Statistical Science
2015, Vol. 30, No. 4, 443–467
DOI: 10.1214/15-STS523

On Russian Roulette Estimates for Bayesian Inference with Doubly-Intractable Likelihoods[1]

Anne-Marie Lyne, Mark Girolami, Yves Atchadé, Heiko Strathmann and Daniel Simpson

Abstract. A large number of statistical models are "doubly-intractable": the likelihood normalising term, which is a function of the model parameters, is intractable, as well as the marginal likelihood (model evidence). This means that standard inference techniques to sample from the posterior, such as Markov chain Monte Carlo (MCMC), cannot be used. Examples include, but are not confined to, massive Gaussian Markov random fields, autologistic models and Exponential random graph models. A number of approximate schemes based on MCMC techniques, Approximate Bayesian computation (ABC) or analytic approximations to the posterior have been suggested, and these are reviewed here. Exact MCMC schemes, which can be applied to a subset of doubly-intractable distributions, have also been developed and are described in this paper. As yet, no general method exists which can be applied to all classes of models with doubly-intractable posteriors.

In addition, taking inspiration from the Physics literature, we study an alternative method based on representing the intractable likelihood as an infinite series. Unbiased estimates of the likelihood can then be obtained by finite time stochastic truncation of the series via Russian Roulette sampling, although the estimates are not necessarily positive. Results from the Quantum Chromodynamics literature are exploited to allow the use of possibly negative estimates in a pseudo-marginal MCMC scheme such that expectations with respect to the posterior distribution are preserved. The methodology is reviewed on well-known examples such as the parameters in Ising models, the posterior for Fisher–Bingham distributions on the d-Sphere and a large-scale Gaussian Markov Random Field model describing the Ozone Column data. This leads to a critical assessment of the strengths and weaknesses of the methodology with pointers to ongoing research.

Key words and phrases: Intractable likelihood, Russian Roulette sampling, Monte Carlo methods, pseudo-marginal MCMC.

Anne-Marie Lyne is a Ph.D. student, Department of Statistical Science, University College London, London WC1E 6BT, United Kingdom (e-mail: lyne.annemarie3@gmail.com). Mark Girolami is Professor, Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom (e-mail: m.girolami@warwick.ac.uk). Yves Atchadé is Associate Professor, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, USA (e-mail: yvesa@umich.edu). Heiko Strathmann is a Ph.D. student, Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, United Kingdom (e-mail: heiko.strathmann@gmail.com). Daniel Simpson is a CRiSM Fellow, Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom (e-mail: dp.simpson@gmail.com).

[1]Code to replicate all results reported can be downloaded from http://www.ucl.ac.uk/roulette.

1. INTRODUCTION

An open problem of growing importance in the application of Markov chain Monte Carlo (MCMC) methods for Bayesian computation is the definition of transition kernels for posterior distributions with intractable data densities. In this paper, we focus on methods for a subset of these distributions known as *doubly-intractable* distributions, a term first coined by Murray, Ghahramani and MacKay (2006). To illustrate what constitutes a doubly-intractable posterior, take some data $\mathbf{y} \in \mathcal{Y}$ used to make posterior inferences about the variables $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that define a statistical model. A prior distribution defined by a density $\pi(\boldsymbol{\theta})$ with respect to Lebesgue measure $d\boldsymbol{\theta}$ is adopted and the data density is given by $p(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y};\boldsymbol{\theta})/\mathcal{Z}(\boldsymbol{\theta})$, where $f(\mathbf{y};\boldsymbol{\theta})$ is an unnormalised function of the data and parameters, and $\mathcal{Z}(\boldsymbol{\theta}) = \int f(\mathbf{x};\boldsymbol{\theta})\,d\mathbf{x}$ is the likelihood normalising term which *cannot be computed*. The posterior density follows in the usual form as

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{y}) &= \frac{p(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})}{p(\mathbf{y})} \\ &= \frac{f(\mathbf{y};\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta})} \times \pi(\boldsymbol{\theta}) \times \frac{1}{p(\mathbf{y})},\end{aligned} \tag{1.1}$$

where $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$. "Doubly-intractable" refers to the fact that not only is $p(\mathbf{y})$ intractable (this is common in Bayesian inference and does not generally present a problem for inference), but $\mathcal{Z}(\boldsymbol{\theta})$ is also intractable.

Bayesian inference proceeds by taking posterior expectations of functions of interest, that is,

$$E_{\pi(\boldsymbol{\theta}|\mathbf{y})}\{h(\boldsymbol{\theta})\} = \int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta} \tag{1.2}$$

and Monte Carlo estimates of the above expectations can be obtained by employing MCMC methods if other exact sampling methods are not available (Gilks, 1996, Robert and Casella, 2010, Liu, 2001, Gelman et al., 1995). To construct a Markov chain with invariant distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$, the Metropolis–Hastings algorithm can be used; a transition kernel is constructed by designing a proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ and accepting the proposed parameter value with probability

$$\begin{aligned}\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) &= \min\left\{1, \frac{\pi(\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\} \\ &= \min\left\{1, \frac{f(\mathbf{y};\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}'|\boldsymbol{\theta})}{f(\mathbf{y};\boldsymbol{\theta})\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \times \frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')}\right\}.\end{aligned} \tag{1.3}$$

Clearly, a problem arises when the value of the normalising term for the data density, $\mathcal{Z}(\boldsymbol{\theta})$, cannot be obtained either due to it being nonanalytic or uncomputable with a finite computational resource. This situation is far more widespread in modern-day statistical applications than a cursory review of the literature would suggest and forms a major challenge to methodology for computational statistics currently (e.g., Møller et al., 2006, Besag and Moran, 1975, Besag, 1974, Green and Richardson, 2002, Møller and Waagepetersen, 2004). We review and study methods which have been published in the Statistics and Physics literature for dealing with such distributions. We then study in detail how to implement a pseudo-marginal MCMC scheme (Beaumont, 2003, Andrieu and Roberts, 2009) in which an unbiased estimate of the target density is required at each iteration, and suggest how these might be realised.

This paper is organised as follows. In Section 2 we describe examples of doubly-intractable distributions along with current inference approaches. These encompass both approximate and exact methods which have been developed in the Statistics, Epidemiology and Image analysis literature. In Section 3 we suggest an alternative approach based on pseudo-marginal MCMC (Beaumont, 2003, Andrieu and Roberts, 2009) in which an unbiased estimate of the intractable target distribution is used in an MCMC scheme to sample from the exact posterior distribution. In Sections 4 and 5 we describe how to realise such unbiased estimates of a likelihood with an intractable normalising term. This is achieved by writing the likelihood as an infinite series in which each term can be estimated unbiasedly. Then Russian Roulette techniques are used to truncate the series such that only a finite number of terms need be estimated whilst maintaining the unbiasedness of the overall estimate. Sections 6 and 7 contain experimental results for posterior inference over doubly-intractable distributions: Ising models, the Fisher–Bingham distribution and a large-scale Gaussian Markov random field. Section 8 contains a discussion of the method and suggests areas for further work.

2. INFERENCE METHODS FOR DOUBLY-INTRACTABLE DISTRIBUTIONS

2.1 Approximate Bayesian Inference

Many models describing data with complex dependency structures are doubly-intractable. Examples which have received attention in the Statistics literature include:

1. The Ising model (Ising, 1925). Originally formulated in the Physics literature as a simple model for interacting magnetic spins on a lattice. Spins are binary random variables which interact with neighbouring spins.
2. The Potts model and autologistic models. Generalisations to the Ising model in which spins can take more than two values and more complex dependencies are introduced. These models are used in image analysis (Besag, 1986, Hughes, Haran and Caragea, 2011), as well as in other fields such as disease mapping (e.g., Green and Richardson, 2002).
3. Spatial point processes. Used to model point pattern data, for example, ecological data (e.g., Silvertown and Antonovics, 2001, Møller and Waagepetersen, 2004) and epidemiological data (e.g., Diggle, 1990).
4. Exponential Random Graph (ERG) models. Used in the field of social networks to analyse global network structures in terms of local graph statistics such as the number of triangles (e.g., Goodreau, Kitts and Morris, 2009).
5. Massive Gaussian Markov random field (GMRF) models. Used in image analysis and spatial statistics, amongst others (e.g., Rue and Held, 2005).

Standard Bayesian inference techniques such as drawing samples from the posterior using MCMC cannot be used due to the intractability of the likelihood normalising term, and hence a number of approximate inference methods have been developed. A common approach when the full likelihood cannot be computed is to use a pseudo-likelihood (Besag, 1974, Besag and Moran, 1975), in which an approximation to the true likelihood is formed using the product of the conditional probabilities for each variable. This can normally be computed efficiently and can therefore replace the full likelihood in an otherwise standard inference strategy to sample from the posterior (e.g., Heikkinen and Hogmander, 1994, Zhou and Schmidler, 2009). This approach scales well with the size of the data and can give a reasonable approximation to the true posterior, but inferences may be significantly biased as long range interactions are not taken into account [this has been shown to be the case for ERG models (Van Duijn, Gile and Handcock, 2009), hidden Markov random fields (Friel et al., 2009) and autologistic models (Friel and Pettitt, 2004)]. Methods based on composite likelihoods have also been used for inference in massive scale GMRF models, in which an approximation to the likelihood is based on the joint density of spatially adjacent blocks (Eidsvik et al., 2014). This has the advantage that the separate parts of the likelihood cannot only be computed more efficiently, but also computed in parallel.

Another pragmatic approach is that of Green and Richardson (2002), in which they discretise the interaction parameter in the Potts model to a grid of closely spaced points and then set a prior over these values. Estimates of the normalising term are then precomputed using thermodynamic integration (as described by Gelman and Meng, 1998) so that no expensive computation is required during the MCMC run. This allowed inference to be carried out over a model for which it would not otherwise have been possible. However, it is not clear what impact this discretisation and use of approximate normalising terms has on parameter inference and it seems preferable, if possible, to retain the continuous nature of the variable and to not use approximations unless justified.

Approximate Bayesian Computation (ABC) (Marin et al., 2012), a technique developed for likelihood free inference (Tavaré et al., 1997, Beaumont, Zhang and Balding, 2002), can also be used. The types of models for which ABC was originally developed are implicit, meaning data can be simulated from the likelihood but the likelihood cannot be written down, and hence neither standard maximum likelihood nor Bayesian methods can be used. For doubly-intractable distributions, it is only the normalising term which cannot be computed, but we can still use the techniques developed in the ABC community. ABC in its simplest form proceeds by proposing an approximate sample from the joint distribution, $p(\mathbf{y}, \boldsymbol{\theta})$, by first proposing $\boldsymbol{\theta}'$ from the prior and then generating a data set from the model likelihood conditional on $\boldsymbol{\theta}'$. This data set is then compared to the observed data and the proposed parameter value accepted if the generated data is "similar" enough to the observed data. An obvious drawback to the method is that it does not sample from the exact posterior, although it has been shown to produce comparable results to other approximate methods and recent advances mean that it can be scaled up to very large data sets (Grelaud, Robert and Marin, 2009, Everitt, 2012, Moores, Mengersen and Robert, 2014).

The "shadow prior" method of Liechty, Liechty and Müller (2009) is an interesting attempt to reduce the computational burden of intractable normalising constants in the case where constraints on the data or parameters cause the intractability. As an example, take data $\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\theta})$ which is constrained to lie in some set A. Depending on the form of A, sampling from the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ can be hindered by an intractable likelihood normalising term. The model is

therefore replaced by $p(\mathbf{y}|\boldsymbol{\delta})I(\mathbf{y} \in A)$, "shadow prior" $p(\boldsymbol{\delta}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \mathcal{N}(\delta_i; \theta_i, \nu)$ and prior $\pi(\boldsymbol{\theta})$, for some ν and where d is the dimension of $\boldsymbol{\theta}$. The conditional posterior $p(\boldsymbol{\theta}|\boldsymbol{\delta}, \mathbf{y}) = p(\boldsymbol{\theta}|\boldsymbol{\delta})$ no longer requires the computation of an intractable normalising term (as dependence on the constrained data has been removed), although updating $\boldsymbol{\delta}$ does. However, this has been reduced to d one-dimensional problems which may be simpler to deal with. The method, of course, only works if the computational burden of the intractable normalising constant is significantly less in the shadow prior format than in the original model, and several examples of when this might be the case are suggested, such as when the parameter in the normalising constant has a complicated hyperprior structure. An approximate version can be implemented in which the normalising constant is ignored in the shadow prior, which can sometimes have very little impact on the final inference. In these cases the computational burden has been eliminated.

Several approximate but consistent algorithms have been developed based on Monte Carlo approximations within MCMC methods. For example, an approach was developed by Atchadé, Lartillot and Robert (2013) in which a sequence of transition kernels are constructed using a consistent estimate of $\mathcal{Z}(\boldsymbol{\theta})$ from the Wang–Landau algorithm (Wang and Landau, 2001). The estimates of the normalising term converge to the true value as the number of iterations increases and the overall algorithm gives a consistent approximation to the posterior. Bayesian Stochastic Approximation Monte Carlo (Jin and Liang, 2014) works in a similar fashion, sampling from a series of approximations to the posterior using the stochastic approximation Monte Carlo algorithm (Liang, Liu and Carroll, 2007), which is based on the Wang–Landau algorithm. These algorithms avoid the need to sample from the model likelihood, but in practice suffer from the curse of dimensionality as the quality of the importance sampling estimate depends on the number and location of the grid points. These points need to grow exponentially with the dimension of the space limiting the applicability of this methodology. They also require a significant amount of tuning to attain good approximations to the normalising term, and hence ensure convergence is achieved.

Alternative methodologies have avoided sampling altogether and instead used deterministic approximations to the posterior distribution. This is particularly the case for GMRF models which often have complex parameter dependencies and are very large in scale, rendering MCMC difficult to apply. INLA (integrated nested Laplace approximations) (Rue, Martino and Chopin, 2009) was designed to analyse latent Gaussian models and has been applied to massive GMRFs in diverse areas such as spatio-temporal disease mapping (Schrödle and Held, 2011) and point processes describing the locations of muskoxen (Illian et al., 2012). By using Laplace approximations to the posterior and an efficient programming implementation, fast Bayesian inference can be carried out for large models. However, this benefit also constitutes a drawback in that users must rely on standard software, and therefore model extensions which could be tested simply when using an MCMC approach are not easy to handle. Further, it is of course necessary to ensure that the assumptions inherent in the method apply so that the approximations used are accurate. It should also be noted that the work of Taylor and Diggle (2014) found that in the case of spatial prediction for log-Gaussian Cox processes, an MCMC method using the Metropolis-adjusted Langevin Algorithm (MALA) algorithm gave comparable results in terms of predictive accuracy and was actually slightly more efficient than the INLA method. Other approximations have also been developed as part of a large body of work in the area, such as iterative methods for approximating the log determinant of large sparse matrices, required to compute the likelihood (Aune, Simpson and Eidsvik, 2014).

2.2 Exact MCMC Methods

As well as approximate inference methods, a small number of exact algorithms have been developed to sample from doubly-intractable posteriors. These are described below as well as advice as to when these algorithms can be used.

2.2.1 *Introducing auxiliary variables*. An exact sampling methodology for doubly-intractable distributions is proposed in Walker (2011), which uses a similar approach to those described in Adams, Murray and MacKay (2009) and Section 9 of Beskos et al. (2006). A Reversible-Jump MCMC (RJMCMC) sampling scheme is developed that cleverly gets around the intractable nature of the normalising term. Consider the univariate distribution $p(y|\boldsymbol{\theta}) = f(y; \boldsymbol{\theta})/\mathcal{Z}(\boldsymbol{\theta})$ where N i.i.d. observations, y_i, are available. In its most general form, it is required that y belongs to some bounded interval $[a, b]$, and that there exists a constant $M < +\infty$ such that $f(y; \boldsymbol{\theta}) < M$ for all $\boldsymbol{\theta}$ and y (it is assumed that $[a, b] = [0, 1]$, and $M = 1$ in the following exposition). The method introduces auxiliary variables $\nu \in (0, \infty)$, $k \in \{0, 1, \ldots\}$, $\{s\}^{(k)} = (s_1, \ldots, s_k)$,

to form the joint density

$$f(\nu, k, \{s\}^{(k)}, \mathbf{y}|\boldsymbol{\theta}) \propto \frac{\exp(-\nu)\nu^{k+N-1}}{k!} \cdot \prod_{j=1}^{k}(1 - f(s_j;\boldsymbol{\theta}))\mathbb{1}(0 < s_j < 1) \cdot \prod_{i=1}^{N} f(y_i;\theta).$$

Integrating out ν and $s^{(k)}$ and summing over all k returns the data distribution $\prod_{i=1}^{N} p(y_i|\boldsymbol{\theta})$. An RJMCMC scheme is proposed to sample from the joint density $f(\nu, k, \{s\}^{(k)}, \mathbf{y}|\boldsymbol{\theta})$ and this successfully gets around the intractable nature of the normalising term. The scheme has been used to sample from the posterior of a Bingham distribution (Walker, 2014).

However, the methodology has some limitations to its generality. Firstly, the unnormalised density function must be strictly bounded from above to ensure the positivity of the terms in the first product. This obviously limits the generality of the methodology to the class of strictly bounded functions; however, this is not overly restrictive, as many functional forms for $f(y_i;\boldsymbol{\theta})$ are bounded, for example, when there is finite support, or when $f(y_i;\boldsymbol{\theta})$ takes an exponential form with strictly negative argument. Even if the function to be sampled is bounded, finding bounds that are tight is extremely difficult and the choice of the bound directly impacts the efficiency of the sampling scheme constructed; see, for example, Ghaoui and Gueye (2009) for bounds on binary lattice models. Ideally we would wish to relax the requirement for the data, $\mathbf{y}$, to belong to a bounded interval, but if we integrate with respect to each s_j over an unbounded interval, then we can no longer return $1 - \mathcal{Z}(\boldsymbol{\theta})$ and the sum over k will therefore no longer define a convergent geometric series equaling $\mathcal{Z}(\boldsymbol{\theta})$. This last requirement particularly restricts the generality and further use of this specific sampling method for intractable distributions.

2.3 Valid Metropolis–Hastings-Type Transition Kernels

An ingenious MCMC solution to the doubly-intractable problem was proposed by Møller et al. (2006) in which the posterior state space is extended as follows:

$$\pi(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) \propto p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta})\frac{f(\mathbf{y};\boldsymbol{\theta})}{\mathcal{Z}(\theta)}.$$

This extended distribution retains the posterior as a marginal. The method proceeds by taking the proposal for $\mathbf{x}, \boldsymbol{\theta}$ to be $q(\mathbf{x}', \boldsymbol{\theta}'|\mathbf{x}, \boldsymbol{\theta}) = \frac{f(\mathbf{x};\boldsymbol{\theta}')}{\mathcal{Z}(\boldsymbol{\theta}')}q(\boldsymbol{\theta}'|\boldsymbol{\theta})$, so that at each iteration the intractable normalising terms cancel in the Metropolis–Hastings acceptance ratio. A drawback of the algorithm is the need to choose the marginal for $\mathbf{x}$, $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, particularly as the authors suggest that ideally this distribution would approximate the likelihood, thereby reintroducing the intractable normalising term.

Murray, Ghahramani and MacKay (2006) simplified and extended the algorithm to the Exchange algorithm, and in the process removed this difficulty by defining a joint distribution as follows:

$$p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\theta}') \propto \frac{f(\mathbf{y};\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta})}\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})\frac{f(\mathbf{x};\boldsymbol{\theta}')}{\mathcal{Z}(\boldsymbol{\theta}')}.$$

At each iteration, MCMC proceeds by first Gibbs sampling $\boldsymbol{\theta}'$ and $\mathbf{x}$, and then proposing to swap the values of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ using Metropolis–Hastings. Again, the intractable normalising terms cancel in the acceptance ratio. Both of these algorithms use only valid MCMC moves and therefore target the *exact* posterior, rendering them a major methodological step forward. However, they both require the capability to sample from the likelihood using a method such as perfect sampling (Propp and Wilson, 1996, Kendall, 2005). This can be considered a restriction to the widespread applicability of this class of methods, as for many models it is not possible, for example, the ERG model in social networks. Even when perfect sampling is possible, for example, for the Ising and Potts models, it becomes prohibitively slow as the size of the model increases. Attempts have been made to relax the requirement to perfectly sample by instead using an auxiliary Markov chain to sample approximately from the model at each iteration (Caimo and Friel, 2011, Liang, 2010, Everitt, 2012, Alquier et al., 2014). In particular, the paper by Alquier et al. (2014) suggests multiple approximate MCMC algorithms for doubly-intractable distributions and then applies results from Markov chain theory to bound the total variation distance between the approximate chains and a hypothetical exact chain. These types of approximate algorithms were in use due to their computational feasibility, and so it is pleasing to see some theoretical justification for their use emerging in the Statistics literature.

3. AN ALTERNATIVE APPROACH USING PSEUDO-MARGINAL MCMC

As has been seen, there are many approximate methods for sampling from doubly-intractable posteriors.

There are also exact methods available, but these can only be applied when it is possible to perfectly sample from the data model. Now we would like to approach the question of whether it is possible to relax this requirement and develop methodology for exact sampling of the posterior when perfect sampling is not possible. To do this, we develop an approach based on the pseudo-marginal methodology (Beaumont, 2003, Andrieu and Roberts, 2009, Doucet, Pitt and Kohn, 2012), and hence we now briefly review the algorithm. The pseudo-marginal class of methods is particularly appealing in that they have the least number of restrictions placed upon them and provide the most general and extensible MCMC methods for intractable distributions. They are sometimes referred to as Exact-approximate methods, based on the property that the invariant distribution of the Markov chain produced is the exact target distribution despite the use of an approximation in the Metropolis–Hastings acceptance probability. To use the scheme, an unbiased and positive estimate of the target density is substituted for the true density, giving an acceptance probability of the form

$$\begin{aligned}\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) &= \min\left\{1, \frac{\hat{\pi}(\boldsymbol{\theta}'|\mathbf{y})}{\hat{\pi}(\boldsymbol{\theta}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\} \\ &= \min\left\{1, \frac{\hat{p}(\mathbf{y}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')}{\hat{p}(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\},\end{aligned} \tag{3.1}$$

where the estimate at each proposal is propagated forward as described in Beaumont (2003), Andrieu and Roberts (2009). For the case of doubly-intractable distributions, assuming the prior is tractable, this equates to a requirement for an unbiased estimate of the likelihood as seen on the right in (3.1) above. The remarkable feature of this scheme is that the corresponding transition kernel has an invariant distribution with $\boldsymbol{\theta}$-marginal given precisely by the desired posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{y})$. To see this, denote all the random variables generated in the construction of the likelihood estimator by the vector $\mathbf{u}$ and its density $p(\mathbf{u})$. These random variables are, for example, those used when generating and accepting a proposal value in a Markov chain as part of a Sequential Monte Carlo estimate. The estimator of the likelihood is denoted $\hat{p}_N(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$, with N symbolising, for example, the number of Monte Carlo samples used in the estimate. The estimator of the likelihood must be unbiased, that is,

$$\int \hat{p}_N(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u}) p(\mathbf{u})\, d\mathbf{u} = p(\mathbf{y}|\boldsymbol{\theta}). \tag{3.2}$$

A joint density for $\boldsymbol{\theta}$ and $\mathbf{u}$ is now defined which returns the posterior distribution after integrating over $\mathbf{u}$:

$$\begin{aligned}\pi_N(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) &\propto \hat{p}_N(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})\pi(\boldsymbol{\theta}) p(\mathbf{u}) \\ &= \frac{\hat{p}_N(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})\pi(\boldsymbol{\theta}) p(\mathbf{u})}{p(\mathbf{y})}.\end{aligned}$$

It is simple to show using equation (3.2) that $\pi_N(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ integrates to 1 and has the desired marginal distribution for $\boldsymbol{\theta}|\mathbf{y}$. Now consider sampling from $\pi_N(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ using the Metropolis–Hastings algorithm, with the proposal distribution for $\mathbf{u}'$ being $p(\mathbf{u}')$. In this case the densities for $\mathbf{u}$ and $\mathbf{u}'$ cancel and we are using the acceptance probability in (3.1). Hence, this algorithm samples from $\pi_N(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ and the samples of $\boldsymbol{\theta}$ obtained are distributed according to the posterior.

This is a result that was highlighted in the statistical genetics literature (Beaumont, 2003), then popularised and formally analysed in Andrieu and Roberts (2009) with important developments such as Particle MCMC (Doucet, Pitt and Kohn, 2012) proving to be extremely powerful and useful in a large class of statistical models. Due to its wide applicability, the pseudo-marginal algorithm has been the subject of several recent papers in the statistical literature, increasing understanding of the methodology. These have covered how to select the number of samples in the unbiased estimate to minimise the computational time (Doucet, Pitt and Kohn, 2012), optimal variance and acceptance rates to maximise efficiency of the chain (Sherlock et al., 2015) and results to order two different pseudo-marginal implementations in terms of the acceptance probability and asymptotic variance (Andrieu and Vihola, 2014). It is interesting to note that the problem of Exact-Approximate inference was first considered in the Quantum Chromodynamics literature almost thirty years ago. This was motivated by the need to reduce the computational effort of obtaining values for the strength of bosonic fields in defining a Markov process to simulate configurations following a specific law; see, for example Kennedy and Kuti (1985), Bhanot and Kennedy (1985), Bakeyev and De Forcrand (2001), Lin, Liu and Sloan (2000), Joo, Horvath and Liu (2003).

3.1 Proposed Methodology

One can exploit the pseudo-marginal algorithm to sample from the posterior, and hence we require unbiased estimates of the likelihood. For each $\boldsymbol{\theta}$ and $\mathbf{y}$, we show that one can construct random variables $\{V_{\boldsymbol{\theta}}^{(j)}, j \geq 0\}$ (where dependence on $\mathbf{y}$ is omitted) such

that the series defined as

$$\pi(\boldsymbol{\theta}, \{V_{\boldsymbol{\theta}}^{(j)}\}|\mathbf{y}) := \sum_{j=0}^{\infty} V_{\boldsymbol{\theta}}^{(j)}$$

is finite almost surely, has finite expectation, and $\mathbb{E}(\pi(\boldsymbol{\theta}, \{V_{\boldsymbol{\theta}}^{(j)}\}|\mathbf{y})) = \pi(\boldsymbol{\theta}|\mathbf{y})$. We propose a number of ways to construct such series. Although unbiased, these estimators are not practical, as they involve infinite series. We therefore employ a computationally feasible truncation of the infinite sum which, crucially, remains unbiased. This is achieved using Russian Roulette procedures well known in the Physics literature (Hendricks and Booth, 1985, Carter and Cashwell, 1975). More precisely, we introduce a random time $\tau_{\boldsymbol{\theta}}$, such that with $\mathbf{u} := (\tau_{\boldsymbol{\theta}}, \{V_{\boldsymbol{\theta}}^{(j)}, 0 \le j \le \tau_{\boldsymbol{\theta}}\})$ the estimate

$$\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) := \sum_{j=0}^{\tau_{\boldsymbol{\theta}}} V_{\boldsymbol{\theta}}^{(j)}$$

satisfies

$$\mathbb{E}\big(\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})|\{V_{\boldsymbol{\theta}}^{(j)}, j \ge 0\}\big) = \sum_{j=0}^{\infty} V_{\boldsymbol{\theta}}^{(j)}.$$

As in the notation used above, $\mathbf{u}$ is a vector of all the random variables used in the unbiased estimate, that is, those used to estimate terms in the series, as well as those used in the roulette methods to truncate the series. As the posterior is only required up to a normalising constant in $\mathbf{y}$ and the prior is assumed tractable, in reality we require an unbiased estimate of the likelihood.

3.2 The Sign Problem

If the known function $f(\mathbf{y}; \boldsymbol{\theta})$ forming the estimate of the target is bounded, then the whole procedure can proceed without difficulty, assuming the bound provides efficiency of sampling. However, in the more general situation where the function is not bounded, there is a complication here in that the unbiased estimate $\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ is not guaranteed to be positive (although its expectation is nonnegative). This issue prevents us from plugging in directly the estimator $\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ in the pseudo-marginal framework for the case of unbounded functions. The problem of such unbiased estimators returning negative valued estimates turns out to be a well-studied issue in the Quantum Monte Carlo literature; see, for example, Lin, Liu and Sloan (2000). The problem is known as the Sign Problem,[2] which in its most general form is NP-hard (nondeterministic polynomial time hard) (Troyer and Wiese, 2005) and at present no general and practical solution is available. Indeed, recent work by Jacob and Thiery (2013) showed that given unbiased estimators of $\lambda \in \mathbb{R}$, no algorithm exists to yield an unbiased estimate of $f(\lambda) \in \mathbb{R}^+$, where f is a nonconstant real-valued function. Therefore, we will need to apply a different approach to this problem.

We follow Lin, Liu and Sloan (2000) and show that with a weighting of expectations it is still possible to compute any integral of the form $\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}$ by Markov chain Monte Carlo.

Suppose that we have an unbiased, but not necessarily positive, estimate of the likelihood $\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$ and we wish to sample from $\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) = \hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})\pi(\boldsymbol{\theta})p(\mathbf{u})/p(\mathbf{y})$, where $p(\mathbf{y}) = \int\int p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})\pi(\boldsymbol{\theta})p(\mathbf{u})\,d\boldsymbol{\theta}\,d\mathbf{u}$ is an intractable normaliser. Although $\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ integrates to one, it is not a probability, as it is not necessarily positive. Define $\sigma(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u}) := \mathsf{sign}(\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u}))$, where $\mathsf{sign}(x) = 1$ when $x > 0$, $\mathsf{sign}(x) = -1$ if $x < 0$ and $\mathsf{sign}(x) = 0$ if $x = 0$. Furthermore, denote $|\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})|$ as the absolute value of the measure, then we have $\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u}) = \sigma(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})|\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})|$.

Suppose that we wish to compute the expectation

$$\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta} = \int\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})\,d\mathbf{u}\,d\boldsymbol{\theta}. \tag{3.3}$$

We can write the above integral as

$$\begin{aligned}
&\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta} \\
&\quad = \int\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})\,d\mathbf{u}\,d\boldsymbol{\theta} \\
&\quad = \frac{1}{p(\mathbf{y})}\int\int h(\boldsymbol{\theta})\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})\pi(\boldsymbol{\theta})p(\mathbf{u})\,d\mathbf{u}\,d\boldsymbol{\theta} \qquad (3.4)\\
&\quad = \frac{\int\int h(\boldsymbol{\theta})\sigma(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})|\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})|\pi(\boldsymbol{\theta})p(\mathbf{u})\,d\mathbf{u}\,d\boldsymbol{\theta}}{\int\int \sigma(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})|\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})|\pi(\boldsymbol{\theta})p(\mathbf{u})\,d\mathbf{u}\,d\boldsymbol{\theta}} \\
&\quad = \frac{\int\int h(\boldsymbol{\theta})\sigma(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})\breve{\pi}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})\,d\mathbf{u}\,d\boldsymbol{\theta}}{\int\int \sigma(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})\breve{\pi}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})\,d\mathbf{u}\,d\boldsymbol{\theta}},
\end{aligned}$$

where $\breve{\pi}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ is the distribution

$$\breve{\pi}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) := \frac{|\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})|\pi(\boldsymbol{\theta})p(\mathbf{u})}{\int\int |\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})|\pi(\boldsymbol{\theta})p(\mathbf{u})\,d\mathbf{u}\,d\boldsymbol{\theta}}.$$

[2] Workshops devoted to the Sign Problem, for example, the International Workshop on the Sign Problem in QCD and Beyond, are held regularly, http://www.physik.uni-regensburg.de/sign2012/.

We can sample from $\breve{\pi}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ using a pseudo-marginal scheme. At each iteration we propose a new value $\boldsymbol{\theta}'$, generate an unbiased estimate of the likelihood $p(\mathbf{y}|\boldsymbol{\theta}', \mathbf{u}')$, and accept it with probability

$$\min\left\{1, \frac{|\hat{p}(\mathbf{y}|\boldsymbol{\theta}', \mathbf{u}')|\pi(\boldsymbol{\theta}')}{|\hat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})|\pi(\boldsymbol{\theta})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\},$$

remembering to save the sign of the accepted estimate. We can then use Monte Carlo to estimate the expectation in (3.3) using (3.4) with

$$\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} = \frac{\sum_{i=1}^{N} h(\boldsymbol{\theta}_i)\sigma(\mathbf{y}|\boldsymbol{\theta}_i, \mathbf{u}_i)}{\sum_{i=1}^{N} \sigma(\mathbf{y}|\boldsymbol{\theta}_i, \mathbf{u}_i)}. \tag{3.5}$$

The output of this MCMC procedure gives an importance-sampling-type estimate for the desired expectation $\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}$, which is consistent but biased (as with estimates from all MCMC methods). Importantly, this methodology gives us freedom to use unbiased estimators which may occasionally return negative estimates. We describe the procedure more systematically in the Appendix (Section B), and we discuss in particular how to compute the effective sample size of the resulting Monte Carlo estimate.

The following section addresses the issue of constructing the unbiased estimator to be used in the overall MCMC scheme.

4. PSEUDO-MARGINAL MCMC FOR DOUBLY-INTRACTABLE DISTRIBUTIONS

The foundational component of pseudo-marginal MCMC is the unbiased and positive estimator of the target density. In the methodology developed here, it is not essential for the estimate of the intractable distribution to be strictly positive and we exploit this characteristic. Note that whilst there are many methods for unbiasedly estimating $\mathcal{Z}(\boldsymbol{\theta})$, such as importance sampling, Sequential Monte Carlo (SMC) (Del Moral, Doucet and Jasra, 2006) and Annealed Importance Sampling (AIS) (Neal, 2001), if we then take some nonlinear function of the estimate, for example, the reciprocal, the overall estimate of the likelihood is no longer unbiased.

It is possible to directly construct an estimator of $1/\mathcal{Z}(\boldsymbol{\theta})$ using an instrumental density $q(\mathbf{y})$ as follows:

$$\frac{1}{\mathcal{Z}(\boldsymbol{\theta})} = \frac{1}{\mathcal{Z}(\boldsymbol{\theta})}\int q(\mathbf{y})\, d\mathbf{y} = \int \frac{q(\mathbf{y})}{f(\mathbf{y};\theta)} p(\mathbf{y}|\boldsymbol{\theta})\, d\mathbf{y}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} \frac{q(\mathbf{y}_i)}{f(\mathbf{y}_i;\theta)}, \quad \mathbf{y}_i \sim p(\cdot|\boldsymbol{\theta});$$

however, this requires the ability to sample from the likelihood, and if we can do this, then we can implement the Exchange algorithm. Further, the variance of the estimate depends strongly on the choice of the instrumental density. A biased estimator can be constructed by sampling the likelihood using MCMC (e.g., Zhang et al., 2012), but a pseudo-marginal scheme based on this estimate will not target the correct posterior distribution. Very few methods to estimate $1/\mathcal{Z}(\boldsymbol{\theta})$ can be found in the Statistics or Physics literature, presumably because in most situations a consistent estimate will suffice. Therefore, we have to look for other ways to generate an unbiased estimate of the likelihood.

In outline, the intractable distribution is first written in terms of a nonlinear function of the nonanalytic/computable normalising term. For example, in equation (1.1), the nonlinear function is the reciprocal $1/\mathcal{Z}(\boldsymbol{\theta})$, and an equivalent representation would be $\exp(-\log \mathcal{Z}(\boldsymbol{\theta}))$. This function is then represented by a convergent Maclaurin expansion which has the property that each term can be estimated unbiasedly using the available unbiased estimates of $\hat{\mathcal{Z}}(\boldsymbol{\theta})$. The infinite series expansion is then stochastically truncated without introducing bias so that only a finite number of terms need be computed. These two components—(1) unbiased independent estimates of the normalising constant, and (2) unbiased stochastic truncation of the infinite series representation—then produce an unbiased, though not strictly positive, estimate of the intractable distribution. The final two components of the overall methodology consist of (3) constructing an MCMC scheme which targets a distribution proportional to the absolute value of the unbiased estimator, and then (4) computing Monte Carlo estimates with respect to the desired posterior distribution as detailed in the previous section.

This method has its roots in several places in the Statistics and Physics literature. In the Physics literature, researchers used a similar method to obtain unbiased estimates of $\exp(-U(x))$ when only unbiased estimates of $U(x)$ were available (Kennedy and Kuti, 1985, Bhanot and Kennedy, 1985). They further showed that even when using such unbiased estimates in place of the true value, detailed balance still held. The method for realising the unbiased estimates at each iteration is also similar to that suggested by Booth (2007), in which he described a method for unbiasedly estimating the reciprocal of an integral, which is of obvious relevance to our case. In the Statistics literature, Douc and Robert (2011) used a geometric

series to estimate an inverse probability, and Beskos et al. (2006), Fearnhead, Papaspiliopoulos and Roberts (2008) also used techniques to truncate a series unbiasedly in their work on likelihood estimation for stochastic diffusions. Finally, both Rhee and Glynn (2012) and McLeish (2011) use roulette methods to realise an unbiased estimate when only biased but consistent estimates are available. This is achieved by writing the quantity to be unbiasedly estimated as an infinite series in which each term is a function of the consistent estimates which can be generated, and then truncating the series using roulette methods.

In the following sections, we study two series expansions of a doubly-intractable likelihood, in which each term can be estimates unbiasedly using unbiased estimates of $\mathcal{Z}(\boldsymbol{\theta})$. Following this comes a description of unbiased truncation methods.

4.1 Geometric Series Estimator

In the following discussion we show how the intractable likelihood can be written as a geometric series in which each term can be estimated unbiasedly. Take a biased estimate of the likelihood $\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y};\boldsymbol{\theta})/\widetilde{\mathcal{Z}}(\boldsymbol{\theta})$, where $\widetilde{\mathcal{Z}}(\boldsymbol{\theta}) > 0$ is ideally an upper bound on $\mathcal{Z}(\boldsymbol{\theta})$ or, alternatively, an unbiased importance sampling estimate or a deterministic approximation. Then, using a multiplicative correction

$$p(\mathbf{y}|\boldsymbol{\theta}) = \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta})\left[1 + \sum_{n=1}^{\infty} \kappa(\boldsymbol{\theta})^n\right], \tag{4.1}$$

where $\kappa(\boldsymbol{\theta}) = 1 - c(\boldsymbol{\theta})\mathcal{Z}(\boldsymbol{\theta})/\widetilde{\mathcal{Z}}(\boldsymbol{\theta})$ and $c(\boldsymbol{\theta})$ ensures $|\kappa(\boldsymbol{\theta})| < 1$, the convergence of a geometric series gives

$$\begin{aligned}\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta})\left[1 + \sum_{n=1}^{\infty} \kappa(\boldsymbol{\theta})^n\right] &= \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times \frac{c(\boldsymbol{\theta})}{1 - \kappa(\boldsymbol{\theta})} \\ &= \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times \frac{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta})} \\ &= p(\mathbf{y}|\boldsymbol{\theta}).\end{aligned}$$

Based on this equality, and with an infinite number of independent unbiased estimates of $\mathcal{Z}(\boldsymbol{\theta})$ each denoted $\hat{\mathcal{Z}}_i(\boldsymbol{\theta})$, an unbiased estimate of the target density is

$$\begin{aligned}\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}) = &\frac{\pi(\boldsymbol{\theta})\tilde{p}(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})} \\ &\cdot c(\boldsymbol{\theta})\left[1 + \sum_{n=1}^{\infty}\prod_{i=1}^{n}\left(1 - c(\boldsymbol{\theta})\frac{\hat{\mathcal{Z}}_i(\boldsymbol{\theta})}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}\right)\right].\end{aligned} \tag{4.2}$$

Notice that the series in (4.2) is finite a.s. and we can interchange summation and expectation if

$$E\left(\left|1 - c(\boldsymbol{\theta})\frac{\hat{\mathcal{Z}}_i(\boldsymbol{\theta})}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}\right|\right) < 1.$$

Since $E(|X|) \le E^{1/2}(|X|^2)$, a sufficient condition for this is $0 < c(\boldsymbol{\theta}) < 2\widetilde{\mathcal{Z}}(\boldsymbol{\theta})\mathcal{Z}(\boldsymbol{\theta})/E(\hat{\mathcal{Z}}_1^2(\boldsymbol{\theta}))$, which is slightly more stringent than $|\kappa(\boldsymbol{\theta})| < 1$. Under this assumption, the expectation of $\hat{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is

$$\begin{aligned}&E\{\hat{\pi}(\boldsymbol{\theta}|\mathbf{y})|\widetilde{\mathcal{Z}}(\boldsymbol{\theta})\} \\ &\quad= \frac{\pi(\boldsymbol{\theta})\tilde{p}(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})} \\ &\qquad\cdot c(\boldsymbol{\theta})\left[1 + \sum_{n=1}^{\infty}\prod_{i=1}^{n}\left(1 - c(\boldsymbol{\theta})\frac{E\{\hat{\mathcal{Z}}_i(\boldsymbol{\theta})\}}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}\right)\right] \\ &\quad= \frac{\pi(\boldsymbol{\theta})\tilde{p}(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})} \times c(\boldsymbol{\theta})\left[1 + \sum_{n=1}^{\infty}\kappa(\boldsymbol{\theta})^n\right] \\ &\quad= \pi(\boldsymbol{\theta}|\mathbf{y}).\end{aligned}$$

Therefore, the essential property $E\{\hat{\pi}(\boldsymbol{\theta}|\mathbf{y})\} = \pi(\boldsymbol{\theta}|\mathbf{y})$ required for Exact-Approximate MCMC is satisfied by this geometric correction. However, there are difficulties with this estimator. It will be difficult in practice to find $c(\boldsymbol{\theta})$ that ensures the series in (4.2) is convergent in the absence of knowledge of the actual value of $\mathcal{Z}(\boldsymbol{\theta})$. By ensuring that $\widetilde{\mathcal{Z}}(\boldsymbol{\theta})/c(\boldsymbol{\theta})$ is a strict upper bound on $\mathcal{Z}(\boldsymbol{\theta})$, denoted by $\mathcal{Z}_U$, guaranteed convergence of the geometric series is established. Even if an upper bound is available, it may not be computationally practical, as upper bounds on normalising constants are typically loose (see, e.g., Ghaoui and Gueye, 2009), making the ratio $\mathcal{Z}(\boldsymbol{\theta})/\mathcal{Z}_U$ extremely small, and, therefore, $\kappa(\boldsymbol{\theta}) \approx 1$; in this case, the convergence of the geometric series will be slow. A more pragmatic approach is to use a pilot run at the start of each iteration to characterise the location and variance of the $\mathcal{Z}(\boldsymbol{\theta})$ estimates, and use this to conservatively select $\widetilde{\mathcal{Z}}(\boldsymbol{\theta})/c(\boldsymbol{\theta})$ such that the series converges. Of course, if the distribution of the estimates is not well enough characterised, then we may not be able to guarantee with probability 1 that $|\kappa(\boldsymbol{\theta})| < 1$, and hence approximation will be introduced into the chain.

In the next section we describe an alternative to the geometric series estimator which does not have the practical issue of ensuring the region of convergence is maintained.

4.2 Unbiased Estimators Using an Exponential Auxilliary Variable

In this section we show how the introduction of an auxiliary variable can enable the posterior density to be written in terms of a Taylor series expansion of the exponential function. The introduction of $\nu \sim \text{Expon}(\mathcal{Z}(\boldsymbol{\theta}))$ defines a joint distribution of the form of

$$\begin{aligned}\pi(\boldsymbol{\theta}, \nu|\mathbf{y}) &= \big[\mathcal{Z}(\boldsymbol{\theta})\exp\big(-\nu\mathcal{Z}(\boldsymbol{\theta})\big)\big] \\ &\quad\cdot \frac{f(\mathbf{y};\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta})} \times \pi(\boldsymbol{\theta}) \times \frac{1}{p(\mathbf{y})} \\ &= \exp\big(-\nu\mathcal{Z}(\boldsymbol{\theta})\big) \times f(\mathbf{y};\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times \frac{1}{p(\mathbf{y})} \\ &= \left[1 + \sum_{n=1}^{\infty} \frac{(-\nu\mathcal{Z}(\boldsymbol{\theta}))^n}{n!}\right] \\ &\quad\cdot f(\mathbf{y};\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times \frac{1}{p(\mathbf{y})}.\end{aligned}$$

Integrating over ν returns the posterior distribution and, therefore, if we sample from this joint distribution, our $\boldsymbol{\theta}$ samples will be distributed according to the posterior. As hinted at in the previous section, the methods used to truncate the series are more computationally feasible if the series converges quickly. Therefore, we introduce $\widetilde{\mathcal{Z}}(\boldsymbol{\theta})$, which is preferably an upper bound on $\mathcal{Z}(\boldsymbol{\theta})$ or, if unavailable, some other approximation. The exponential can then be expanded as follows:

$$\begin{aligned}\exp\big(-\nu\mathcal{Z}(\boldsymbol{\theta})\big) &= \exp\big(-\nu\widetilde{\mathcal{Z}}(\boldsymbol{\theta})\big) \\ &\quad\cdot \exp\big(\nu\big(\widetilde{\mathcal{Z}}(\boldsymbol{\theta}) - \mathcal{Z}(\boldsymbol{\theta})\big)\big) \\ &= \exp\big(-\nu\widetilde{\mathcal{Z}}(\boldsymbol{\theta})\big) \\ &\quad\cdot \left(1 + \sum_{n=1}^{\infty} \frac{\nu^n}{n!}\big(\widetilde{\mathcal{Z}}(\boldsymbol{\theta}) - \mathcal{Z}(\boldsymbol{\theta})\big)^n\right).\end{aligned}$$

If $\widetilde{\mathcal{Z}}(\boldsymbol{\theta})$ is an upper bound on $\mathcal{Z}(\boldsymbol{\theta})$, then its introduction prevents the terms in the Taylor series from alternating in sign by ensuring the exponent is positive; this helps to reduce the impact of returning negative estimates. Even if $\widetilde{\mathcal{Z}}(\boldsymbol{\theta})$ is not a strict upper bound, its presence reduces the absolute value of the exponent, which improves the convergence properties of the series, and therefore makes the truncation methods described in the next section more efficient.

An unbiased estimator of the series is

$$\begin{aligned}&\widehat{\exp\big(-\nu\mathcal{Z}(\boldsymbol{\theta})\big)} \\ &= \exp\big(-\nu\widetilde{\mathcal{Z}}(\boldsymbol{\theta})\big) \\ &\quad\cdot \left[1 + \sum_{n=1}^{\infty} \frac{\nu^n}{n!} \prod_{i=1}^{n}\big(\widetilde{\mathcal{Z}}(\boldsymbol{\theta}) - \hat{\mathcal{Z}}_i(\boldsymbol{\theta})\big)\right],\end{aligned} \tag{4.3}$$

where $\{\hat{\mathcal{Z}}_i(\boldsymbol{\theta}), i \geq 1\}$ are i.i.d. random variables with expectation equal to $\mathcal{Z}(\boldsymbol{\theta})$. The magnitude of the exponent can present computational barriers to the implementation of this scheme; if $\mathcal{Z}(\boldsymbol{\theta})$ is very large, it is easier to carry out the division $\hat{\mathcal{Z}}(\boldsymbol{\theta})/\mathcal{Z}(\boldsymbol{\theta})$ in (4.2) (which can be computed in log space) than the subtraction $\mathcal{Z}(\boldsymbol{\theta}) - \hat{\mathcal{Z}}(\boldsymbol{\theta})$ in (4.3). On the other hand, since $n!$ grows faster than the exponential, this series is always well defined (finite almost surely).

In Fearnhead, Papaspiliopoulos and Roberts (2008), the *Generalised Poisson Estimator*, originally proposed in Beskos et al. (2006), is employed to estimate transition functions that are similar to (4.3). Here again, this series is finite almost surely with finite expectation. The choice of which estimator to employ will be problem dependent and, in situations where it is difficult to guarantee convergence of the geometric series, this form of estimator may be more suitable.

In the following section, we discuss the final element of the proposed methodology: unbiased truncation of the infinite series estimators.

5. UNBIASED TRUNCATION OF INFINITE SUMS: RUSSIAN ROULETTE

Two unbiased estimators of nonlinear functions of a normalising constant have been considered. Both of them rely on the availability of an unbiased estimator for $\mathcal{Z}(\boldsymbol{\theta})$ and a series representation of the nonlinear function. We now require a computationally feasible means of obtaining the desired estimator without explicitly computing the infinite sum and without introducing any bias into the final estimate. It transpires that there are a number of ways to randomly truncate the convergent infinite sum $\mathcal{S}(\boldsymbol{\theta}) = \sum_{i=0}^{\infty}\phi_i(\boldsymbol{\theta})$ in an unbiased manner. These stem from work by von Neumann and Ulam in the 1940s; see Papaspiliopoulos (2011) for a good review of such methods.

5.1 Single Term Weighted Truncation

The simplest unbiased truncation method is to define a set of probabilities and draw an integer index k with probability q_k, then return $\phi_k(\theta)/q_k$ as the estimator. It is easy to see that the estimator is unbiased as $E\{\hat{\mathcal{S}}(\boldsymbol{\theta})\} = \sum_k q_k\phi_k(\boldsymbol{\theta})/q_k = \mathcal{S}(\boldsymbol{\theta})$. The definition of the probabilities should be chosen to minimise the variance of the estimator; see, for example, Fearnhead, Papaspiliopoulos and Roberts (2008). An example could be that each index is drawn from a Poisson distribution $k \sim \text{Poiss}(\lambda)$ with $q_k = \lambda^k\exp(-\lambda)/k!$. However, in the case of a geometric series where $\phi_k(\boldsymbol{\theta}) =$

$\phi^k(\boldsymbol{\theta})$, the variance of the estimator will be infinite with this choice since the combinatorial function $k!$ grows faster than the exponential. Using the geometric distribution as our importance distribution, the variance is finite subject to some conditions on the choice of p, the parameter of the geometric distribution. To see this, note that, as k is chosen with probability $q_k = p^k(1-p)$, the second moment $\mathbb{E}[\hat{S}^2] = \sum_{k=0}^{\infty} \hat{S}_k^2 q_k = \sum_{k=0}^{\infty} \phi_k^2 / p^k(1-p)$ is finite if $\lim_{k\to\infty} |\phi_{k+1}^2 / p\phi_k^2| < 1$.

5.2 Russian Roulette

An alternative unbiased truncation that exhibits superior performance in practice is based on a classic Monte Carlo scheme, known as Russian Roulette in the Physics literature (Lux and Koblinger, 1991, Carter and Cashwell, 1975). The procedure is based on the simulation of a finite random variable (stopping time) τ according to some probabilities $p_n = \mathbb{P}(\tau \geq n) > 0$ for all $n \geq 0$ with $p_0 = 1$. Define the weighted partial sums as $S_0 = \phi_0$ and for $k \geq 1$

$$S_k = \phi_0 + \sum_{j=1}^{k} \frac{\phi_j}{p_j}.$$

The Russian Roulette estimate of S is $\hat{S} = S_\tau$. Russian Roulette implementations in the Physics literature commonly choose a stopping time of the form

$$\tau = \inf\{k \geq 1 : U_k \geq q_k\},$$

where $\{U_j, j \geq 1\}$ are i.i.d. $\mathcal{U}(0,1)$, $q_j \in (0,1]$ and $\hat{S} = S_{\tau-1}$. In this case $p_n = \prod_{j=1}^{n-1} q_j$.

It can be shown that the expectation of the estimate is as required:

$$\begin{aligned}\sum_{k=0}^{n} S_k \mathbb{P}(\tau = k) &= \sum_{k=0}^{n} S_k (p_k - p_{k+1}) \\ &= \phi_0 + \sum_{k=0}^{n-1} S_{k+1} p_{k+1} - \sum_{k=0}^{n} S_k p_{k+1} \\ &= \sum_{k=0}^{n} \phi_k - S_n p_{n+1}.\end{aligned}$$

By Kronecker's lemma, $\lim_{n\to\infty} p_n S_n = 0$, and $|p_{n+1} S_n| = (p_{n+1}/p_n) p_n |S_n| \leq p_n |S_n| \to 0$, as $n \to \infty$. We conclude that $\mathbb{E}[\hat{S}(\boldsymbol{\theta})] = \sum_{k=0}^{\infty} S_k \mathbb{P}(\tau = k) = \sum_{k=0}^{\infty} \phi_k = S(\boldsymbol{\theta})$. We refer the reader to the Appendix for a more detailed discussion relating to the variance of such an estimator and how to design the sequence of probabilities (p_n).

Based on results presented in the Appendix, for a geometric series where $\phi_k(\boldsymbol{\theta}) = \phi^k(\boldsymbol{\theta})$, if one chooses $q_j = q$, then the variance will be finite provided $q > \phi(\boldsymbol{\theta})^2$. In general, there is a trade-off between the computing time of the scheme and the variance of the returned estimate. If the selected q_j's are close to unity, the variance is small, but the computing time is high. But if q_j's are close to zero, the computing time is fast, but the variance can be very high, possibly infinite. In the case of the geometric series, $\phi_k(\boldsymbol{\theta}) = \phi^k(\boldsymbol{\theta})$, choosing $q_j = q = \phi(\boldsymbol{\theta})$ works reasonably well in practice.

As an illustrative example, consider the joint density

$$\begin{aligned}p(\boldsymbol{\theta}, \nu, \mathbf{u}|\mathbf{y}) &= \exp\bigl(-\nu \widetilde{\mathcal{Z}}(\boldsymbol{\theta})\bigr) \\ &\quad \cdot \left(1 + \sum_{n=1}^{\tau_\theta} \frac{\nu^n}{q^n n!} \prod_{i=1}^{n} \bigl(\widetilde{\mathcal{Z}}(\boldsymbol{\theta}) - \hat{\mathcal{Z}}_i(\boldsymbol{\theta})\bigr)\right) \\ &\quad \cdot \frac{f(\mathbf{y};\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{y})},\end{aligned} \tag{5.1}$$

where the random variable $\mathbf{u}$ represents the random variables in the estimates $\hat{\mathcal{Z}}_i(\boldsymbol{\theta})$ and the random variable used in Russian Roulette truncation, and $q^n = \prod_{l=1}^{n} q_l$ denotes the probabilities in the Russian Roulette truncation. If we define a proposal for ν' as $q(\nu'|\boldsymbol{\theta}') = \widetilde{\mathcal{Z}}(\boldsymbol{\theta}') \exp(-\nu' \widetilde{\mathcal{Z}}(\boldsymbol{\theta}'))$ and a proposal for $\boldsymbol{\theta}'$ as $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$, then the Hastings ratio for a transition kernel with invariant density $\pi(\boldsymbol{\theta}, \nu, \mathbf{u}|\mathbf{y})$ follows as

$$\begin{aligned}&\frac{f(\mathbf{y};\boldsymbol{\theta}')}{f(\mathbf{y};\boldsymbol{\theta})} \times \frac{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta}')} \times \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})} \\ &\quad \cdot \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \times \phi(\nu, \nu', \boldsymbol{\theta}, \boldsymbol{\theta}'),\end{aligned} \tag{5.2}$$

where

$$\begin{aligned}&\phi(\nu, \nu', \boldsymbol{\theta}, \boldsymbol{\theta}') \\ &= \frac{1 + \sum_{m=1}^{\tau_{\theta'}} \frac{(\nu')^m}{q^m m!} \prod_{j=1}^{m} (\widetilde{\mathcal{Z}}(\boldsymbol{\theta}') - \hat{\mathcal{Z}}_j(\boldsymbol{\theta}'))}{1 + \sum_{n=1}^{\tau_\theta} \frac{\nu^n}{q^n n!} \prod_{i=1}^{n} (\widetilde{\mathcal{Z}}(\boldsymbol{\theta}) - \hat{\mathcal{Z}}_i(\boldsymbol{\theta}))}.\end{aligned} \tag{5.3}$$

It is interesting to note that $\phi(\nu, \nu', \boldsymbol{\theta}, \boldsymbol{\theta}')$ acts as a multiplicative correction for the Hastings ratio that uses the approximate normalising term $\widetilde{\mathcal{Z}}(\boldsymbol{\theta})$ rather than the actual $\mathcal{Z}(\boldsymbol{\theta})$. The required marginal $\pi(\boldsymbol{\theta}|\mathbf{y})$ follows due to the unbiased nature of the estimator.

The Russian Roulette methodology has been used in various places in the literature. McLeish (2011) and Rhee and Glynn (2012), Glynn and Rhee (2014) cleverly use the Russian Roulette estimator to "debias" a biased but consistent estimator. We would like to unbiasedly estimate X, for which we have available

only a sequence of approximations, X_i, with $\mathrm{E}[X_i] \to \mathrm{E}[X]$ as $i \to \infty$. Define an infinite series, $S = \mathrm{X}_0 + \sum_{n=1}^{\infty}(X_n - X_{n-1})$; an unbiased estimate of S is an unbiased estimate of X, assuming that the estimates are good enough to interchange expectation and summation. To achieve a computationally feasible and unbiased estimator of X, the Roulette or Poisson truncation schemes can then be applied. In the context of our work, this provides an alternative to the geometric or exponential series described above, in which only a consistent estimator is required. One drawback to this debiasing scheme for use in pseudo-marginal MCMC is that there is no obvious way to reduce the probability of the final estimate being negative. Russian Roulette is also employed extensively in the modelling of Neutron Scattering in Nuclear Physics and Ray Tracing in Computer Graphics (Hendricks and Booth, 1985, Carter and Cashwell, 1975).

Now that the complete Exact-Approximate MCMC scheme has been detailed, the following section illustrates the methodology on some models that are doubly-intractable, considering the strengths and weaknesses.

6. EXPERIMENTAL EVALUATION

6.1 Ising Lattice Spin Models

Ising models are examples of doubly-intractable distributions over which it is challenging to perform inference. They form a prototype for priors for image segmentation and autologistic models, for example, Hughes, Haran and Caragea (2011), Gu and Zhu (2001), Møller et al. (2006). Current exact methods such as the Exchange algorithm (Murray, Ghahramani and MacKay, 2006) require access to a perfect sampler (Propp and Wilson, 1996), which, while feasible for small grids, cannot be scaled up. A practical alternative is employed in Caimo and Friel (2011), where an auxiliary MCMC run is used to approximately simulate from the model. This is inexact and introduces bias, but it is hoped that the bias has little practical impact. We compare this approximate scheme with our exact methodology in this section.

For an $N \times N$ grid of spins, $\mathbf{y} = (y_1, \ldots, y_{N^2})$, $y \in \{+1, -1\}$, the Ising model has likelihood

$$p(\mathbf{y}; \alpha, \beta) = \frac{1}{\mathcal{Z}(\alpha, \beta)} \exp\left(\alpha \sum_{i}^{N^2} y_i + \beta \sum_{i \sim j} y_i y_j\right), \tag{6.1}$$

where i and j index the rows and column of the lattice and the notation $i \sim j$ denotes summation over nearest neighbours. Periodic boundary conditions are used in all subsequent computation. The parameters α and β indicate the strength of the external field and the interactions between neighbours, respectively. The normalising constant,

$$\mathcal{Z}(\alpha, \beta) = \sum_{\mathcal{Y}} \exp\left(\alpha \sum_{i}^{N^2} y_i + \beta \sum_{i \sim j} y_i y_j\right), \tag{6.2}$$

requires summation over all 2^{N^2} possible configurations of the model, which is computationally infeasible even for moderately sized lattices. This is, in fact, a naive bound as the transfer matrix method (see, e.g., MacKay, 2003), which has complexity $N2^N$ that can also be used to compute the partition function.

Experiments were carried out on a small 10×10 lattice to enable a detailed comparison of the various algorithms. A configuration was simulated using a perfect sampler with parameters set at $\alpha = 0$ and $\beta = 0.2$. Inference was carried out over the posterior distribution $p(\beta|\mathbf{y})$ ($\alpha = 0$ was fixed). A standard Metropolis–Hastings sampling scheme was used to sample the posterior, with a normal proposal distribution centred at the current value and acceptance rates tuned to around 40%. A uniform prior on $[0, 1]$ was set over β. As no tight upper bound is available on the normalising term $\mathcal{Z}(\boldsymbol{\theta})$, the debiasing series construction of McLeish (2011) and Glynn and Rhee (2014), described at the end of Section 5.2, was used to construct an unbiased estimate of the likelihood. The sequence of biased but consistent estimates of $1/\mathcal{Z}(\boldsymbol{\theta})$ was produced by taking the reciprocal of unbiased SMC estimates of $\mathcal{Z}(\boldsymbol{\theta})$ with an increasing number of importance samples and temperatures [see Del Moral, Doucet and Jasra (2006) for a good introduction to SMC]. SMC proceeds by defining a high-dimensional importance density which is sampled sequentially, and in this case we used a geometric schedule (Gelman and Meng, 1998, Neal, 2001) to define the sequence of distributions

$$p(\mathbf{y}|\boldsymbol{\theta})_n \propto p(\mathbf{y}|\boldsymbol{\theta})^{\phi_n} U(\mathbf{y})^{1-\phi_n},$$

with $0 \leq \phi_1 < \cdots < \phi_p = 1$ and $U(\cdot)$ a uniform distribution over all the grids in $\mathcal{Y}$. A Gibbs transition kernel, in which one spin was randomly selected and updated according to its conditional distribution, was used to sequentially sample the high-dimensional space. The initial estimate, $1/\mathcal{Z}(\boldsymbol{\theta})_0$, used 100 temperatures and 100 importance samples; the ith estimate used 100×2^i temperatures and importance samples.

TABLE 1

Monte Carlo estimates of the mean and standard deviation of the posterior distribution $p(\beta|\mathbf{y})$ using the five algorithms described. The debiasing series estimates have been corrected for negative estimates. The exact chain was run for 100,000 *iterations and then the second half of samples used to achieve a "gold standard" estimate. An estimate of the effective sample size (ESS) is also shown based on* 10,000 *MCMC samples*

	Roulette	**Poisson**	**Exchange (approx)**	**Exchange (exact)**	**Exact**
Mean	0.2004	0.2005	0.2013	0.2010	0.2008
Standard deviation	0.0625	0.0626	0.0626	0.0626	0.0625
ESS	2538	2660	1727	1732	3058

The infinite series was truncated unbiasedly using both Poisson truncation and Russian Roulette. For comparison, the posterior distribution was also sampled using the Exchange algorithm, the approximate form of the Exchange algorithm (Caimo and Friel, 2011) with an auxiliary Gibbs sampler run for 50,000 steps at each iteration, and an "exact" MCMC chain using the matrix transfer method to calculate the partition function at each iteration. All chains were run for 20,000 iterations and the second half of the samples used for Monte Carlo estimates.

The exact posterior mean and standard deviation are not available for comparison, but the estimates from the five methods agree well (Table 1). The traces in Figure 1 show that the algorithms mix well and Figures 2 and 3 show that the estimates of the mean and standard deviation agree well. Estimates of the Effective sample size (ESS) are also included in Table 1, which give an idea of how many independent samples are obtained from each method per 10,000 samples.

Approximately 5% of estimates were negative when using roulette truncation and 10% when using Poisson truncation; however, using the correction in equation (3.5), expectations with respect to the posterior still converge to the correct values. If we had opted to implement the geometric series construction of Section 4.1 in order to reduce the number of negative estimates, we have available only a naive upper bound for the partition function corresponding to setting all spins to $+1$. This bound is very loose and therefore impractical, as the series converges very slowly. Hence, the availability of a method to deal with negative estimates frees us from atrocious upper bounds that would explode the asymptotic variance of the chains.

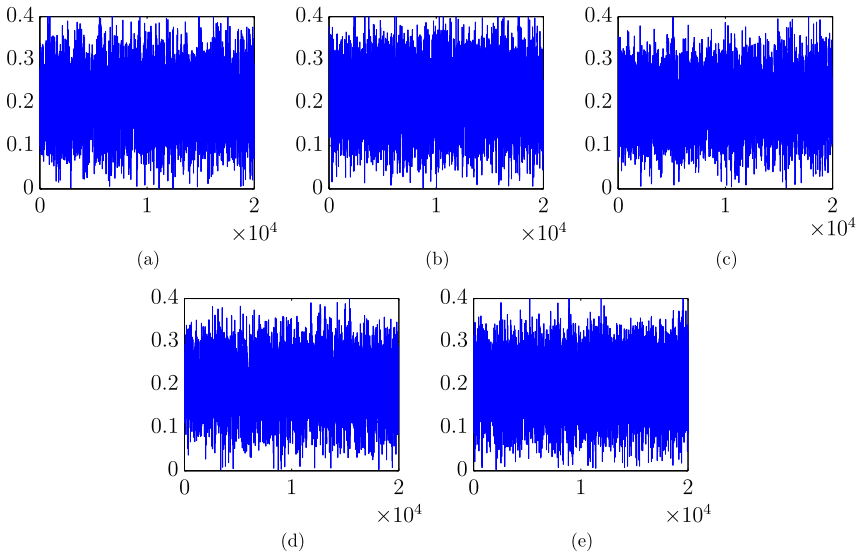

FIG. 1. *Traces of samples using the debiasing infinite series with* (a) *Russian Roulette*, (b) *Poisson truncation*, *and* (c) *the approximate Exchange algorithm*, (d) *the Exchange algorithm using perfect samples and* (e) *an MCMC chain with the partition function calculated using the matrix transfer method. Note in* (a) *and* (b) *the samples are not drawn from the posterior distribution*, $p(\beta|\mathbf{y})$, *but from the* (*normalised*) *absolute value of the estimated density.*

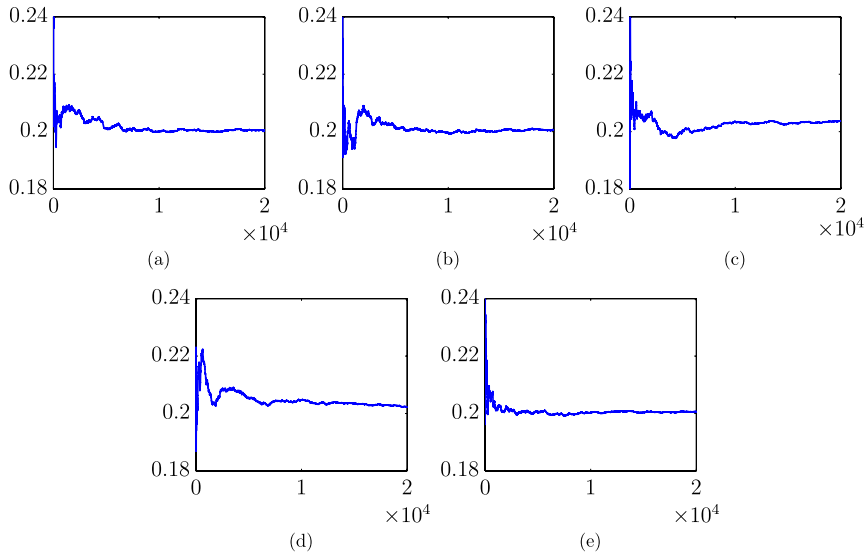

FIG. 2. *Plots of the running mean for the posterior distribution* $p(\beta|\mathbf{y})$ *of a* 10×10 *Ising model using three methods*: (a) *debiasing series with roulette truncation*, (b) *debiasing series with Poisson truncation*, (c) *approximate Exchange*, (d) *the Exchange algorithm using perfect samples and* (e) *an MCMC chain with the partition function calculated using the matrix transfer method.*

The autocorrelation functions (Figure 4) and the effective sample size (Table 1) of both Russian Roulette and Poisson truncation outperform the approximate and exact Exchange algorithm in this example and are comparable to the exact implementation; of course, it is possible to improve the performance of our algorithm by using more computation, whereas this is not possible with the Exchange algorithm. It should be noted that the Exchange algorithm in this guise is less computationally intensive. However, it becomes impossible to perfectly sample as the size of the lattice increases, whereas our algorithm can still be implemented, albeit with considerable computational expense. Note that even at this small lattice size, the approximate version of Exchange looks noticeably less stable.

We have further experimented on larger lattices, for example, we have used both the Exchange algorithm and our methodology to carry out inference over a

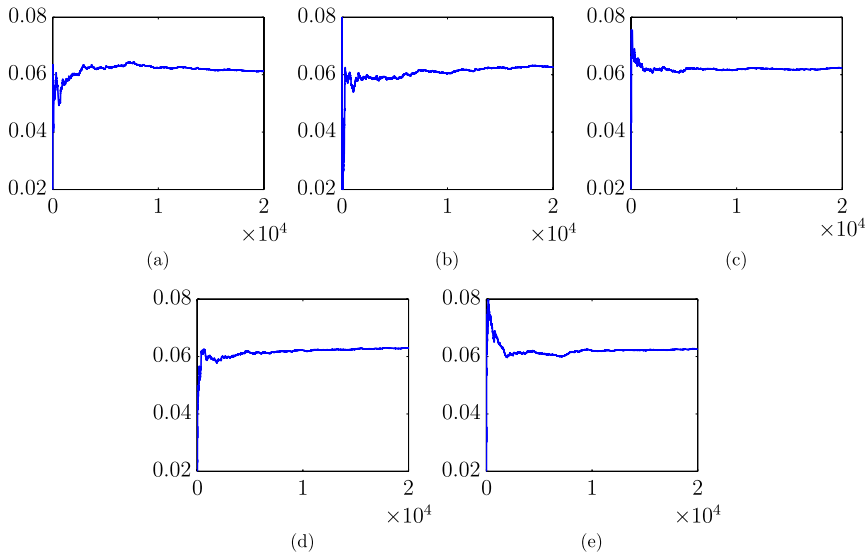

FIG. 3. *Plots of the running standard deviation for the posterior distribution* $p(\beta|\mathbf{y})$ *of a* 10×10 *Ising model using three methods*: (a) *debiasing series with roulette truncation*, (b) *debiasing series with Poisson truncation*, (c) *approximate Exchange*, (d) *the Exchange algorithm using perfect samples and* (e) *an MCMC chain with the partition function calculated using the matrix transfer method.*

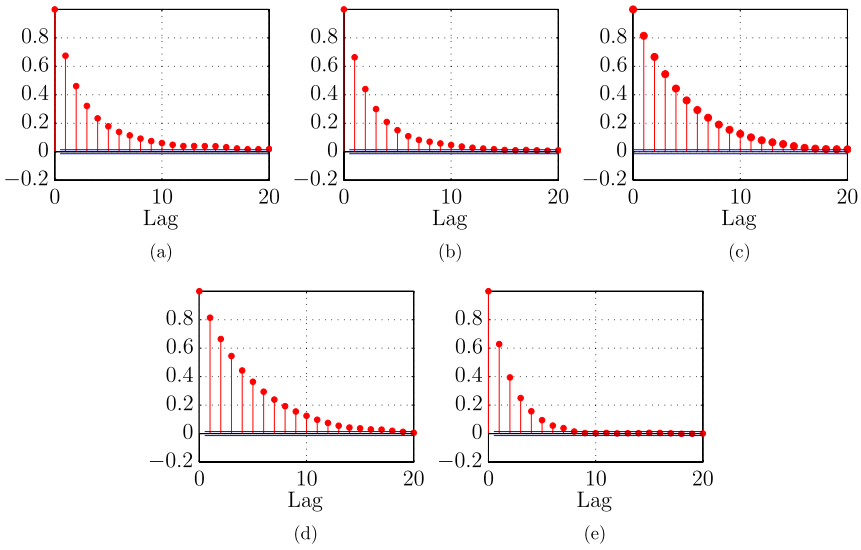

FIG. 4. *Autocorrelation plots for samples drawn from the posterior distribution* $p(\beta|\mathbf{y})$ *of a* 10×10 *Ising model using five methods*: (a) *debiasing series with roulette truncation*, (b) *debiasing series with Poisson truncation*, (c) *approximate Exchange*, (d) *the Exchange algorithm using perfect samples and* (e) *an MCMC chain with the partition function calculated using the matrix transfer method.*

40×40 grid. At this size it is not possible to use the matrix transfer method to run an "exact" chain. Sequential Monte Carlo (SMC) was used to estimate $\mathcal{Z}_i(\theta)$ at each iteration in the Roulette implementation. The estimates of the means and the standard deviations from both methods again agreed well (to the third or fourth decimal place). We have also carried out inference over a 60×60 grid; however, it is no longer possible to perfectly sample at this size, particularly for parameter values near the critical value.

6.2 The Fisher–Bingham Distribution on a Sphere

The Fisher–Bingham distribution (Kent, 1982) is constructed by constraining a multivariate Gaussian vector to lie on the surface of a d-dimensional unit radius sphere, S_d. Its form is

$$p(\mathbf{y}|\mathbf{A}) \propto \exp\{\mathbf{y}'\mathbf{A}\mathbf{y}\},$$

where $\mathbf{A}$ is a $d \times d$ symmetric matrix and, from here on, we take $d = 3$. After rotation to principle axes, $\mathbf{A}$ is diagonal and so the probability density can be written as

$$p(\mathbf{y}|\boldsymbol{\lambda}) \propto \exp\left\{\sum_{i=1}^{d} \lambda_i y_i^2\right\}.$$

This is invariant under addition of a constant factor to each λ_i, so for identifiability we take $0 = \lambda_1 \geq \lambda_2 \geq \lambda_3$. The normalising constant, $\mathcal{Z}(\boldsymbol{\lambda})$, is given by

$$\mathcal{Z}(\boldsymbol{\lambda}) = \int_{\mathcal{S}} \exp\left\{\sum_{i=1}^{d} \lambda_i y_i^2\right\} \mu(d\mathbf{y}),$$

where $\mu(d\mathbf{y})$ represents the Hausdorff measure on the surface of a sphere. Very few papers have presented Bayesian posterior inference over the distribution due to the intractable nature of $\mathcal{Z}(\boldsymbol{\lambda})$. However, in a recent paper, Walker uses an auxiliary variable method (Walker, 2011) outlined in the Introduction to sample from $p(\boldsymbol{\lambda}|\mathbf{y})$. We can apply our version of the Exact-Approximate methodology, as we can use importance sampling to get unbiased estimates of the normalising constant.

Twenty data points were simulated using an MCMC sampler with $\boldsymbol{\lambda} = [0, 0, -2]$ and posterior inference was carried out by drawing samples from $p(\lambda_3|\mathbf{y})$, that is, it was assumed $\lambda_1 = \lambda_2 = 0$. Our Exact-Approximate methodology was applied using the geometric construction with Russian Roulette truncation. A uniform distribution on the surface of a sphere was used to draw importance samples for the estimates of $\mathcal{Z}(\boldsymbol{\lambda})$. The proposal distribution for the parameters was Gaussian with mean given by the current value, a uniform prior on $[-5, 0]$ was set over λ_3, and the chain was run for 20,000 iterations. Walker's auxiliary variable technique was also implemented for comparison using the same prior but with the chain run for 200,000 samples and then the chain thinned by taking every 10th sample to reduce strong autocorrelations between samples. In each case the final 10,000 samples were then used for Monte Carlo estimates.

In the Russian Roulette method, six negative estimates were observed in 10,000 estimates. The estimates of the mean and standard deviation of the

TABLE 2
Estimates of the posterior mean and standard deviation of the posterior distribution using roulette and Walker's method for the Fisher–Bingham distribution. An estimate of the effective sample size (ESS) is also shown based on 10,000 *MCMC samples*

	Roulette	**Walker**
Estimate of mean	−2.377	−2.334
Estimate of standard deviation	1.0622	1.024
ESS	1356	212

posterior agree well (Table 2), however, the effective sample size and autocorrelation of the Russian Roulette method are superior as seen in Figure 5. Note that it is also possible to get an upper bound on the importance sampling estimates for the Fisher–Bingham distribution. If we change our identifiability constraint to be $0 = \lambda_1 \leq \lambda_2 \leq \lambda_3$, we now have a convex sum in the exponent which can be maximised by giving unity weight to the largest λ, that is, $\sum_{i=1}^{d} \lambda_i y_i^2 < \lambda_{\max}$. We can compute $\widetilde{\mathcal{Z}}(\boldsymbol{\theta})$ as $1/N \sum_n \exp(\lambda_{\max})/g(y_n)$, where $g(y)$ is the importance distribution.

7. THE LIMITS OF EXACT APPROXIMATE METHODS: THE OZONE DATA SET

In the previous sections of this paper we have combined various ideas from both the Statistics and Physics literature to suggest a pseudo-marginal MCMC scheme for doubly-intractable distributions. Further, we have shown experimentally that this method can be implemented in a range of Bayesian inference problems. We now turn our attention to a case where this methodology runs into difficulty.

It is tempting to think that the method could be used to tackle very large problems in which, for example, the likelihood requires the computation of the determinant of a very large matrix. For many problems the matrix in question is so large that it is not possible to compute its Cholesky decomposition, and hence not possible to compute the determinant. As methods are available to produce unbiased estimates of the log determinant (Bai, Fahey and Golub, 1996, Aune, Simpson and Eidsvik, 2014), the idea would be to write the determinant, $\mathcal{D}(\theta)$, as $\mathcal{D}(\theta) = \exp(\log \mathcal{D}(\theta))$ and then use the Maclaurin series expansion of the exponential function in which each term can be estimated unbiasedly. The infinite series can then be unbiasedly truncated using Russian Roulette methods and the overall estimate plugged into a pseudo-marginal MCMC scheme. Theoretically, this is an exact scheme to sample from the posterior of such a model; however, upon closer inspection, there are several practical difficulties associated with such an approach, namely, that it is not possible to realise a fully unbiased estimate of the log determinant. For exposition purposes, we now

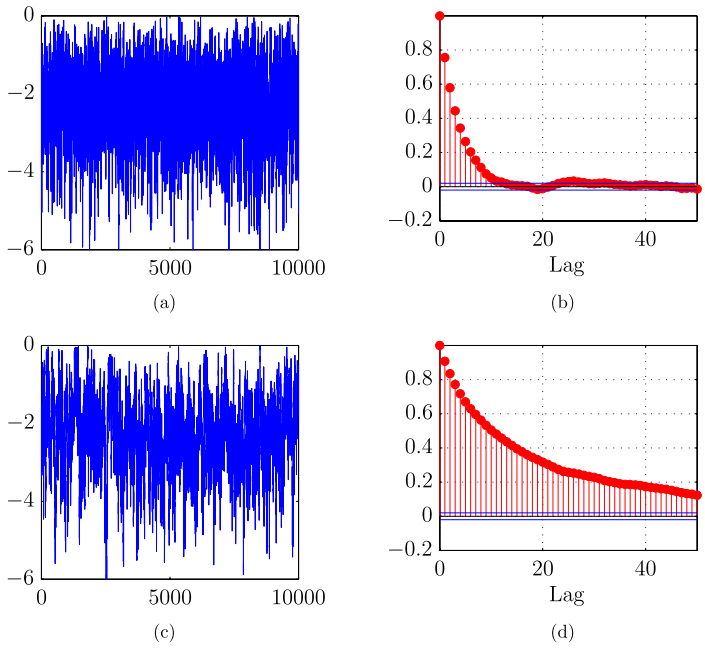

FIG. 5. *Sample traces and autocorrelation plots for the Fisher–Bingham distribution for the geometric tilting with Russian Roulette truncation* [(a) *and* (b)] *and Walker's auxiliary variable method* [(c) *and* (d)].

describe a specific example of a posterior for which it is difficult if not impossible to realise an unbiased estimate of the likelihood. In particular, we consider the total column ozone data set that has been used many times in the literature to test algorithms for large spatial problems (Cressie and Johannesson, 2008, Jun and Stein, 2008, Bolin and Lindgren, 2011, Aune, Simpson and Eidsvik, 2014, Eidsvik et al., 2014). This data set is representative of the types of problems for which exact Markov chain Monte Carlo is considered infeasible. While large, this data set is still of a size to run exact inference on and it serves as an interesting example of a problem in which the methods discussed in this paper break down. Full details and an implementation can be found at http://www.ucl.ac.uk/roulette.

We begin by describing the model and inference problem, and then suggest reasons why an application of the pseudo-marginal approach may run into difficulties. We close by describing results we were able to obtain and giving pointers to alternative approaches for similar problems.

7.1 The Model

The data, which is shown in Figure 6, consists of $N = 173{,}405$ ozone measurements gathered by a satellite with a passive sensor that measures back-scattered light (Cressie and Johannesson, 2008). While a full analysis of this data set would require careful modelling of both the observation process and the uncertainty of the field, for the sake of simplicity, we will focus on fitting a stationary model.

We model the data using the following three-stage hierarchical model:

$$\begin{aligned} y_i|\mathbf{x},\kappa,\tau &\sim \mathcal{N}(\mathbf{A}\mathbf{x},\tau^{-1}\mathbf{I}),\\ \mathbf{x}|\kappa &\sim \mathcal{N}(\mathbf{0},\mathbf{Q}(\kappa)^{-1}),\\ \kappa &\sim \log_2\mathcal{N}(0,100),\quad \tau\sim\log_2\mathcal{N}(0,100), \end{aligned} \tag{7.1}$$

where $\mathbf{Q}(\kappa)$ is the precision matrix of a Matérn stochastic partial differential equation (SPDE) model defined on a fixed triangulation of the globe and $\mathbf{A}$ is a matrix that evaluates the piecewise linear basis functions in such a way that $x(s_i) = [\mathbf{A}\mathbf{x}]_i$. The parameter κ controls the range over which the correlation between two values of the field is essentially zero (Lindgren, Rue and Lindström, 2011). The precision matrix $\mathbf{Q}(\kappa)$ is sparse, which allows both for low-memory storage and for fast matrix-vector products.

In this paper, the triangulation over which the SPDE model is defined has $n = 196{,}002$ vertices that are spaced regularly around the globe, allowing piecewise linear spatial prediction. As the observation process is Gaussian, a straightforward calculation shows that

$$\begin{aligned} \mathbf{x}|\mathbf{y},\kappa,\tau \sim N\big(&\tau\big(\mathbf{Q}(\kappa)+\tau\mathbf{A}^T\mathbf{A}\big)^{-1}\mathbf{A}^T\mathbf{y},\\ &\big(\mathbf{Q}(\kappa)+\tau\mathbf{A}^T\mathbf{A}\big)^{-1}\big). \end{aligned} \tag{7.2}$$

Given the hierarchical model in (7.1), we are interested in the parameters κ and τ only. To this end, we sample their joint posterior distribution given the observations $\mathbf{y}$, marginalised over the latent field $\mathbf{x}$,

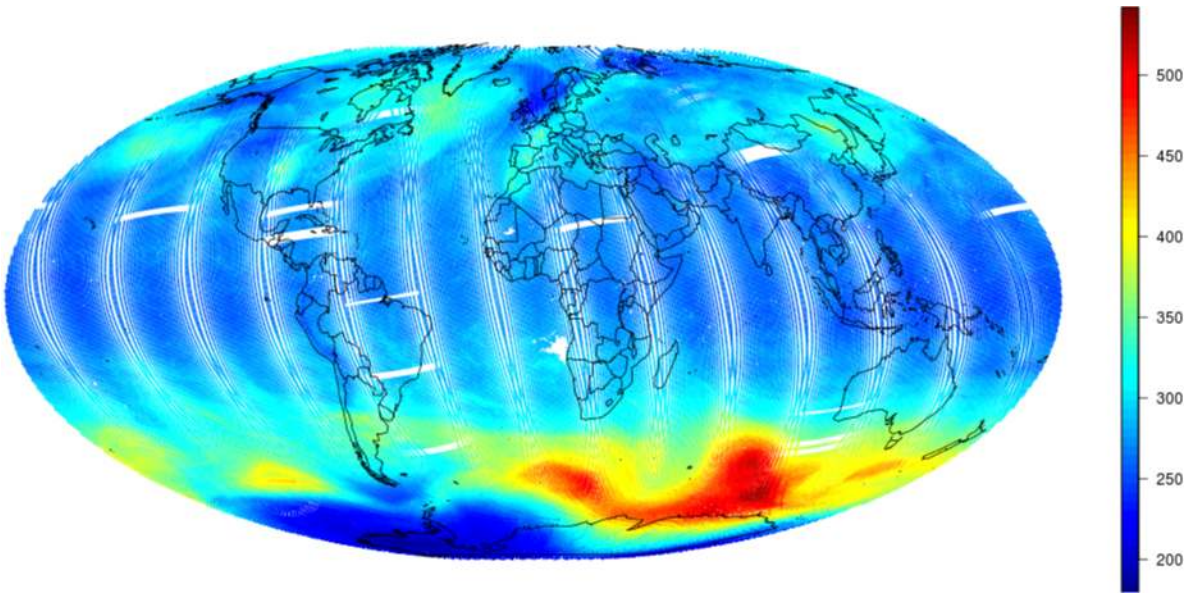

FIG. 6. *The total ozone column data set, aligned with a map of the world.*

which gives $\pi(\kappa, \tau|\mathbf{y}) \propto \pi(\mathbf{y}|\kappa, \tau)\pi(\kappa)\pi(\tau)$. To compute this expression, we need the marginal likelihood $\pi(\mathbf{y}|\kappa, \tau)$, which in this case is available analytically since $\pi(\mathbf{y}|\mathbf{x}, \tau)$ and $\pi(\mathbf{x}|\kappa)$ are both Gaussian,

$$\begin{aligned}\pi(\mathbf{y}|\kappa, \tau) &= \int \pi(\mathbf{y}|\mathbf{x}, \kappa, \tau)\pi(\mathbf{x}|\kappa)\, d\mathbf{x} \\ &= \mathcal{N}\big(\mathbf{0}, \tau^{-1}\mathbf{I} + \mathbf{A}\mathbf{Q}(\kappa)^{-1}\mathbf{A}^T\big).\end{aligned} \tag{7.3}$$

Using the matrix inversion lemma to avoid storing nonsparse matrices, the log marginal likelihood is

$$\begin{aligned}2\mathcal{L}(\theta) &:= 2\log\pi(\mathbf{y}|\kappa, \tau) \\ &= C + \log\big(\det\big(\mathbf{Q}(\kappa)\big)\big) + N\log(\tau) \\ &\quad - \log\big(\det\big(\mathbf{Q}(\kappa) + \tau\mathbf{A}^T\mathbf{A}\big)\big) \\ &\quad - \tau\mathbf{y}^T\mathbf{y} + \tau^2\mathbf{y}^T\mathbf{A}\big(\mathbf{Q}(\kappa) + \tau\mathbf{A}^T\mathbf{A}\big)^{-1}\mathbf{A}^T\mathbf{y}.\end{aligned} \tag{7.4}$$

7.2 Likelihood Estimation and Russian Roulette

In order to apply a pseudo-marginal MCMC scheme, we require an unbiased estimate of (7.3), for which we first need to compute unbiased estimates of the log-likelihood (7.4). Those are then plugged into a Russian Roulette truncated Maclaurin expansion of the exponential function, $\exp(\mathcal{L}(\theta)) = \sum_{n=0}^{\infty} \frac{\mathcal{L}(\theta)^n}{n!}$ [after replacing each $\mathcal{L}(\theta)$ with an unbiased estimate], to obtain the required unbiased estimate of the overall Gaussian likelihood (7.3).

To construct an unbiased estimator of (7.4), the main challenge is to estimate $\log(\det(\mathbf{Q}))$. We note that

$$\log\big(\det(\mathbf{Q})\big) = \operatorname{tr}\big(\log(\mathbf{Q})\big) = \mathbb{E}_{\mathbf{z}}\big(\mathbf{z}^T\log(\mathbf{Q})\mathbf{z}\big), \tag{7.5}$$

where $\mathbf{z}$ is a vector of i.i.d. centred, unit variance random variables (Bai, Fahey and Golub, 1996). Therefore, an unbiased estimator of the log-determinant can be constructed through Monte Carlo estimates of the expectation with respect to the distribution of $\mathbf{z}$. Aune, Simpson and Eidsvik (2014) used rational approximations and Krylov subspace methods to compute each $\log(\mathbf{Q})\mathbf{z}$ in (7.5) to machine precision, and they introduced a graph colouring method that massively reduces the variance in the Monte Carlo estimator. This approach is both massively parallel and requires a low-memory overhead, as only $\mathcal{O}(1)$ large vectors need to be stored on each processor.

However, as already mentioned, several issues are foreseeable when applying the pseudo-marginal MCMC scheme to the posterior. We emphasize two main points here:

1. *Numerical linear algebra*: In order to compute estimates of the log-likelihood (7.4), we need to solve a number of sparse linear systems. More precisely, we apply the methodology of Aune, Simpson and Eidsvik (2014), which reduces computing each log-determinant to solving a family of shifted linear equations for each of the $\log(\mathbf{Q})\mathbf{z}$ in (7.5). In addition, we need to solve the matrix inversions in (7.4). Note that each sparse linear system is independent and may be solved on its own separate computing node. Speed of convergence for solving these sparse linear systems largely depends on the condition number of the underlying matrix—the ratio of the largest and the smallest eigenvalues. In this example, the smallest eigenvalue of $\mathbf{Q}(\kappa)$ is arbitrarily close to zero, which catastrophically affects convergence of the methods described in Aune, Simpson and Eidsvik (2014). We can partially overcome these practical issues by regularising the matrix's smallest eigenvalue via adding a small number to the diagonal, shrinking the condition number using preconditioning matrices for the conjugate gradient, and setting a large iteration limit for the linear solvers. These convergence problems are typical when considering spatial models, as the eigenvalues of the continuous precision operator are unbounded. This suggests a fundamental limitation to exact-approximate methods for these models: it is impossible to attain full floating point precision when solving these linear systems, and hence the resulting Markov chain cannot exactly target the marginal posterior density $\pi(\kappa, \tau|\mathbf{y})$.

2. *Scaling*: A big challenge for practically implementing the Russian Roulette step is the large amount of variability in the estimator for (7.4), which is amplified by Russian Roulette. Denote by $\widehat{\mathcal{L}(\theta)} - U$ the unbiased estimator of the log-likelihood in equation (7.4), shifted towards a lower bound (see below) to reduce its absolute value. When the variance of the log-determinant estimator is large, the exponential series expansion will converge slowly and we will need to keep a large number of terms in order to keep the variance of the overall estimate low. We can get around this by borrowing the idea of "scaling-and-squaring" from numerical analysis (Golub and Van Loan, 1996).

We find an integer $E \in \mathbb{N}$ with $E \approx \mathcal{L}(\theta) - U$, for example, by averaging over a number of estimates. We then write

$$\exp\big(\mathcal{L}(\theta) - U\big) = \left(\exp\left(\frac{\mathcal{L}(\theta) - U}{E}\right)\right)^E. \tag{7.6}$$

In order to compute an unbiased estimate for this expression, we need to multiply E unbiased estimates of

$\exp\{(\mathcal{L}(\theta)-U)/E\}$, each of which we can obtain using Russian Roulette. This is now an easier problem since $(\mathcal{L}(\theta)-U)/E \approx 1$ is close to one. Therefore, the exponential series converges rapidly so that we only need a few estimates for $(\mathcal{L}(\theta)-U)/E$ in order to obtain one estimate of $\exp\{(\mathcal{L}(\theta)-U)/E\}$. The fact that a lower bound for $\mathcal{L}(\theta)$ is unavailable compromises unbiasedness of the estimator. In practice, this, however, was not measurable and drastically improved run-time.

7.3 Results and Remarks on Approximate Schemes

As this model is sufficiently small to (with some effort) perform exact inference, we began by finding the exact marginal posterior $\pi(\kappa, \tau|\mathbf{y})$, which is shown in Figure 7. The resulting density is relatively simple, which suggests that an appropriately scaled random walk Metropolis algorithm is sufficient for exploring it. As expected, in contrast to the other cases examined in this paper, we found that the Russian Roulette random walk Metropolis chain failed to converge for this problem: the chain exhibited catastrophic sticking and therefore extremely high autocorrelation. This is likely due to a combination of (a) our approximations in the log-likelihood estimator (due to the ill-posed linear systems), (b) the variation of the log-determinant estimator due to slow convergence of the linear solvers, and (c) the bias due to introducing the above scaling trick to the Russian Roulette scheme.

The fact that it is infeasible to realise a genuinely unbiased estimate of the normalising term for a model of this nature and size may mean that the Russian Roulette framework (and perhaps the entire concept of exact-approximate methods) is not the right approach for this type of model. We note that it has been shown previously in the literature that there are limitations to the efficiency of the pseudo-marginal scheme. For example, Sherlock et al. (2015) established results on optimal scaling and acceptance rate which indicate compromised efficiency when using the scheme. We close with the remark that compromising the ambitious goal of performing full and exact Bayesian inference on this problem might be a reasonable approach for practitioners who are interested in using models of the above type for solving large-scale problems. Recently, there has been an increased interest in approximate Markov transition kernels that allow such trade-off between computing time and introduced bias. Most of those methods are based on subsampling available observations in the Big Data case (Bardenet, Doucet and Holmes, 2014, Korattikara, Chen and Welling, 2014, Welling and Teh, 2011), and are therefore not available for the described ozone model, where we aim to do inference for a *single* observation. Similarly, the Exchange algorithm (Murray, Ghahramani and MacKay, 2006) is unavailable due to sampling from the likelihood being infeasible.

Using an approximate Markov transition kernel, induced from any approximation to the likelihood, leads to a chain whose invariant distribution is not equal to the true marginal posterior. Recently, Alquier et al. (2014) reviewed and analysed many cases of such approximate MCMC algorithms. A weak form of convergence is given by Alquier et al. (2014) (Theorem 2.1), which states that a Markov chain induced by an approximate transition kernel which approaches its exact version in the limit has an invariant distribution and this converges to the desired distribution as the kernel converges to the exact kernel under certain conditions. Theoretically, it is possible to apply this approach to the ozone example, as we can get a biased estimator for the log-likelihood via avoiding the Russian Roulette, and this kernel becomes exact when we use a very large number of iterations in the linear solvers. However, the slow convergence of the solvers remains a problem and in fact leads to such large variation in

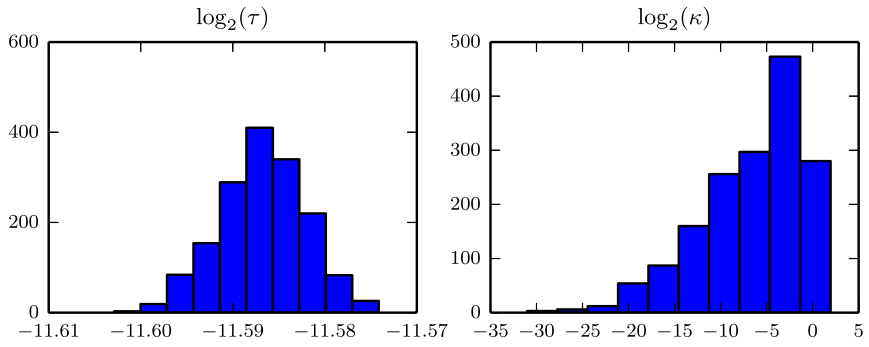

FIG. 7. *Histograms of the marginals* $p(\log_2(\tau)|\mathbf{y})$ *(left) and* $p(\log_2(\kappa)|\mathbf{y})$ *(right).*

the log-likelihood estimate that again the chain catastrophically sticks. It would seem, for the moment, that further approximation is required in order to carry out Bayesian inference. For example, in Shaby (2014), it is suggested that a function other than the likelihood, for example, a composite likelihood, can be used in an MCMC scheme to obtain samples from a "quasi-posterior" which can then be rotated and scaled to give asymptotically valid estimates.

8. DISCUSSION AND CONCLUSION

The capability to perform pseudo-marginal MCMC on a wide class of doubly-intractable distributions has been reviewed and established in this paper. The methods described are not reliant on the ability to simulate exactly from the underlying model, only on the availability of unbiased estimates of the inverse of a normalising term, which makes them applicable to a wider range of problems than has been the case to date.

The development of this method, which returns an unbiased estimate of the target distribution, is based on the stochastic truncation of a series expansion of the desired density. If the intractable likelihood is composed of a bounded function and nonanalytic normalising term, then the proposed methodology can proceed to full MCMC with no further restriction. However, in the more general case, where an unbounded function forms the likelihood, then the almost sure guarantee of *positive* unbiased estimates is lost. The potential bias induced due to this lack of strict positivity is dealt with by adopting a scheme employed in the QCD literature where an absolute measure target distribution is used in the MCMC and the final Monte Carlo estimate is "sign corrected" to ensure that expectations with respect to the posterior are preserved. The inflation of the Monte Carlo error in such estimates is a function of the severity of the sign problem and this has been characterised in our work. What has been observed in the experimental evaluation is that, for the examples considered, the sign problem is not such a practical issue when the variance of the estimates of the normalising terms is well controlled and this has been achieved by employing Sequential Monte Carlo Sampling in some of the examples. Hence, one of the areas for future work is efficient estimators of the normalising term, which can be either unbiased or merely consistent. Indeed, for the total column ozone data set, it is not possible at present to realise a completely unbiased estimate of $\log(\det(\mathbf{Q}))$, as is required for the pseudo-marginal methodology. The inherent computational parallelism of the methodology, due to it only requiring a number of independent estimates of normalising constants, indicates that it should be possible to implement this form of inference on larger models than currently possible, however it is also clear that there is some limit to how much the method can be scaled up. For the time being, approximate methods described in Section 2 can be used for very large-scale models, for example, analytic approximations to the posterior (Rue, Martino and Chopin, 2009) or ABC (Moores, Mengersen and Robert, 2014) could be used.

It has been shown (Jacob and Thiery, 2013) that it is not possible to realise strictly positive estimates of the target distribution using the series expansions described in this paper, unless the estimates of the normalising term lie in a bounded interval. In its most general representation it is recognised that the sign problem is NP-hard, implying that a practical and elegant solution may remain elusive for some time to come. However, other ideas from the literature, such as the absolute measure approach (Lin, Liu and Sloan, 2000), can be used to tackle the sign problem. The methodology described in this paper provides a general scheme with which Exact-Approximate MCMC for Bayesian inference can be deployed on a large class of statistical models. This opens up further opportunities in statistical science and the related areas of science and engineering that are dependent on simulation-based inference schemes.

APPENDIX A: RUSSIAN ROULETTE

Consider approximating the sum $S = \sum_{k\geq 0} \alpha_k$ assumed finite. Let τ denote a finite random time taking positive integer values such that $p_n \stackrel{\text{def}}{=} \mathbb{P}(\tau \geq n) > 0$ for all $n \geq 0$. The fact that τ is finite almost surely means that

$$\mathbb{P}(\tau = \infty) = \lim_{n\to\infty} p_n = 0. \tag{A.1}$$

We consider the weighted partial sums $S_0 = \alpha_0$, and for $k \geq 1$,

$$S_k = \alpha_0 + \sum_{j=1}^{k} \frac{\alpha_j}{p_j}.$$

For completeness, we set $S_\infty = \infty$. The Russian Roulette random truncation approximation of S is

$$\hat{S} = S_\tau.$$

If τ can be easily simulated and the probabilities p_n are available then $\hat{S}$ can be computed. The next result states that $\hat{S}$ is an unbiased estimator of S.

PROPOSITION A.1. *The random variable $\hat{S}$ has finite expectation, and $\mathbb{E}(\hat{S}) = S$.*

PROOF. Set $\bar{S}_0 = |\alpha_0|$, and $\bar{S}_k = |\alpha_0| + \sum_{j=1}^{k} |\alpha_j|/p_j$. Then for all $n \geq 1$

$$\begin{aligned}
&\sum_{k=0}^{n} |S_k| \mathbb{P}(\tau = k) \\
&\quad \leq \sum_{k=0}^{n} \bar{S}_k \mathbb{P}(\tau = k) = \sum_{k=0}^{n} \bar{S}_k (p_k - p_{k+1}) \\
&\quad = \bar{S}_0 p_0 + \sum_{k=1}^{n} (\bar{S}_k - \bar{S}_{k-1}) p_k \\
&\qquad + \sum_{k=1}^{n} \bar{S}_{k-1} p_k - \sum_{k=0}^{n} \bar{S}_k p_{k+1} \\
&\quad = \sum_{k=0}^{n} |\alpha_k| - \bar{S}_n p_{n+1} \leq \sum_{k=0}^{n} |\alpha_k|.
\end{aligned}$$

Since $\sum_n |\alpha_n| < \infty$, we conclude that $\sum_n |S_n| \mathbb{P}(\tau = n) < \infty$, hence $\mathbb{E}(|\hat{S}|) < \infty$. A similar calculation as above gives for all $n \geq 1$,

$$\sum_{k=0}^{n} S_k \mathbb{P}(\tau = k) = \sum_{k=0}^{n} \alpha_k - S_n p_{n+1}.$$

By Kronecker's lemma $\lim_{n\to\infty} p_n S_n = 0$, and $|p_{n+1} S_n| = (p_{n+1}/p_n) p_n |S_n| \leq p_n |S_n| \to 0$, as $n \to \infty$. We conclude that $\mathbb{E}(\hat{S}) = \sum_{k=0}^{\infty} S_k \mathbb{P}(\tau = k) = \sum_{k=0}^{\infty} \alpha_k$. □

This random truncation approximation of the series $\sum_n \alpha_n$ is known in the Physics literature as Russian Roulette. It has been re-derived apparently independently by McLeish (2011). In the Physics literature it is common to choose τ as a stopping time of the form

$$\tau = \inf\{k \geq 1 : U_k \geq q_k\},$$

where $\{U_j, j \geq 1\}$ are i.i.d. $\mathcal{U}(0,1)$, $q_j \in (0,1]$ and $\hat{S} = S_{\tau-1}$. In this case $p_n = \prod_{j=1}^{n-1} q_j$. The random time τ can be thought as the running time of the algorithm. It is tempting to choose τ such that the Russian Roulette terminates very quickly. The next result shows that the resulting variance will be high, possibly infinite.

PROPOSITION A.2. *If*

$$\sum_{n\geq 1} \frac{|\alpha_n|}{p_n} \sup_{j \geq n} \left| \sum_{\ell=n}^{j} \alpha_\ell \right| < \infty,$$

then $\mathsf{Var}(\hat{S}) < \infty$ *and*

$$\mathsf{Var}(\hat{S}) = \alpha_0^2 + \sum_{n\geq 1} \frac{\alpha_n^2}{p_n} + 2 \sum_{n \geq 1} \alpha_n S_{n-1} - S^2.$$

If $\{\alpha_n\}$ is a sequence of nonnegative numbers and $\sum_{n\geq 1} \alpha_n S_{n-1} = \infty$, then $\mathsf{Var}(\hat{S}) = \infty$.

PROOF. $\mathsf{Var}(\hat{S}) = \mathbb{E}(\hat{S}^2) - S^2$. So it suffices to work with $\mathbb{E}(\hat{S}^2)$. $\mathbb{E}(\hat{S}^2) = \sum_{k=0}^{\infty} S_k^2 \mathbb{P}(\tau = k) = \lim_{n\to\infty} \sum_{k=0}^{n} S_k^2 \mathbb{P}(\tau = k)$. For any $n \geq 1$, we use the same telescoping trick used in Proposition A.1 to get

$$\begin{aligned}
&\sum_{k=0}^{n} S_k^2 \mathbb{P}(\tau = k) \\
&\quad = \sum_{k=0}^{n} S_{k-1}^2 (p_k - p_{k+1}) \qquad \text{(A.2)} \\
&\quad = \alpha_0^2 + \sum_{k=1}^{n} \frac{\alpha_k^2}{p_k} + 2 \sum_{k=1}^{n} \alpha_k S_{k-1} - S_n^2 p_{n+1}.
\end{aligned}$$

By Jensen's inequality $S_n^2 \leq (\sum_{k=1}^{n} p_k^{-1}) \times (\sum_{k=1}^{n} p_k^{-1} \alpha_k^2)$. Hence, using Kronecker's lemma, we see that

$$\begin{aligned}
p_{n+1} S_n^2 &\leq p_n S_n^2 \\
&\leq \left(p_n \sum_{k=1}^{n} \frac{1}{p_k} \right) \left(\sum_{k=1}^{n} \frac{\alpha_k^2}{p_k} \right) \qquad \text{(A.3)} \\
&= o\left(\sum_{k=1}^{n} \frac{\alpha_k^2}{p_k} \right), \quad \text{as } n \to \infty,
\end{aligned}$$

so it suffices to show that the sequence $\sum_{k=1}^{n} \frac{\alpha_k^2}{p_k} + \sum_{k=1}^{n} \alpha_k S_{k-1}$ is bounded. But

$$\begin{aligned}
&\left| \sum_{j=1}^{n} \frac{\alpha_k^2}{p_k} + \sum_{k=1}^{n} \alpha_k S_{k-1} \right| \\
&\quad = \left| \alpha_0 \sum_{j=0}^{n} \alpha_j + \sum_{j=1}^{n} \frac{\alpha_j}{p_j} \left(\sum_{k=j}^{n} \alpha_k \right) \right| \\
&\quad \leq |\alpha_0| \sum_{j\geq 0} |\alpha_j| + \sup_n \sum_{j=1}^{n} \frac{|\alpha_j|}{p_j} \left| \sum_{k=j}^{n} \alpha_k \right|,
\end{aligned}$$

and the two terms on the right-hand side are bounded under the stated assumptions. Therefore the series $\sum_n S_n^2 \mathbb{P}(\tau = n)$ is summable and the variance formula follows by taking the limit as $n \to \infty$ in (A.2).

To establish the rest of the proposition, we deduce from (A.3) that for n large enough

$$\sum_{k=1}^{n} S_k^2 \mathbb{P}(\tau = k) \geq \alpha_0^2 + 2\sum_{k=1}^{n} \alpha_k S_{k-1},$$

which easily implies the statement. □

REMARK A.1. As an example, for a geometric sequence $\alpha_i = \alpha^i$ for $\alpha \in (0,1)$, and we choose $q_i = q$ for some $q \in (0,1)$, then for $\alpha^2/q < 1$, the condition of Proposition A.2 are satisfied and $\mathsf{Var}(\hat{S}) < \infty$. If $q > \alpha^2$ the variance is infinite. The average computing time of the algorithm is $\mathbb{E}(\hat{\tau}) = \frac{1}{1-q}$. Although this variance/computing speed trade-off can be investigate analytically, a rule of thumb that works well in simulations is to choose $q = \alpha$.

APPENDIX B: COMPUTING ABSOLUTE MEASURE EXPECTATIONS

Let $(\mathsf{X}, \mathcal{B})$ denotes a general measure space with a reference sigma-finite measure dx. Let $\pi : \mathsf{X} \to \mathbb{R}$ a function taking possibly negative values such that $\int |\pi(x)|\,dx < \infty$. We assume that $\int \pi(x)\,dx > 0$ and we wish to compute the quantity

$$I = \frac{\int h(x)\pi(x)\,dx}{\int \pi(x)\,dx},$$

for some measurable function $h : \mathsf{X} \to \mathbb{R}$ such that $\int |h(x)\pi(x)|\,dx < \infty$. We introduce $\sigma(x) = \mathsf{sign}(\pi(x))$, and $p(x) = \frac{|\pi(x)|}{\int |\pi(x)|\,dx}$. Thus p is a probability density on X. Suppose that we can construct an ergodic Markov chain $\{X_n, n \geq 0\}$ with invariant distribution p, for instance using the Metropolis–Hastings algorithm. An importance sampling-type estimate for I is given by

$$\hat{I}_n = \frac{\sum_{k=1}^{n} \sigma(X_k)h(X_k)}{\sum_{k=1}^{n} \sigma(X_k)}.$$

$\hat{I}_n$ has the following properties.

PROPOSITION B.1.

1. *If the Markov chain $\{X_n, n \geq 0\}$ is phi-irreducible and aperiodic, then $\hat{I}_n$ converges almost surely to I as $n \to \infty$.*

2. *Suppose that $\{X_n, n \geq 0\}$ is geometrically ergodic and $\int |h(x)|^{2+\varepsilon} p(x)\,dx < \infty$ for some $\varepsilon > 0$. Then*

$$\sqrt{n}(\hat{I}_n - I) \xrightarrow{\mathsf{w}} \mathbf{N}\big(0, \sigma^2(h)\big),$$

where

$$\sigma^2(h) = \frac{C_{11} + I^2 C_{22} - 2IC_{12}}{r^2},$$

and

$$C_{11} = \mathsf{Var}_p\big(\{h\sigma\}(X)\big) \cdot \sum_{j=-\infty}^{\infty} \mathsf{Corr}_p\big(\{h\sigma\}(X), P^{|j|}\{h\sigma\}(X)\big),$$

$$C_{22} = \mathsf{Var}_p\big(\sigma(X)\big) \sum_{j=-\infty}^{\infty} \mathsf{Corr}_p\big(\sigma(X), P^{|j|}\sigma(X)\big),$$

$$C_{12} = \frac{1}{2}\sqrt{\mathsf{Var}_p\big(\{h\sigma\}(X)\big)\mathsf{Var}_p\big(\sigma(X)\big)} \cdot \Bigg[\sum_{j=-\infty}^{\infty} \mathsf{Corr}_p\big(\{h\sigma\}(X), P^{|j|}\sigma(X)\big) + \sum_{j=-\infty}^{\infty} \mathsf{Corr}_p\big(\sigma(X), P^{|j|}\{h\sigma\}(X)\big)\Bigg].$$

PROOF. Part (1) is a straightforward application of the law of large numbers for the Markov chain $\{X_n, n \geq 0\}$: as $n \to \infty$, $\hat{I}_n$ converges almost surely to

$$\frac{\int \sigma(x)h(x)p(x)\,dx}{\int \sigma(x)p(x)\,dx} = \frac{\int h(x)\pi(x)\,dx}{\int \pi(x)\,dx} = I.$$

A bivariate central limit theorem using the Cramer–Wold device gives that

$$\sqrt{n}\begin{pmatrix} \frac{1}{n}\sum_{k=1}^{n} \sigma(X_k)h(X_k) - rI \\ \frac{1}{n}\sum_{k=1}^{n} \sigma(X_k) - r \end{pmatrix} \xrightarrow{\mathsf{w}} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathbf{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{pmatrix}\right],$$

where C_{11}, C_{12} and C_{22} are as given above.

By the delta method, it follows that $\sqrt{n}(\hat{I}_n - I) \xrightarrow{\mathsf{w}} \frac{Z_1 - IZ_2}{r} \sim \mathbf{N}(0, \frac{C_{11}+I^2C_{22}-2IC_{12}}{r^2})$. □

We can roughly approximate the asymptotic variance $\sigma^2(h)$ as follows. Suppose for simplicity that the Markov chain is reversible, so that

$$C_{12} = \sqrt{\mathsf{Var}_p\big(\{h\sigma\}(X)\big)\mathsf{Var}_p\big(\sigma(X)\big)} \cdot \sum_{j=-\infty}^{\infty} \mathsf{Corr}_p\big(\{h\sigma\}(X), P^{|j|}\sigma(X)\big).$$

Assume also that the mixing of the Markov chain is roughly the same across all the functions:

$$\begin{aligned}&\sum_{j=-\infty}^{\infty} \mathsf{Corr}_p\big(\{h\sigma\}(X), P^{|j|}\{h\sigma\}(X)\big)\\&\quad=\sum_{j=-\infty}^{\infty} \mathsf{Corr}_p\big(\sigma(X), P^{|j|}\sigma(X)\big)\\&\quad=\sum_{j=-\infty}^{\infty} \frac{\mathsf{Corr}_p(\sigma(X), P^{|j|}\{h\sigma\}(X))}{\mathsf{Corr}_p(\sigma(X), \{h\sigma\}(X))}\equiv V,\end{aligned}$$

where we also assume that $\mathsf{Corr}_p(\{h\sigma\}(X), \sigma(X)) \neq 0$. Therefore

$$\begin{aligned}\frac{\sigma^2(h)}{V} &\approx \big(\mathsf{Var}_p\big(\{h\sigma\}(X)\big) + I^2\mathsf{Var}_p\big(\sigma(X)\big)\\&\qquad - 2I\mathsf{Cov}_p\big(\{h\sigma\}(X), \sigma(X)\big)\big)/r^2\\&= \frac{r\check{\pi}(h^2\sigma) + I^2 - 2Ir\check{\pi}(h\sigma)}{r^2},\end{aligned}$$

where $\check{\pi} = \pi/\int\pi$, and $\check{\pi}(f) = \int f(x)\check{\pi}(x)\,dx$. By a Taylor approximation of $(h,\sigma)\mapsto h^2\sigma$ around $(\check{\pi}(h), \check{\pi}(\sigma))$, it comes easily that $r\check{\pi}(h^2\sigma) = \check{\pi}(h^2) + 2Ir\check{\pi}(h\sigma) - 2I^2$, so that

$$\sigma^2(h) \approx \big(\check{\pi}\big(h^2\big) - I^2\big) \times \frac{V}{r^2}.$$

Thus a quick approximation of the Monte Carlo variance of $\hat{I}_n$ is given by

$$\begin{aligned}&\frac{1}{n} \times \left\{\frac{\sum_{k=1}^n h^2(X_k)\sigma(X_k)}{\sum_{k=1}^n \sigma(X_k)} - \left(\frac{\sum_{k=1}^n h(X_k)\sigma(X_k)}{\sum_{k=1}^n \sigma(X_k)}\right)^2\right\}\\&\quad\cdot \frac{\hat{V}}{\{1/n\sum_{k=1}^n \sigma(X_k)\}^2},\end{aligned}$$

where $\hat{V}$ is an estimate of the common autocorrelation sum. For example $\hat{V}$ can be taken as the lag-window estimate of $\sum_{j=-\infty}^{\infty} \mathsf{Corr}_p(\{h\sigma\}(X), P^{|j|}\{h\sigma\}(X))$.

The quantity $\frac{1}{n}\sum_{k=1}^n \sigma(X_k)$ which estimates r is indicative of the severity of the issue of returning negative estimates. The smaller r, the harder it is to estimate I accurately.

ACKNOWLEDGEMENTS

Anne-Marie Lyne is supported by UCL Systems Biology. Mark Girolami is most grateful to Arnaud Doucet, David Barber, Christian Robert, Nicolas Chopin and Gareth Roberts for numerous motivating discussions regarding this work. Mark Girolami is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) via the Established Career Research Fellowship EP/J016934/1 and the Programme Grant *Enabling Quantification of Uncertainty for Large-Scale Inverse Problems*, EP/K034154/1, http://www.warwick.ac.uk/equip. He also gratefully acknowledges support from a Royal Society Wolfson Research Merit Award. Yves Atchadé is supported by the NSF on grant NSF-SES 1229261. Heiko Strathmann is supported by the Gatsby Charitable Foundation. Daniel Simpson is supported by CRiSM (Warwick).

REFERENCES

ADAMS, R. P., MURRAY, I. and MACKAY, D. J. (2009). Nonparametric bayesian density modeling with gaussian processes. Preprint. Available at arXiv:0912.4896.

ALQUIER, P., FRIEL, N., EVERITT, R. and BOLAND, A. (2014). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. Preprint. Available at arXiv:1403.5496.

ANDRIEU, C. and ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37** 697–725. MR2502648

ANDRIEU, C. and VIHOLA, M. (2014). Establishing some order amongst exact approximations of mcmcs. Preprint. Available at arXiv:1404.6909.

ATCHADÉ, Y. F., LARTILLOT, N. and ROBERT, C. (2013). Bayesian computation for statistical models with intractable normalizing constants. *Braz. J. Probab. Stat.* **27** 416–436. MR3105037

AUNE, E., SIMPSON, D. P. and EIDSVIK, J. (2014). Parameter estimation in high dimensional Gaussian distributions. *Stat. Comput.* **24** 247–263. MR3165552

BAI, Z., FAHEY, M. and GOLUB, G. (1996). Some large-scale matrix computation problems. *J. Comput. Appl. Math.* **74** 71–89. MR1430368

BAKEYEV, T. and DE FORCRAND, P. (2001). Noisy Monte Carlo algorithm reexamined. *Phys. Rev. D* **63** 54505.

BARDENET, R., DOUCET, A. and HOLMES, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of the* 31*st International Conference on Machine Learning* 405–413. JMLR Workshop and Conference Proceedings.

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

BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162** 2025–2035.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BESAG, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302. MR0876840

BESAG, J. E. and MORAN, P. A. P. (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika* **62** 555–562. MR0391451

BESKOS, A., PAPASPILIOPOULOS, O., ROBERTS, G. O. and FEARNHEAD, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 333–382. MR2278331

BHANOT, G. and KENNEDY, A. (1985). Bosonic lattice gauge theory with noise. *Phys. Lett. B* **157** 70–76.

BOLIN, D. and LINDGREN, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Ann. Appl. Stat.* **5** 523–550. MR2810408

BOOTH, T. (2007). Unbiased Monte Carlo estimation of the reciprocal of an integral. *Nucl. Sci. Eng.* **156** 403–407.

CAIMO, A. and FRIEL, N. (2011). Bayesian inference for exponential random graph models. *Soc. Netw.* **33** 41–55.

CARTER, L. L. and CASHWELL, E. D. (1975). Particle-transport simulation with the Monte Carlo method. Technical report, Los Alamos Scientific Lab., N. Mex. (USA). MR0416421

CRESSIE, N. and JOHANNESSON, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 209–226. MR2412639

DEL MORAL, P., DOUCET, A. and JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 411–436. MR2278333

DIGGLE, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *J. Roy. Statist. Soc. Ser. A* 349–362.

DOUC, R. and ROBERT, C. P. (2011). A vanilla Rao–Blackwellization of Metropolis–Hastings algorithms. *Ann. Statist.* **39** 261–277. MR2797846

DOUCET, A., PITT, M. and KOHN, R. (2012). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. Preprint. Available at arXiv:1210.1871.

EIDSVIK, J., SHABY, B. A., REICH, B. J., WHEELER, M. and NIEMI, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *J. Comput. Graph. Statist.* **23** 295–315. MR3215812

EVERITT, R. G. (2012). Bayesian parameter estimation for latent Markov random fields and social networks. *J. Comput. Graph. Statist.* **21** 940–960. MR3005805

FEARNHEAD, P., PAPASPILIOPOULOS, O. and ROBERTS, G. O. (2008). Particle filters for partially observed diffusions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 755–777. MR2523903

FRIEL, N. and PETTITT, A. N. (2004). Likelihood estimation and inference for the autologistic model. *J. Comput. Graph. Statist.* **13** 232–246. MR2044879

FRIEL, N., PETTITT, A. N., REEVES, R. and WIT, E. (2009). Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *J. Comput. Graph. Statist.* **18** 243–261. MR2749831

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London. MR1385925

GHAOUI, L. E. and GUEYE, A. (2009). A convex upper bound on the log-partition function for binary distributions. In *Advances in Neural Information Processing Systems* (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) **21** 409–416. Neural Information Processing Systems (NIPS).

GILKS, W. R. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London. MR1397966

GLYNN, P. W. and RHEE, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.* **51A** 377–389. MR3317370

GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720

GOODREAU, S. M., KITTS, J. A. and MORRIS, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* **46** 103–125.

GREEN, P. J. and RICHARDSON, S. (2002). Hidden Markov models and disease mapping. *J. Amer. Statist. Assoc.* **97** 1055–1070. MR1951259

GRELAUD, A., ROBERT, C. P. and MARIN, J.-M. (2009). ABC methods for model choice in Gibbs random fields. *C. R. Math. Acad. Sci. Paris* **347** 205–210. MR2538114

GU, M. G. and ZHU, H.-T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 339–355. MR1841419

HEIKKINEN, J. and HOGMANDER, H. (1994). Fully Bayesian approach to image restoration with an application in biogeography. *Applied Statistics* **43** 569–582.

HENDRICKS, J. and BOOTH, T. (1985). Mcnp variance reduction overview. In *Monte-Carlo Methods and Applications in Neutronics, Photonics and Statistical Physics* 83–92. Springer, Berlin.

HUGHES, J., HARAN, M. and CARAGEA, P. C. (2011). Autologistic models for binary data on a lattice. *Environmetrics* **22** 857–871. MR2861051

ILLIAN, J., SØRBYE, S., RUE, H. and HENDRICHSEN, D. (2012). Using INLA to fit a complex point process model with temporally varying effects–a case study. *J. Environ. Statist.* **3** 1–25.

ISING, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift Für Physik A Hadrons and Nuclei* **31** 253–258.

JACOB, P. E. and THIERY, A. H. (2013). On non-negative unbiased estimators. Preprint. Available at arXiv:1309.6473.

JIN, I. H. and LIANG, F. (2014). Use of SAMC for Bayesian analysis of statistical models with intractable normalizing constants. *Comput. Statist. Data Anal.* **71** 402–416. MR3131979

JOO, B., HORVATH, I. and LIU, K. (2003). The Kentucky noisy Monte Carlo algorithm for Wilson dynamical fermions. *Phys. Rev. D* **67** 074505.

JUN, M. and STEIN, M. L. (2008). Nonstationary covariance models for global data. *Ann. Appl. Stat.* **2** 1271–1289. MR2655659

KENDALL, W. S. (2005). Notes on perfect simulation. *Markov Chain Monte Carlo*: *Innovations and Applications* **7**. World Scientific, Singapore.

KENNEDY, A. and KUTI, J. (1985). Noise without noise: A new Monte Carlo method. *Phys. Rev. Lett.* **54** 2473–2476.

KENT, J. T. (1982). The Fisher–Bingham distribution on the sphere. *J. Roy. Statist. Soc. Ser. B* **44** 71–80. MR0655376

KORATTIKARA, A., CHEN, Y. and WELLING, M. (2014). Austerity in MCMC land: Cutting the Metropolis–Hastings budget. In *Proceedings of the* 31*st International Conference on Machine Learning* 181–189. JMLR Workshop and Conference Proceedings.

LIANG, F. (2010). A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *J. Stat. Comput. Simul.* **80** 1007–1022. MR2742519

LIANG, F., LIU, C. and CARROLL, R. J. (2007). Stochastic approximation in Monte Carlo computation. *J. Amer. Statist. Assoc.* **102** 305–320. MR2345544

LIECHTY, M. W., LIECHTY, J. C. and MÜLLER, P. (2009). The shadow prior. *J. Comput. Graph. Statist.* **18** 368–383. MR2749837

LIN, L., LIU, K. and SLOAN, J. (2000). A noisy Monte Carlo algorithm. *Phys. Rev. D* **61** 074505.

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. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727

LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR1842342

LUX, I. and KOBLINGER, L. (1991). *Monte Carlo Particle Transport Methods*: *Neutron and Photon Calculations*, *Vol.* 102. CRC press, Boca Raton.

MACKAY, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge Univ. Press, New York. MR2012999

MARIN, J.-M., PUDLO, P., ROBERT, C. P. and RYDER, R. J. (2012). Approximate Bayesian computational methods. *Stat. Comput.* **22** 1167–1180. MR2992292

MCLEISH, D. (2011). A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods Appl.* **17** 301–315. MR2890424

MØLLER, J. and WAAGEPETERSEN, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC, Boca Raton, FL. MR2004226

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

MOORES, M. T., MENGERSEN, K., ROBERT, C. P. (2014). Preprocessing for approximate bayesian computation in image analysis. Preprint. Available at arXiv:1403.4359.

MURRAY, I., GHAHRAMANI, Z. and MACKAY, D. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the* 22*nd Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*-06) 359–366. AUAI Press, Arlington, VI.

NEAL, R. M. (2001). Annealed importance sampling. *Stat. Comput.* **11** 125–139. MR1837132

PAPASPILIOPOULOS, O. (2011). Monte Carlo probabilistic inference for diffusion processes: A methodological framework. In *Bayesian Time Series Models* 82–103. Cambridge Univ. Press, Cambridge. MR2894234

PROPP, J. G. and WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9** 223–252.

RHEE, C.-H. and GLYNN, P. W. (2012). A new approach to unbiased estimation for SDE's. In *Proceedings of the Winter Simulation Conference*, *WSC*'12, *Berlin*, *Germany* 17:1–17:7. Winter Simulation Conference.

ROBERT, C. P. and CASELLA, G. (2010). *Introducing Monte Carlo Methods with R*. Springer, New York. MR2572239

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. Chapman & Hall/CRC, Boca Raton, FL. MR2130347

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602

SCHRÖDLE, B. and HELD, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics* **22** 725–734. MR2843139

SHABY, B. A. (2014). The open-faced sandwich adjustment for MCMC using estimating functions. *J. Comput. Graph. Statist.* **23** 853–876. MR3224659

SHERLOCK, C., THIERY, A. H., ROBERTS, G. O. and ROSENTHAL, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43** 238–275. MR3285606

SILVERTOWN, J. and ANTONOVICS, J. (2001). *Integrating Ecology and Evolution in a Spatial Context*: 14*th Special Symposium of the British Ecological Society* **14**. Cambridge Univ. Press, Cambridge.

TAVARÉ, S., BALDING, D. J., GRIFFITHS, R. C. and DONNELLY, P. (1997). Inferring coalescence times from dna sequence data. *Genetics* **145** 505–518.

TAYLOR, B. M. and DIGGLE, P. J. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *J. Stat. Comput. Simul.* **84** 2266–2284. MR3223624

TROYER, M. and WIESE, U.-J. (2005). Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations. *Phys. Rev. Lett.* **94** 170201.

VAN DUIJN, M. A., GILE, K. J. and HANDCOCK, M. S. (2009). A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks* **31** 52–62.

WALKER, S. G. (2011). Posterior sampling when the normalizing constant is unknown. *Comm. Statist. Simulation Comput.* **40** 784–792. MR2783887

WALKER, S. G. (2014). A Bayesian analysis of the Bingham distribution. *Braz. J. Probab. Stat.* **28** 61–72. MR3165428

WANG, F. and LANDAU, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86** 2050.

WELLING, M. and TEH, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the* 28*th International Conference on Machine Learning* 681–688. Omnipress, Madison, WI.

ZHANG, Y., GHAHRAMANI, Z., STORKEY, A. J. and SUTTON, C. A. (2012). Continuous relaxations for discrete Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems* **4** 3194–3202.

ZHOU, X. and SCHMIDLER, S. (2009). Bayesian parameter estimation in Ising and Potts models: A comparative study with applications to protein modeling. Technical report.