Bayesian Analysis (2025) **TBA**, Number TBA, pp. 1–30

A Correlated Pseudo-Marginal Approach to Doubly Intractable Problems

Yu Yang[*,**], Matias Quiroz[‡,§], Robert Kohn[†,‖] and Scott A. Sisson[¶,**]

Abstract. Doubly intractable models are encountered in a number of fields, e.g. social networks, ecology and epidemiology. Inference for such models requires the evaluation of a likelihood function, whose normalising factor depends on the model parameters and is assumed to be computationally intractable. The normalising constant of the posterior distribution and the additional normalising factor of the likelihood function result in a so-called doubly intractable posterior, for which it is difficult to directly apply Markov chain Monte Carlo methods (MCMC). We propose a signed pseudo-marginal Metropolis-Hastings algorithm with an unbiased block-Poisson estimator to sample from the posterior distribution of doubly intractable models. As the estimator can be negative, the algorithm targets the absolute value of the estimated posterior and uses an importance sampling estimator to ensure simulation-consistent estimates of the posterior mean of a function of the parameters. The importance sampling estimator can perform poorly when its denominator is close to zero. We derive a finite-sample concentration inequality that ensures, with high probability, that this pathological case does not occur. Our estimator for doubly intractable problems has three advantages over existing estimators. First, the estimator is well-suited for efficient parallelisation and vectorisation. Second, its structure is ideal for correlated pseudo-marginal methods, which are well known to dramatically increase sampling efficiency. Third, the estimator enables the derivation of heuristic guidelines for tuning its hyperparameters under simplifying assumptions. We demonstrate the superior performance of our method in the standard benchmark example that models correlated spatial data using the Ising model, as well as the Kent distribution model for spherical data.

Keywords: pseudo-marginal MCMC, doubly intractable posterior, Ising model, spherical data.

1 Introduction

Markov chain Monte Carlo (MCMC) methods (see, e.g., Brooks et al., 2011, for an overview) sample from a posterior distribution without evaluating its normalising constant, also known as the marginal likelihood. However, in some settings, the likelihood

arXiv: 2210.02734

[*]School of Economics, University of New South Wales, yu.yang2@unsw.edu.au
[†]School of Economics, University of New South Wales, r.kohn@unsw.edu.au
[‡]School of Mathematical and Physical Sciences, University of Technology Sydney, quiroz.matias@gmail.com
[§]Human Technology Institute, University of Technology Sydney
[¶]School of Mathematics & Statistics, University of New South Wales, scott.sisson@unsw.edu.au
[‖]Data Analytics for Resources and Environments (DARE), University of Sydney
[**]UNSW Data Science Hub, UNSW Sydney, Australia

 https://doi.org/10.1214/25-BA1573

function itself contains an additional normalising constant that depends on the model parameters, and the resulting so-called doubly intractable posterior distribution falls outside the standard MCMC framework. To distinguish these normalisation quantities, we refer to the first as a normalising constant and the latter as a normalising function. Many well-known models give rise to doubly intractable posteriors, such as the exponential random graph models for social networks (Hunter and Handcock, 2006) and non-Gaussian Markov random field models in spatial statistics, including the Ising model and its variants (Lenz, 1920; Ising, 1925; Hughes et al., 2011). Doubly intractable models are recognised as among the most challenging problems in statistics, as they involve an intractable likelihood and also make it difficult to simulate from the model for fixed parameter values (Rudolf et al., 2024).

Several algorithms are available to tackle the doubly intractable problem in Bayesian statistics; see Park and Haran (2018) for a review. These algorithms are classified into two main categories, with some overlap between them. The first category of methods introduces cleverly chosen auxiliary variables that cancel the normalising function when carrying out the MCMC sampling, and standard MCMC such as the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) can thus be applied. This approach is model-dependent and cannot always be applied. The second category of methods, which applies more generally, approximates the likelihood function (including the normalising function) and substitutes the approximation in place of the exact likelihood in the estimation procedure. The pseudo-marginal (PM) method (Beaumont, 2003; Andrieu and Roberts, 2009) is often used when a positive and unbiased estimator of the likelihood is available through Monte Carlo simulation. However, in some problems, including doubly intractable models, forming an unbiased estimator that is almost surely positive is only possible under unrealistic assumptions (Jacob and Thiery, 2015). The so-called Russian roulette estimator (Lyne et al., 2015) is an example of a method that can be used to unbiasedly estimate the likelihood function in doubly intractable models, although the estimate is not necessarily positive. Our paper focuses on methods that, asymptotically (in terms of chain length), yield samples from the exact doubly intractable posterior or provide estimates of expectations with respect to it.

We propose a method for exact inference on posterior expectations in doubly intractable problems based on the approach in Lyne et al. (2015), where an unbiased, but not necessarily positive, estimator of the likelihood function is used. The algorithm targets a posterior density that uses the absolute value of the likelihood, resulting in iterates from a perturbed target density. We follow Lyne et al. (2015) and reweight the samples from the perturbed target density using importance sampling to obtain simulation-consistent estimates of the expectation of a function of the parameters with respect to the true posterior density. By simulation-consistent, we mean that the posterior expectation can be estimated to an arbitrary precision by increasing the number of iterations of the algorithm. While our method does not sample from the target of interest, we refer to it as exact due to its simulation-consistent property.

Our main contribution is to explore the use of the block-Poisson (BP) estimator (Quiroz et al., 2021) in the context of estimating doubly intractable models us-

ing the signed pseudo-marginal Metropolis-Hastings (PMMH) approach. Moreover, we contribute to the signed PMMH literature by deriving a finite-sample (in terms of chain length) result that guarantees, with high probability, that the denominator of the importance sampling estimator used to compute expectations under the doubly intractable posterior remains bounded away from zero. A near–zero denominator of the importance sampling estimator is known to be detrimental for signed PMMH methods (Lyne et al., 2015; Quiroz et al., 2021). Our method provides the following advantages over the Russian roulette method. First, the BP estimator's simpler structure enables greater computational efficiency through parallelisation and vectorisation. Second, the block form of our estimator makes it possible to correlate the estimators of the doubly intractable posterior at the current and proposed draws in the MH algorithm. Introducing such correlation dramatically improves the efficiency of PM algorithms (Tran et al., 2016; Deligiannidis et al., 2018). Finally, under simplifying assumptions, some statistical properties of the logarithm of the absolute value of our estimator are derived and used to obtain heuristic guidelines to tune the hyperparameters of the estimator. We demonstrate empirically that our method outperforms Lyne et al. (2015) when estimating the parameter of an Ising model following the settings in the review paper Park and Haran (2018). In a real data application for directional data, our method has a significantly shorter computing time compared to competing Bayesian methods. To the best of our knowledge, our method and that of Lyne et al. (2015) with its extensions are the only alternatives in the PM framework to perform exact inference (in the sense of simulation-consistent estimates of posterior expectations) for general doubly intractable problems. Compared to algorithms using auxiliary variables to avoid evaluating the normalising function, signed PMMH algorithms are more widely applicable and generic as they do not require exact sampling from the likelihood, which may be hard to implement. Exact sampling refers to the ability to draw independent (data) samples exactly distributed according to the likelihood.

The rest of the paper is organised as follows. Section 2 introduces the doubly intractable problem and discusses previous research. Section 3 introduces our methodology and presents a theoretical result (Theorem 1) that helps avoid a pathological case arising in finite-length MCMC chains when implementing signed pseudo-marginal methods, a practical concern that appears to have received little attention in the literature. Section 3 establishes the guidelines for tuning the hyperparameters of the estimator. Section 4 reports on a replicated simulation study from the review paper Park and Haran (2018) for the Ising model and, additionally, the Kent distribution for modelling directional data (with the details in Yang et al., 2025, Section S9.3 due to space restrictions). Section 5 analyses four real-world datasets using the Kent distribution. Section 6 concludes and outlines future research. The paper has a supplement that contains all proofs and details of the simulation studies, as well as the details of the other methods applied. We refer to equations, sections, lemmas in the main paper as (1.1), Section 1, Lemma 1 etc., and to equations, sections and lemmas, etc. in the supplement as (S1.1), Section S1 and Lemma S1, etc.

2 Doubly intractable problems

2.1 Doubly intractable posterior distributions

Let $p(\mathbf{y}|\boldsymbol{\theta})$ denote the density of the data vector $\mathbf{y}$, where $\boldsymbol{\theta}$ is the vector of model parameters. Suppose $p(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta})/Z(\boldsymbol{\theta})$, where $f(\mathbf{y}|\boldsymbol{\theta})$ is computable while the normalising function $Z(\boldsymbol{\theta})$ is not. The reason that $Z(\boldsymbol{\theta})$ is intractable may be that it is prohibitively expensive to evaluate numerically, or lacks a closed form because $Z(\boldsymbol{\theta})$ is defined as an integral over a complex or high-dimensional space which is hard to evaluate, or it involves summing over an intractably large number of terms. Two examples are given below to demonstrate the intractability for both discrete and continuous observations $\mathbf{y}$.

Example 1 (The Ising model (Ising, 1925))**.** *Consider an $L \times L$ lattice with binary observation $y_{ij} \in \{-1, 1\}$ in row i and column j. The likelihood of $\theta \in \mathbb{R}$ is*

$$p(\mathbf{y}|\theta) = \frac{1}{Z(\theta)} \exp(\theta S(\mathbf{y})); \quad S(\mathbf{y}) = \sum_{i=1}^{L} \sum_{j=1}^{L-1} y_{i,j} y_{i,j+1} + \sum_{i=1}^{L-1} \sum_{j=1}^{L} y_{i,j} y_{i+1,j}; \tag{2.1}$$
$$\text{with } Z(\theta) = \sum_{\mathbf{y}} \exp(\theta S(\mathbf{y})).$$

The normalising function $Z(\theta)$ in the Ising model is a sum over 2^{L^2} terms, making it computationally intractable even for moderate values of L. See Section 4.1 for a further discussion.

Example 2 (The Kent distribution (Kent, 1982))**.** *The density of the Kent distribution for $\mathbf{y} \in \mathbb{R}^3, \|\mathbf{y}\| = 1$, is*

$$f(\mathbf{y}|\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\gamma_3}, \beta, \kappa) = \frac{1}{c(\kappa, \beta)} \exp\left\{\kappa \boldsymbol{\gamma_1}^\top \cdot \mathbf{y} + \beta\left[(\boldsymbol{\gamma_2}^\top \cdot \mathbf{y})^2 - (\boldsymbol{\gamma_3}^\top \cdot \mathbf{y})^2\right]\right\}; \tag{2.2}$$
$$\text{with } c(\kappa, \beta) = 2\pi \sum_{j=0}^{\infty} \frac{\Gamma(j + 0.5)}{\Gamma(j+1)} \beta^{2j} (0.5\kappa)^{-2j-0.5} I_{2j+0.5}(\kappa),$$

where $I_\nu(.)$ is the modified Bessel function of the first kind and $\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\gamma_3}$ form a set of 3-dimensional orthonormal vectors. The normalising function $c(\kappa, \beta)$ is an infinite sum and thus intractable to evaluate. See Yang et al. (2025, Section S9.3) for a further discussion.

The doubly intractable posterior density of $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})p(\mathbf{y})} \propto \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}, \tag{2.3}$$

where $\pi(\boldsymbol{\theta})$ is the prior for $\boldsymbol{\theta}$ and

$$p(\mathbf{y}) = \int \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{2.4}$$

is the normalising constant for the posterior. Suppose we devise a Metropolis-Hastings algorithm to sample from (2.3) with a proposal density $q(\cdot|\boldsymbol{\theta})$. The probability of accepting a proposed sample $\boldsymbol{\theta}'$ is

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')f(\mathbf{y}|\boldsymbol{\theta}')/Z(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta}))/Z(\boldsymbol{\theta})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}. \tag{2.5}$$

The marginal likelihood in (2.4) cancels in (2.5), but the normalising function does not. Since $Z(\boldsymbol{\theta})/Z(\boldsymbol{\theta}')$ is computationally intractable, (2.5) cannot be evaluated and thus MCMC sampling via the Metropolis-Hastings algorithm is impossible.

2.2 Previous research

Previous research on doubly intractable problems is mainly divided into the auxiliary variable approach and the likelihood approximation approach; see Park and Haran (2018) for an excellent review of both approaches.

The auxiliary variable approach cleverly chooses the joint transition kernel of the parameters and the auxiliary variables so that the normalising function cancels in the resulting MH acceptance ratio. The most well-known algorithms in this space are the exchange algorithm (Murray et al., 2006) and the auxiliary variable method (Møller et al., 2006). Andrieu et al. (2020) extend the exchange algorithm by leveraging averaged acceptance ratios for improved stability. These algorithms are model-dependent and, crucially, rely on the sampling technique used to draw observations from the likelihood function. Perfect sampling (Propp and Wilson, 1996) is often used to generate data samples from the model without knowing the normalising function. However, for some complex models, such as the Ising model on a large grid, perfect sampling is prohibitively expensive. To overcome this issue, Liang (2010) and Liang et al. (2016) relax the requirement of exact sampling and propose the double MH sampler and the adaptive exchange algorithm. However, the former generates inexact inference results, while the latter suffers from memory issues as many intermediate variables need to be stored within each iteration.

Methods belonging to the likelihood approximation approach can be simulation-consistent. One example is Atchadé et al. (2013), who directly approximate $Z(\boldsymbol{\theta})$ through multiple importance sampling. Their approach also depends on an auxiliary variable technique, but does not require perfect sampling. The downside is similar to that of the adaptive exchange algorithm; a large memory is usually required to store the intermediate variables generated in each iteration. An alternative method is to approximate $1/Z(\boldsymbol{\theta})$ directly using the signed PMMH algorithm to replace the likelihood function by its unbiased estimator as proposed in Lyne et al. (2015). To obtain the unbiased estimator, $1/Z(\boldsymbol{\theta})$ is expressed as a geometric series, which is truncated using a Russian roulette (RR) approach. The RR method first appeared in the physics literature (Carter and Cashwell, 1975) and is useful for obtaining an unbiased estimator through a finite stochastic truncation of an infinite series. To implement RR, a tight upper bound for $Z(\boldsymbol{\theta})$ is required; otherwise, the convergence of the geometric series is slow and makes the algorithm inefficient. In practice, an upper bound is usually unavailable, which may

lead to negative estimates of the likelihood, and thus a signed PMMH approach is necessary, although this inflates the asymptotic variance of the resulting importance sampling estimator of posterior expectations. In particular, having nearly half of the estimates being negative is detrimental for the asymptotic variance (Lyne et al., 2015; Quiroz et al., 2021). It is therefore crucial to control the probability of a negative estimate, which is difficult for the RR estimator as, to the best of our knowledge, no analytical expression exists for this quantity. In contrast, our estimator is more tractable, and the probability of a positive estimate is analytically derived under simplifying assumptions. Besides the upper bound, a few other hyperparameters of the RR estimator need to be determined. We are unaware of any guidelines for selecting these based on analytical expressions. We provide such guidelines for the hyperparameters in the estimator proposed in our paper. Wei and Murray (2017) combine RR with Markov chain coupling to produce an estimator with lower variance and a larger probability of producing positive estimates. However, their estimator is not tractable enough to readily provide tuning guidelines. Cai and Adams (2022) propose a multi-fidelity MCMC method to approximate the doubly intractable target density, which, like the Russian roulette method, stochastically truncates an infinite series and uses slice sampling (Murray and Graham, 2016). However, similarly to Lyne et al. (2015), the method lacks guidelines for tuning the hyperparameters.

Finally, although our focus is on exact approaches, we note that several approximate methods for estimating doubly intractable posteriors are available. Alquier et al. (2016) propose the noisy exchange algorithm, which uses multiple auxiliary variables to obtain an estimate of the acceptance ratio in the exchange algorithm. As the number of auxiliary variables goes to infinity, the algorithm samples from the exact doubly intractable posterior. However, for a finite number of auxiliary variables, the method is approximate. Variational Bayes approaches include Tan and Friel (2020) and Lee et al. (2024). Park and Haran (2020) propose approximating the normalising function at several parameter values using importance sampling, and interpolating it at other parameter values via Gaussian process-based emulation.

3 Methodology

3.1 The block-Poisson estimator

The block-Poisson estimator (Quiroz et al., 2021) was proposed for estimating the likelihood unbiasedly given an unbiased estimator of the log-likelihood obtained by data subsampling. The BP estimator builds on the Poisson estimator (Wagner, 1988; Papaspiliopoulos, 2011), which is useful for estimating $\exp(B)$ unbiasedly given an unbiased estimator $\widehat{B}$ of B, i.e. $E(\widehat{B}) = B$. The estimator $\widehat{B}$ can be obtained by Monte Carlo integration based on M samples. The Poisson estimator of $\exp(B)$ is

$$\exp(m+a)\prod_{h=1}^{\chi}\frac{\widehat{B}^{(h)}-a}{m},\ \chi\sim\text{Pois}(m),\ m\in\mathbb{Z},\ a\in\mathbb{R}, \tag{3.1}$$

where $\widehat{B}^{(h)}$ are independent copies of $\widehat{B}$.

The block-Poisson estimator also estimates $\exp(B)$ unbiasedly and consists of λ Poisson estimators similar to (3.1), however, estimating $\exp(B/\lambda)$ unbiasedly. The idea behind using blocks of Poisson estimators, instead of a single one as in (3.1), is to allow for correlation between successive iterates in the PM algorithm as described in Section 3.2. Similarly to the likelihood approximation approaches discussed above, the BP estimator is implemented in combination with an auxiliary variable ν, and an estimator of the normalising function. Omitting details of the auxiliary variable method that are explained in Section 3.2, assume $B(\boldsymbol{\theta}) = -\nu Z(\boldsymbol{\theta})$ where $\nu \sim \text{Exp}(Z(\boldsymbol{\theta}))$. Our procedure estimates the likelihood for given ν, and thus ν is treated as fixed (non-random) in the results derived for our estimator that we use for the tuning guidelines. The BP estimator produces (for a fixed ν) an unbiased estimator of $\exp(-\nu Z(\boldsymbol{\theta}))$ using unbiased estimators of the normalising function $Z(\boldsymbol{\theta})$. One advantage of the BP estimator over the RR estimators is that its simple form, together with simplifying assumptions, e.g. that the estimator of the normalising function is normal, enables hyperparameter tuning based on analytical expressions. As a result, the BP estimator is more likely to produce positive estimates if tuned following our guidelines in Section 3.3. Controlling the signs of the estimates is desirable for efficient estimation based on MCMC output as discussed in Sections 3.2 and 3.3.

Definition 1 describes the BP estimator $\widehat{L}_B$ of the likelihood we use for doubly intractable problems. Lemma 1 gives the expectation and variance of $\widehat{L}_B$. Lemmas 2 and 3 establish useful results for tuning the hyperparameters of the estimator (see Section 3.3). The proofs are in Yang et al. (2025, Section S1) Section S1.

Definition 1. *The block-Poisson estimator of* $\exp(B(\boldsymbol{\theta}))$ *is defined as*

$$\widehat{L}_B(\boldsymbol{\theta}) = \prod_{l=1}^{\lambda} \xi_l(\boldsymbol{\theta}), \;\; \xi_l(\boldsymbol{\theta}) = \exp(a/\lambda + m) \prod_{h=1}^{\chi_l} \frac{\widehat{B}^{(h,l)}(\boldsymbol{\theta}) - a}{m\lambda}, \tag{3.2}$$

where λ *is the number of blocks,* $\chi_l \sim \text{Pois}(m)$*, a Poisson distribution with mean* m*,* a *is an arbitrary constant and* m *is the expected number of estimators used within each block. The unbiased estimates of* B*,* $\widehat{B}^{(h,l)}$ *are independent with the indexes* $h = 1, \ldots, \chi_l, l = 1, \ldots, \lambda$*.*

Remark 1. *For estimating* $\exp(B(\boldsymbol{\theta}))$*, only the product* $m\lambda$ *matters, which represents the total computational effort. The specific allocation between* m *and* λ *becomes relevant only under the correlated pseudo-marginal approach, where increasing* λ *allows stronger correlation between successive likelihood estimates.*

Remark 2. *If* $m = 1$ *(so that the computational effort is* λ*) then* (3.2) *is the estimator in Quiroz et al. (2021). To ensure a positive estimator with probability 1,* a *needs to be a lower bound for all* $\widehat{B}^{(h,l)}$*.*

Lemma 1. *Denote* $\sigma^2_{\widehat{B}} = \text{Var}(\widehat{B}(\boldsymbol{\theta}))$*, and assume* $\sigma^2_{\widehat{B}} < \infty$ *and* $E(\widehat{B}(\boldsymbol{\theta})) = B(\boldsymbol{\theta})$*. The following properties hold for* $\widehat{L}_B(\boldsymbol{\theta})$ *in* (3.2)*:*

(i) $E(\widehat{L}_B(\boldsymbol{\theta})) = \exp(B(\boldsymbol{\theta}))$*.*

(ii) $\mathrm{Var}(\widehat{L}_B(\boldsymbol{\theta})) = \exp\left[\dfrac{(B(\boldsymbol{\theta})-a)^2+\sigma^2_{\widehat{B}}}{m\lambda} + 2a + m\lambda\right] - \exp(2B(\boldsymbol{\theta}))$.

(iii) $\mathrm{Var}(\widehat{L}_B(\boldsymbol{\theta}))$ *is minimised at* $a = B(\boldsymbol{\theta}) - m\lambda$, *given fixed* m *and* λ.

Remark 3. *As noted by an anonymous reviewer, Lemma 1 implies that, to estimate* $\exp(B(\boldsymbol{\theta}))$ *with a relative variance of constant order,* $m\lambda \asymp \sigma^2_{\widehat{B}}$ *is required; see Yang et al. (2025, Section S1) for details.*

Part (i) of Lemma 1 shows that given an unbiased estimator $\widehat{B}(\boldsymbol{\theta})$ of $B(\boldsymbol{\theta})$, the BP estimator is unbiased for $\exp(B(\boldsymbol{\theta}))$ for any values of the hyperparameters. Let a_{opt} denote the value of a that minimises the variance of the estimator in Part (ii) (for a fixed m and λ). Part (iii) of Lemma 1 shows that $a_{\mathrm{opt}} = B(\boldsymbol{\theta}) - m\lambda$.

Similarly to the RR estimator, the BP estimator is not necessarily positive unless a is a lower bound of $\widehat{B}(\boldsymbol{\theta})$, i.e. $\widehat{B}(\boldsymbol{\theta}) > a$. This implies that $a < 0$ if $\widehat{Z}(\boldsymbol{\theta}) > 0$ (recall that $\widehat{B}(\boldsymbol{\theta}) = -\nu\widehat{Z}(\boldsymbol{\theta})$, $\nu \geq 0$) which usually holds in doubly intractable problems. Selecting a large negative value of a (to account for extreme outcomes of $\widehat{B}(\boldsymbol{\theta})$), Part (iii) suggests that $m\lambda$ should also be large to keep the variance small (typically, $|\widehat{B}(\boldsymbol{\theta})| \ll m\lambda$). This translates to a computationally costly estimator due to many products in the BP estimator. We follow Quiroz et al. (2021) and advocate the use of a soft lower bound, i.e., one that may lead to negative estimates, but still gives a $\Pr(\widehat{L}_B(\boldsymbol{\theta}) \geq 0)$ close to one. Lemma 2 shows that the probability $\Pr(\widehat{L}_B(\boldsymbol{\theta}) \geq 0)$ has an analytical expression. It is crucial to have this probability close to one for the algorithm to be efficient.

Lemma 2. *Suppose that* $a = a_{\mathrm{opt}} = B(\boldsymbol{\theta}) - m\lambda$. *Then,*

$$\Pr(\widehat{L}_B(\boldsymbol{\theta}) \geq 0) = \frac{1}{2}\Big(1 + (1 - 2\Psi(a, m, \lambda, M))^{\lambda}\Big), \tag{3.3}$$

with $\Psi(a, m, \lambda, M) = \Pr(\xi < 0) = \frac{1}{2}\sum_{j=1}^{\infty}\left(1 - (1 - 2\Pr(A_m \leq 0))^j\right)\Pr(\chi_l = j)$, $\chi_l \sim \mathrm{Pois}(m)$ *and* $A_m = [\widehat{B}(\boldsymbol{\theta}) - B(\boldsymbol{\theta})]/(m\lambda) + 1$, *and* M *is the number of Monte Carlo samples to estimate a single* $\widehat{B}(\boldsymbol{\theta})$.

Remark 4. *To compute the probability* $\Pr(A_m \leq 0)$ *in practice, we assume that* $\widehat{B}(\boldsymbol{\theta})$ *is normal with variance* $\sigma^2_{\widehat{B}}$ *as in Lemma 3. When the Monte Carlo samples* M *are independent, then* $\sigma^2_{\widehat{B}} \propto 1/M$.

Remark 5. *An anonymous referee noted that a tractable lower bound for the probability in (3.3) can be derived to guide the choice of* $m\lambda$ *ensuring that* $\Pr(\widehat{L}_B \geq 0) \geq 1 - \varepsilon$ *for any* $\varepsilon > 0$. *Under a sub-Gaussian tail bound for* $\widehat{B}$ *(which holds if* $\widehat{B}$ *is assumed normal, as in Lemma 3), it follows that* $m\lambda \geq \sigma_{\widehat{B}}\sqrt{2\log(1/\varepsilon)}$; *see Yang et al. (2025, Section S1) for details.*

Figure 1 illustrates some terms in Lemma 2 under the assumptions in Remark 4 and $m = 1$. The left panel shows the probability of an individual term in the block-Poisson estimator, i.e. $\Pr(\xi < 0)$ in (3.2), being negative as a function of $\sigma^2_{\widehat{B}}$ ($\log_{10}$ scale)

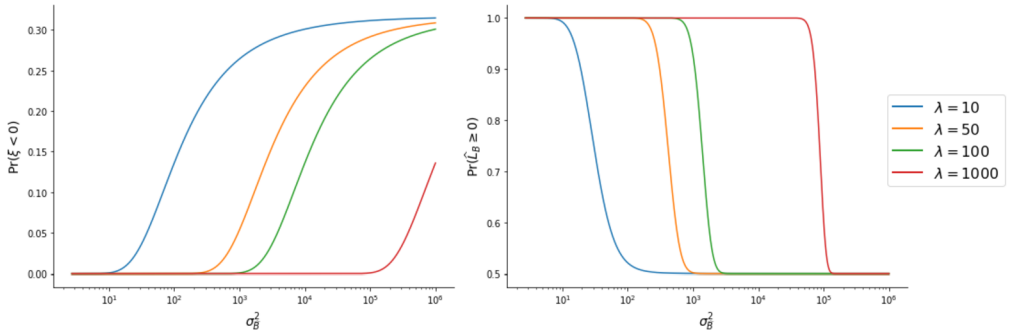


Figure 1: Plots of $\Pr(\xi < 0)$ (left panel) and $\Pr(\widehat{L} \geq 0)$ (right panel) in Lemma 2 when $m = 1$ as a function of $\sigma^2_{\widehat{B}}$ ($\log_{10}$ scale) for various λ values; see legend.

for various λ values. The right panel shows the probability of the overall block-Poisson estimator being positive, i.e. $\Pr(\widehat{L}_B \geq 0)$ in (3.3), for the same $\sigma^2_{\widehat{B}}$ and λ. As λ increases, $\Pr(\xi < 0)$ remains near zero over a wider span of $\sigma^2_{\widehat{B}}$ (left panel). Consequently, the pathological regime (detailed later) where $\Pr(\widehat{L}_B \geq 0) \approx 0.5$ is pushed to much larger $\sigma^2_{\widehat{B}}$. Thus, this regime can be avoided by increasing λ or by reducing $\sigma^2_{\widehat{B}}$ (the latter via increasing M). Section 3.3 develops a tuning strategy that maximises an objective function including $\Pr(\widehat{L}_B \geq 0)$.

Lemma 3 derives the variance of the logarithm of the absolute value of the block-Poisson estimator by assuming that $\widehat{B}^{(h,l)}(\boldsymbol{\theta})$ is normal.

Lemma 3. *If* $\widehat{B}^{(h,l)}(\boldsymbol{\theta}) \overset{\text{iid}}{\sim} N(B(\boldsymbol{\theta}), \sigma^2_{\widehat{B}})$ *for all* h *and* l*, when* $a = a_{\text{opt}} = B(\boldsymbol{\theta}) - m\lambda$*, then the variance of* $\log|\widehat{L}_B|$ *is*

$$\sigma^2_{\log|\widehat{L}_B|} = m\lambda(\nu_B^2 + \eta_B^2),$$

where

$$\eta_B = \log(\sigma_{\widehat{B}}/(m\lambda)) + 0.5\left(\log 2 + E_J(\psi^{(0)}(0.5 + J))\right)$$

and

$$\nu_B^2 = 0.25\left(E_J(\psi^{(1)}(0.5 + J)) + \text{Var}_J(\psi^{(0)}(0.5 + J))\right),$$

with $J \sim \text{Pois}((m\lambda)^2/(2\sigma^2_{\widehat{B}}))$ *and* $\psi^{(q)}$ *is the polygamma function of order* q*.*

Remark 6. *In our method,* $\widehat{B}^{(h,l)}(\boldsymbol{\theta}) = -\nu\widehat{Z}^{(h,l)}(\boldsymbol{\theta})$*. The assumption holds if* $\widehat{Z}^{(h,l)}(\boldsymbol{\theta})$ *is normal (recall that* ν *is treated as non-random in the tuning procedure).*

Section 3.3 tunes the hyperparameters using the results above. However, we have not yet dealt with the fact that $a_{\text{opt}} = B(\boldsymbol{\theta}) - m\lambda$ is itself intractable as it includes the normalisation function. A sensible approach is to estimate the normalising function, i.e. $\widehat{a}_{\text{opt}} = \widehat{B}(\boldsymbol{\theta}) - m\lambda$. One can show that Part (i) still holds if a is random, however,

the extra randomness in $\widehat{L}_B(\boldsymbol{\theta})$ may cause problems such as an infinite $\text{Var}(\widehat{L}_B(\boldsymbol{\theta}))$. We therefore consider the non-random value $a_{\text{sub}} = -1 - m\lambda$, where the subscript highlights that it is a sub-optimal choice. The choice is motivated by $E(\widehat{B}(\boldsymbol{\theta})) = -\nu Z(\boldsymbol{\theta})$, with ν replaced by its expected value $1/Z(\boldsymbol{\theta})$. We note that this sub-optimal choice does not have implications for the exactness of the algorithm. Moreover, Yang et al. (2025, Section S3) includes a simulation study showing that given our assumptions, the sub-optimal soft lower bound a_{sub} often results in similar quantities used by the tuning procedure as those of the optimal choice a_{opt} and its estimated version $\widehat{a}_{\text{opt}}$.

3.2 Signed block PMMH with the BP estimator

Lyne et al. (2015) use an auxiliary variable ν to cancel the reciprocal of the normalising function in (2.3) and end up with $\exp(-\nu Z(\boldsymbol{\theta}))$ (instead of the reciprocal). Specifically, assuming that $\nu \sim \text{Exp}(Z(\boldsymbol{\theta}))$, the joint density of $\boldsymbol{\theta}$ and the auxiliary variable ν is

$$\begin{aligned}\pi(\boldsymbol{\theta}, \nu|\mathbf{y}) &= Z(\boldsymbol{\theta}) \exp(-\nu Z(\boldsymbol{\theta})) \frac{f(\mathbf{y}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \pi(\boldsymbol{\theta}) \frac{1}{p(\mathbf{y})} \\ &\propto \exp(-\nu Z(\boldsymbol{\theta})) f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}). \end{aligned} \tag{3.4}$$

We can use the BP estimator in Section 3.1 to obtain an unbiased estimator of the augmented posterior in (3.4), up to a normalising constant for a fixed (non-random) ν. Denote the unbiased estimator of $\exp(-\nu Z(\boldsymbol{\theta}))$ by $\widehat{\exp}(-\nu Z(\boldsymbol{\theta}))$, where the hat is placed over the exponential for readability, though it represents an estimator of the entire expression. To emphasise the source of randomness in the estimator, let $\mathbf{u}$ be a set of random numbers with density $\pi(\mathbf{u})$ (assumed independent of $\boldsymbol{\theta}$) and write the estimator (with a small abuse of notation) as $\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u})$ for a fixed ν. In the block-Poisson estimator, $\mathbf{u}$ includes all the random numbers used to generate the estimates $\widehat{B}$ and the Poisson variables χ. The unbiasedness of the estimator is with respect to the density $\pi(\mathbf{u})$, i.e.

$$\exp(-\nu Z(\boldsymbol{\theta})) = \int_{\mathbf{u}} \widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u}) \pi(\mathbf{u}) d\mathbf{u}. \tag{3.5}$$

The augmented version of the posterior density in (3.4) is

$$\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}, \nu|\mathbf{y}) = \widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u}) \pi(\mathbf{u}) f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \frac{1}{p(\mathbf{y})}. \tag{3.6}$$

It is easy to show that, under the unbiasedness condition in (3.5), integrating out $\mathbf{u}$ in (3.6) gives the marginal density of interest in (3.4) for $\boldsymbol{\theta}, \nu$. However, we cannot sample from (3.6) using a pseudo-marginal algorithm as the BP estimates may be negative and hence it is not a valid density. We follow Lyne et al. (2015) and consider the target density

$$\overline{\pi}(\boldsymbol{\theta}, \mathbf{u}, \nu) = |\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}, \nu|\mathbf{y})| = |\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u})| \pi(\mathbf{u}) f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \frac{1}{p(\mathbf{y})}. \tag{3.7}$$

Integrating out $\mathbf{u}$ in (3.7) does not give the marginal density of interest in (3.4) because $|\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u})|$ is biased. Lyne et al. (2015) propose reweighting the MCMC iterates using importance sampling to obtain a simulation-consistent estimate of the expectation of an arbitrary function $\psi(\boldsymbol{\theta})$ (assuming the expectation exists) with respect to the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$, i.e.

$$E_\pi(\psi(\boldsymbol{\theta})) = \int_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \tag{3.8}$$

which is outlined in detail in Yang et al. (2025, Section S5). Let the expectation with respect to the augmented target posterior $|\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}, \nu|\mathbf{y})|$ in (3.7) be denoted by $E_{\overline{\pi}}$. The importance sampling estimator is then computed as

$$\widehat{E}_\pi(\psi(\boldsymbol{\theta})) = \frac{\widehat{E}_{\overline{\pi}}(\psi(\boldsymbol{\theta})S(\boldsymbol{\theta}, \mathbf{u}, \nu))}{\widehat{E}_{\overline{\pi}}(S(\boldsymbol{\theta}, \mathbf{u}, \nu))} = \frac{\sum_{i=1}^N \psi(\boldsymbol{\theta}^{(i)})s^{(i)}}{\sum_{i=1}^N s^{(i)}}, \tag{3.9}$$

where $S(\boldsymbol{\theta}, \mathbf{u}, \nu) = \text{sign}(\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}, \nu|\mathbf{y}))$ and $s^{(i)} = S(\boldsymbol{\theta}^{(i)}, \mathbf{u}^{(i)}, \nu^{(i)})$. We often use the shorthand $S = S(\boldsymbol{\theta}, \mathbf{u}, \nu)$ for the sign of the posterior estimate (inherited from the likelihood estimate).

Quiroz et al. (2021) and Lyne et al. (2015) prove a central limit theorem for the estimator in (3.9). The resulting asymptotic variance is finite if (i) $\mathrm{E}_{\overline{\pi}}(S) \neq 0$, corresponding to $\Pr_{\overline{\pi}}(S = +1) \neq 0.5$, which is akin to $\Pr\left(\widehat{L}_B(\boldsymbol{\theta}) \geq 0\right) \neq 0.5$ in Lemma 2, and (ii) the variance and inefficiency factor (under $\overline{\pi}$) of ψS are finite; see (Quiroz et al., 2021, Theorem S1) for details. Although a finite asymptotic variance is reassuring, it does not reflect the performance of the estimator in (3.9) for a finite number of samples, which may still be poor despite the favorable asymptotic behavior. This issue was not addressed in Lyne et al. (2015); Quiroz et al. (2021). In particular, the variance might be large if the denominator in (3.9) is close to zero. The minimal condition for a finite variance of (3.9) is that $\left|\sum_{i=1}^N s^{(i)}\right| > 0$; however, for reliable performance, we would prefer $\left|\sum_{i=1}^N s^{(i)}\right| \gg 0$, say

$$\left|\sum_{i=1}^N s^{(i)}\right| > cN, \qquad \text{for some } 0 < c < |\mu|, \tag{3.10}$$

where $\mu = E_{\overline{\pi}}(S) = 2\tau - 1$, with $\tau = \Pr_{\overline{\pi}}(S = +1)$. To ensure that the sum in (3.10) is bounded away from zero, it is necessary that $\tau \neq 0.5$, i.e. $\mu \neq 0$ (otherwise $c = 0$).

Theorem 1 below shows that the probability of (3.10) can be made arbitrarily close to 1 by increasing N. The proof of the theorem uses the Bernstein-type concentration inequality for Markov chains in Paulin (2015). The following assumptions are made.

Assumption 1. *We assume that:*

(i) $\{\boldsymbol{\theta}^{(i)}, \mathbf{u}^{(i)}, \nu^{(i)}\}_{i=1}^N$ *is a realisation of a stationary, reversible Markov chain on* $\Omega = \mathbb{R}^{\dim(\boldsymbol{\theta})} \times \mathbb{R}^{\dim(\mathbf{u})} \times \mathbb{R}_{>0}$ *with stationary distribution* $\overline{\pi}$ *in* (3.7).

(ii) *The Markov chain in (i) has spectral gap* $0 < \delta < 1$.

(iii) *Define* $S : \Omega \to \Omega'$, $S = \text{sign}(\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}, \nu|\mathbf{y}))$ *with* $\widehat{\pi}$ *in* (3.6) *and* $\Omega' = \{-1, +1\}$. *Let* η *be a measure on* Ω' *with*

$$\eta\left(\{+1\}\right) = \int_{\boldsymbol{\theta}} \int_{\nu} \int_{\mathbf{u}} \mathbb{1}\left(S(\boldsymbol{\theta}, \mathbf{u}, \nu) = +1\right) \overline{\pi}(\boldsymbol{\theta}, \mathbf{u}, \nu) d\mathbf{u} d\nu d\boldsymbol{\theta}, \tag{3.11}$$

where $\mathbb{1}(\cdot)$ *is the indicator function. We assume that* $\eta\left(\{+1\}\right) \neq \eta\left(\{-1\}\right)$, *i.e.* $\eta\left(\{+1\}\right) = \Pr_{\overline{\pi}}(S = +1) \neq 0.5$, *and hence*

$$\mu = E_{\overline{\pi}}(S) = 2\eta\left(\{+1\}\right) - 1 \neq 0. \tag{3.12}$$

Remark 7. *Note that* τ *defined after* (3.10) *and* $\eta\left(\{+1\}\right)$ *are the same,* $\tau = \eta\left(\{+1\}\right)$; (3.11) *shows how the measure* η *is induced from* $\overline{\pi}$.

Assumption 1 is justified as follows. Theorem S1(ii) in Quiroz et al. (2021) states that the signed block pseudo-marginal (block introduced below) algorithm converges in total variation norm to $\overline{\pi}$, which justifies the assumed stationary distribution in Assumption 1(i). For Assumption 1(ii), loosely speaking, the spectral gap quantifies the rate at which a Markov chain mixes, that is, how rapidly it converges to its stationary distribution; see Levin et al. (2009)[Ch. 12] for details. The assumption $0 < \delta < 1$ ensures that we exclude Markov chains that do not mix ($\delta = 0$) or are independent ($\delta = 1$). The justification for excluding the first case is that all chains in our experiments are empirically observed to mix well. The second case is excluded since this implies independent sampling, which is unrealistic. Assumption 1(iii), $\tau = \eta(\{+1\}) = \Pr_{\overline{\pi}}(S = +1) \neq 0.5$, is reasonable because it can be imposed via the analytical expression in Lemma 2; Section 3.3 shows that the tuning procedure gives $\tau \gg 0.5$.

Theorem 1. *Suppose that Assumption 1 holds. Then, for any* $\varepsilon > 0$ *and any* $0 < c < |\mu|$, *with* μ *in* (3.12), *there exists a constant* N_0 *such that,*

$$\Pr_{\overline{\pi}}\left(\left|\sum_{i=1}^{N} s^{(i)}\right| > cN\right) \geq 1 - \varepsilon, \quad \textit{for all } N > N_0. \tag{3.13}$$

Moreover, the convergence of the probability in (3.13) *to 1 is exponentially fast in* N.

Remark 8. *The constant* N_0 *is computable and depends on* ε, c, τ, *and* δ; *see the proof of the theorem in Yang et al. (2025, Section S2) for an expression.*

The concentration result in Theorem 1 ensures that, with high probability, the denominator of the estimator in (3.9) remains bounded away from zero. Consequently, the finite-N estimator $\widehat{E}_{\pi}(\psi(\boldsymbol{\theta}))$ avoids divisions by near-zero values with high probability, and exhibits controlled variability in practice. The proof of the theorem is in Yang et al. (2025, Section S2). Figure 2 illustrates how N_0 varies with c for $\varepsilon = 0.001$ (ensuring the probability in (3.13) is at least 0.999). The left panel corresponds to a well-tuned case ($\tau \gg 0.5$), while the right panel represents a case near the pathological scenario ($\tau \approx 0.5$). Each panel shows the results under three mixing scenarios; see the caption for details. In the well-tuned case, choosing $N \geq 1{,}000$ ($N \gg 1{,}000$ in our examples)

comfortably avoids the pathological regime under moderate mixing $\delta = 0.30$ (chains mix moderately well in our examples with a random walk proposal). In contrast, in the near-pathological case, the required N to remain far from the pathological regime becomes prohibitively large (note the log-scale on the y-axis), even under strong mixing. This highlights the importance of our tuning strategy to avoid such cases (we ensure $\tau \approx 0.99$ in our examples).

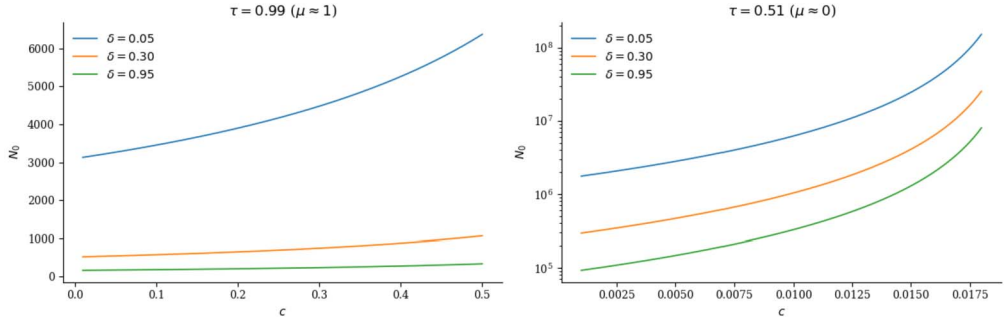


Figure 2: Values of N_0 from Theorem 1 as a function of c ($0 < c < |\mu|$) when $\varepsilon = 0.001$. The left and right panels correspond to $\tau = 0.99$ and $\tau = 0.51$, respectively. Three mixing scenarios for the Markov chain $s^{(i)}$ are shown: $\delta = 0.05$ (slow), $\delta = 0.3$ (moderate), and $\delta = 0.95$ (strong). The right panel uses a log-scale for the vertical axis.

Finally, to make the signed pseudo-marginal algorithm for sampling from (3.7) more efficient, we correlate the estimators at the current and proposed draws to decrease the variability of the difference of the log of the likelihood estimators. This provides a substantial advantage over the standard pseudo-marginal method that proposes $\mathbf{u}$ independently in each iteration (Deligiannidis et al., 2018; Tran et al., 2016). We follow the approach in Tran et al. (2016), where the correlation is induced by blocking the random numbers and only updating one of the blocks when evaluating the likelihood at the proposed value, while keeping the rest of the blocks fixed. The BP estimator uses the random numbers $\mathbf{u}_l$ to estimate ξ_l, $l = 1, \ldots, \lambda$, and group them as $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_\lambda) = \mathbf{u}_{1:\lambda}$. Note that each $\mathbf{u}_l$ includes random numbers of different sizes depending on the value of $\chi_l \sim \text{Pois}(m)$. If the number of blocks λ is sufficiently large, the correlation ρ between the logarithms of the likelihood estimators evaluated at the current and proposed draws is approximately $1 - 1/\lambda$ (Quiroz et al., 2021). We can adjust the number of blocks λ to achieve a prespecified correlation between the log of the likelihood estimates.

Algorithm 1 outlines one iteration of our method when using an exponential proposal for the auxiliary variable ν.[1] Rewriting (3.14) as[2]

$$\frac{\pi(\boldsymbol{\theta}')f(\mathbf{y}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \times \frac{\widehat{Z}_P^{-1}(\boldsymbol{\theta}')}{\widehat{Z}_P^{-1}(\boldsymbol{\theta})} \times \frac{|\widehat{\exp}(-\nu' Z(\boldsymbol{\theta}')|\mathbf{u}'_{1:\lambda})|/\exp(-\nu'\widehat{Z}_P(\boldsymbol{\theta}'))}{|\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u}_{1:\lambda})|/\exp(-\nu\widehat{Z}_P(\boldsymbol{\theta}))},$$

[1] An anonymous reviewer suggested an alternative deterministic proposal for ν, which is outlined in Yang et al. (2025, Section S5)Section S5.

[2] Tran et al. (2016) show that $\pi(\mathbf{u}'_{1:\lambda})q(\mathbf{u}_{1:\lambda}|\mathbf{u}'_{1:\lambda}) = \pi(\mathbf{u}_{1:\lambda})q(\mathbf{u}'_{1:\lambda}|\mathbf{u}_{1:\lambda})$ for a block proposal.

Algorithm 1 One iteration of the signed block PMMH update with the BP estimator.

1: **Input:** Current values of $\nu, \boldsymbol{\theta}, \mathbf{u}_{1:\lambda}$.
2: **Output:** Updated values of $\nu, \boldsymbol{\theta}, \mathbf{u}_{1:\lambda}$ and $\text{sign}(\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}_{1:\lambda}, \nu|\mathbf{y}))$.
3: Generate $\mathbf{u}'_{1:\lambda} \leftarrow \mathbf{u}_{1:\lambda}$ from $q(\mathbf{u}'_{1:\lambda}|\mathbf{u}_{1:\lambda})$ by updating one block of random numbers.
4: Generate $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$.
5: Compute the unbiased estimates, $\widehat{Z}(\boldsymbol{\theta}')$, and use them to construct the BP estimator via (3.2). The proposal distribution of the auxiliary variable ν' is an exponential distribution with mean $1/\widehat{Z}_P(\boldsymbol{\theta}')$:

$$q(\nu'|\boldsymbol{\theta}', \mathbf{u}') = \widehat{Z}_P(\boldsymbol{\theta}') \exp(-\nu' \widehat{Z}_P(\boldsymbol{\theta}')),$$

where $\widehat{Z}_P(\boldsymbol{\theta}')$ is the average of the $\widehat{Z}(\boldsymbol{\theta}')$s used in the BP estimator.
6: Set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$, $\nu \leftarrow \nu'$ and $\mathbf{u}_{1:\lambda} \leftarrow \mathbf{u}'_{1:\lambda}$ with probability

$$\min\left\{1, \frac{|\widehat{\pi}(\boldsymbol{\theta}', \nu', \mathbf{u}'_{1:\lambda}|\mathbf{y})|}{|\widehat{\pi}(\boldsymbol{\theta}, \nu, \mathbf{u}_{1:\lambda}|\mathbf{y})|} \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')q(\mathbf{u}_{1:\lambda}|\mathbf{u}'_{1:\lambda})}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})q(\mathbf{u}'_{1:\lambda}|\mathbf{u}_{1:\lambda})} \frac{\widehat{Z}_P(\boldsymbol{\theta})}{\widehat{Z}_P(\boldsymbol{\theta}')} \frac{\exp(-\nu \widehat{Z}_P(\boldsymbol{\theta}))}{\exp(-\nu' \widehat{Z}_P(\boldsymbol{\theta}'))}\right\}, \quad (3.14)$$

where

$$\widehat{\pi}(\boldsymbol{\theta}, \nu, \mathbf{u}_{1:\lambda}|\mathbf{y}) = \widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u}_{1:\lambda}) f(\mathbf{y}|\boldsymbol{\theta}) \pi(\mathbf{u}_{1:\lambda}) \pi(\boldsymbol{\theta}) p^{-1}(\mathbf{y}),$$

and $\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u}_{1:\lambda})$ is obtained by the BP estimator.
7: Record $s = \text{sign}(\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}_{1:\lambda}, \nu|\mathbf{y}))$ which is also the sign of $\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u}_{1:\lambda})$.

we observe that

$$\frac{|\widehat{\exp}(-\nu' Z(\boldsymbol{\theta}')|\mathbf{u}'_{1:\lambda})|/\exp(-\nu' \widehat{Z}_P(\boldsymbol{\theta}'))}{|\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u}_{1:\lambda})|/\exp(-\nu \widehat{Z}_P(\boldsymbol{\theta}))}$$

acts as a bias-correction for the bias induced when estimating $Z^{-1}(\boldsymbol{\theta}')/Z^{-1}(\boldsymbol{\theta})$ by $\widehat{Z}_P^{-1}(\boldsymbol{\theta}')/\widehat{Z}_P^{-1}(\boldsymbol{\theta})$. When forming the $\widehat{Z}_P$ estimators, we recommend using the average of the corresponding $\widehat{Z}(\boldsymbol{\theta})$s used in the BP estimator. This does not affect the unbiasedness property of the BP estimator and is computationally efficient, as the $\widehat{Z}(\boldsymbol{\theta})$s are already computed and the extra cost in obtaining the average is negligible.

3.3 Tuning the signed block PMMH with the BP estimator

Pitt et al. (2012) provide guidelines to tune the number of particles, i.e. the number of samples used in the likelihood estimation procedure, in a pseudo-marginal algorithm with a positive unbiased estimator to achieve an optimal trade-off between computing time and MCMC efficiency as measured by the integrated autocorrelation time (IACT), also known as the inefficiency factor (IF). Suppose that $\theta^{(j)}$, $j = 1, 2, \ldots$, are the iterates after convergence of the MCMC and let $\vartheta^{(j)} = \psi(\theta^{(j)})$ be a scalar function of the iterates. Let r_τ be the correlation between $\vartheta^{(j)}$ and $\vartheta^{(j+\tau)}$. In pseudo-marginal algorithms, r_τ depends on the variance of the log of the likelihood estimator $\widehat{L}$, which we denote by

$\sigma^2_{\log \widehat{L}}$. The inefficiency factor is defined as

$$\mathrm{IF}(\sigma^2_{\log \widehat{L}}) = 1 + 2\sum_{\tau=1}^{\infty} r_\tau(\sigma^2_{\log \widehat{L}}).$$

A larger $\sigma^2_{\log \widehat{L}}$ results in a stickier chain and thus $\mathrm{IF}(\sigma^2_{\log \widehat{L}})$ is an increasing function of $\sigma^2_{\log \widehat{L}}$; see Pitt et al. (2012) for details. To also take the computing time into account when determining the number of particles to use in the estimation of $\log \widehat{L}$, Pitt et al. (2012) show that the number of particles is inversely proportional to $\sigma^2_{\log \widehat{L}}$ and define the computational time $\mathrm{CT}(\sigma^2_{\log \widehat{L}}) = \mathrm{IF}(\sigma^2_{\log \widehat{L}})/\sigma^2_{\log \widehat{L}}$. This measure takes into account both the mixing of the chain (through IF) and the cost of computing the estimator (through the number of particles, which is inversely proportional to $\sigma^2_{\log \widehat{L}}$). Pitt et al. (2012) show that, under certain simplifying assumptions, $\sigma^2_{\log \widehat{L}} \approx 1$ is optimal, and thus the guideline is to choose the number of particles to achieve this.

Quiroz et al. (2021) extend the guidelines in Pitt et al. (2012) to cases when the likelihood estimator is not necessarily positive. The derivation of our guidelines follows those in Quiroz et al. (2021), with modifications that account for a different estimator. Following Quiroz et al. (2021), we tune the hyperparameters by minimising the following computational time (CT)

$$\mathrm{CT} = m\lambda M \frac{\mathrm{IF}_{|\widehat{\pi}|,\psi s}\left(\sigma^2_{\log|\widehat{L}_B|}(m,\lambda,M|\gamma)\right)}{(2\tau(m,\lambda,M)-1)^2}, \tag{3.15}$$

where the dependence on $\boldsymbol{\theta}$ is omitted for γ (defined in (3.16) below) and $\tau(\cdot) = \Pr(\widehat{L}_B(\boldsymbol{\theta}) > 0)$ in Lemma 2. The first term $m\lambda M$ in (3.15) is proportional to the expected cost per iteration since there are λ blocks in total and each block includes m estimates on average with M Monte Carlo samples in each. The latter refers to the number of samples used to produce a single $\widehat{B}^{(h,l)}$ in (3.2). The denominator in (3.15) shows that it is important to have a large proportion of estimates of the same sign and that having close to half of the estimates is detrimental for the CT. The numerator in (3.15) is the inefficiency factor, which measures the MCMC sampling efficiency of drawing ψ's from the target distribution $|\widehat{\pi}|$. Quiroz et al. (2021, Section S2) derives the specific form of the IF. The IF is determined by the variance of the log of the absolute likelihood estimator $\sigma^2_{\log|\widehat{L}_B|}$ (recall the discussion when tuning using a positive likelihood estimator above) in Lemma 3, which in turn depends on the hyperparameters m, λ, M, and in addition γ. We define $\gamma(\boldsymbol{\theta})$ as the variance of a single Monte Carlo sample, say $-\nu\widehat{Z}_i(\boldsymbol{\theta})$, i.e. such that $\sigma^2_{\widehat{B}} = \gamma(\boldsymbol{\theta})/M$ for the Monte Carlo estimate $\widehat{B} = \sum_{i=1}^{M} -\nu\widehat{Z}_i(\boldsymbol{\theta})/M$ based on M independent Monte Carlo samples. Note that $\gamma(\boldsymbol{\theta})$ does not depend on M; however, it depends on ν (recall ν is treated as non-random) as

$$\begin{aligned}\gamma(\boldsymbol{\theta}) = \mathrm{Var}(-\nu\widehat{Z_i}(\boldsymbol{\theta})) &= \nu^2\mathrm{Var}(\widehat{Z_i}(\boldsymbol{\theta})) \\ &= \frac{2}{Z(\boldsymbol{\theta})^2}\mathrm{Var}(\widehat{Z_i}(\boldsymbol{\theta})).\end{aligned} \tag{3.16}$$

Similarly to when we replaced ν by its expected value when determining a_{sub} in Section 3.1, ν^2 in (3.16) is replaced by its second moment when tuning the hyperparameters.

Figure 3 shows the effects of the number of blocks (λ) and Monte Carlo samples (M) on the logarithm of CT, τ and $\sigma^2_{\log|\widehat{L}_B|}$. We consider the three cases $\gamma = 10^2, 100^2, 500^2$ (left to right columns respectively) which show that the optimal λ (corresponding to the minimal CT) varies with different values of M and increases with γ (top row). The minimum CT is associated with a high probability of a positive estimator (τ) (middle row). The last row indicates that $\sigma^2_{\log|\widehat{L}_B|}$ decreases as a function of λ for large λ. Comparing the top nine panels with the bottom nine, a high correlation $\rho = 0.99$, reduces λ_{opt} from 295 (no correlation, $\rho = 0$) to 195 for $\gamma = 500^2$. Conversely, $\rho = 0.99$ requires at least 100 blocks. So when the variance γ is small, introducing a high correlation increases the CT as more blocks are required compared to the uncorrelated case. Our implementation follows the approach in Tran et al. (2016) which sets the correlation ρ to a value close to 1. Comparing the first row of the top panel (a) in Figure 3 with that of the bottom panel (b), shows that a high correlation significantly reduces the CT (y-axis is in log-scale) per iteration for large γ.

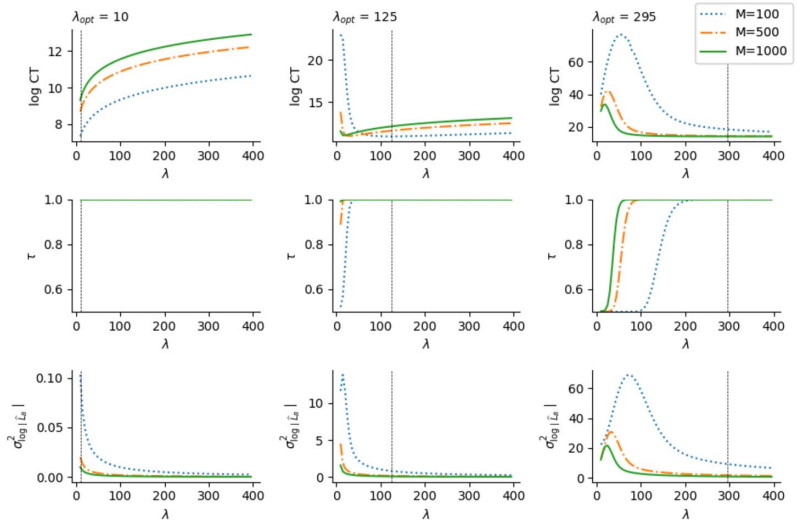
We conclude that the tuning depends on γ in (3.16). For conservative tuning, we set γ to a large value $\gamma_{\max}$ by using a grid search over possible $\boldsymbol{\theta}$, meaning the process is calibrated to guard against worst-case scenarios. Note that grid search scales poorly beyond two dimensions; in such cases, Bayesian optimisation (Shahriari et al., 2016) offers a practical alternative for the parameter spaces typical of doubly intractable problems. The tuning process starts with fixed values of λ and m to find the optimal value for M by minimising (3.15). In Figure 4, we fix the values of λ and m, with $\lambda = 50, 100$ (the corresponding ρ are 0.98 and 0.99 respectively), and $m = 1$. A standard optimiser is used to find the optimal value M_{opt} for each of the γ. The dots in the left panel of Figure 4 plot M_{opt} for various values of $\sqrt{\gamma}$.[3] The figure shows that M_{opt} increases as a function of $\sqrt{\gamma}$. The right panel shows how the minimised log CT increases as a function of $\sqrt{\gamma}$. To estimate the relationship between M_{opt} and $\sqrt{\gamma}$, a quadratic polynomial is fitted to the points in the left panel.

The tuning below is based on $\gamma_{\max}$, leading to a conservative tuning of M_{opt}.

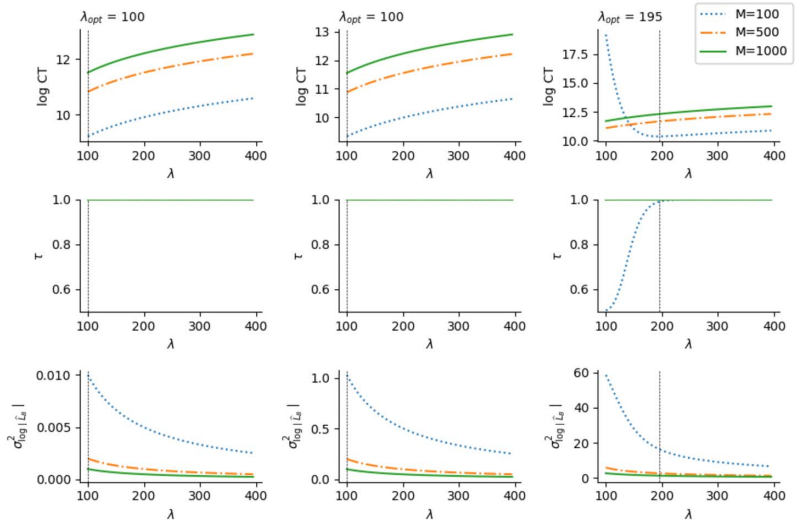
1. Obtain a rough understanding of the support of the posterior distribution of $\boldsymbol{\theta}$. Yang et al. (2025, Section S4) proposes an approximate method that relies on a normality assumption of $\widehat{Z}(\boldsymbol{\theta})$, which can be used in the tuning phase. Alternatively, we can apply another approximate method tailored for the specific problem (e.g. the variational mean-field approximation of the Ising model in Jain et al., 2018). It is also possible to optimise the posterior distribution by plugging the biased estimator $(1/\widehat{Z}(\boldsymbol{\theta}))$.

2. Estimate $\gamma(\boldsymbol{\theta})$ in (3.16) for the $\boldsymbol{\theta}$ candidates from Step 1. The estimator $\widehat{Z_i}(\boldsymbol{\theta})$ replaces the unknown $Z(\boldsymbol{\theta})$.

[3] Taking the square root puts the quantity on the standard deviation scale.



(a) $\rho = 0$



(b) $\rho = 0.99$

Figure 3: The effect of the number of blocks λ on the logarithm of CT, τ and $\sigma^2_{\log|\widehat{L}_B|}$. The Poisson parameter m is fixed at 1 for each set of panels (a,b). The correlation term is set to $\rho = 0$ (upper panel), 0.99 (bottom panel). Columns from left to right correspond to three different settings of $\gamma = 10^2$, 100^2, and 500^2. The top, middle and last rows of each panel show the log of the CT in (3.15), the probability of obtaining a positive estimator $\tau(m, \lambda, M)$ (see Lemma 2) and the variance of log of the absolute value of the likelihood estimate (see Lemma 3). The vertical line on each plot represents λ_{opt}, the optimal λ, which minimises the log of the CT within each of the settings.

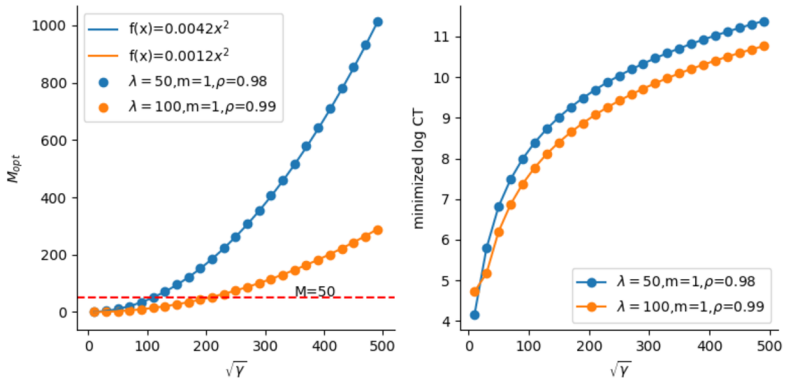


Figure 4: Left panel: The optimal value M_{opt} as a function of $\sqrt{\gamma}$. The lines are quadratic polynomials fitted to the scattered dots. The horizontal dashed line represents the threshold $M = 50$, which is the minimal number of blocks required in the algorithm. Right panel: The minimised log CT as a function of $\sqrt{\gamma}$.

3. Obtain the maximum value $\gamma_{\max}(\boldsymbol{\theta})$ of $\gamma(\boldsymbol{\theta})$ from Step 2. A sensible first tuning attempt sets $\lambda = 100, m = 1, \rho = 0.99$ and $M_{\text{opt}} = \max\{50, 0.0012 \times \gamma_{\max}(\boldsymbol{\theta})\}$.

   If $\gamma_{\max}(\boldsymbol{\theta})$ is small to moderately large, e.g. $\gamma_{\max}(\boldsymbol{\theta}) < 100^2$, having many blocks increases CT (e.g. top left and top middle panels in Figure 3). In this case, a weaker correlation also produces an efficient algorithm with smaller CT. A suitable setting in this scenario is $\lambda = 50, m = 1, \rho = 0.98$ and $M_{\text{opt}} = \max\{50, 0.0042 \times \gamma_{\max}(\boldsymbol{\theta})\}$.

   For an even smaller $\gamma(\boldsymbol{\theta})$, the correlation can be relaxed further. In the Ising model example, $\lambda = 10$ is sufficient when the variability is low; see Section 4.1 and Yang et al. (2025, Section S8).

The tuning is theoretically sub-optimal as it is based on simplifying assumptions in its derivation and practical implementation. Yang et al. (2025, Section S3) presents a simulation study demonstrating that tuning based on the simplified lower bound closely approximates the intractable one, which is reassuring. In our experience, the main objective of the tuning is met by our guidelines, namely: avoiding the pseudo-marginal chain getting stuck. Finally, we note that the algorithm is still exact under the simplified tuning.

4 Simulation study

This section empirically compares the performance of our method on a simulated Ising model example from Park and Haran (2018) with some competing methods. An additional simulated example for the Kent distribution is found in Section S9.3. All our comparisons are on an empirical level rather than a theoretical level. Theoretical comparisons are either impossible because the competing methods are analytically intractable

or outside the scope of this paper. Examples of the former are that expressions of the variance of the log of the absolute value or the probability of a positive estimator are not readily available for Russian roulette estimators. An example of the latter includes extending the ordering of pseudo-marginal MCMC chains (Andrieu and Vihola, 2016) to a signed pseudo-marginal setting.

The example in the main paper is the Ising model, which is the usual benchmark example for doubly intractable problems, as perfect sampling is efficient for this model on small grids, and thus the exchange algorithm (Murray et al., 2006) can provide a ground truth. We use the same settings as the survey paper Park and Haran (2018), in which the Ising model serves as the main benchmark example. The second example, which is included in Yang et al. (2025, Section S9.3), considers the Kent distribution, where the intractable normalising function is an infinite sum. Unlike the Ising model, efficient perfect sampling is hard for this model, and thus the exchange algorithm fails. Section Yang et al. (2025, Section S6) provides the implementation details for the exchange algorithm. To the best of our knowledge, exact Bayesian inference has not been considered for the Kent distribution due to its intractability.

Our method, abbreviated BP, is a correlated signed pseudo-marginal method that utilises the block-Poisson (BP) estimator and is compared to the signed pseudo-marginal methods introduced in Lyne et al. (2015). Two implementations of the latter are considered, which differ by their use of auxiliary variables and what form of the normalising function they estimate. RR-aux, uses a Russian roulette (RR) estimator of the exponent of the normalising function and requires auxiliary variables (-aux) to turn the reciprocal into an exponent as in (3.4). RR, on the other hand, estimates the reciprocal directly and does not require auxiliary variables. Yang et al. (2025, Section S6) provides the implementation details for both Russian roulette methods.

Table 1 lists the requirements and features of the exact methods considered in our paper. The multiple observations case is relevant for the Kent distribution example, and Section S9.1 in the supplement discusses its scalability for the different estimators. Yang et al. (2025, Section S7) provides a detailed computational analysis, including the structural differences between the BP, RR, and RR-aux estimators, and the derivation of their arithmetic complexities. In summary, the BP estimator offers computational advantages due to its independent product structure that enables efficient parallelisation and vectorisation via single instruction, multiple data (SIMD) instructions (Warne et al., 2022; Hennessy and Patterson, 2011, Ch. 4). The Russian roulette estimators RR and RR-aux, in contrast, involve a sum of dependent, nested products, and thus cannot exploit parallelisation or vectorisation as effectively, due to their sequential computation pattern. As shown in Section S7 in the supplement, achieving a similar parallelism as BP by recomputing each nested product, thereby creating an independent sum, incurs a quadratic computational cost in the number of sum terms, compared to the linear cost of the BP product.

Finally, we emphasise again that the signed pseudo-marginal methods are exact in providing simulation-consistent estimates of expectations under the exact doubly intractable posterior. By contrast, the exchange algorithm provides the usual exact inference MCMC methods do, i.e. samples from the invariant distribution (the posterior

density) after burn-in. However, for a finite number of MCMC iterations, the exchange algorithm might not have properly converged due to mixing problems, which can arise if sampling from the likelihood is inefficient. In the Ising example, this occurs for large grids, and in such cases, signed pseudo-marginal algorithms are tractable alternatives (Lyne et al., 2015). For the Kent distribution example, we find that the accept/reject method in Kent et al. (2013) does not give efficient sampling from the likelihood in any of our settings.

Requirements and features / Method	BP	RR	RR-aux	Exchange
Auxiliary variable(s) required	Yes	No	Yes	Yes
Sampling from the likelihood without knowing the normalising function	No	No	No	Yes
Correlated pseudo-marginal (PM) (only PM methods)	Yes	No	No	NA
Estimator scales with multiple observations	Yes	No	Yes	NA
Estimator utilises vectorisation and parallelisation	Yes	No	No	NA

Table 1: Requirements (first 2) and features (bottom 3) for the methods. BP = block-Poisson, RR = Russian roulette, RR-aux = Russian roulette with an auxiliary variable. NA stands for not available.

4.1 The Ising model

The Ising model (Lenz, 1920; Ising, 1925) has widespread applications, such as understanding phase transitions in thermodynamic systems (Fredrickson and Andersen, 1984), interactive image segmentation in vision problems (Kolmogorov and Zabin, 2004) and modelling small-world networks (Herrero, 2002). It is the typical benchmark example in the literature to evaluate methods for tackling the doubly intractable problem; see e.g. Møller et al. (2006); Lyne et al. (2015); Atchadé et al. (2013); Park and Haran (2018). However, most of the existing methods use auxiliary variable approaches, as it is feasible to draw observations from the likelihood function perfectly, so-called perfect sampling, for moderately small grids. The signed pseudo-marginal methods, such as RR and our approach, do not require perfect sampling, which makes them applicable to more general problems. Lyne et al. (2015) report that their pseudo-marginal approach handles larger grids than the exchange algorithm.

Recall Example 1 in Section 2.1, where θ is a scalar parameter and $S(\mathbf{y})$ imposes spatial dependence; a stronger interaction between observations is associated with a larger θ. The data simulations are conducted using perfect sampling (Propp and Wilson, 1996), which samples exactly without evaluating the normalising function. Perfect sampling uses coupling to guarantee that the samples are generated from a Markov chain which has already converged to its equilibrium distribution. Following the set-

tings in Park and Haran (2018), two scenarios are considered on a 10×10 grid, with $\theta = 0.2$ and 0.43; see Figure S3 in the supplement for an illustration.

For all the algorithms considered, a uniform distribution on $[0, 1]$ is selected as the prior for θ. We adopt a random walk proposal centred at the current θ with a step size 0.07. The signed pseudo-marginal methods (RR, RR-aux, BP) require an unbiased estimator for $Z(\theta)$. We use annealed importance sampling (AIS) (Neal, 2001) to obtain the estimate of $Z(\theta)$. The method starts by sampling from a tractable distribution (the prior) and ends at the intractable target (the posterior) via a sequence of intermediate distributions. The transitions between the distributions are completed via Gibbs updates, and the weights associated with the transitions finally constitute the normalising function of interest; see Neal (2001) for details of AIS in general and Yang et al. (2025, Section S8) for its implementation for the Ising model.

To obtain the "gold" standard to evaluate the accuracy of the results, we follow Park and Haran (2018), where an exchange algorithm with 1,010,000 iterations is performed. The first 10,000 iterations are discarded for burn-in and the remaining iterates are thinned so that 10,000 posterior samples remain. We use the same set of hyperparameters in RR and RR-aux; see details in Yang et al. (2025, Section S8). For the tuning of the BP method, the two scenarios $\theta = 0.2$ and $\theta = 0.43$ result in different $\gamma_{\max}$ values, with the former yielding a smaller and the latter a larger value for the guidelines. We find that when $\theta = 0.2$, the AIS method gives a sufficiently low value for $\gamma_{\max}$ so that $\lambda = 10$ is appropriate. When $\theta = 0.43$, on the other hand, the strong dependence leads to higher variability in $\widehat{Z}(\theta)$ (see Yang et al., 2025, Section S8). We therefore increased the number of blocks to 50 for the BP estimator as per the tuning guidelines. To ensure a fair comparison, we also increased the number of particles (number of samples) in the importance samplers of AIS from 100 to 500 for RR and RR-aux to decrease the variance.

Table 2 summarises the simulation results. When $\theta = 0.2$, all the estimates are close to those of the gold standard, which is expected since the methods are simulation-consistent and no major mixing problems are encountered. The BP method has the smallest computing time and the second smallest IACT, and obtains a factor of roughly $14\times$ improvement in terms of effective sample size per unit of computing time (in sec) compared to the competing Russian roulette approaches. When $\theta = 0.43$, the results of the BP and RR match well with those of the gold standard; however, not for RR-aux due to severe mixing issues; the chain gets stuck after around 1,000 iterations as reported in Yang et al. (2025, Figure S4). We were unable to find settings for RR-aux to work when $\theta = 0.43$, which emphasizes the importance of having tuning guidelines. In this example, BP is a factor of roughly $2\times$ more efficient than RR in terms of effective sample size per computing time.

4.2 The Kent distribution

This example is outlined in detail in Yang et al. (2025, Section S9.3). We here only summarise the main findings.

$\theta_{\text{true}} = 0.2$							
Method	Mean	95%HPD	IACT	Time(s)	ESS/s	λ	# particles
Gold (Exchange)	0.205	(0.075, 0.337)	1	–	–	–	–
BP	0.203	(0.075, 0.334)	8.39	606	3.9	10	100
RR	0.199	(0.072, 0.324)	7.59	9,825	0.27	–	100
RR-aux	0.199	(0.069, 0.332)	9.16	7,447	0.32	–	100
$\theta_{\text{true}} = 0.43$							
Method	mean	95%HPD	IACT	time(s)	ESS/s	λ	# particles
Gold (Exchange)	0.433	(0.330, 0.533)	1.04	–	–	–	–
BP	0.435	(0.324, 0.540)	7.30	4,072	0.67	50	100
RR	0.435	(0.330, 0.548)	7.16	10,016	0.29	–	500
RR-aux	0.633	(0.633, 0.633)	NA	87,525	NA	–	500

Table 2: Results for the Ising model. All the chains, except for the "gold standard", ran for 20,000 iterations using the algorithms (Gold = exchange algorithm, BP = block-Poisson, RR = Russian roulette, RR-aux = Russian roulette with an auxiliary variable). For BP, RR and RR-aux, the mean estimates are corrected for the negative estimates using (3.9). The highest posterior density (HPD) intervals are calculated by the `coda` package in `R`. The IACT calculation is based on all the samples as the chains start at the true value. For BP, RR and RR-aux, the calculation of the IACT accounts for the negative estimates via (3.15). Time denotes the running time in seconds. ESS/s is the effective sample size per second. For BP, λ refers to the number of blocks; # particles is the number of particles (number of samples) used in the AIS. NA indicates not available due to sampler nonconvergence.

The results indicate that the block-Poisson (BP) estimator consistently delivers the most accurate and computationally efficient inference. The RR-auxiliary (RR-aux) method attains comparable accuracy but at a considerably higher computational cost, while the plain Russian-roulette (RR) and exchange methods perform poorly owing to slow mixing and limited scalability. As the sample size increases, the ESS_κ/s gap between BP and the RR methods becomes more pronounced, with factors of roughly 17×–35× improvement compared to RR-aux, and more than 70×–140× compared to RR when $n = 1{,}000$. Among the frequentist approaches, the moment estimator fails near the boundary $\beta/\kappa \approx 0.5$, whereas the maximum likelihood estimate (MLE) improves with larger samples but remains less accurate than the Bayesian methods.

5 An empirical study on spherical data

We now analyse four real spherical datasets using the Kent distribution and our method. Each data set contains samples from two groups that are formed naturally from the

sample collection process. Figure 5 plots the spherical datasets.

1. **Palaeomagnetic** (Palaeo) (Wood, 1982): Thirty-three estimates of previous magnetic pole positions were obtained using palaeomagnetic techniques. Each estimate is associated with a different site in Tasmania. The data is originally from Schmidt (1976) and the author points out that the data is likely to fall mainly into two groups of distinct geographical regions. Following Figueiredo (2009), the first group contains 9 observations with the indices 9, 10, 11, 12, 14, 16, 23, 24, 30. The second group has 24 observations.

2. **Magnetic** (Fisher et al., 1993, Table B8): Measurements of magnetic remanence from a set of 62 specimens is obtained. The specimens are from Mesozoic Dolerite from Prospect, New South Wales, after successive partial demagnetisation stages (200° and 350°). An experiment was conducted to determine the blocking temperature spectrum of the magnetisation components.

3. **Sandstone** (Fisher et al., 1993, Table B23): Measurements of natural remanent magnetisation in Old Red Sandstone rocks in Pembrokeshire, Wales. The measurements consist of specimens from two sites with the number of observations 35 and 13, respectively.

4. **Stone** (Fisher et al., 1993, Table B25): Measurements of the longest axis and shortest axis (101 observations) orientations of tabular stones on a slope at Windy Hills, Scotland.

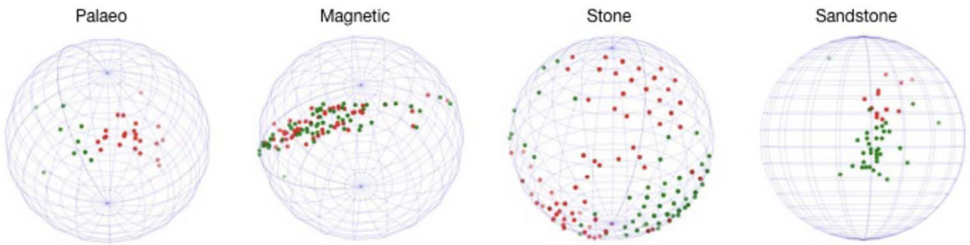


Figure 5: Illustration of the datasets. Green points and red points refer to the observations from groups 1 and 2, respectively.

The two groups are modelled separately by assuming a non-hierarchical structure on the prior for all the parameters. The data is modelled in the same way as in Section 4.2 using the density function in (2.2).

Table 3 shows the posterior mean/estimate, ESS and ESS per second for β/κ. The results for β and κ are in Yang et al. (2025, Section S9.4). The table shows that BP, RR and RR-aux have similar estimates and credible intervals for all four datasets. Exchange has a significantly larger posterior mean than those of the other Bayesian methods. It also has wider credible intervals for all datasets. The abnormal performance of Exchange is consistent with the results in Section 4.2, where we attribute this issue to the sampling

Paleo	Group 1 (n = 9)			Group 2 (n = 24)		
	β/κ	$\text{ESS}_{\beta/\kappa}$	$\text{ESS}_{\beta/\kappa}$/s	β/κ	$\text{ESS}_{\beta/\kappa}$	$\text{ESS}_{\beta/\kappa}$/s
BP	0.15	235.49	9.33	0.14	201.70	21.12
RR	0.15	49.47	1.37	0.12	201.60	2.36
RR-aux	0.19	158.03	11.11	0.10	165.48	9.14
Exchange	0.27	118.14	8.80	0.19	75.02	5.76
Moment	0.15	–	–	0.15	–	–
MLE	0.17	–	–	0.16	–	–
Magnetic	**Group 1 (n = 62)**			**Group 2 (n = 62)**		
	β/κ	$\text{ESS}_{\beta/\kappa}$	$\text{ESS}_{\beta/\kappa}$/s	β/κ	$\text{ESS}_{\beta/\kappa}$	$\text{ESS}_{\beta/\kappa}$/s
BP	0.48	184.78	18.50	0.49	140.27	14.43
RR	0.48	44.38	0.26	0.50	1.90	< 0.01
RR-aux	0.48	146.95	5.37	0.49	153.40	5.26
Exchange	0.49	155.64	10.11	0.49	117.07	7.63
Moment	0.35	–	–	0.38	–	–
MLE	0.50	–	–	0.50	–	–
Sandstone	**Group 1 (n = 35)**			**Group 2 (n = 13)**		
	β/κ	$\text{ESS}_{\beta/\kappa}$	$\text{ESS}_{\beta/\kappa}$/s	β/κ	$\text{ESS}_{\beta/\kappa}$	$\text{ESS}_{\beta/\kappa}$/s
BP	0.08	147.92	5.87	0.20	97.11	10.28
RR	0.06	98.09	0.88	0.17	149.16	2.16
RR-aux	0.08	218.60	10.43	0.17	111.36	7.38
Exchange	0.17	95.56	6.65	0.28	7.24	0.56
Moment	0.09	–	–	0.27	–	–
MLE	0.11	–	–	0.29	–	–
Stone	**Group 1 (n = 101)**			**Group 2 (n = 101)**		
	β/κ	$\text{ESS}_{\beta/\kappa}$	$\text{ESS}_{\beta/\kappa}$/s	β/κ	$\text{ESS}_{\beta/\kappa}$	$\text{ESS}_{\beta/\kappa}$/s
BP	0.13	82.67	3.31	0.49	197.48	20.54
RR	0.11	10.81	0.03	0.49	11.79	0.04
RR-aux	0.14	75.69	2.05	0.49	138.70	3.78
Exchange	0.43	71.26	5.01	0.49	239.20	18.49
Moment	0.06	–	–	0.21	–	–
MLE	0.14	–	–	0.5	–	–

Table 3: Results for the Kent model for the four datasets when estimating β/κ. All the chains ran for 10,000 iterations. The term β/κ refers to the posterior mean for the Bayesian methods and estimates for the frequentist methods. The sign correction is applied for BP, RR and RR-aux.

inefficiency from a FB_5 distribution. In addition, while BP and RR-aux occasionally trade places when ranked by ESS, the ESS/s metric yields a more decisive ordering: BP outperforms RR-aux in all cases but two (Group 1 for Sandstone and Paleo). The factors of improvement span a wide range of roughly 1.6× to 7×. In contrast, RR yields the

lowest ESS as well as ESS/s in every scenario, demonstrating its poor performance on the Kent distribution and its lack of scalability to multiple observations. The moment estimation demonstrates significant deviations from both the Bayesian results and MLE estimates in the limiting case where $\beta/\kappa \approx 0.5$ occurs (Magnetic and Stone, Group 2). See Yang et al. (2025, Section S9.4) for posterior mean/estimates with 95% credible intervals (confidence interval for Moment and MLE).

6 Conclusions and future research

We propose the signed block PMMH with the block-Poisson estimator for exact inference in doubly intractable problems. The method requires only an unbiased estimator of the normalising function, making it applicable to a broader class of models than competitors such as the exchange algorithm, which relies on perfect sampling. We also derive a finite-sample result ensuring, with high probability, that the denominator of the importance-sampling estimator remains bounded away from zero, removing a pathological case in signed PMMH methods overlooked in the literature.

Compared with the Russian roulette approach of Lyne et al. (2015), the block-Poisson estimator yields a smaller variance of the log-likelihood difference in the MH acceptance ratio through correlated pseudo-marginal updates. The Russian roulette method lacks clear hyperparameter-tuning guidance; we provide heuristic rules based on analytical properties. In the Ising model, our approach is 2–14 times more efficient (time-normalised), while in the Kent model with $n = 1{,}000$, efficiency gains range from 17–35 times over RR-aux and 70–140 over RR.

The signed PMMH algorithm can be computationally demanding when unbiased estimation of the normalising function is expensive. In the Ising model, each iteration requires several annealed importance-sampling runs that dominate runtime, making the BP–RR difference less pronounced. When unbiased estimates are inexpensive, as in the Kent model, BP's advantages are substantial.

Future work includes extending the method to other doubly intractable problems, developing tuning strategies for non-Gaussian estimators of the normalising function, and obtaining tighter bounds for the concentration inequality in Theorem 1, where the Bernstein-type improvements via iterated Poincaré inequalities in Huang and Li (2024) may prove useful. Finally, an anonymous reviewer suggested the Rao–Blackwellised Metropolis–Hastings with Averaged Acceptance Ratios algorithm of Andrieu et al. (2020) as a potential alternative to the correlated pseudo-marginal approach. We leave the investigation of a signed extension of this method for future work.

Acknowledgements

We thank Chris Sherlock for helpful comments on an earlier version of this manuscript. We thank the Associate Editor and two referees for helpful comments that significantly improved the manuscript.

Funding

Yu Yang was financially supported by a University International Postgraduate Award from UNSW Sydney. Matias Quiroz was partially supported by the Marine Ecosystems Research Mobilising AI and Data (MERMAID) Collaboration (CLB-3127). Robert Kohn was partially supported by the Australian Research Council (IC190100031, DP210103873). Scott Sisson was supported by the Australian Research Council (FT170100079).

Supplementary Material

Supplementary Material to "A correlated pseudo-marginal approach to doubly intractable problems" (DOI: 10.1214/25-BA1573SUPP; .pdf). The supplement contains all proofs and full details of the simulation studies, including the Kent distribution example.

References

Alquier, P., Friel, N., Everitt, R. G., and Boland, A. (2016). "Noisy Monte Carlo: convergence of Markov chains with approximate transition kernels." *Statistics and Computing*, 26: 29–47. MR3439357. doi: https://doi.org/10.1007/s11222-014-9521-x. 6

Andrieu, C. and Roberts, G. O. (2009). "The pseudo-marginal approach for efficient Monte Carlo computations." *The Annals of Statistics*, 37(2): 697–725. MR2502648. doi: https://doi.org/10.1214/07-AOS574. 2

Andrieu, C. and Vihola, M. (2016). "Establishing some order amongst exact approximations of MCMC." *The Annals of Applied Probability*, 26(5): 2661–2696. MR3563190. doi: https://doi.org/10.1214/15-AAP1158. 19

Andrieu, C., Yıldırım, S., Doucet, A., and Chopin, N. (2020). "Metropolis-Hastings with averaged acceptance ratios." *arXiv preprint arXiv:2101.01253*. 5, 25

Atchadé, Y. F., Lartillot, N., and Robert, C. (2013). "Bayesian computation for statistical models with intractable normalizing constants." *Brazilian Journal of Probability and Statistics*, 27(4): 416–436. MR3105037. doi: https://doi.org/10.1214/11-BJPS174. 5, 20

Beaumont, M. A. (2003). "Estimation of population growth or decline in genetically monitored populations." *Genetics*, 164(3): 1139–1160. 2

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press. MR2742422. doi: https://doi.org/10.1201/b10905. 1

Cai, D. and Adams, R. P. (2022). "Multi-fidelity Monte Carlo: A pseudo-marginal approach." *Advances in Neural Information Processing Systems*. 6

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. 5

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018). "The correlated pseudomarginal method." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5): 839–870. MR3874301. doi: https://doi.org/10.1111/rssb.12280. 3, 13

Figueiredo, A. (2009). "Discriminant analysis for the von Mises-Fisher distribution." *Communications in Statistics-Simulation and Computation*, 38(9): 1991–2003. MR2751184. doi: https://doi.org/10.1080/03610910903200281. 23

Fisher, N. I., Lewis, T., and Embleton, B. J. (1993). *Statistical Analysis of Spherical Data*. Cambridge University Press. MR1247695. 23

Fredrickson, G. H. and Andersen, H. C. (1984). "Kinetic Ising model of the glass transition." *Physical Review Letters*, 53(13): 1244. 20

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57(1): 97–109. MR3363437. doi: https://doi.org/10.1093/biomet/57.1.97. 2

Hennessy, J. L. and Patterson, D. A. (2011). *Computer Architecture: A Quantitative Approach*. Elsevier. 19

Herrero, C. P. (2002). "Ising model in small-world networks." *Physical Review E*, 65(6): 066110. MR2037587. doi: https://doi.org/10.1103/PhysRevE.68.036106. 20

Huang, D. and Li, X. (2024). "Bernstein-type inequalities for Markov chains and Markov Processes: A Simple and robust proof." *arXiv preprint arXiv:2408.04930*. 25

Hughes, J., Haran, M., and Caragea, P. C. (2011). "Autologistic models for binary data on a lattice." *Environmetrics*, 22(7): 857–871. MR2861051. doi: https://doi.org/10.1002/env.1102. 2

Hunter, D. R. and Handcock, M. S. (2006). "Inference in curved exponential family models for networks." *Journal of Computational and Graphical Statistics*, 15(3): 565–583. MR2291264. doi: https://doi.org/10.1198/106186006X133069. 2

Ising, E. (1925). "Beitrag zur theorie des ferromagnetismus." *Zeitschrift für Physik*, 31(1): 253–258. 2, 4, 20

Jacob, P. E. and Thiery, A. H. (2015). "On nonnegative unbiased estimators." *The Annals of Statistics*, 43(2): 769–784. MR3319143. doi: https://doi.org/10.1214/15-AOS1311. 2

Jain, V., Koehler, F., and Mossel, E. (2018). "The mean-field approximation: Information inequalities, algorithms, and complexity." In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, 1326–1347. PMLR. 16

Kent, J. T. (1982). "The Fisher-Bingham distribution on the sphere." *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1): 71–80. MR0655376. 4

Kent, J. T., Ganeiber, A. M., and Mardia, K. V. (2013). "A new method to simulate the Bingham and related distributions in directional data analysis with applications." *arXiv preprint arXiv:1310.8110*. 20

Kolmogorov, V. and Zabin, R. (2004). "What energy functions can be minimized via graph cuts?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2): 147–159. 20

Lee, H., Kim, S., Kang, B., and Park, J. (2024). "A Stein Gradient Descent Approach for Doubly Intractable Distributions." *arXiv preprint arXiv:2410.21021*. 6

Lenz, W. (1920). "Beitršge zum verstšndnis der magnetischen eigenschaften in festen kšrpern." *Physikalische Z*, 21: 613–615. 2, 20

Levin, D. A., Peres, Y., and Wilmer, E. L. (2009). *Markov Chains and Mixing Times*. American Mathematical Society, 1st edition. MR2466937. doi: https://doi.org/10.1090/mbk/058. 12

Liang, F. (2010). "A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants." *Journal of Statistical Computation and Simulation*, 80(9): 1007–1022. MR2742519. doi: https://doi.org/10.1080/00949650902882162. 5

Liang, F., Jin, I. H., Song, Q., and Liu, J. S. (2016). "An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants." *Journal of the American Statistical Association*, 111(513): 377–393. MR3494666. doi: https://doi.org/10.1080/01621459.2015.1009072. 5

Lyne, A.-M., Girolami, M., Atchadé, Y., Strathmann, H., and Simpson, D. (2015). "On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods." *Statistical Science*, 30(4): 443–467. MR3432836. doi: https://doi.org/10.1214/15-STS523. 2, 3, 5, 6, 10, 11, 19, 20, 25

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equation of state calculations by fast computing machines." *The Journal of Chemical Physics*, 21(6): 1087–1092. 2

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(2): 451–458. MR2278096. doi: https://doi.org/10.1093/biomet/93.2.451. 5, 20

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). "MCMC for doubly-intractable distributions." In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, 359–366. Arlington, Virginia, USA: AUAI Press. 5, 19

Murray, I. and Graham, M. (2016). "Pseudo-marginal slice sampling." In *Artificial Intelligence and Statistics*, 911–919. PMLR. 6

Neal, R. M. (2001). "Annealed importance sampling." *Statistics and Computing*, 11(2): 125–139. MR1837132. doi: https://doi.org/10.1023/A:1008923215028. 21

Papaspiliopoulos, O. (2011). "Monte Carlo probabilistic inference for diffusion processes: A methodological framework." *Bayesian Time Series Models*, 82–103. MR2894234. 6

Park, J. and Haran, M. (2018). "Bayesian inference in the presence of intractable normalizing functions." *Journal of the American Statistical Association*, 113(523): 1372–1390. MR3862364. doi: https://doi.org/10.1080/01621459.2018.1448824. 2, 3, 5, 18, 19, 20, 21

Park, J. and Haran, M. (2020). "A Function Emulation Approach for Doubly Intractable Distributions." *Journal of Computational and Graphical Statistics*, 29(1): 66–77. MR4085864. doi: https://doi.org/10.1080/10618600.2019.1629941. 6

Paulin, D. (2015). "Concentration inequalities for Markov chains by Marton couplings and spectral methods." *Electronic Journal of Probability*, 20: 1–32. MR3383563. doi: https://doi.org/10.1214/EJP.v20-4039. 11

Pitt, M. K., dos Santos Silva, R., Giordani, P., and Kohn, R. (2012). "On some properties of Markov chain Monte Carlo simulation methods based on the particle filter." *Journal of Econometrics*, 171(2): 134–151. MR2991856. doi: https://doi.org/10.1016/j.jeconom.2012.06.004. 14, 15

Propp, J. G. and Wilson, D. B. (1996). "Exact sampling with coupled Markov chains and applications to statistical mechanics." *Random Structures & Algorithms*, 9(1-2): 223–252. MR1611693. doi: https://doi.org/10.1002/(sici)1098-2418(199608/09)9:1/2<223::aid-rsa14>3.0.co;2-o. 5, 20

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., and Dang, K.-D. (2021). "The block-Poisson estimator for optimally tuned exact subsampling MCMC." *Journal of Computational and Graphical Statistics*, 30(4): 877–888. MR4356592. doi: https://doi.org/10.1080/10618600.2021.1917420. 2, 3, 6, 7, 8, 11, 12, 13, 15

Rudolf, D., Smith, A., and Quiroz, M. (2024). "Perturbations of Markov Chains." *arXiv preprint arXiv:2404.10251*. 2

Schmidt, P. (1976). "The non-uniqueness of the Australian Mesozoic palaeomagnetic pole position." *Geophysical Journal International*, 47(2): 285–300. 23

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2016). "Taking the Human Out of the Loop: A Review of Bayesian Optimization." *Proceedings of the IEEE*, 104(1): 148–175. 16

Tan, L. S. and Friel, N. (2020). "Bayesian variational inference for exponential random graph models." *Journal of Computational and Graphical Statistics*, 29(4): 910–928. MR4191251. doi: https://doi.org/10.1080/10618600.2020.1740714. 6

Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. (2016). "The block pseudo-marginal sampler." *arXiv preprint arXiv:1603.02485*. 3, 13, 16

Wagner, W. (1988). "Unbiased multi-step estimators for the Monte Carlo evaluation of certain functional integrals." *Journal of Computational Physics*, 79(2): 336–352. MR0973333. doi: https://doi.org/10.1016/0021-9991(88)90020-4. 6

Warne, D., Sisson, S. A., and Drovandi, C. (2022). "Vector operations for accelerating expensive Bayesian computations – A tutorial guide." *Bayesian Analysis*, 17(2): 593–622. MR4483232. doi: https://doi.org/10.1214/21-ba1265. 19

Wei, C. and Murray, I. (2017). "Markov chain truncation for doubly-intractable inference." In *Artificial Intelligence and Statistics*, 776–784. PMLR. 6

Wood, A. (1982). "A bimodal distribution on the sphere." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(1): 52–58. MR0665678. doi: https://doi.org/10.2307/2347074. 23

Yang, Y., Quiroz, M., Kohn, R., and Sisson, S. A. (2025). "Supplementary Material to "A correlated pseudo-marginal approach to doubly intractable problems"." *Bayesian Analysis*. doi: https://doi.org/10.1214/25-BA1573SUPP. 3, 4, 7, 8, 10, 11, 12, 13, 16, 18, 19, 21, 23, 25