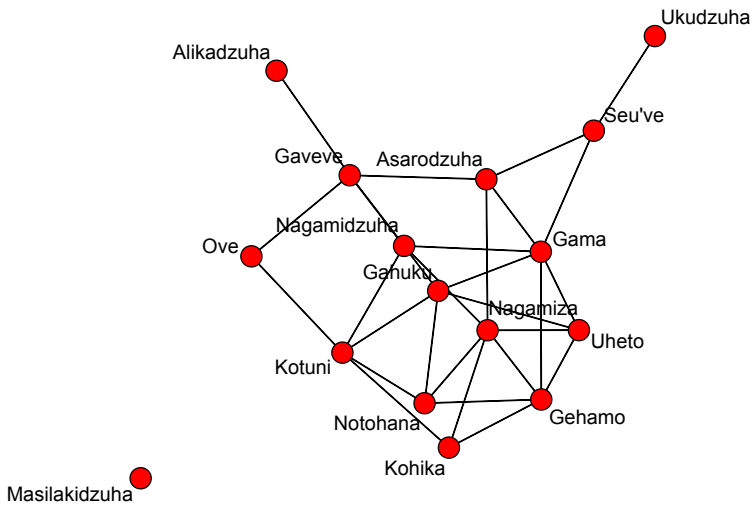

Figure 4: Gamaneg graph. The vertices represent 16 sub-tribes of the Eastern central highlands of New Guinea and edges represent an antagonistic relationship between two sub-tribes.

of observed edges and the two-star statistic $s_2(y)$ is also as defined above. The prior distributions for θ_1 and θ_2 were both set to be independent $N(0, 5^2)$ distributions, from which is 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 in this setting (Friel, 2013), making it well-suited to our purposes.

Again we focus on the challenge of estimating the posterior mean $\boldsymbol{\mu} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[\boldsymbol{\theta}]$, in this case performing independent estimation with $g(\boldsymbol{\theta}) = \theta_j$ for $j = 1$ and $j = 2$. 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 reduced variance 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.[2] Fig. 5 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. 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 2 times variance reduction; this is fully compatible with our methodology and if combined, should yield a further reduction in variance.

[2]For 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.

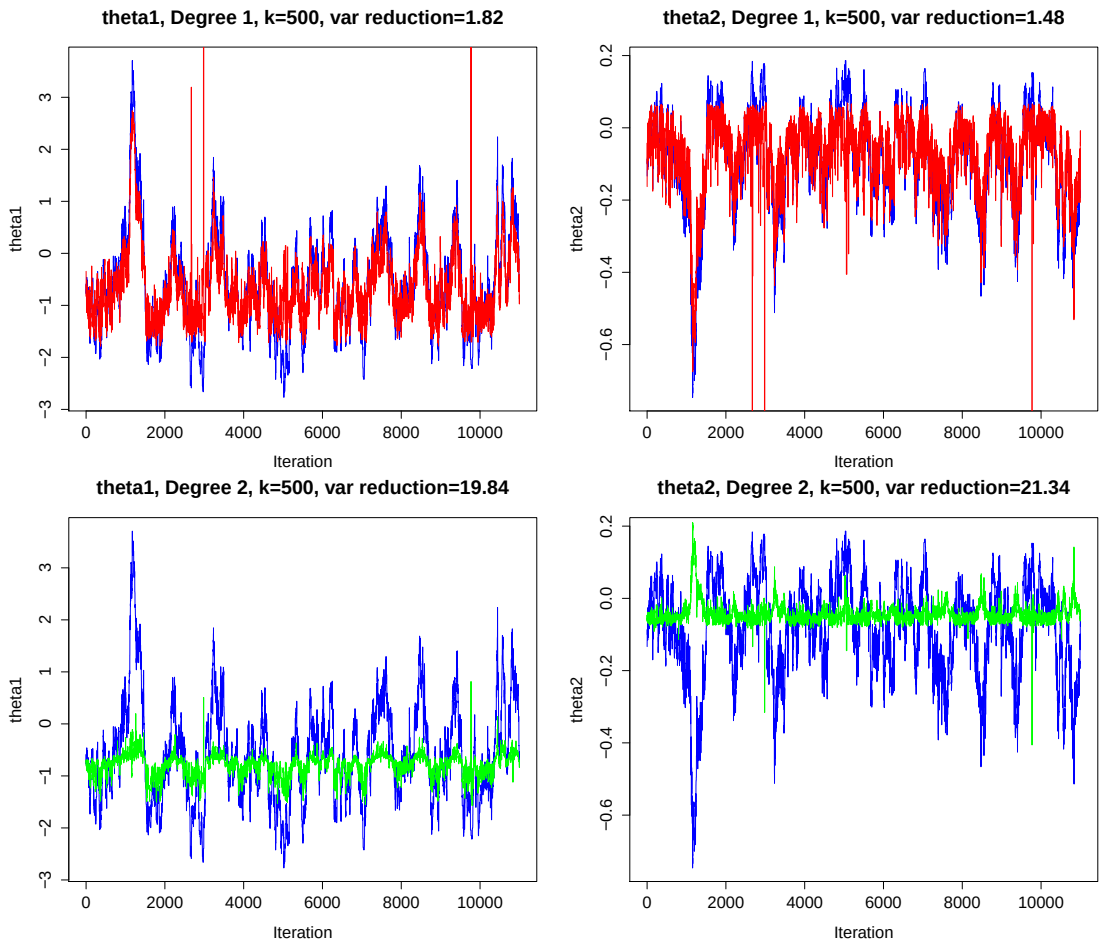

Figure 5: Exponential random graph model: The top row displays the trace plot for θ_1 and θ_2 in uncontrolled (blue) and controlled versions for a degree-one (red) polynomial, while the bottom row is similar but for a degree-two (green) polynomial.

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). 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). A general stochastic diffusion is defined as

$$d\boldsymbol{X}(t) = \boldsymbol{\alpha}(\boldsymbol{X}(t);\boldsymbol{\theta})dt + \boldsymbol{\beta}^{1/2}(\boldsymbol{X}(t);\boldsymbol{\theta})d\boldsymbol{W}(t), \qquad \boldsymbol{X}(0) = \boldsymbol{X}_0, \tag{55}$$

where $\boldsymbol{X}(t)$ is a stochastic process taking values in $\mathbb{R}^d$, $\boldsymbol{\alpha} : \mathbb{R}^d \times \Theta \to \mathbb{R}^d$ is a drift function, $\boldsymbol{\beta} : \mathbb{R}^d \times \Theta \to \mathbb{R}^d \times \mathbb{R}^d$ is a diffusion function, $\boldsymbol{W}(t)$ is a d-dimensional Weiner process, $\boldsymbol{\theta} \in \Theta$ are unknown model parameters and $\boldsymbol{X}_0 \in \mathbb{R}^d$ is a known initial state.

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, we introduce a fine discretisation $t_1, \ldots, t_T$ of time with mesh size δt. Write $\boldsymbol{X}_i = \boldsymbol{X}(t_i)$. We use the Euler-Maruyama approximation to the SDE likelihood, so that

$$p(\boldsymbol{X}|\boldsymbol{\theta}) \propto \prod_{i=2}^{T} \mathcal{N}(\boldsymbol{X}_i|\boldsymbol{X}_{i-1} + \boldsymbol{\alpha}_i\delta t, \boldsymbol{\beta}_i\delta t) \tag{56}$$

where we have used the shorthand $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}(\boldsymbol{X}_{i-1};\boldsymbol{\theta})$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}(\boldsymbol{X}_{i-1};\boldsymbol{\theta})$. We partition $\boldsymbol{X} = [\boldsymbol{X}^o\ \boldsymbol{X}^u]$ such that $\boldsymbol{y} = \boldsymbol{X}^o$ are observed (for simplicity without noise) and $\boldsymbol{x} = \boldsymbol{X}^u$ are unobserved. This is a model that exhibits a Type II intractability and we thus estimate the score using

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{y}) \approx \frac{1}{K}\sum_{k=1}^{K} \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \boldsymbol{x}^{(k)}|\boldsymbol{y}) \tag{57}$$

where $\boldsymbol{x}^{(k)}$ are independent samples from $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. 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 $p(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, we have that

$$\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y}) = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{X}|\boldsymbol{\theta}). \tag{58}$$

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

$$\nabla_{\theta_j} \log p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{i=2}^{T} \begin{array}{l} -\frac{1}{2}\mathrm{tr}(\boldsymbol{\beta}_i^{-1}\nabla_{\theta_j}\boldsymbol{\beta}_i) + (\nabla_{\theta_j}\boldsymbol{\alpha}_i)^T\boldsymbol{\beta}_i^{-1}(\boldsymbol{X}_i - \boldsymbol{X}_{i-1} - \boldsymbol{\alpha}_i\delta t) \\ +\frac{1}{2\delta t}(\boldsymbol{X}_i - \boldsymbol{X}_{i-1} - \boldsymbol{\alpha}_i\delta t)^T\boldsymbol{\beta}_i^{-1}(\nabla_{\theta_j}\boldsymbol{\beta}_i)\boldsymbol{\beta}_i^{-1}(\boldsymbol{X}_i - \boldsymbol{X}_{i-1} - \boldsymbol{\alpha}_i\delta t) \end{array} \tag{59}$$

Consider the specific example of the Susceptible-Infected-Recovered (SIR) model from epidemiology, that has a stochastic representation given by

$$\boldsymbol{\alpha}(\boldsymbol{X};\boldsymbol{\theta}) = \begin{bmatrix} -\theta_1 X_1 X_2 \\ \theta_1 X_1 X_2 - \theta_2 X_2 \end{bmatrix}, \qquad \boldsymbol{\beta}(\boldsymbol{X};\boldsymbol{\theta}) = \frac{1}{N}\begin{bmatrix} \theta_1 X_1 X_2 & -\theta_1 X_1 X_2 \\ -\theta_1 X_1 X_2 & \theta_1 X_1 X_2 + \theta_2 X_2 \end{bmatrix} \tag{60}$$

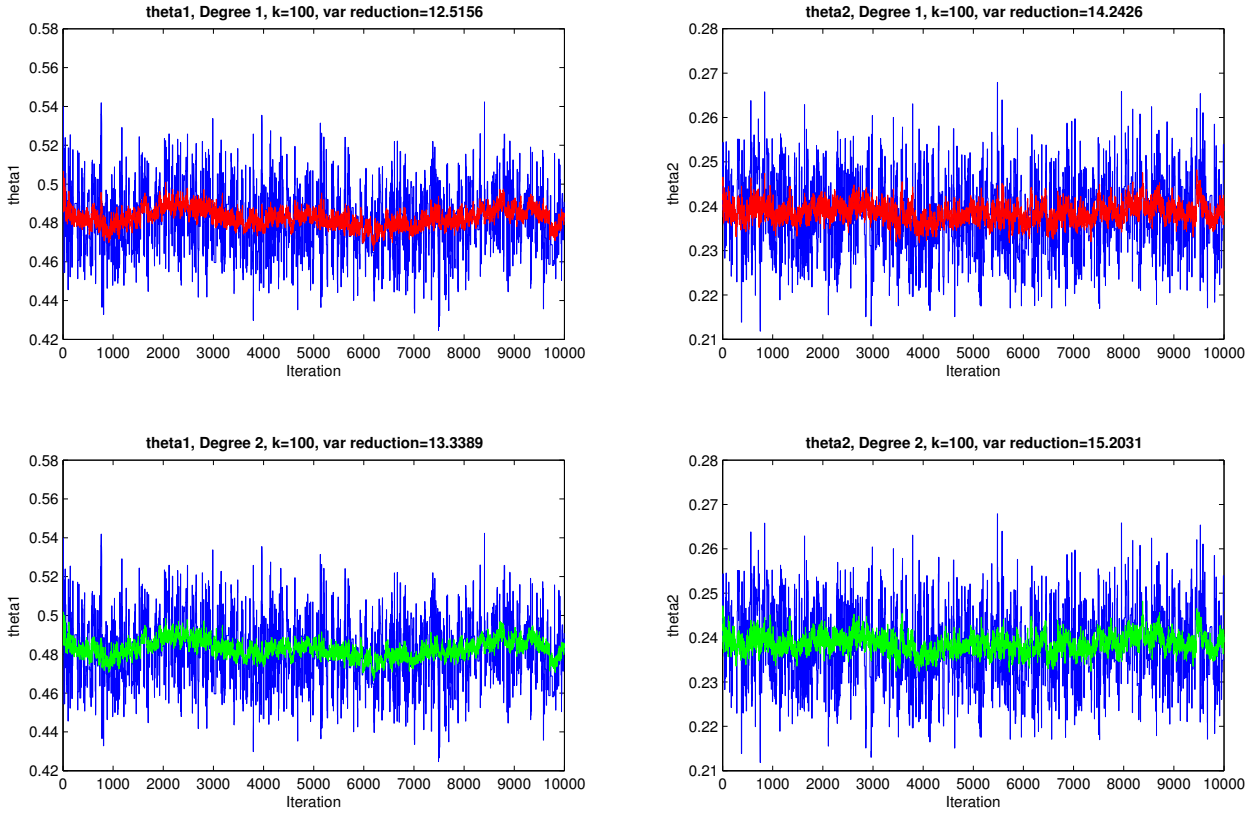

Figure 6: SIR model. The top row displays the trace plots for θ_1 and θ_2 in uncontrolled (blue) and controlled (red) versions for a degree-one polynomial. The bottom row is similar, but for degree-two polynomials.

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 $\boldsymbol{X}_0 = [0.99, 0.01]$, 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 $\boldsymbol{X}^u$.

Fig. 6 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-score 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 the proposed methodology in spatial statistics, social network analysis and inference for latent data models such as SDEs. Importantly, the "reduced variance" methodology provides a straight-forward means to leverage multi-core architectures for Bayesian estimation, that compliments recent work for MCMC in this direction by Alquier *et al.* (2014); Maclaurin and Adams (2014); Angelino *et al.* (2014); Korattikara *et al.* (2014); Bardenet *et al.* (2014).

Our theoretical analysis revealed that just $K = 1$ forward-simulation provides the optimal variance reduction per unit (serial) computation. Thus a reduced-variance estimator can leverage the simulation stage of the exchange algorithm, or the sampling stage of the pseudo-marginal algorithm, to achieve variance reduction with essentially no additional computational effort, making it a default procedure. Furthermore, when multi-core processing architectures are available, additional variance reduction can be achieved (per unit time) and it was shown that the proposed 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.

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 e.g. Girolami and Calderhead, 2011). This would retain the inherently parallel nature of the simulation procedure and also have interesting connections with the recent work of Alquier *et al.* (2014) where it is shown, in a similar manner to this paper, that forward-simulation from the likelihood model can yield useful approximate Monte Carlo schemes that converge to the target posterior 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) without any additional computational cost. 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. Another 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 $\hat{\boldsymbol{u}}(\boldsymbol{\theta}|\boldsymbol{y})$ is likely to have a low variance. A final direction for further research would be to move beyond independent estimation of the score $\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y})$ for each value of $\boldsymbol{\theta}$; here nonparametric regression techniques could play a role and this should yield further reductions in estimator variance.

Acknowledgements 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]. CJO was supported by the EPSRC Centre for Research in Statistical Methodology [EP/D002060/1]. The authors thank Mark Girolami for encouraging us to pursue this line of research.

References

Alquier, P., Friel, N., Everitt, R., and Boland, A. (2014). "Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels." arXiv:1403.5496.

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.

Andrieu, C., and Roberts, G. O. (2009). "The pseudo-marginal approach for efficient Monte Carlo computations." *The Annals of Statistics*, 37(2):697-725.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). "Particle Markov chain Monte Carlo (with Discussion)." *Journal of the Royal Statistical Society, Series B*, 72(3):269-342.

Angelino, E., Kohler, E., Waterland, A., Seltzer, M., and Adams, R. P. (2014). "Accelerating MCMC via Parallel Predictive Prefetching." arXiv:1403.7265.

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.

Atchadé, Y., Lartillot, N., and Robert, C. (2013). "Bayesian computation for intractable normalizing constants." *Brazilian Journal of Probability and Statistics*, 27(3):417-436.

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.

Besag, J. E. (1972). "Nearest-neighbour systems and the auto-logistic model for binary data." *Journal of the Royal Statistical Society, Series B*, 34(1):697-725.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006). "Exact and computationally efficient likelihoodbased estimation for discretely observed diffusion processes (with discussion)." *Journal of the Royal Statistical Society, Series B*, 68(3):333-382.

Caimo, A., and Friel, N. (2011). "Bayesian inference for exponential random graph models." *Social Networks*, 33:41-55.

Caimo, A., and Friel, N. (2013). "Bayesian model selection for exponential random graph models." *Social Networks*, 35:11-24.

Caimo, A., and Friel, N. (2014). "Bergm: Bayesian inference for exponential random graphs using R." *Journal of Statistical Software*, to appear.

Caimo, A., and Mira, A. (2014). "Efficient computational strategies for Bayesian social networks." *Statistics and Computing*, to appear.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2009). "Statistical modelling of spatial extremes." *Statistical Science*, 27:161-186.

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*, 74(1):133-161.

Everitt, R. (2012). "Bayesian parameter estimation for latent Markov random fields and social networks." *Journal of Computational and graphical Statistics.*, 21(4):940-960.

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*, 50(2):201-220.

Friel, N., and Rue, H. (2007). "Recursive computing and simulation-free inference for general factorizable models." *Biometrika*, 94:661-672.

Friel, N. (2013). "Estimating the evidence for Gibbs random fields." *Journal of Computational and Graphical Statistics*, 22:518-532.

Fuchs, C. (2013). *Inference for Diffusion Processes with Applications in Life Sciences.* Springer, Heidelberg.

Girolami, M., and Calderhead, B. (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods." *Journal of the Royal Statistical Society, Series B*, 73(2):1-37.

Glasserman, P. (2004). *Monte Carlo methods in financial engineering.* Springer, New York.

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.

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*, 54(3):657-699.

Hammer, H., and Tjelmeland, H. (2008). "Control variates for the Metropolis-Hastings algorithm." *Scandinavian Journal of Statistics* 35(3):400-414.

Kendall, P. C., and Bourne, D. E. (1992). "Vector analysis and Cartesian tensors (3rd ed.)." CRC Press, Florida.

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.

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.

Liang, F. (2010). "A double MetropolisHastings sampler for spatial models with intractable normalizing constants." *Journal of Statistical Computation and Simulation* 80(9):1007-1022.

Lyne, A. M., Girolami, M., Atchade, Y., Strathmann, H., and Simpson, D. (2013). "Playing Russian Roulette with Intractable Likelihoods." arXiv:1306.4032.

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.

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.

Mira, A., Möller, J., and Roberts, G. O. (2001). "Perfect Slice Samplers." *Journal of the Royal Statistical Society, Series B*, 63(3):593-606.

Mira, A., Tenconi, P., and Bressanini, D. (2003). "Variance reduction for MCMC." Technical Report 2003/29, Universitá degli Studi dell' Insubria, Italy.

Mira, A., Solgi, R., and Imparato, D. (2013). "Zero Variance Markov Chain Monte Carlo for Bayesian Estimators." *Statistics and Computing* 23(5):653-662.

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.

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.

Oates, C. J., Papamarkou, T., and Girolami, M. (2014). "The Controlled Thermodynamic Integral for Bayesian Model Comparison." *Journal of the American Statistical Association*, to appear.

Oates, C. J., Dondelinger, F., Bayani, N., Korkola, J., Gray, J. W., and Mukherjee, S. (2014). "Causal network inference using biochemical kinetics." *Bioinformatics*, to appear.

Papamarkou, T., Mira, A., and Girolami, M. (2014). "Zero Variance Differential Geometric Markov Chain Monte Carlo Algorithms." *Bayesian Analysis*, 9(1):97-128.

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.

Rao, V., Lin, L., and Dunson, D. (2014). "Data augmentation for models based on rejection sampling." arXiv:1406.6652.

Read, K. E. (1954). "Cultures of the Central Highlands, New Guinea." *Southwestern Journal of Anthropology* 10(1):1-43.

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.

Rubinstein, R. Y., and Marcus, R. (1985). "Efficiency of Multivariate Control Variates in Monte Carlo Simulation." *Operations Research*, 33(3):661-677.

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*, 71(2):319-392.

Sherlock, C., Thiery, A., Roberts, G. O., and Rosenthal, J. S. (2014). "On the efficiency of pseudo-marginal random walk Metropolis algorithm." arXiv 1309.7209.

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.

West, M., and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models (2nd ed.)*. Springer-Verlag, New York.