THE BLOCK-POISSON ESTIMATOR FOR OPTIMALLY TUNED EXACT SUBSAMPLING MCMC

MATIAS QUIROZ^{1,2}, MINH-NGOC TRAN³, MATTIAS VILLANI⁴, ROBERT KOHN¹ AND KHUE-DUNG DANG¹

ABSTRACT. Speeding up Markov Chain Monte Carlo (MCMC) for datasets with many observations by data subsampling has recently received considerable attention in the literature. The currently available methods are either approximate, highly inefficient or limited to small dimensional models. We propose a pseudo-marginal MCMC method that estimates the likelihood by data subsampling using a block-Poisson estimator. The estimator is a product of Poisson estimators, each based on an independent subset of the observations. The construction allows us to update a subset of the blocks in each MCMC iteration, thereby inducing a controllable correlation between the estimates at the current and proposed draw in the Metropolis-Hastings ratio. This makes it possible to use highly variable likelihood estimators without adversely affecting the sampling efficiency. Poisson estimators are unbiased but not necessarily positive. We therefore follow Lyne et al. (2015) and run the MCMC on the absolute value of the estimator and use an importance sampling correction for occasionally negative likelihood estimates to estimate expectations of any function of the parameters. We provide analytically derived guidelines to select the optimal tuning parameters for the algorithm by minimizing the variance of the importance sampling corrected estimator per unit of computing time. The guidelines are derived under idealized conditions, but are demonstrated to be quite accurate in empirical experiments. The guidelines apply to any pseudo-marginal algorithm if the likelihood is estimated by the block-Poisson estimator, including the class of doubly intractable problems in Lyne et al. (2015). We illustrate the method in a logistic regression example and find dramatic improvements compared to regular MCMC without subsampling and a popular exact subsampling approach recently proposed in the literature.

KEYWORDS: Bayesian inference, Control variates, Data subsampling, Exact inference, Poisson Estimator, Pseudo-marginal MCMC.

¹School of Economics, UNSW Business School, University of New South Wales. ²Research Division, Sveriges Riksbank. ³Discipline of Business Analytics, University of Sydney. ⁴Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University.

1. Introduction

Standard Markov Chain Monte Carlo (MCMC) algorithms require evaluating the likelihood function for the full dataset and are therefore prohibitively expensive for so-called tall datasets with many observations. One recent strand of literature attempts to speed up MCMC algorithms by using random subsets of the data, see Korattikara et al. (2014); Bardenet et al. (2014, 2017); Maclaurin and Adams (2014); Liu et al. (2015, 2017); Quiroz et al. (2018); Bierkens et al. (2016). Section 2 briefly reviews these approaches and highlights possible pitfalls.

Bardenet et al. (2017) provide an excellent review of subsampling approaches and propose a positive unbiased estimator of the likelihood in a pseudo-marginal framework (Beaumont, 2003; Andrieu and Roberts, 2009) to accelerate the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970). Their unbiased likelihood estimator is constructed from a sequence of unbiased log-likelihood estimates from small batches of observations used in a Rhee-Glynn type debiasing estimator (Rhee and Glynn, 2015); see also Strathmann et al. (2015) for an alternative use of debiasing to estimate posterior expectations by combining estimates from a sequence of partial posteriors. To ensure positiveness, Bardenet et al. (2017) use a lower bound for the log-likelihood estimates in all batches (Jacob and Thiery, 2015). Bardenet et al. (2017) quickly dismiss this approach as the estimator's large variability causes the pseudo-marginal chain to get stuck for long spells.

We propose an alternative unbiased estimator of the likelihood, called the block-Poisson estimator, with several attractive properties and demonstrate that it can be successfully used for subsampling MCMC. The block-Poisson estimator is a product of Poisson estimators (Wagner 1988; 1989; Beskos et al., 2006; Papaspiliopoulos, 2009; Fearnhead et al., 2010) with the following features. First, the product form makes it possible to only update the subsamples in some of the blocks in each iteration and thereby generate a controllable correlation between the log of the estimated likelihood at the current and the proposed MH draw. Such dependent pseudo-marginal schemes are known to be very efficient as they can use substantially noisier likelihood estimators (smaller subsamples) without adversely affecting

the sampling efficiency of the chain (Deligiannidis et al., 2017). Second, the block-Poisson estimator has a lower variance than the Rhee-Glynn estimator in Bardenet et al. (2017). Third, the block-Poisson estimator uses the very effective variance reducing control variates proposed in Quiroz et al. (2018) and Bardenet et al. (2017). Fourth, the block-Poisson estimator uses a soft lower bound for its constituent log-likelihood batch estimators, rather than a strict lower bound as in Bardenet et al. (2017). Section 3.3 explains that this is computationally more efficient than a strict bound, but also makes it possible to get negative likelihood estimates, which is not allowed in the usual pseudo-marginal MCMC framework in Andrieu and Roberts (2009). Lyne et al. (2015) propose an ingenious solution to this problem by running the pseudo-marginal sampler on the absolute value of the likelihood estimate followed by an importance sampling correction step to estimate any function of the model parameters in a simulation consistent way. We will refer to this pseudo-marginal Metropolis Hastings with importance sampling sign correction as signed PMMH.

It is well known that pseudo-marginal algorithms need careful tuning of the number of particles (subsamples) and several recent papers develop practical guidelines for this choice when the likelihood estimator is strictly positive; see in particular Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015). A major contribution of our article is that we derive easily implemented guidelines on the optimal number of subsamples (or more generally particles) for the signed PMMH based on the block-Poisson estimator. The optimal number of subsamples is chosen to minimize the asymptotic variance of the importance sampling estimator for a given computational budget and therefore balances i) the inefficiency in the MCMC on the absolute measure, ii) the computational cost of the likelihood estimator and iii) the probability of negative estimates. We show that the asymptotic variance of the importance sampling estimator can be obtained in closed form when the likelihood is estimated by the block-Poisson estimator. The guidelines are derived under idealized conditions, but are demonstrated to be quite accurate in empirical experiments. These optimality results apply much more generally than subsampling, for example in any of the

doubly intractable problems listed in Lyne et al. (2015) by using their Exponential auxiliary variable construction.

We demonstrate that subsampling MCMC using the block-Poisson estimator gives an efficient sampler that does not get stuck and generates many more efficient draws for a given computational budget compared to MH on the full sample. We also show empirically that our exact subsampling MCMC approach is dramatically more efficient than Firefly Monte Carlo (Maclaurin and Adams, 2014), a highly cited exact subsampling algorithm.

Our article is organized as follows. Section 2 briefly reviews the main subsampling approaches proposed in the recent literature. Section 3 introduces our block-Poisson estimator and derives its key properties. Section 4 outlines our proposed sampling algorithm, proves its convergence, and provides guidelines on the selection of tuning parameters to obtain an optimal implementation. Section 5 demonstrates the methodology and Section 6 concludes and outlines future research. There is a supplement with four sections. Section S1 derives the guidelines for the optimal tuning of the signed PMMH with the block-Poisson estimator. Section S2 provides more details and also validates the guidelines versus empirical experiments. Section S3 considers the case where the likelihood estimate is positive and introduces the block pseudo marginal sampler. Section S4 contains all the proofs. We refer to equations, sections, lemmas in the main paper as Equation 1, Section 1, Lemma 1 etc., and to equations, sections and lemmas, etc in the supplement as Equation (S1), Section S1 and Lemma S1, etc.

2. Previous research

Previous research in scalable MCMC by data subsampling is either approximate (Korattikara et al., 2014; Bardenet et al., 2014, 2017; Quiroz et al., 2018) or exact (Maclaurin and Adams, 2014; Liu et al., 2015, 2017; Bierkens et al., 2016). We also note that exact subsampling approaches using a delayed acceptance MCMC framework (Christen and Fox, 2005) have been proposed (Banterle et al., 2014; Payne and Mallick, 2017; Quiroz et al., 2017),

but since the full dataset must be evaluated for any accepted sample, these methods are not fully subsampling approaches.

The algorithms in Korattikara et al. (2014); Bardenet et al. (2014, 2017) all replace the computationally costly MH ratio with a hypothesis test based on a fraction of the data, thereby significantly speeding up computations. Bardenet et al. (2017) evaluate these methods and conclude that their method with concentration inequalities and control variates clearly outperforms the algorithms in Korattikara et al. (2014) and Bardenet et al. (2014). A drawback of their method is that it relies on a bound for the difference between the log-likelihood contributions at the proposed and current sample, and that of the control variates. Their proposed Taylor-Lagrange bound can result in an upper bound that is too rough, which then has to be compensated by a very large subsample before the accept (or reject) decision can be taken (Quiroz et al., 2018).

Quiroz et al. (2018) estimate the MH ratio based on a random subsample and use a dependent pseudo-marginal approach to sample from the posterior. They derive an efficient unbiased log-likelihood estimator with control variates and apply a bias-correction to get an approximately unbiased likelihood estimator. Their target distribution in the MCMC is therefore a perturbation of the posterior, but is shown to be within $O(n^{-1}m^{-2})$ of the true posterior, where n is the sample size and m the subsample size. However, their bias-correction implies that the approximation error is a function of the variance of the log-likelihood estimator, and targeting a large variance may therefore degrade the posterior approximation. The approximation error is small in their examples even when the variance of the log-likelihood estimator is large, but there is no guarantee that this will carry over to other applications. In contrast, our proposed method is simulation consistent for the posterior expectation of any function of the parameters, regardless of the estimator's variability. See Section 4.3 and Appendix S2 for a comparison with the approximate approach in Quiroz et al. (2018).

The Firefly Monte Carlo algorithm in Maclaurin and Adams (2014) (see also Liu et al., 2015 and Liu et al., 2017 for alternative implementations) introduces an auxiliary variable for

each observation which determines if it should be included in the evaluation of the posterior in a given MCMC iteration. A lower bound for each likelihood term caters for the excluded observations. The authors suggest using Gibbs sampling with the original parameters and the auxiliary variables in two different blocks. The method has been documented to be very inefficient, see e.g. Bardenet et al. (2017) and Quiroz et al. (2018), because of the strong dependence between the model parameters and the auxiliary variables, with only a fraction of the auxiliary variables allowed to be updated in a given iteration. We even experiment with updating all auxiliary variables in our examples in Section 5 but, remarkably, find that the high inefficiency persists.

Finally, there have been some recently proposed MCMC algorithms based on piecewise deterministic Markov processes that have been shown to preserve the correct target distribution when subsampling the data, the most prominent example being the algorithm based on the zig-zag process (Bierkens and Roberts, 2017; Bierkens et al., 2016). The implementation of the zig-zag process method, in its current form, requires an upper bound of the absolute value of the gradient of the log-likelihood, and how tight this bound is determines the probability of accepting a Markov move. Hence, similarly to acceptance sampling, the zig-zag sampling method is likely to not scale well with the dimension of the parameter space, which is arguably the scenario in which tall datasets are most useful.

3. The block-poisson estimator

3.1. **The estimator.** Suppose that we have conditional independence between the observations $y = (y_1, \dots, y_n)$ given a parameter vector $\theta \in \Theta \subset \mathbb{R}^p$, so that the likelihood decomposes as

(3.1)
$$L(\theta) := p(y|\theta) = \exp(\ell(\theta)), \quad \ell(\theta) = \sum_{k=1}^{n} \ell_k(\theta), \quad \text{where } \ell_k(\theta) := \log p(y_k|\theta, x_k)$$

is the log-likelihood contribution of the kth observation with covariates x_k . Our block-Poisson estimator of the likelihood in Definition 1 below relies on an unbiased estimator of the log-likelihood, $\ell(\theta)$. Quiroz et al. (2018) propose to estimate $\ell(\theta)$ by a difference estimator based on control variates that are especially tailored for the computationally cheap repeated estimates needed in pseudo-marginal MCMC. The control variate $q_k(\theta)$ for the kth observation is such that the differences $d_k(\theta) = \ell_k(\theta) - q_k(\theta)$ are small. The difference estimator then uses the trivial identity

(3.2)
$$\ell(\theta) = q(\theta) + d(\theta), \text{ where } q(\theta) = \sum_{k=1}^{n} q_k(\theta) \text{ and } d(\theta) = \sum_{k=1}^{n} d_k(\theta)$$

to unbiasedly estimate $\ell(\theta)$ via an unbiased estimator of $d(\theta)$. Since the control variates homogenize the population, we can for example use the sample mean estimator from a subsample of size m drawn with replacement to estimate d:

$$\widehat{d}_m = \frac{n}{m} \sum_{i=1}^m d_{u_i},$$

where the u_i are iid from the distribution $\Pr(u_i = k) = 1/n$ for k = 1, ..., n, and we suppress the dependence on θ to simplify notation. The variance of the estimator is

$$\sigma_{\widehat{d}}^2 := V[\widehat{d}_m] = \frac{\gamma}{m}$$
, where $\gamma := n^2 \sigma_{d_{u_i}}^2$ with $\sigma_{d_{u_i}}^2 := V[d_{u_i}]$.

There have been several proposals for unbiased estimators of the likelihood in the literature, in particular the Rhee-Glynn estimator in Rhee and Glynn (2015), Russian roulette estimators in Lyne et al. (2015) and the Poisson estimator in Wagner (1988). We now propose a modified Poisson estimator which is constructed to be particularly useful for the block pseudo-marginal algorithm developed in Section 4.

Definition 1. The block-Poisson likelihood estimator is defined as

(3.4)
$$\widehat{L}_B(\theta) = \exp(q(\theta)) \prod_{l=1}^{\lambda} \xi_l, \, \xi_l = \exp\left(\frac{a+\lambda}{\lambda}\right) \prod_{h=1}^{\mathcal{X}_l} \left(\frac{\widehat{d}_m^{(h,l)} - a}{\lambda}\right),$$

where λ is a positive integer, a is a real number, $\widehat{d}_m^{(h,l)}$ is some unbiased estimator of $d = \ell - q$ from a small batch of m observations, and $\mathcal{X}_1, \ldots, \mathcal{X}_{\lambda}$ are independent Pois(1) variables. The rightmost product is defined to be 1 whenever $\mathcal{X}_l = 0$.

The block-Poisson estimator in Eq. (3.4) is a product over λ Poisson estimators. This construction makes it possible to update the subsampled observations only in a subset of the products, thereby inducing a correlation ρ between the log of the absolute value of the estimated likelihood at the current and proposed parameter value in the MH ratio, see Section 4. Inducing dependence between estimates in subsequent MCMC iterations is known to be very advantageous for pseudo-marginal MCMC since it allows us to use a much more variable likelihood estimator (Deligiannidis et al., 2017). A more variable estimator translates into using smaller subsample sizes and therefore faster MCMC iterations. We emphasize that the idea with the product of Poisson estimators is to induce correlation, not to reduce the variance of the estimator, see Section 4.2. In fact, Part (v) of Lemma 1 shows that the variance of the block-Poisson estimator is the same as the traditional unbiased Poisson estimator (Papaspiliopoulos, 2009)

(3.5)
$$\widehat{L}_P(\theta) = \exp(q(\theta)) \exp(a+\lambda) \prod_{h=1}^G \left(\frac{\widehat{d}_m^{(h)} - a}{\lambda}\right), \quad G \sim \operatorname{Pois}(\lambda),$$

where the product is 1 if G = 0. The following lemma proves some important properties of the $\widehat{L}_B(\theta)$ estimator.

Lemma 1. Assume $\sigma_{\widehat{d}}^2 < \infty$. Then, for any $\theta \in \Theta$,

- (i) $E[\widehat{L}_B] = L$.
- (ii) \widehat{L}_B is almost surely positive only if $\widehat{d}_m^{(h,l)} \geq a$ almost surely for all h and l.
- (iii) $V[|\widehat{L}_B|] < \infty$
- (iv) For a fixed λ , $V[\widehat{L}_B]$ is minimized for $a = d \lambda$.
- (v) $V[\widehat{L}_B(\theta)] = V[\widehat{L}_P(\theta)].$

Although \widehat{L}_B is unbiased by Part (i) of Lemma 1, Part (ii) shows that the estimator is only positive with probability 1 if a is a lower bound of all $\widehat{d}_m^{(h,l)}$. We argue in Section 3.3 that it is prohibitively expensive to choose a in this way. We therefore adopt the approach in Lyne et al. (2015) and carry out the pseudo-marginal on $|\widehat{L}_B|$. Part (iii) of Lemma 1 ensures that $|\widehat{L}_B|$ has a finite variance. Part (iv) of Lemma 1 motivates the simplification of setting $a = d - \lambda$ and we optimize only with respect to m and λ . Under this simplifying assumption

the estimator becomes

(3.6)
$$\widehat{L}_B(\theta) = \exp(q(\theta)) \prod_{l=1}^{\lambda} \xi_l, \ \xi_l = \exp\left(\frac{d}{\lambda}\right) \prod_{h=1}^{\mathcal{X}_l} \left(\frac{\widehat{d}_m^{(h,l)} - d}{\lambda} + 1\right).$$

In Section 4 we follow Pitt et al. (2012) and derive the optimal tuning parameters m and λ by studying a performance measure that involves the variance of the log of the estimator. Since our MCMC is actually run on $|\hat{L}_B|$, we need an expression for the variance of $\log |\hat{L}_B|$. Lemma 2 gives this variance under a Gaussian assumption for the batch means $\hat{d}_m^{(h,l)}$, which can be motivated by the Central Limit Theorem (CLT). For small m, Lemma S5 gives the corresponding result when the $\hat{d}_m^{(h,l)}$ follow a finite mixture of normals distributions.

Lemma 2. Assume that $\widehat{d}_m^{(h,l)} \stackrel{iid}{\sim} \mathcal{N}(d, \frac{\gamma}{m})$ for all h and l. The variance of $\log |\widehat{L}_B|$ when $a = d - \lambda$ is then

$$\sigma_{\log|\widehat{L}_B|}^2 = \lambda(\nu^2 + \eta^2),$$

where

$$\eta := \log \left(\sqrt{\frac{\gamma}{m\lambda^2}} \right) + \frac{1}{2} \left(\log 2 + \mathcal{E}_J \left[\psi^{(0)} (1/2 + J) \right] \right)$$

and

$$\nu^2 := \frac{1}{4} \left(E_J \left[\psi^{(1)} (1/2 + J) \right] + V_J \left[\psi^{(0)} (1/2 + J) \right] \right),$$

with $J \sim \text{Pois}\left(\frac{m\lambda^2}{2\gamma}\right)$ and $\psi^{(q)}$ is the polygamma function of order q. Furthermore, $\sigma^2_{\log|\widehat{L}_B|} < \infty$ for all m > 0, $\lambda > 0$ and $\gamma > 0$.

We note that the variance expression in Lemma 2 contains an infinite sum through the expectation and variance of the Poisson random variable J, but in practice we can obtain accurate approximations by truncation, because $\Pr(J=j)$ decreases very quickly to zero as j increases and the polygamma functions are either bounded or grow much slower than the rate of decrease of $\Pr(J=j)$ (see the proof of Lemma 2 in Appendix S4). We discuss the optimal choice of tuning parameters m and λ in Section 4.3.

3.2. Control variates. We use two different types of control variates $q_k(\theta)$ to approximate $\ell_k(\theta)$, both based on a Taylor expansion of the log density $\ell_k(\theta) := \log p(y_k|x_k, \theta)$.

The first control variate, suggested in Bardenet et al. (2017), expands $\ell_k(\theta)$ around some reference value θ^* . This parameter expanded control variate has a computational complexity of O(1) (Bardenet et al., 2017) and therefore the overall computational cost for \widehat{L}_B is $m \sum_{l=1}^{\lambda} \mathcal{X}_l$.

Quiroz et al. (2018) present an alternative control variate, which does not require a reference value of the parameter. Instead, the expansion is with respect to $\eta_k = (y_k, x_k)$ around a reference value of the data η^* . In order to make the approximation local, a sparse set of the data is obtained by clustering the data into \mathcal{K} clusters before the MCMC. For each data point η_k that belongs to cluster c, $q_k(\theta)$ is an expansion around the centroid η_c^* . This data expanded control variate has a computational complexity of $O(\mathcal{K})$ (Quiroz et al., 2018) and an overall computational cost for \widehat{L}_B of $m \sum_{l=1}^{\lambda} \mathcal{X}_l + O(\mathcal{K})$.

See Quiroz et al. (2018) for the asymptotic properties of these two types of control variates with respect to n.

3.3. **Soft lower bound.** While the lower bound of all $\widehat{d}_m^{(h,l)}$ ensures that \widehat{L}_B is positive (Lemma 1 part (ii)), it is impractical for two reasons. First, for most models we typically need to evaluate $d_k = \ell_k - q_k$ for all data points to find a lower bound. Second, the optimal implementation outlined in Section 3.1 requires that $\lambda = d - a$. If the control variates are accurate then d is small relative to a, implying that $\lambda \approx -a$. Hence, a large -a implies a large number of products in the block-Poisson estimator, and a large computational cost.

We therefore advocate using a soft lower bound, i.e. one that is not necessarily a lower bound for all outcomes of $\widehat{d}_m^{(h,l)}$ but still gives a $\Pr(\widehat{L}_B \geq 0)$ close to one. The soft lower bound makes it possible to obtain negative likelihood estimates, and $\Pr(\widehat{L}_B \geq 0)$ is in Section 4.3 shown to be a crucial quantity for the efficiency of our method. Lemma 3 gives an analytically tractable expression for this probability.

Lemma 3.

$$\Pr(\widehat{L}_B \ge 0) = \frac{1}{2} \left(1 + \left(1 - 2\Psi(a, m, \lambda, \gamma) \right)^{\lambda} \right)$$

where

$$\Psi(a, m, \lambda, \gamma) := \Pr(\xi_l < 0) = \frac{1}{2} \sum_{j=1}^{\infty} \left(1 - \left(1 - 2 \Pr\left(A_m < 0 \right) \right)^j \right) \Pr(\mathcal{X}_l = j), \quad \mathcal{X}_l \sim \operatorname{Pois}(1),$$

and
$$A_m := \frac{\widehat{d}_m - d}{\lambda} + 1$$
.

The probability of a positive estimator \hat{L}_B can be computed whenever $\Pr(A_m < 0)$ is analytically available. By the CLT, we can expect A_m to be normally distributed when m is sufficiently large as \hat{d}_m will be normal in this case.

4. Methodology

Section 4.1 first outlines the signed PMMH algorithm proposed by Lyne et al. (2015) and our signed block PMMH extension that induces dependence between estimators at subsequent iterations. PMMH algorithms are known to be sensitive to tuning parameters like the number of particles, and Lyne et al. (2015) do not provide any guidelines nor optimality conditions for these choices. In Section 4.3 we derive analytical guidelines for optimal tuning of the signed block PMMH using the block-Poisson estimator. The guidelines are derived under idealized assumptions, but we give some empirical evidence and sensitivity analysis in Appendix S2.3 suggesting that the guidelines are accurate in practice. This analysis extends that of Pitt et al. (2012), who derive guidelines for optimal implementing of regular PMMH with a strictly positive likelihood estimator. Optimal tuning of the signed block PMMH is very important since it opens up a whole range of new applications beyond subsampling, including the doubly intractable problems discussed in Lyne et al. (2015).

4.1. The signed block PMMH algorithm. We start by reviewing the signed PMMH algorithm in Lyne et al. (2015) with independent likelihood estimators at the current and proposed values of the Markov chain, and subsequently present the signed block PMMH based on the block-Poisson estimator, which instead gives dependent likelihood estimators. In Appendix S3 we demonstrate that when the likelihood estimator is almost surely positive and factorizes into blocks, we obtain the so called block PMMH algorithm. The optimal tuning of the block PMMH is given in Appendix S3.

Let $p_{\Theta}(d\theta)$ and $\pi(d\theta) := p(d\theta|y) \propto L(\theta)p_{\Theta}(d\theta)$ denote the prior and the posterior measure of θ . Let $\widehat{L}(\theta, U)$ be an unbiased but not necessarily positive estimator of the likelihood, for example the block-Poisson in Eq. (3.4). Here, U is the set of all random variables involved when constructing \widehat{L} and we usually take U to be a set of uniform random numbers. Write p(du) for the probability measure of U. The unbiasedness of $\widehat{L}(\theta, U)$ means that

(4.1)
$$L(\theta) = \int \widehat{L}(\theta, u) p(du), \text{ for any } \theta.$$

It is invalid to define a target posterior measure using the estimator $\widehat{L}(\theta, u)$ because it can be negative. Instead, we define the joint target measure

$$(4.2) \ \overline{\pi}(d\theta, du) := \frac{1}{\overline{C}} |\widehat{L}(\theta, u)| p_{\Theta}(d\theta) p(du), \ \overline{C} := \int_{\Theta} C(\theta) p_{\Theta}(d\theta), \ C(\theta) := \int |\widehat{L}(\theta, u)| p(du).$$

This is a proper Lebesgue product measure on $\Theta \otimes \mathcal{U}$, and admits the posterior $\pi(d\theta)$ as its marginal measure only if $\widehat{L}(\theta, u) \geq 0$ almost surely. Define

(4.3)
$$\nu(d\theta) := \int \overline{\pi}(d\theta, du) = \frac{C(\theta)p_{\Theta}(d\theta)}{\overline{C}}.$$

Let $S(\theta, u) = \operatorname{sign}(\widehat{L}(\theta, u)) \in \mathcal{S} := \{-1, 1\}$ where $\operatorname{sign}(\cdot)$ is 1 if its input is positive and -1 otherwise. Often, the ultimate goal in Bayesian inference is to estimate an integral of the form

$$E_{\pi}[\psi] = \int_{\Theta} \psi(\theta) \pi(d\theta)$$

for some function $\psi(\theta)$ on Θ . Lyne et al. (2015) cleverly note that

Hence, we can use a pseudo-marginal scheme (Beaumont, 2003; Andrieu and Roberts, 2009) to obtain N samples $\{\theta^{(i)}, u^{(i)}, i = 1, ..., N\}$ from $\overline{\pi}(d\theta, du)$ in Eq. (4.2) and then estimate Eq. (4.4). Note that it is necessary to store only the $\theta^{(i)}$ and the signs $s^{(i)} = S(\theta^{(i)}, u^{(i)})$, but

not the $u^{(i)}$. The estimator of $E_{\pi}[\psi]$ is

(4.5)
$$\widehat{E}_{\pi}[\psi] = \frac{\sum_{i=1}^{N} \psi(\theta^{(i)}) s^{(i)}}{\sum_{i=1}^{N} s^{(i)}}.$$

This elegant observation allows exact inference in the sense that this estimator is guaranteed to converge $\overline{\pi}$ -almost surely to the true value $E_{\pi}[\psi]$ as $N \to \infty$. Notice that the θ iterates themselves are not distributed according to $\pi(d\theta)$, but rather $\nu(d\theta)$ in Eq. (4.3).

As mentioned before, using a conventional pseudo-marginal scheme to generate from $\overline{\pi}$ can be inefficient because of highly variable estimates $\widehat{L}(\theta, u)$. Deligiannidis et al. (2017) therefore proposed, in the case with positive likelihood estimators, to correlate the random numbers underlying the estimators $\widehat{L}(\theta', u')$ at the proposed and $\widehat{L}(\theta, u)$ at the current draws; see also Dahlin et al. (2015). This so called correlated pseudo-marginal method is a substantial advance for pseudo-marginal algorithms since a much larger $\sigma_{\log \widehat{L}}^2$ can be targeted without adversely affecting the sampling efficiency.

We now propose an alternative way to correlate the estimates by partitioning U into G blocks and update only a single random block together with θ in each iteration, keeping the other blocks fixed. In the block-Poisson, we use \widetilde{U}_l as the set of random numbers to compute ξ_l , $l = 1, \ldots, \lambda$, and group them as

$$(4.6) U = (U_1, \dots, U_G) := U_{1:G},$$

so that each U_i , i = 1, ..., G, contains λ/G of the \widetilde{U}_l . Note that we do not require that $G = \lambda$, so the blocks in the block-Poisson estimator do not need to correspond to the blocking of U in Eq. (4.6). The extended target in Eq. (4.2) becomes

(4.7)
$$\overline{\pi}(d\theta, du_{1:G}) := \frac{p_{\Theta}(d\theta)}{\overline{C}} \prod_{i=1}^{G} |\widehat{L}^{(i)}(\theta, u_i)| p_U(du_i),$$

where each $\widehat{L}^{(i)}$ with $\mathrm{E}[\widehat{L}^{(i)}] = L(\theta)^{1/G}$ corresponds to the likelihood estimate when using U_i , so that $\mathrm{E}\left[\prod_{i=1}^G \widehat{L}^{(i)}\right] = L(\theta)$. When blocking, we will write $u_i \in \mathcal{U}$ for $i = 1, \ldots, G$, so that $u_{1:G} \in \mathcal{U}^G$.

Let $u := (u_1, \dots, u_G)$ and $u' := (u'_1, \dots, u'_G)$ denote the current and proposed values of U, respectively. The proposal density which updates a single block at random is

(4.8)
$$q_{U}(u;du') := \frac{1}{G} \sum_{i=1}^{G} p_{U_{i}}(du'_{i}) \prod_{j \neq i} \delta_{u_{j}}(du'_{j}),$$

where $\delta_a(\cdot)$ is the Dirac delta measure centered at a. Let $q_{\Theta}(\theta; d\theta')$ be the proposal for θ so that the joint proposal for θ and u is $q_{\Theta}(\theta; d\theta')q_U(u; du')$. Consider now the signed block Metropolis-Hastings sampling scheme for θ and u.

Algorithm 1 (Block PMMH sampling for the absolute measure). Generate θ', u' using the proposal $q_{\Theta}(\theta; d\theta')q_U(u; du')$ and accept the proposal with probability $\alpha_{\Theta,U}(\theta, u, \theta', u') = 1 \wedge r_{\Theta,U}(\theta, u; \theta', u')$, where

$$r_{\Theta,U}(\theta, u; \theta', u') = \frac{\overline{\pi}(d\theta', du')}{\overline{\pi}(d\theta, du)} \times \frac{q_{\Theta}(\theta'; d\theta)q_{U}(u'; du)}{q_{\Theta}(\theta; d\theta')q_{U}(u; du')}.$$

The following lemma gives a workable expression for the acceptance probability of Algorithm 1.

Lemma 4.

(4.9)
$$\frac{\overline{\pi}(d\theta',du')}{\overline{\pi}(d\theta,du)} \times \frac{q_{\Theta}(\theta';d\theta)q_{U}(u';du)}{q_{\Theta}(\theta;d\theta')q_{U}(u;du')} = \frac{|\widehat{L}(\theta',u')|p_{\Theta}(d\theta')}{|\widehat{L}(\theta,u)|p_{\Theta}(d\theta)} \frac{q_{\Theta}(\theta';d\theta)}{q_{\Theta}(\theta;d\theta')}.$$

Theorem 1 below proves the ergodic properties of Algorithm 1 based on assumptions about the following sampling scheme targeting the intractable $\nu(d\theta)$ and some additional conditions stated in Assumption 1.

Algorithm 2 (MH sampling for the θ -marginal absolute measure). Generate θ' using the proposal $q_{\Theta}(\theta; d\theta')$ and accept with probability $\alpha_{\Theta}(\theta, \theta') = 1 \wedge r_{\Theta}(\theta, \theta')$, where

$$r_{\Theta}(\theta, \theta') = \frac{\nu(d\theta')q(\theta'; d\theta)}{\nu(d\theta)q(\theta; d\theta')}.$$

Assumption 1.

(i) For all $\theta \in \Theta$ and $u \in \mathcal{U}^G$, $-\infty < \widehat{L}(\theta, u) < \infty$ and $C(\theta) > 0$, where $C(\theta)$ is defined in Eq. (4.2).

(ii) Let $P_{\Theta}(\theta; d\theta') = K_{\Theta}(\theta; d\theta') + \delta_{\theta}(d\theta') \left(1 - \int K_{\Theta}(\theta; d\theta')\right)$ and $K_{\Theta}(\theta; d\theta') = \alpha_{\Theta}(\theta, \theta')q_{\Theta}(\theta; d\theta')$ be the Markov transition kernel and sub-stochastic kernel of Algorithm 2. We assume that if $P_{\Theta}(\theta; d\theta') > 0$ then $K_{\Theta}(\theta; d\theta') > 0$.

The following theorem gives the convergence properties of the block PMMH sampling scheme for the absolute measure in Algorithm 1. We note that Part (iii) of Theorem 1 was also obtained by Lyne et al. (2015).

Theorem 1. Suppose that the Markov chain in Algorithm 2 is irreducible and aperiodic and that Assumption 1 holds.

- (i) Algorithm 1 is reversible.
- (ii) Samples from Algorithm 1 converges to $\overline{\pi}$ in total variation norm.
- (iii) Suppose also that $E_{\pi}[|\psi|] < \infty$ and $E_{\overline{\pi}}(S) \neq 0$. Then, $\widehat{E}_{\pi}[\psi] \to E_{\pi}[\psi] \overline{\pi}$ -almost surely.
- (iv) Define

(4.10)
$$\operatorname{IF}_{\overline{\pi},\psi S} = \frac{V_{\overline{\pi}}(\psi S) + 2\sum_{j=1}^{\infty} \Gamma_j}{V_{\overline{\pi}}(\psi S)},$$

where $\Gamma_j = \text{Cov}_{\overline{\pi}}\Big(S_0\psi(\theta_0), S_j\psi(\theta_j)\Big)$ and $\Big\{(\theta_0, S_0), \dots, (\theta_j, S_j), \dots\Big\}$ are the MCMC iterates generated by the sampling scheme in Algorithm 1. If $V_{\overline{\pi}}[\psi S]IF_{\overline{\pi},\psi S} < \infty$ and $E_{\overline{\pi}}(S) \neq 0$, then

(4.11)
$$\sqrt{N} \left(\widehat{E}_{\pi}(\psi) - E_{\pi}(\psi) \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{V_{\overline{\pi}}(\psi S) IF_{\overline{\pi}, \psi S}}{E_{\overline{\pi}}(S)^2} \right).$$

4.2. Correlation implied by the block PMMH. We first note that

Lemma 5. U_1, \ldots, U_G are independent $(\overline{\pi})$ conditional on θ , i.e., $\overline{\pi}(u_{1:G}|\theta) = \prod_{i=1}^G \pi(u_i|\theta)$. Let $\ell(\theta', u') = \log |\widehat{L}_B(\theta', u')|$ and $\ell(\theta, u) = \log |\widehat{L}_B(\theta', u')|$. Then,

$$\ell(\theta', u') = \sum_{i=1}^{G} \log |L^{(i)}(\theta', u'_i)|$$
 and $\ell(\theta, u) = \sum_{i=1}^{G} \log |L^{(i)}(\theta, u_i)|$

so that $\ell(\theta', u')$ and $\ell(\theta, u)$ have G - 1 out of G u_i terms in common and each of the u_i are independent by Lemma 5. Hence, for G large and θ close to θ' , we argue that $\ell(\theta', u')$ and $\ell(\theta, u)$ will be approximately normal with a correlation close to $\rho = 1 - \frac{1}{G}$. Furthermore, if

the sample is large, then θ and θ' are likely to be close by the Bernstein von Mises theorem (Van der Vaart, 1998, Chapter 10.2). We note that in our case G = 100 and the sample size n is large. The above demonstrates why the form of the block-Poisson estimator in Eq. (3.4) allows us to induce a simple and controllable correlation between the log of the estimator at the current and proposed draws.

4.3. Tuning the signed block PMMH algorithm. We now outline in detail how to choose the tuning parameters m, λ and a for the signed block PMMH with the block-Poisson estimator. For brevity and clarity, we place all assumptions and technical derivations in Appendix S1, and focus here on the problem formulation, implementation of guidelines and illustration of the results. Further details are in Appendix S2. The theoretical framework is based on Pitt et al. (2012), which we here extend to signed block PMMH algorithms. As in Pitt et al. (2012), the derived guidelines are based on stylized assumptions to make the analysis tractable and transparent. Appendix S2.3 verifies that the guidelines are accurate and useful in practice.

Using the block-Poisson estimator with lower bound $a = d - \lambda$, we can define the Computational Time (CT) of Algorithm 1 as $m\lambda$ (expected computing cost) times this inefficiency, i.e.

(4.12)
$$\operatorname{CT}_{B}(\lambda, m|\gamma) := m\lambda \frac{\operatorname{IF}_{\overline{\pi}, \psi S}\left(\sigma_{\log|\widehat{L}_{B}|}^{2}(\lambda, m|\gamma)\right)}{(2\tau(\lambda, m|\gamma) - 1)^{2}}.$$

where the inefficiency $IF_{\overline{\pi},\psi S}$ is defined by Part (iv) of Theorem 1 and $E_{\overline{\pi}}[S] = 2 \Pr(\widehat{L}_B > 0) - 1$ with $\Pr(\widehat{L}_B > 0) := \tau$. To motivate this definition, we note that by Part (iv) of Theorem 1,

$$\frac{\mathrm{IF}_{\overline{\pi},\psi S}\left(\sigma_{\log|\widehat{L}_B|}^2(\lambda,m|\gamma)\right)}{\mathrm{E}_{\overline{\pi}}(S)^2} = \lim_{N \to \infty} \frac{N\mathrm{V}_{\overline{\pi}}\left(\mathrm{E}_{\pi}(\psi)\right)}{\mathrm{V}_{\overline{\pi}}(\psi)}$$

is the *true* inefficiency of the estimator $\widehat{E}_{\pi}(\psi)$. The CT captures the computational cost/time of generating the equivalent of a single independent Monte Carlo draw, and it is this objective function that we seek to minimize. Note the following regarding the CT expression in

Eq. (4.12). First, as described above, we use blocking with $\rho = 1 - \frac{1}{G} = 0.99$ (G = 100) throughout the paper and we have therefore for notational clarity suppressed the dependence on ρ in the CT expression. Second, since the MCMC is run on $|\hat{L}_B|$, it is the variance of $\log |\hat{L}_B|$ that enters the IF. Third, both $\sigma^2_{\log |\hat{L}_B|}$ and τ depend on $\gamma := n^2 \sigma^2_{du_i}$, which is the intrinsic variability in the estimator and is therefore fixed for a given dataset and choice of control variates. Fourth, setting $a = d - \lambda$ is infeasible in practice since knowing d requires evaluating the log-likelihood contribution for all observations, and we below outline a strategy to approximate d prior to the MCMC. Fifth, we can minimize CT with respect to m and λ as follows. For any choice of m and λ , and, assuming that \hat{d}_m is normally distributed, we can compute $\sigma^2_{\log |\hat{L}_B|}$ by Lemma 2, and τ by Lemma 3, and IF is obtained as in Pitt et al. (2012) (but for a correlated sampler as described in Appendix S1) and can be computed by one-dimensional numerical integration; see Lemma S3 in Appendix S1.

We can now evaluate CT over a grid of (m, λ) pairs and choose the minimizing pair. We conjecture however, based on numerical experiments, that the optimal solution to Eq. (4.12) is obtained with $m_{\text{opt}} = 1$ when \widehat{d}_m is normal. This can be intuitively understood as follows. First, the computational cost in the CT is proportional to the product $m\lambda$, so the individual m and λ do not matter for the cost. But λ has a much larger effect on increasing τ than m, which explains why one wants to spend computational resources on more batches (larger λ) rather than increasing the batch sizes (larger m). However, even though m=1 is likely to be optimal in the setting above, it is important to remember that this conclusion is based on the assumption of normality. Therefore, we recommend setting m large enough for the CLT to guarantee that the \widehat{d}_m are close to normal. We can then use the simplified expressions for $\sigma^2_{\log |\widehat{L}_B|}$ in Lemma 2, and τ in Lemma 3, rather than the more complex expression for these quantities under the mixture of normals assumption in Appendix S2. The efficiency loss from using for example m=30 rather than m=1 is modest, even when the tails of the d_k are fairly heavy, see Appendix S2.

Our recommended approach is to set m=30 and to minimize CT with respect to λ under the assumption of normal \hat{d}_m . Figure 1 shows how the optimal λ depends on γ . The

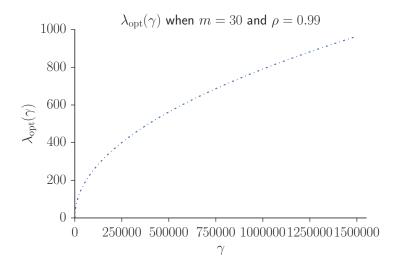


FIGURE 1. Empirical relationship between the intrinsic population variability γ and the λ that minimizes CT when m=30 and $\rho=0.99$. The relationship is given by $\lambda_{\rm opt}=\exp\left(-0.1022+0.4904\log(\gamma)\right)$.

estimated relationship is given in the caption of the figure. The following summarizes our optimal implementation strategy.

- 1. Sample θ values $\mathcal{T} = \{\theta^{(1)}, ..., \theta^{(M)}\}$ from a multivariate student t approximation of the posterior obtained by optimizing the log posterior based on $\ell(\theta)$ estimated from a single subsample.
- 2. Using a subsample of \widetilde{m} observations, estimate $\gamma(\theta)$ by $\widehat{\gamma}(\theta) = n^2 \widehat{\sigma}_{d_{u_i}}^2(\theta)$ for each $\theta \in \mathcal{T}$, and set $\gamma_{\max} \coloneqq \operatorname{argmax}_{\theta \in \mathcal{T}} \widehat{\gamma}(\theta)$.
- 3. Set the lower bound to $a = \bar{d} \lambda$, where $\bar{d} = M^{-1} \sum_{j=1}^{M} \widehat{d}_{\widetilde{m}}(\theta^{(j)})$ using the evaluations from Step 2 to compute the batch means $\widehat{d}_{\widetilde{m}}$.
- 4. With $\rho = 0.99$ and using m = 30, set λ to its optimal value (see Figure 1):

(4.13)
$$\lambda_{\text{opt}} = \exp(-0.1022 + 0.4904 \log(\gamma_{\text{max}})).$$

Step 1 can be replaced by a pilot MCMC run on a small subset of the data or any other crude approximation. The extra cost in terms of evaluations used in both Step 2 and 3 above is included in our algorithm whenever we compare against another algorithm. Note that Step 3 is performed only once before the MCMC, since we can not re-estimate d in each iteration as that would make the random variates U dependent across the products in the

block-Poisson estimator. If d is highly variable as a function of θ , Step 3 can be refined by estimating a regression model $d = f(\theta)$, but none of our applications have required this. In Section 5, we look closer into how $\gamma(\theta)$ and $d(\theta)$ behave in our applications. For other values of ρ and m, Step 4 can be replaced by direct minimization of the CT with respect to λ . Note that the optimal λ is based on the largest γ , which results in a conservatively low estimator variance, which is well known to be a good strategy in pseudo-marginal MCMC (Pitt et al., 2012).

Figure 2 shows the CT, $\Pr(\widehat{L} \geq 0)$ and $\sigma_{\log |\widehat{L}_B|}^2$ as a function of λ . The figure marks out the optimal value λ_{opt} . The figure shows that the optimal λ_{opt} results in a high probability of a positive estimator, regardless of γ .

Finally, Appendix S2 documents that the CT of the signed PMMH with the block-Poisson estimator is 2-3 times larger than for the approximate approach in Quiroz et al. (2018) for most γ , and has up to 8 times larger CT when γ is very small. This shows that there is a trade-off between exactness and computational time. The next section demonstrates empirically that our exact approach outperforms both MH on the full data set and the well-known Firefly MC algorithm for exact subsampling MCMC in Maclaurin and Adams (2014).

5. Application

5.1. **Model and data.** Our experiments consider the logistic regression for $y_i \in \{0, 1\}$ on covariates $x_i \in \mathbb{R}^p$, with density

$$p(y_i|x_i,\theta) = \left(\frac{1}{1 + \exp(x_i^T \theta)}\right)^{y_i} \left(\frac{1}{1 + \exp(-x_i^T \theta)}\right)^{1 - y_i}, \text{ with } p_{\Theta}(\theta) = \mathcal{N}(\theta|0, 10I),$$

which we fit to the same three datasets used in Quiroz et al. (2018): i) the CovType data as used in Collobert et al. (2002), with n = 550,087 observations and p = 11 variables, ii) the firm Bankruptcy data as used in Giordani et al. (2014), containing n = 4,748,089 observations and eight covariates, and iii) the HIGGS dataset (Baldi et al., 2014) with n = 1,100,000 observations and 21 covariates as in Quiroz et al. (2018).

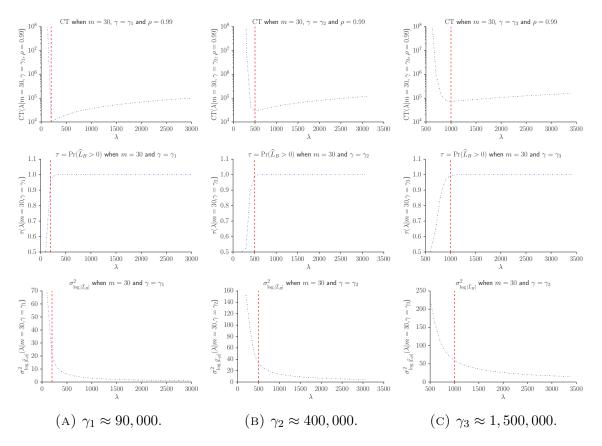


FIGURE 2. Optimality illustrations for three different γ , assuming $\rho=0.99$, m=30 and that \widehat{d}_m are normal. The columns correspond to three different γ . The top row shows the Computational Time (CT) in Eq. (4.12) as a function of λ . The middle row shows the probability of a positive estimator τ in Lemma 3 as a function of λ . The bottom rows shows the variance of the log of the absolute value of block-Poisson estimator in Eq. (2). The vertical red line marks the optimal value of λ , i.e. $\lambda_{\rm opt}$.

5.2. **Empirical studies.** Our first experiment compares the block-Poisson estimator to both Firefly Monte Carlo and standard MH on the full dataset. This experiment uses the parameter expanded control variates to be on par with Firefly Monte Carlo, which also uses a central measure of θ to construct its lower bound. Parameter expanded control variates results in a small $\gamma(\theta)$ if θ^* is close to the posterior mode. Figure 3 illustrates the magnitude of the (estimated) γ values over different θ values. Recall that we propose finding the optimal λ based on the maximal γ .

Our second objective is to show that our block-Poisson estimator can tolerate a large γ (by increasing λ , following the recommended guidelines in Section 4.3). To generate larger

 γ , we use the data expanded control variates in Section 3.2 with a small number of clusters. Moreover, the d_k population can in this case contain severe outliers and hence this scenario provides a serious challenge for our method. Figure 3 illustrates the magnitude of the (estimated) γ over different values of θ for the data expanded control variates; the values are considerably larger than the corresponding parameter expanded control variates in Figure 3.

In all our examples we simulate N=55,000 samples from the posterior and discard 5,000 as burn-in. We use a random walk Metropolis proposal with a scaling factor for the posterior covariance at the mode of $2.38/\sqrt{p}$ for MCMC (Roberts et al., 1997) and $2.5/\sqrt{p}$ for subsampling MCMC (Sherlock et al., 2015). We use the same scaling for the proposal of Firefly Monte Carlo as the MCMC.

5.3. Experiment 1: Comparisons against Firefly Monte Carlo and MCMC. Maclaurin and Adams (2014) also consider a logistic regression to demonstrate the performance of their Firefly algorithm, because the lower bound of the log-likelihood contribution is easily obtained. We choose the optimally tuned lower bound described in Maclaurin and Adams (2014), which makes the lower bound extremely tight when θ is close to the posterior mode θ^* . Following Maclaurin and Adams (2014), we allow 10% of the observations to change indicator in each iteration. For our block-Poisson estimator we use the parameter expanded control variates, expanded around θ^* . We confirm that the normality assumption for \hat{d}_m when m=30 is reasonable and, based on the values of γ from Figure 3, we find that $\lambda=100$ is optimal according to Eq. (4.13) rounded to the nearest allowable λ . We set the optimal lower bound as $a=-\lambda$, as the mean of $\hat{d}_{\tilde{m}}$ in Figure 3 is very small. All algorithms are started at θ^* .

We measure the performance of our subsampling MCMC using an empirical version of the CT in Eq. (4.12), in which the IF for the chain $\{s_i\theta_i, i=1,2,\ldots,N\}$ is estimated with the coda package in R (Plummer et al., 2006). Moreover, the cost is taken as the average number of evaluations over the MCMC iterations used when forming the estimator (the number of terms within a factor in the product is random). Finally, τ is replaced by its empirical estimate. For MCMC we use a similar measure but set $\tau = 1$ and the number of evaluations

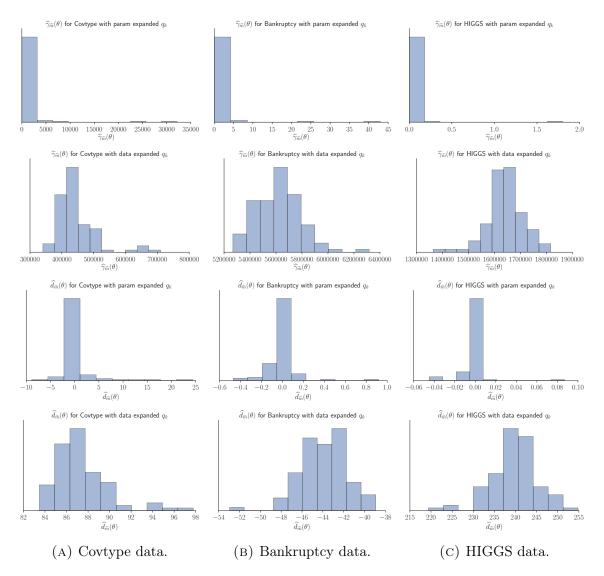


FIGURE 3. Estimating $\gamma(\theta)$ and $d(\theta)$ for different θ on the three datasets. Estimates of $\gamma(\theta) = n^2 \sigma_{d_{u_i}}^2$ and $d(\theta)$ based on $\widetilde{m} = 0.10n$. The histograms are obtained by, for a fixed u, estimating the quantities unbiasedly for 100 values of θ sampled from a multivariate Student t approximation of the posterior with 5 degrees of freedom. This generates over-dispersed θ values and hence we obtain a γ_{max} which is conservative. The results are shown for three datasets, Covtype (A), Bankruptcy (B) and HIGGS (C). Note that this figure is not useful for evaluating the normality assumption of $\widehat{d}_{\widetilde{m}}$, because u is fixed.

to n. We define the estimated Relative Computational Time (RCT) for the block-Poisson \mathcal{B} against any algorithm \mathcal{A} with $\tau = 1$ as

(5.1)
$$\widehat{RCT}_{\mathcal{A}} := \frac{\overline{CC}_{\mathcal{A}}\widehat{IF}_{\mathcal{A}}}{\overline{CC}_{\mathcal{B}}\widehat{IF}_{\mathcal{B}}/(2\widehat{\tau}_{\mathcal{B}} - 1)^{2}},$$

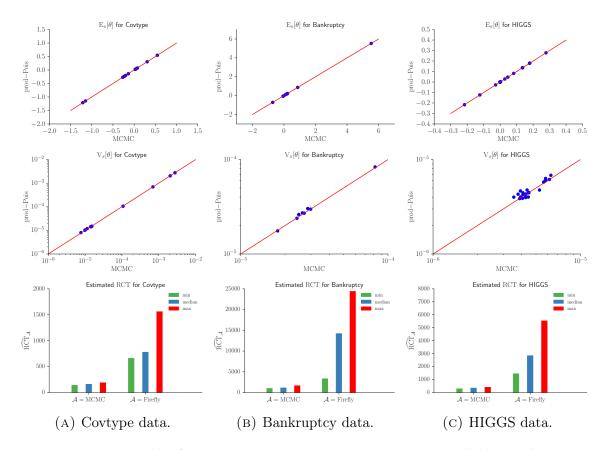


FIGURE 4. Results from Experiment 1 using parameter expanded control variates. Estimates of posterior expectation (upper), variance (middle) and Relative Computational Time (RCT) in Eq. (5.1) (lower) of correlated product-Poisson with $\lambda=100$ relative to MCMC and Firefly Monte Carlo. The results are shown for three datasets, Covtype (A), Bankruptcy (B) and HIGGS (C).

where CC is the computational cost introduced in Section 3.1 and the bars denote averages over MCMC iterations.

Figure 4 shows, for the three datasets, the accuracy of estimating the posterior expectations (upper panels) and variances (middle panels) using our method. We note that some estimates are visually off the 45-degree line, but all deviations are within the usual Monte Carlo error, and we conclude that the estimates are very accurate. The bottom row of Figure 4 also shows the relative computational times in Eq. (5.1) compared to MCMC and Firefly Monte Carlo, respectively. The results show that significant gains (in the order of several hundreds) are achieved with our block-Poisson approach compared to MCMC. More strikingly, the figure shows gains in the order of several thousands against Firefly Monte Carlo, implying that Firefly Monte Carlo is performing worse than MCMC. Indeed, Firefly

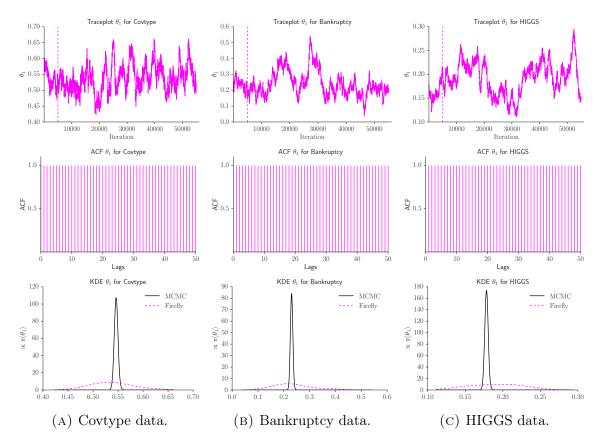


FIGURE 5. Results for Firefly Monte Carlo. Trace plots (upper), estimate of autocorrelation function (middle) and kernel density estimate (lower) of the posterior samples using FireFly Monte Carlo. The vertical line in the upper panels marks the burn-in 5,000.

Monte Carlo has a lower computational cost than MCMC (because it uses less density evaluations), but gives extremely autocorrelated samples. Figure 5 illustrates this for θ_1 (the performance is similar for all parameters), which also shows that Firefly Monte Carlo gives heavily distorted posteriors for these datasets in finite time, even if it is known to target the true posterior. Somewhat surprisingly, the poor performance of FireFly Monte Carlo remains even when all the indicators are updated in a given iteration.

5.4. Experiment 2: Performance when γ is large. Our next experiment is a serious test of our methodology with larger γ . To this end, we use the data expanded control variates in Quiroz et al. (2018) and cluster y = 0 and y = 1 separately, with a very small number of clusters in relation to n. We follow Quiroz et al. (2018) and choose K = 1042, 16374, 355 for, respectively, the Covtype, Bankruptcy and HIGGS datasets. This results in a d_k population

that may contain severe outliers, and therefore normality is not guaranteed for m=30. We found that m=100,100,600 are large enough for assuming normality in the Covtype, Bankruptcy and HIGGS datasets, respectively, and the corresponding optimal λ are $\lambda=500,1100,300$.

Figure 6 shows the results. Although our control variates are significantly less accurate than in Experiment 1, our algorithm is still more efficient than MCMC and dramatically more efficient that Firefly Monte Carlo. Note that Firefly Monte Carlo is based on the same extremely tight lower bound as in Experiment 1, yet our algorithm with poor control variates still performs much better. This example illustrates that even when γ is large and varies a lot over the parameter space, our guidelines provides a λ which results in an efficient MCMC chain that still outperforms MH on the full dataset and Firefly Monte Carlo.

6. Conclusions and Future Research

We propose an algorithm for fast exact simulation-based inference where the likelihood is estimated cheaply by efficient data subsampling. At the core of the algorithm is a novel block-Poisson estimator that estimates the likelihood unbiasedly while inducing a control-lable dependence between estimates at successive MCMC iterations. Such dependence over the iterations has been established to be very benefical for the efficiency of pseudo marginal algorithms (Deligiannidis et al., 2017). We argue that using a strict lower bound in the estimator is computationally wasteful and instead advocate using a soft lower bound such that the estimates can be occasionally negative. The negative estimates are handled with the signed PMMH approach in Lyne et al. (2015) where the pseudo-marginal MCMC is based on the absolute value of the likelihood estimate followed by a sign-correcting importance sampling step to estimate any function of the parameters consistently. A major contribution is that we derive practical guidelines for the tuning parameters of the estimator in signed PMMH by minimizing the asymptotic variance of the importance sampling estimator per unit of computing time, thereby taking into account the computing cost of the likelihood estimator, the inefficiency of the MCMC and the probability of a negative sign. The guidelines

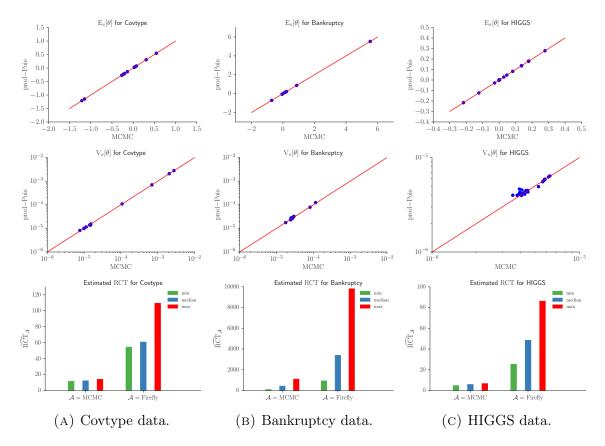


FIGURE 6. Results from Experiment 2 using data expanded control variates with large γ . Estimates of posterior expectation (upper), variance (middle) and Relative Computational Time (RCT) in Eq. (5.1) (lower) of block-Poisson with $\lambda=500,1100,300$ (Covtype, Bankruptcy, HIGGS) relative to MCMC and Firefly Monte Carlo. The results are shown for three datasets, Covtype (A), Bankruptcy (B) and HIGGS (C).

are based on idealized assumptions, but we demonstrate that the guidelines are accurate, effective and robust to features of the data. The methodology proposed here is applied to subsampling, but applies to a much larger set of problems, and in particular models with intractable normalizing constants.

We demonstrate the performance of our algorithm for logistic regression on three commonly used datasets. The proposed algorithm dramatically outperforms both MH on the full dataset, and the widely cited Firefly Monte Carlo subsampling algorithm.

An attractive feature of our approach is that it provides a consistent estimator for any function of the parameters regardless of the variance of the likelihood estimator, and is therefore suitable for high dimensional problems where, inevitably, the variance becomes

large. We are currently investigating the role of the block-Poisson estimator in constructing efficient high-dimensional proposals. Some work in this direction is found in Dang et al. (2017), and the block-Poisson estimator opens up the possibility for optimally tuning these algorithms.

7. ACKNOWLEDGMENTS

Matias Quiroz, Robert Kohn and Khue-Dung Dang were partially financially supported by Australian Research Council Center of Excellence grant CE140100049. Mattias Villani was partially financially supported by Swedish Foundation for Strategic Research (Smart Systems: RIT 15-0097).

REFERENCES

- Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, pages 697–725.
- Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5.
- Banterle, M., Grazian, C., and Robert, C. P. (2014). Accelerating Metropolis-Hastings algorithms: Delayed acceptance with prefetching. arXiv preprint arXiv:1406.2660.
- Bardenet, R., Doucet, A., and Holmes, C. (2014). Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *Proceedings of The 31st International Conference on Machine Learning*, pages 405–413.
- Bardenet, R., Doucet, A., and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43.
- Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.
- Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):333–382.

- Bierkens, J., Fearnhead, P., and Roberts, G. (2016). The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. arXiv preprint arXiv:1607.03188.
- Bierkens, J. and Roberts, G. (2017). A piecewise deterministic scaling limit of lifted Metropolis–Hastings in the Curie–Weiss model. *The Annals of Applied Probability*, 27(2):846–882.
- Christen, J. A. and Fox, C. (2005). MCMC using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810.
- Collobert, R., Bengio, S., and Bengio, Y. (2002). A parallel mixture of SVMs for very large scale problems. *Neural Computation*, 14(5):1105–1114.
- Dahlin, J., Lindsten, F., Kronander, J., and Schön, T. B. (2015). Accelerating pseudo-marginal Metropolis-Hastings by correlating auxiliary variables. arXiv preprint arXiv:1511.05483.
- Dang, K.-D., Quiroz, M., Kohn, R., Tran, M.-N., and Villani, M. (2017). Hamiltonian Monte Carlo with energy conserving subsampling. arXiv preprint arXiv:1708.00955.
- Deligiannidis, G., Doucet, A., and Pitt, M. K. (2017). The correlated pseudo-marginal method. arXiv preprint arXiv:1511.04992v4.
- Doucet, A., Pitt, M., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313.
- Fearnhead, P., Papaspiliopoulos, O., Roberts, G. O., and Stuart, A. (2010). Random-weight particle filtering of continuous time processes. *Journal of the Royal Statistical Society:* Series B (Statistical Methodology), 72(4):497–512.
- Feller, W. (2008). An introduction to probability theory and its applications, volume 2. John Wiley & Sons.
- Giordani, P., Jacobson, T., Von Schedvin, E., and Villani, M. (2014). Taking the twists into account: Predicting firm bankruptcy risk with splines of financial ratios. *Journal of Financial and Quantitative Analysis*, 49(4):1071–1099.

- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.
- Jacob, P. E. and Thiery, A. H. (2015). On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769–784.
- Korattikara, A., Chen, Y., and Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 181–189.
- Liu, S., Mingas, G., and Bouganis, C.-S. (2015). An exact MCMC accelerator under custom precision regimes. In Field Programmable Technology (FPT), 2015 International Conference on, pages 120–127. IEEE.
- Liu, S., Mingas, G., and Bouganis, C.-S. (2017). An unbiased MCMC FPGA-based accelerator in the land of custom precision arithmetic. *IEEE Transactions on Computers*, 66(5):745–758.
- Lyne, A.-M., Girolami, M., Atchade, Y., Strathmann, H., and Simpson, D. (2015). On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. Statistical Science, 30(4):443–467.
- Maclaurin, D. and Adams, R. P. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*.
- 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.
- Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.
- Papaspiliopoulos, O. (2009). A methodological framework for Monte Carlo probabilistic inference for diffusion processes. *Manuscript*. Available at http://wrap.warwick.ac.uk/35220/1/WRAP_Papaspiliopoulos_09-31w.pdf.

- Pav, S. E. (2015). Moments of the log non-central chi-square distribution. arXiv preprint arXiv:1503.06266.
- Payne, R. D. and Mallick, B. K. (2017). Two-stage Metropolis-Hastings for tall data. *Journal of Classification*, (To appear).
- Pitt, M. K., Silva, R. d. S., 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.
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.
- Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2018). Speeding up MCMC by efficient data subsampling. *Journal of American Statistical Association*, (To appear).
- Quiroz, M., Tran, M.-N., Villani, M., and Kohn, R. (2017). Speeding up MCMC by delayed acceptance and data subsampling. *Journal of Computational and Graphical Statistics*, (To appear).
- Rhee, C. and Glynn, P. W. (2015). Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043.
- Roberts, G. O., Gelman, A., Gilks, W. R., et al. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.
- Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.
- Sherlock, C., Thiery, A. H., Roberts, G. O., and Rosenthal, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *The Annals of Statistics*, 43(1):238–275.
- Strathmann, H., Sejdinovic, D., and Girolami, M. (2015). Unbiased Bayes for big data: Paths of partial posteriors. arXiv preprint arXiv:1501.03326.
- Van der Vaart, A. W. (1998). Asymptotic statistics, volume 3. Cambridge university press.

Wagner, W. (1988). Unbiased multi-step estimators for the Monte Carlo evaluation of certain functional integrals. *Journal of Computational Physics*, 79(2):336–352.

Wagner, W. (1989). Unbiased Monte Carlo estimators for functionals of weak solutions of stochastic differential equations. Stochastics: An International Journal of Probability and Stochastic Processes, 28(1):1–20.

Walck, C. (1996). Hand-book on statistical distributions for experimentalists. Technical report. http://inspirehep.net/record/1389910/files/suf9601.pdf.

S1. DERIVATIONS FOR OPTIMAL IMPLEMENTATION

This appendix derives the framework for obtaining the heuristic guidelines for choosing λ and m in Section 4. In particular, we obtain a tractable expression for the inefficiency factor for the signed block PMMH sampler, which we use for minimizing the computational time CT_B . The analysis follows Pitt et al. (2012) and our simplifying assumptions are in the same spirit.

Define,

(S1)
$$C_i(\theta) = \int_{\mathcal{U}} |\widehat{L}^{(i)}(\theta, u_i)| p_U(du_i), \widetilde{Z}_i(\theta, u_i) = \log |\widehat{L}^{(i)}(\theta, u_i)| - \log C_i(\theta), \quad i = 1, \dots, G,$$

$$Z(\theta, u) = \sum_{i=1}^{G} \widetilde{Z}_i(\theta, u_i)$$
 and $\widetilde{Z}_{1:G} = (\widetilde{Z}_1, \dots, \widetilde{Z}_G).$

We then have the following lemma whose proof is straightforward and is omitted.

Lemma S1.

(S2)
$$\overline{\pi}(d\theta, du) = \exp(z)\nu(d\theta) \prod_{i=1}^{G} p_U(du_i), \quad \overline{\pi}(d\widetilde{z}_{1:G}|\theta) = \prod_{i=1}^{G} \exp(\widetilde{z}_i) p_{Z_i}(d\widetilde{z}_i),$$

$$\mathbb{E}_{u_i \sim p_U} \left(\exp\left(\widetilde{Z}_i\right) |\theta\right) = 1 \quad and \quad \mathbb{E}_{u \sim p_U} \left(\exp(Z) |\theta\right) = 1,$$
with $\nu(d\theta)$ in Eq. (4.3).

Eq. (S2) shows that the \tilde{z}_i are $\bar{\pi}$ -independent conditional on θ . Let $v = \hat{L}(\theta, u)$ and take w such that (θ, v, z, w) is diffeomorphism of (θ, u) . Then, $p(d\theta, dv, dz, dw) = p(d\theta, du)$. This allows us to transform the measure to a workable expression in Lemma S2. Next, we assume that,

Assumption S1. $S := sign(\widehat{L}(\theta, U))$ and $Z(\theta, U)$ are $\overline{\pi}$ -independent given θ .

Assumption S1 is reasonable as Z is defined in terms of the absolute value of the estimator and therefore ignores the sign.

Lemma S2. Suppose that Assumption S1 holds. Then,

(S3)
$$\overline{\pi}(d\theta, ds, dz) = \exp(z)\nu(d\theta)p(ds|\theta)p(dz|\theta).$$

Similarly to Pitt et al. (2012), we consider a hypothetical chain targeting Eq. (S3). The following assumption presents the idealized proposal which makes the derivation of the inefficiency tractable.

Assumption S2.

- (i) If $U_i \sim p_U(\cdot)$, i = 1, ..., G, then the distribution of $Z = \sum_{i=1}^G \widetilde{Z}_i(\theta, U_i)$ conditional on θ is $\mathcal{N}(-\sigma^2/2, \sigma^2)$ with $\sigma^2 := V_{u \sim p_U}[Z]$, which is independent of θ .
- (ii) $q(\theta,s,z;d\theta',ds',dz') = \nu(d\theta')p(ds'|\theta')q(z;dz'|\sigma^2,\rho), \ with \ \rho=1-\frac{1}{G} \ and$

(S4)
$$q(z; dz'|\rho, \sigma^2) = \mathcal{N}\left(dz' - \frac{\sigma^2}{2}(1-\rho) + \rho z, \sigma^2(1-\rho^2)\right) \text{ for } \rho = 1 - \frac{1}{G}.$$

The mean in Part (i) of Assumption S2 is -1/2 of the variance which is consistent with the fact that $E_{u\sim p_U}[e^Z]=1$. This implies that, if $Z\sim \overline{\pi}$, then $Z\sim \mathcal{N}(\sigma^2/2,\sigma^2)$. The proposal $q(z;dz'|\rho,\sigma^2)$ in Part (ii) of Assumption S2 implies that the correlation between the current z and proposal z' is $\rho=1-1/G$. The assumption that $\rho=1-1/G$ is plausible because the current Z and proposed Z' differ only by one block; see also the discussion at the end of Section 4.2. The inefficiency factor for the independent pseudo-marginal method in Pitt et al. (2012) is derived using $\rho=0$. The next lemma uses their proof, but with the proposal in Eq. (S4).

Lemma S3. Suppose that Assumption S2 holds. Then,

(i) The Metropolis-Hastings acceptance probability of the proposal $q(\theta, s, z; d\theta', ds', dz')$ is given by

$$\min\{1, \exp(z'-z)\}.$$

(ii) The acceptance probability conditional on z of the idealized sampling scheme is

$$k(z|\sigma^2, \rho) = \int \min\{1, \exp(z'-z)\} q(z; dz'|\sigma^2, \rho).$$

(iii) The inefficiency of the sampling scheme is

$$\operatorname{IF}_{\overline{\pi}}(\sigma^2, \rho) = 1 + 2\operatorname{E}_{\overline{\pi}(z|\theta)}\left(\frac{1 - k(z|\sigma^2, \rho)}{k(z|\sigma^2, \rho)}\right),$$

where

(S5)
$$k(z|\sigma^2, \rho) = \exp(-x + w^2/2)\Phi\left(\frac{x}{w} - w\right) + \Phi\left(\frac{-x}{w}\right),$$

with $x := \left(z + \frac{\sigma^2}{2}\right)(1-\rho)$, $w := \sigma\sqrt{1-\rho^2}$ and Φ denotes the standard normal cumulative density function.

The inefficiency $IF_{\overline{\pi}}(\sigma^2, \rho)$ can be computed accurately using one-dimensional numerical integration as $\overline{\pi}(z|\theta) \sim \mathcal{N}(\sigma^2/2, \sigma^2)$. We end this section by presenting an alternative set of more restrictive assumptions that imply Part (i) of Assumption S2 and the proposal in Part (ii) of Assumption S2, i.e. Eq. (S4). This assumption serves the purpose of providing greater understanding of the results above.

Assumption S3. If $U_i \sim p_U(\cdot)$, i = 1, ..., G, then the distribution of $\widetilde{Z}_i(\theta, U_i)$ conditional on θ is $\mathcal{N}(-\sigma^2/2G, \sigma^2/G)$ with $\sigma^2 := V_{u \sim p_U}[Z]$ for $Z = \sum_{i=1}^G \widetilde{Z}_i(\theta, U_i)$, which is independent of θ .

The following lemma then implies the desired result.

Lemma S4. Suppose that Assumption S3 holds and let

$$Z = \sum_{i=1}^{G} \widetilde{Z}_i(\theta, U_i) \text{ and } Z' = \sum_{i \neq j, i=1}^{G} \widetilde{Z}_i(\theta', U_i) + \widetilde{Z}_j(\theta', U_j'),$$

with $U_i \sim \overline{\pi}$ and $U'_j \sim p_{U_j}(\cdot)$. Then,

(S6)
$$\begin{bmatrix} z \\ z' \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \frac{\sigma^2}{2} \\ -\frac{\sigma^2}{2} (1 - 2\rho) \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \text{ with } \rho = 1 - \frac{1}{G}.$$

S2. Details on optimally implementing the block-Poisson estimator

This appendix provides a comprehensive study of the optimal guidelines, addressing the following points.

- We consider an alternative approach, based on assuming that \widehat{d}_m is a mixture of normals. We demonstrate that, when the d_k follow a Student t distribution, then the guidelines provided by this alternative approach do not differ much compared to that of \widehat{d}_m following a normal distribution when m=30. In contrast, for small values of m, the guidelines do not coincide. We conclude that m should be chosen large enough for \widehat{d}_m to be normally distributed.
- We compare the performance of signed PMMH using the block-Poisson estimator to the approximate subsampling approach in Quiroz et al. (2018).
- We evaluate the accuracy of the guidelines under both idealized conditions, i.e. when Assumptions S1 and S2 hold. We also test the guidelines when σ^2 depends on θ , i.e. when Part (i) of Assumption S2 does not hold. We find that, even when the assumptions are not fulfilled, the recommended λ from the guidelines is close to the λ that gives the best empirical performance. In particular, our guidelines do not recommend a λ which results in a very inefficient chain for θ . This is also true for the applications considered in the paper.
- S2.1. Optimal tuning when \widehat{d}_m follows a mixture of normals. Suppose the following hypothetical and ideal situation, in which the d_k are normally distributed. Then, \widehat{d}_m is normally distributed even when m=1, and hence the guidelines in our article are still valid. When \widehat{d}_m is not normal, we can model it by a finite mixture of normals, which is known to approximate any distribution arbitrarily well with enough components. We propose to fit a mixture to \widehat{d}_m using characteristic functions, which is described below. The following lemma

generalizes Lemma 2. Its proof is in Appendix S4. The notation $\text{Mix}-\mathcal{N}(\mu, \sigma^2, \omega)$ means that we have a standardized finite mixture with C components. Its parameters are $\mu \coloneqq \mu_{1:C}$, $\sigma \coloneqq \sigma_{1:C}^2$, $\omega \coloneqq \omega_{1:C}$ with $\sum_j \omega_j = 1$ and, by standardization, its mean and variance are $\sum_j \omega_j \mu_j = 0$ and $\sum_j \omega_j (\sigma_j^2 + \mu_j^2) - 1 = 1$.

Lemma S5. Let $\bar{d}_m^{(h,l)} = \sqrt{\frac{m}{\gamma}} (\hat{d}_m^{(h,l)} - d) \stackrel{iid}{\sim} \text{Mix} - \mathcal{N}(\mu, \sigma^2, \omega)$ follow mixture of normals for all h and l such that $\mathrm{E}\left[\bar{d}_m^{(h,l)}\right] = 0$ and $\mathrm{V}\left[\bar{d}_m^{(h,l)}\right] = 1$. The variance of $\log \left|\hat{L}_B\right|$ when $a = d - \lambda$ is then

$$V\left[\log\left|\widehat{L}_{B}\right|\right] = \lambda \sum_{c=1}^{C} \omega_{c}(\nu_{c}^{2} + (\eta_{c} - \eta)^{2}) + \lambda \eta^{2},$$

where

$$\eta_c = \log\left(\frac{\sigma_c}{\lambda}\sqrt{\frac{\gamma}{m}}\right) + \frac{1}{2}\left(\log 2 + \mathcal{E}_{J_c}\left[\psi^{(0)}(1/2 + J_c)\right]\right),$$

 $\eta = \sum_{c=1}^{C} \omega_c \eta_c$ and

$$\nu_c^2 := \frac{1}{4} \left(\mathcal{E}_{J_c} \left[\psi^{(1)} (1/2 + J_c) \right] + \mathcal{V}_{J_c} \left[\psi^{(0)} (1/2 + J_c) \right] \right)$$

with $J_c \sim \operatorname{Pois}\left(\frac{(\mu_c + \sqrt{\frac{m}{\gamma}}\lambda)^2}{2\sigma_c^2}\right)$ and $\psi^{(q)}$ is the polygamma function of order q. Furthermore, $\operatorname{V}\left[\log\left|\widehat{L}_B\right|\right] < \infty$ for all m > 0, $\lambda > 0$ and μ, σ, ω .

Given the finite mixture distribution for \widehat{d}_m , it is straightforward to compute the probability that the estimator is positive, as this becomes a mixture of normal cumulative distribution functions. Hence, we can also numerically evaluate the CT in Eq. (4.12) for the finite mixture of normal case, and can optimize λ for a given m. Figure S7 plots the optimal $\lambda_{\rm opt}$ as a function of γ for several values of m in the mixture of normals case, and also displays the corresponding $\lambda_{\rm opt}$ for the case of a normal \widehat{d}_m . The figure illustrates that, when m=30 the two approaches result in nearly indistinguishable optimal values for all γ . In contrast, when m is small, the optimal guidelines can differ substantially, especially for larger values of γ .

We now outline in detail how to find the parameters $(\omega_{1:G}, \sigma_{1:G}^2, \mu_{1:G})$ in the finite mixture distribution. Let $X_1, ..., X_n | \theta \stackrel{iid}{\sim} f_X(x|\theta)$ with a finite mean μ and variance σ^2 . We are interested in approximating the distribution of the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ by a

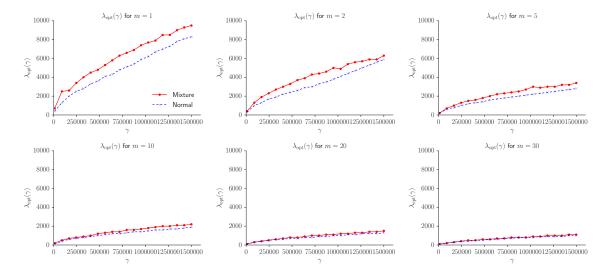


FIGURE S7. Optimal λ as a function of γ for the normal approach vs the finite mixture of normal approach. Each figure shows, for a specific value of m (see title), the λ obtained by minimizing the CT in Eq. (4.12) under a normal assumption for \widehat{d}_m (blue dashed line) and a finite mixture of normal assumption for \widehat{d}_m (red solid line).

mixture of normals

$$f_{\bar{X}_n}(x) \approx g_{\bar{X}_n}(x|\eta, \psi^2, \omega) = \sum_{c=1}^C \omega_c N(x|\eta_c, \psi_c^2),$$

where $N(x|\eta, \psi^2)$ denotes a normal density with mean η and variance ψ^2 . Let $\beta = \{\omega, \eta, \psi\}$ to be all parameters of the mixture and let $g_{\bar{X}_n}^{(\beta)}(x)$ denote the density for the sample mean from the mixture with parameters β .

The aim is to find the necessary number of mixture components C and the parameters β of the mixture that approximates $f_{\bar{X}_n}(x)$ well. We will do so by minimizing the L_2 distance between $f_{\bar{X}_n}(x)$ and $g(x|\theta,\psi^2,\omega)$

$$d\left(f_{\bar{X}_n}, g_{\bar{X}_n}^{(\beta)}\right) = \int \left(f_{\bar{X}_n}(x) - g_{\bar{X}_n}^{(\beta)}(x)\right)^2 dx.$$

The L_2 distance is very convenient since by Plancherel's theorem we can turn the density matching problem into a Characteristic Function (CF) matching problem,

$$d\left(f_{\bar{X}_n},g_{\bar{X}_n}^{(\beta)}\right) = \int \left(f_{\bar{X}_n}(x) - g_{\bar{X}_n}^{(\beta)}(x)\right)^2 dx = \int \left(\varphi_{\bar{X}_n}(t) - \varphi_{\bar{X}_n}^{(\beta)}(t)\right)^2 dt = d\left(\varphi_{\bar{X}_n}(t),\varphi_{\bar{X}_n}^{(\beta)}(t)\right),$$

where $\varphi_X(t)$ is the characteristic function $\varphi_X(t) := \mathbb{E}\left[e^{itX}\right]$ for a random variable X. Matching CFs is especially attractive here since the density of the sample mean $f_{\bar{X}_n}$ may be intractable, but its CF is straightforward to obtain,

$$\varphi_{\bar{X}_n}(t) = (\varphi_X(t/n))^n$$
,

where $\varphi_X(t)$ is the CF of $f_X(x|\theta)$. We will match CFs for the standardized mean $\bar{Z}_n = (\sqrt{n}/\sigma)(\bar{X}_n - \mu)$. Using the property $\varphi_{a+bX}(t) = e^{ita}\varphi_X(bt)$ we obtain the CF for the standardized mean as

$$\varphi_{\bar{Z}_n}(t) = e^{-it\mu\sqrt{n}/\sigma} \left(\varphi_X(t/(\sqrt{n}\sigma))\right)^n.$$

The CF for a normal mixture $\sum_{c=1}^{C} \omega_c N(x|\eta_c, \psi_c^2)$ is $\varphi_X(t) = \sum_{c=1}^{C} \omega_c \varphi_{X_c}(t)$, where $\varphi_{X_c}(t) = \exp(i\eta_c t - \psi_c^2 t^2/2)$ is the CF of the cth mixture component.

We minimize $d\left(\varphi_{\bar{Z}_n}(t), \varphi_{\bar{Z}_n}^{(\beta)}(t)\right)$ with respect to β for a given C by reparameterizing the standard deviations in the mixture components in exponential form and the weights ω using the softmax function. The minimization is subject to the restrictions that the mixture has zero mean and unit variance: $\sum_{c=1}^{C} \omega_c \eta_c = 0$ and $\sum_{c=1}^{C} \omega_c \left(\kappa_c^2 + (\eta_c - \eta)^2\right) = 1$. These restrictions can be enforced directly or indirectly via penalties.

S2.2. Comparison against the approximate approach in Quiroz et al. (2018). Quiroz et al. (2018) use the bias-corrected likelihood estimator,

$$\widehat{L}_A := \exp\left(q + \widehat{d}_M - \frac{n^2}{2M}\widehat{\sigma}_{d_{u_i}}^2\right),$$

based on a subsample of size M. It can be shown that if the d_k are normally distributed, then

$$\sigma_{\log \widehat{L}_A}^2 := V[\log \widehat{L}_A] = \frac{\gamma}{M} + \frac{\gamma^2}{2M^3},$$

and we can then define the computational time

$$\operatorname{CT}_{\mathcal{A}}(M|\gamma, \rho = 0.99) = M \cdot \operatorname{IF}\left(\sigma_{\log \widehat{L}_{A}}^{2}(M|\gamma, \rho = 0.99)\right).$$

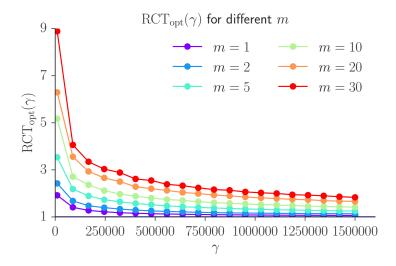


FIGURE S8. RCT in Eq. (S1) as a function of λ for different values of m.

Similarly to the block-Poisson estimator in Section 4.3, we can, for each γ , minimize this CT with respect to M. Then, We can obtain the optimal Relative Computational Time (RCT) as a function of γ ,

(S1)
$$RCT_{opt}(\gamma) = \frac{CT_B(\lambda_{opt}|\gamma, m, \rho = 0.99)}{CT_A(M_{opt}|\gamma, \rho = 0.99)}.$$

Figure S8 plots this for several values of m and shows that the approximate approach has lower CT, by a factor of 2 and 9 for m = 30, depending of the value of γ .

- S2.3. Testing the guidelines under several scenarios. We conclude this section by testing the optimal guidelines against empirical performance to see if they are sensible. To this end, we consider a simple Poisson regression model with one covariate and no intercept, $y_k \sim \text{Poisson}(\exp(\theta x_k))$. We use m=30 and obtain the differences $d_k = \ell_k q_k$ in the following two ways:
 - (1) We sample a population d_k and obtain $\sigma_d^2 = V[d_k] = 1/n \sum_{k=1}^n (d_k \overline{d})^2$. Since d_k is a fixed population for any θ (by construction), $\gamma = n^2 \sigma_d^2$ is a known number which does not depend on θ , which is the main assumption behind the derivation of the guidelines in Appendix S1.
 - (2) For each proposed θ , we compute $d_k(\theta) = \ell_k(\theta) q_k(\theta)$, taking $q_k(\theta)$ to be the data expanded control variate in Section 3.2. We then compute $\gamma(\theta) = n^2 \sigma_d^2(\theta)$, so γ

depends on θ and, unlike Step (1), the assumptions in our theory are not fulfilled. For optimal tuning, we take the largest $\gamma(\theta)$ and determine the optimal λ based on this, as outlined in Section 4.3.

The guidelines are tested as follows. We run pseudo-marginal chains to sample θ for a several different λ values. We choose a perfect proposal by computing $\nu(\theta) \propto \int |\widehat{L}(\theta,u)| p_U(u) du$ on a grid of θ values (by Monte Carlo simulation). Since θ is one dimensional, this distribution is easily sampled by the inverse cdf method. Running a long MCMC chain of θ , we can estimate the integrated auto-correlation time of $s\theta$ and compute the empirical version of the CT similarly to the description in Section 5. Repeating this procedure for each of the MCMC chains we run (each one corresponding to a specific value of λ), we can choose the λ that gives the smallest empirical computational time and compare it to that recommended by the guidelines. Figure S9 shows that when $\lambda_{\rm opt}$ provided by the guidelines does not agree with the empirically obtained one, the difference in computational time between the different λ is very small. We therefore conclude that the guidelines are sensible, and more importantly, they never suggest a λ which is too small resulting in a catastrophically large CT.

S3. The block PMMH with
$$\widehat{L} \geq 0$$

In this appendix we derive a special case of our algorithm in which the sign is $1 \, \overline{\pi}$ -almost surely, which we refer to as the block PMMH. We provide guidelines for its optimal tuning. Suppose that the $\widehat{L}^{(i)}(\theta, u) \geq 0, i = 1, \ldots, G$ for all $\theta \in \Theta, u_{1:G} \in \mathcal{U}^G$, so that the likelihood estimator is $\overline{\pi}$ -almost surely positive, i.e. $\Pr_{\overline{\pi}}(S = 1) = 1$. Then, $\nu(d\theta) = \pi(d\theta)$ (the posterior), $C(\theta) = L(\theta)$ (the likelihood), $\overline{C} = p(y)$ (the marginal likelihood), Eq. (4.7) becomes

$$\overline{\pi}(d\theta, du_{1:G}) = \frac{p_{\Theta}(d\theta)}{\overline{C}} \prod_{i=1}^{G} \widehat{L}^{(i)}(\theta, u_i) p_U(du_i),$$

and Eq. (4.5) becomes

(S1)
$$\widehat{\mathcal{E}}_{\pi}[\psi] = \frac{1}{N} \sum_{i=1}^{N} \psi(\theta^{(i)}).$$

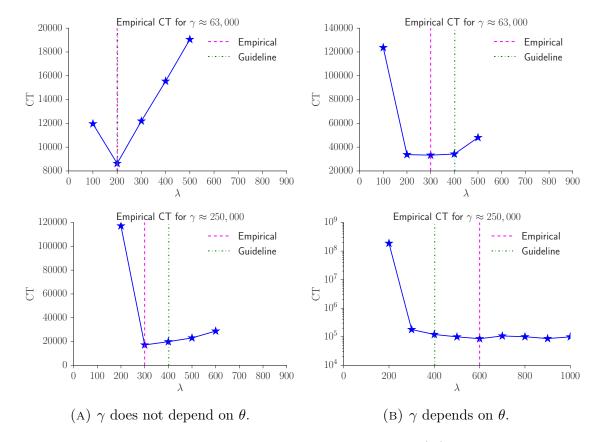


FIGURE S9. Empirical choice of λ vs guidelines. Panel (A) shows the empirical CT for two different γ values, under the scenario that γ is independent of θ . Panel (B) shows the empirical CT for two different γ values, but under the scenario that γ depends on θ . In this case, we have tuned the algorithm according to the largest $\gamma(\theta)$, as described in Section 4.3. In all figures, the vertical lines correspond to the optimal λ according to the guidelines (dotted green) and the empirically observed optimum (dashed magenta).

Then the following convergence result holds. Its proof follows from the proof of Theorem 1. Corollary 1. Suppose that $\widehat{L}(\theta, u) \geq 0$. Suppose furthermore that the Markov chain in Algorithm 2 is irreducible and aperiodic. Then,

- (i) Algorithm 1 is reversible.
- (ii) Algorithm 1 converges to $\overline{\pi}$ in total variation norm, with marginal $\pi(d\theta) = \int_{\mathcal{U}} \overline{\pi}(d\theta, du)$.
- (iii) Suppose that $E_{\pi}[|\psi|] < \infty$. Then, $\widehat{E}_{\pi}[\psi] \to E_{\pi}[\psi]$ a.s. $(\overline{\pi})$.
- (iv) Define

$$IF_{\overline{\pi},\psi} = \frac{V_{\overline{\pi}}(\psi) + 2\sum_{j=1}^{\infty} \overline{\Gamma}_j}{V_{\overline{\pi}}(\psi)},$$

where $\overline{\Gamma} = Cov_{\overline{\pi}} \Big(\psi(\theta_0), \psi(\theta_j) \Big)$ and $(\theta_0, \dots, \theta_j, \dots)$ are the MCMC iterates of θ . If $V_{\overline{\pi}}[\psi]IF_{\overline{\pi},\psi} = V_{\pi}[\psi]IF_{\overline{\pi},\psi} < \infty$, then

$$\sqrt{N} \Big(\widehat{\mathcal{E}}_{\pi}(\psi) - \mathcal{E}_{\pi}(\psi) \Big) \to \mathcal{N} \Big(0, \mathcal{V}_{\pi}(\psi) \mathrm{IF}_{\overline{\pi}, \psi} \Big).$$

We shall call the algorithm with a nonnegative likelihood estimator in Corollary 1 the block PMMH. We note that by Part (ii) of Corollary 1, unlike the signed block PMMH, the iterates of θ for the block PMMH are distributed according to $\pi(d\theta)$.

S3.1. Tuning the block PMMH. We now consider tuning the block PMMH, which is a special case of the signed block PMMH with only positive signs and therefore $\tau := 1$; see Section 4.1.

Pitt et al. (2012) show that the optimal number of particles m, when $\sigma_{\log \widehat{L}}^2 \propto 1/m$, is given implicitly by the $\sigma_{\log \widehat{L}}^2$ that minimizes the computational time

(S2)
$$\operatorname{CT}(\sigma_{\log \widehat{L}}^2) \coloneqq m \cdot \operatorname{IF}(\sigma_{\log \widehat{L}}^2) \propto \frac{\operatorname{IF}(\sigma_{\log \widehat{L}}^2)}{\sigma_{\log \widehat{L}}^2},$$

where IF denotes the inefficiency factor of the pseudo-marginal chain; see also Doucet et al. (2015). We note that this computational time only depends on $\sigma_{\log \widehat{L}}^2$, which in turn is proportional to 1/m, hence simplifying the expression compared to Eq. (4.12). Pitt et al. (2012) derive, based on several assumptions which we also invoke in Appendix S1, an analytic expression for the IF and shows that $\mathrm{CT}(\sigma_{\log \widehat{L}}^2)$ in Eq. (S2) is minimized when $\sigma_{\log \widehat{L}}^2 \approx 1$. Under less restrictive assumptions Doucet et al. (2015); Sherlock et al. (2015) show that the optimal value of $\sigma_{\log \widehat{L}}^2$ ranges between approximately 1 and 3.3.

The next lemma shows the optimal $\sigma_{\log \widehat{L}}^2$ for our block PMMH and is proved using the IF derived in Pitt et al. (2012), but incorporating our block scheme.

Lemma S6. Given the assumptions in Appendix S1, $\sigma_{\log \widehat{L}}^2 \approx 2.16^2/(1-\rho^2)$ minimizes Eq. (S2) when $\rho = 1 - 1/G$ is close to 1.

Lemma S6 shows that our block PMMH with G = 100 speeds up the independent pseudomarginal significantly, as the optimal value of $\sigma_{\log \hat{L}}^2$ is 234, which is much larger than 1-3.3.

S4. Proofs

S4.1. **Proof of Lemma 1**. Our estimator is given by Eq. (3.4) and we note that since $\widehat{d}_m^{(h,l)}$ are independent for all h and l, it follows that $\xi_1, \ldots, \xi_{\lambda}$ are independent. The following lemma is useful for the proof. Its proof is straightforward and omitted.

Lemma S7. Suppose that $X \sim \text{Pois}(1)$ and that $A < \infty$. Then

- (i) $E_X[A^X] = \exp(A 1)$.
- (ii) $V_X[A^X] = \exp(-1)(\exp(A^2) \exp(2A 1)).$

Proof of Lemma 1. Proof of Part (i). By the law of iterated expectations,

$$E[\xi] = E_{\mathcal{X}}[E_{\widehat{d}|\mathcal{X}}[\xi|\mathcal{X}]] = \exp(a/\lambda + 1)E_{\mathcal{X}}\left[\left(\frac{d-a}{\lambda}\right)^{\mathcal{X}}\right]$$
$$= \exp(d/\lambda),$$

where the last equality follows from part (i) of Lemma S7. Since the ξ_l are independent for $l = 1, ..., \lambda$,

$$E[\widehat{L}_B] = \exp(q) \prod_{l=1}^{\lambda} E[\xi_l] = \exp(q+d) = L.$$

Proof of Part (ii) By Eq. (3.4),

$$\{\widehat{d}_m^{(h,l)} \ge a, \, \forall h, l\} \subset \{\widehat{L}_B \ge 0\}.$$

Hence $\Pr(\widehat{L}_B \geq 0) \geq \Pr(\widehat{d}_m^{(h,l)} \geq a, \forall h, l) = 1.$

Proof of Part (iii) Since $E[|\widehat{L}_B|] \ge |E[\widehat{L}_B]|$ we obtain

$$V[|\widehat{L}_{B}|] = E[|\widehat{L}_{B}|^{2}] - E[|\widehat{L}_{B}|]^{2}$$

$$\leq E[\widehat{L}_{B}^{2}] - E[\widehat{L}_{B}]^{2}$$

$$= V[\widehat{L}_{B}].$$

We now derive $V[\widehat{L}_B]$ and show that it is finite which proves the result. By the law of total variance

(S1)
$$V[\xi] = E_{\mathcal{X}}[V_{\widehat{d}|\mathcal{X}}[\xi|\mathcal{X}]] + V_{\mathcal{X}}[E_{\widehat{d}|\mathcal{X}}[\xi|\mathcal{X}]].$$

To compute $V_{\widehat{d}|\mathcal{X}}[\xi|\mathcal{X}]$, note that for a collection of independent random variables $X_1, \dots X_J$,

(S2)
$$V\left[\prod_{j=1}^{J} X_{j}\right] = \prod_{j=1}^{J} \left(V[X_{j}] + E[X_{j}]^{2}\right) - \prod_{j=1}^{J} E[X_{j}]^{2}.$$

Hence,

$$V_{\widehat{d}|\mathcal{X}}[\xi|\mathcal{X}] = \exp\left(2\frac{a+\lambda}{\lambda}\right) \left(\left(\frac{\sigma_{\widehat{d}}^2}{\lambda^2} + \frac{(d-a)^2}{\lambda^2}\right)^{\mathcal{X}} - \left(\frac{(d-a)^2}{\lambda^2}\right)^{\mathcal{X}}\right)$$

and taking the outer expectations and applying Lemma S7,

$$E_{\mathcal{X}}[V_{\widehat{d}|\mathcal{X}}[\xi|\mathcal{X}]] = \exp\left(\frac{2a}{\lambda} + \frac{(d-a)^2}{\lambda^2} + 1\right) \left(\exp\left(\frac{\sigma_{\widehat{d}}^2}{\lambda^2}\right) - 1\right).$$

Next,

$$V_{\mathcal{X}}[E_{\widehat{d}|\mathcal{X}}[\xi|\mathcal{X}]] = \exp\left(2\frac{a+\lambda}{\lambda}\right)V_{\mathcal{X}}\left[\left(\frac{d-a}{\lambda}\right)^{\mathcal{X}}\right]$$
$$= \exp\left(\frac{2a}{\lambda}+1\right)\left(\exp\left(\frac{(d-a)^2}{\lambda^2}\right)-\exp\left(2\frac{d-a}{\lambda}-1\right)\right),$$

by part (ii) of Lemma S7 which, by Eq. (S1) and simplification, yields

$$V[\xi] = \exp\left(\frac{1}{\lambda^2} \left(\sigma_{\widehat{d}}^2 + (d-a)^2\right) + \frac{2a}{\lambda} + 1\right) - \exp\left(\frac{2d}{\lambda}\right).$$

To compute $V[\widehat{L}_B] = \exp(2q)V\left[\prod_{l=1}^{\lambda} \xi_l\right]$, we use Eq. (S2) with $E[\xi]^2 = \exp(2d/\lambda)$ to obtain

$$V\left[\prod_{l=1}^{\lambda} \xi_{l}\right] = \exp\left(\sum_{l=1}^{\lambda} \frac{1}{\lambda^{2}} \left(\sigma_{\widehat{d}}^{2} + (d-a)^{2}\right) + \frac{2a}{\lambda} + 1\right) - \exp\left(\sum_{l=1}^{\lambda} \frac{2d}{\lambda}\right)$$
$$= \exp\left(\frac{1}{\lambda} \left(\sigma_{\widehat{d}}^{2} + (d-a)^{2}\right) + 2a + \lambda\right) - \exp\left(2d\right).$$

We note that this expression exists as long as $\sigma_{\hat{d}}^2 < \infty$.

Proof of Part (iv). To minimize $V[\hat{L}_B]$ above for a fixed $\lambda > 0$, it is sufficient to minimize the exponent

$$f(a) = \frac{1}{\lambda} \left(\sigma_{\widehat{d}}^2 + (d-a)^2 + 2a + \lambda \right).$$

Since $f'(a) = -2(d-a)/\lambda + 2$ and $f''(a) = 2/\lambda > 0$, $a = d - \lambda$ is the minimum.

Proof of Part (v). This follows from the expression of the variance $V[\widehat{L}_B]$ derived in the proof of part (iii) of Lemma 1 and that of Eq. (3.4) derived in Papaspiliopoulos (2009).

Proof of Lemma 2. Follows straightforwardly from the more general Lemma S5, which is proved in Appendix S4.3. \Box

Proof of Lemma 3. We first note that

$$\Pr\left(\widehat{L}_B \ge 0\right) = \Pr\left(\prod_{l=1}^{\lambda} \xi_l \ge 0\right)$$

and that $\{\hat{L}_B < 0\}$ can only occur if there is an odd number of negative terms ξ_l , $l = 1, \ldots \lambda$. (Feller, 2008, p. 277) gives an expression for the probability of an odd number of negatives terms among a total of λ terms that depends on the probability of a single ξ_l being negative, i.e. $\Pr(\xi_l < 0)$. From this we obtain

$$\Pr(\widehat{L}_B \ge 0) = \frac{1}{2} \left(1 + (1 - 2\Pr(\xi_l < 0))^{\lambda} \right).$$

Notice that

$$\Pr(\xi_l < 0) = \sum_{j=1}^{\infty} \Pr\left(\prod_{l=1}^{j} A_m^{(l)} < 0\right) \Pr(\mathcal{X}_l = j), \quad \mathcal{X}_l \sim \operatorname{Pois}(1),$$

and we can again apply the result in Feller (2008) to obtain

$$\Pr\left(\prod_{l=1}^{j} A_m^{(l)} < 0\right) = \frac{1}{2} \left(1 - \left(1 - 2\Pr(\xi_l < 0)\right)^j\right),\,$$

which concludes the proof.

Proof of Lemma 4. We need to show that $p_U(du)q_U(du'|u) = p_U(du')q_U(du|u')$. Now,

$$p_U(du)q_U(du'|u) = \prod_{k=1}^G p_{U_k}(du_k) \frac{1}{G} \sum_{i=1}^G p_U(u_i) \prod_{j \neq i} \delta_{u_j}(du'_j)$$
$$= \prod_{k=1}^G p_{U_k}(du'_k) \frac{1}{G} \sum_{i=1}^G p_U(u'_i) \prod_{j \neq i} \delta_{u'_j}(du_j)$$

because $g(du')\delta_{u'}(du) = g(du)\delta_u(du')$ for any measure $g(\cdot)$.

Proof of Theorem 1. Part (i). Reversibility follows because we are dealing with a MH sampler.

Part (ii). The proof is essentially that of Theorem 1 of Andrieu and Roberts (2009), but under slightly different conditions. Consider first the case G = 1. Let $\mathcal{B}(\Theta)$ be the Borel sets of Θ and $\mathcal{B}(\mathcal{U})$ the Borel sets of \mathcal{U} . We will first show that if Algorithm 2 can reach the set $A \in \mathcal{B}(\Theta)$ from $\theta \in \Theta$ in one step, i.e. $P_{\Theta}(\theta; A) > 0$, then $\overline{P}_{\Theta,\mathcal{U}}(\theta, u; A \times B) > 0$ for any $u \in \mathcal{U}$ and $B \in \mathcal{B}(\mathcal{U})$ with $P_{\mathcal{U}}(B) > 0$, where $\overline{P}_{\Theta,\mathcal{U}}$ is the transition kernel of Algorithm 1 and is given by

$$\overline{P}_{\Theta,U}(\theta,u;d\theta',du') = \overline{K}_{\Theta,U}(\theta,u;d\theta',du') + \delta_{\theta,u}(d\theta',du') \left(1 - \int \overline{K}_{\Theta,U}(\theta,u;d\theta',du')\right),$$

with $\overline{K}_{\Theta,U}(\theta, u; d\theta', du') = \alpha_{\Theta,U}(\theta, u, \theta', u')q_{\Theta}(\theta; d\theta')q_{U}(u; du')$.

Let $\alpha_Z(z,z') = 1 \wedge \exp(z'-z)$. We first note that $r_{\Theta,U}(\theta,u;\theta',u') = e^{z'-z}r_{\Theta}(\theta;\theta')$, and that $1 \wedge (xy) \geq (1 \wedge x)(1 \wedge y)$. Hence, $\alpha_{\Theta,U}(\theta,u,\theta',u') \geq \alpha_{\Theta}(\theta,\theta')\alpha_Z(z,z')$, so that $\overline{K}_{\Theta,U}(\theta,u;d\theta',du') \geq K_{\Theta}(\theta;d\theta')\alpha_Z(z,z')q_U(u;du')$ and that $\overline{P}_{\Theta,U}(\theta,u;d\theta',du') \geq \overline{K}_{\Theta,U}(\theta,u;d\theta',du')$. Thus, if $\overline{P}_{\Theta,U}(\theta,u;A \times B) = 0$, then $K_{\Theta}(\theta;d\theta')\alpha_Z(z,z')q_U(u;du') = 0$ almost everywhere $\theta' \in A$ as $\alpha_Z(z,z')q_U(u;du') > 0$ by Part (i) of Assumption 1. But this contradicts Part (ii) of Assumption 1. This proves the one step result. Now we can similarly show by induction that if $P_{\Theta}^i(\theta;A) > 0$ for $i = 1, \ldots, k$ implies that $\overline{P}_{\Theta,U}(\theta,u;A \times B) > 0$ for any $u \in \mathcal{U}$ and $B \in \mathcal{B}(\mathcal{U})$ with $P_U(B) > 0$, then the same holds true for i = k + 1. This completes the proof for G = 1.

We now consider the G=2 case. We will show that for $k \geq 2$, if $P_{\Theta}^k(\theta;A) > 0$ then $\overline{P}_{\Theta,U}^k(\theta,u_{1:2};A\times(B_1\times B_2))>0$ for $u_{1:2}\in \overline{\mathcal{U}}^2$ with $p_U(B_1)>0$ and $p_U(B_2)>0$. Let $\widetilde{B}_{-2}=B_1\times\{u_2\}$ and $\widetilde{B}_{-1}=\{u_1\}\times B_2$. Then, we can show similarly to the G=1 case that if $P_{\Theta}(\theta;A)>0$, then $\overline{P}_{\Theta,U}(\theta,u_{1:2};A\times(\widetilde{B}_{-1}\cup\widetilde{B}_{-2}))>0$. The result for $k\geq 2$ follows as in the G=1 case.

We can similarly obtain the result that for a general G. If $k \geq G$ and $P_{\Theta}^k(\theta; A) > 0$ then $\overline{P}_{\Theta,U}^k(\theta, u_{1:G}; A \times (B_1 \times B_2 \times \cdots \times B_G)) > 0$, where $B_i \in \mathcal{B}(\mathcal{U})$ with $p_U(B_i) > 0$ for all $i = 1, \ldots, G$. Suppose that the result is true for $G = 1, \ldots, g$. Now define $\widetilde{B}_{-i} = B_1 \times B_2 \times B_{i-1} \times \{u_i\} \times B_{i+1} \times \cdots \times B_{g+1}$. Then, by the induction hypothesis, if $P_{\Theta}^g(\theta; A) > 0$, then $\overline{P}_{\Theta,U}^g(\theta, u_{1:g+1}; A \times (B_{-1} \cup \cdots \cup B_{-g-1}) > 0$ assuming that $p_U(B_i) > 0$ for $i = 1, \ldots, g+1$. The required result now follows as in the G = 1 case.

The irreducibility and aperiodicity of the sampling scheme now follows from that of Algorithm 2.

Part (ii) follows from the strong law of large numbers for Markov chain, see e.g. Meyn and Tweedie (2012, Theorem 17.0.1).

To prove Part (iii), we first consider the numerator of Eq. (4.5). By Theorem 27 in Roberts and Rosenthal (2004),

$$\sqrt{N} \Big(N^{-1} \sum_{i=1}^{N} S_i \psi_i - \mathcal{E}_{\overline{\pi}}(S\psi) \Big) \to N \Big(0, \mathcal{V}_{\overline{\pi}}(\psi S) \mathrm{IF}_{\overline{\pi}, \psi S} \Big).$$

Next, by the strong law of large numbers for Markov chains $N^{-1} \sum_{i=1}^{N} S_i \to E_{\overline{\pi}}(S) \overline{\pi}$ -almost surely, i.e. the denominator of Eq. (4.5) converges to $E_{\overline{\pi}}(S) \neq 0$. The result now follows from Slutsky's theorem.

Proof of Lemma 5. Using the notation in Eq. (S1), it is straightforward to show that

$$\overline{\pi}(du_{1:G}|\theta) = \prod_{i=1}^{G} \exp\left(\widetilde{z}_i(u_i, \theta)\right) p_U(u_i).$$

Proof of Lemma S2. By Assumption S1 and $\overline{\pi}(d\theta, du) = \nu(\theta)\overline{\pi}(du|\theta)$,

$$\overline{\pi}(\theta, dv, dz, dw) = \overline{\pi}(d\theta, du)|J(\theta, u)| = \exp(z)C(\theta)p(d\theta, dv, dz, dw).$$

Hence, $\overline{\pi}(d\theta, dv, dz) = \exp(z)C(\theta)p(d\theta, dv, dz)$ and

$$\overline{\pi}(d\theta, ds, dz) = \exp(z)C(\theta) \int_{s=\operatorname{sign}(v)} p(d\theta, dv, dz)$$
$$= \exp(z)\nu(\theta)p(ds|\theta)p(dz|\theta),$$

with
$$p(ds|\theta) := \int_{s=\text{sign}(v)} p(dv|\theta)$$
.

Proof of Lemma S3. Part (i) follows because

$$\frac{\overline{\pi}(d\theta',ds',dz')}{\overline{\pi}(d\theta,ds,dz)}\frac{\nu(d\theta)p(ds|\theta)q(z';dz|\sigma^2,\rho)}{\nu(d\theta')p(ds'|\theta')q(z;dz'|\sigma^2,\rho)},$$

where the perfect proposals for θ and s cancel the corresponding terms in $\overline{\pi}$. Moreover, we can show that

$$p(dz)q(z;dz'|\sigma^2,\rho) = p(dz')q(z';dz|\sigma^2,\rho).$$

Part (ii) follows from Part (i) and the fact that $q(z; dz'|\sigma^2, \rho)$ does not depend on θ and s. Part (iii) then follows from Lemma 4 in Pitt et al. (2012), but with the probability of accepting a proposal conditional on z which arises from the correlated proposal.

Proof of Lemma S4. $E[Z'] = (G-2)\sigma^2/2G$ since, under Assumption S3,

$$\sum_{i \neq j, i=1}^{G} Z_i(\theta', U_i) \sim \mathcal{N}\left(\frac{(G-1)\sigma^2}{2G}, \frac{(G-1)\sigma^2}{G}\right) \quad \text{and} \quad Z_j(\theta', U_j') \sim \mathcal{N}\left(-\frac{\sigma^2}{2G}, \frac{\sigma^2}{G}\right).$$

It also follows that $V[Z'] = \sigma^2$. Moreover, $Z \sim \mathcal{N}(\sigma^2/2, \sigma^2)$, which concludes the proof. \square

S4.3. **Proofs for Appendix S2.** To prove Lemma S5, we first need the preliminary Lemmas S8 to S10.

Lemma S8. (Non-central χ^2 is a Poisson mixture of central χ^2). If $J \sim \text{Pois}(\mu/2)$ and $W|J \sim \chi^2(k+2J)$, then marginally $W \sim \chi^2(k,\mu)$.

Proof. See Walck (1996).

Lemma S9. (Moments of log central χ^2 .(Pav, 2015)). If $W \sim \chi^2(k)$ and $Y = \log W$, then $EY = \log 2 + \psi(k/2)$ and $VY = \psi^{(1)}(k/2)$.

Lemma S10. (Moments of log non-central χ^2). If $W \sim \chi^2(k,\mu)$ and $Y = \log W$, then

$$E[Y] = \log 2 + E_J \left(\psi^{(0)}(k/2 + J) \right)$$

$$V[Y] = E_J \left[\psi^{(1)}(k/2 + J) \right] + V_J \left[\psi^{(0)}(k/2 + J) \right],$$

where $J \sim \text{Pois}(\mu/2)$ and $\psi^{(q)}$ is the polygamma function of order q.

Proof. Follows from Pav (2015). From the mixture representation in Lemma S8 we know that we can represent $W \sim \chi^2(k,\mu)$ as $J \sim \text{Pois}(\mu/2)$ and $W|J \sim \chi^2(k+2J)$. By the law of iterated expectations and Lemma S9

$$E[Y] = E_J[E_{W|J}[Y]] = \log 2 + E_J[\psi^{(0)}(k/2 + J)].$$

Also, from the law of total variance and Lemma S9

$$V[Y] = E_J[V_{W|J}[Y]] + V_J[E_{W|J}[Y]] = E_J[\psi^{(1)}(k/2 + J)] + V_J[\psi^{(0)}(k/2 + J)].$$

Proof of Lemma S5. When $a = d - \lambda$ we have

$$\log \left| \widehat{L}_B \right| = q + d + \sum_{l=1}^{\lambda} \sum_{h=1}^{\mathcal{X}_l} \log \left(\left| \frac{\widehat{d}_m^{(h,l)} - d}{\lambda} + 1 \right| \right) = q + d + \sum_{l=1}^{\lambda} \sum_{h=1}^{\mathcal{X}_l} \log \left(\left| \frac{\sqrt{\frac{\gamma}{m}} \overline{d}_m^{(h,l)}}{\lambda} + 1 \right| \right),$$

where $\mathcal{X}_l \sim \text{Pois}(1)$, $l=1,...,\lambda$. Let $I^{(h,l)} \in \{1,...,C\}$ be indicators such that $I^{(h,l)}=c$ means that observation $\bar{d}_m^{(h,l)}$ comes from the cth mixture component $N(\mu_c, \sigma_c^2)$ with $\Pr(I^{(h,l)}=c)=\omega_c$. Now, since the $\bar{d}_m^{(h,l)}$ are iid and the total number of $\bar{d}_m^{(h,l)}$ is $\sum_{l=1}^{\lambda} \mathcal{X}_l$ we have

$$V\left(\log\left|\widehat{L}_{B}\right||\mathcal{X}_{1:\lambda}\right) = \left(\sum_{l=1}^{\lambda} \mathcal{X}_{l}\right) V\log\left(\left|\frac{\sqrt{\frac{\gamma}{m}} \bar{d}_{m}^{(h,l)}}{\lambda} + 1\right|\right).$$

Define

$$X^{(h,l)} = \frac{\sqrt{\frac{\gamma}{m}} \left(\bar{d}_m^{(h,l)} - \mu_{I^{(h,l)}} \right)}{\lambda},$$

and note that

$$X^{(h,l)}|\left(I^{(h,l)}=c\right)\sim N\left(0,\tilde{\sigma}_c^2\right),$$

where $\tilde{\sigma}_c^2 = \frac{\sigma_c^2}{\lambda^2} \frac{\gamma}{m}$. Now, conditional on $I^{(h,l)} = c$, we have

$$\begin{split} \log\left(\left|\frac{\sqrt{\frac{\gamma}{m}}\overline{d}_{m}^{(h,l)}}{\lambda}+1\right|\right) &= \log\left(\left|X^{(h,l)}+\frac{\sqrt{\frac{\gamma}{m}}\mu_{c}+\lambda}{\lambda}\right|\right) \\ &\stackrel{d}{=} \log\left(\left|\tilde{\sigma}_{c}Z+\frac{\sqrt{\frac{\gamma}{m}}\mu_{c}+\lambda}{\lambda}\right|\right), \text{ where } Z \sim N(0,1) \\ &= \log\tilde{\sigma}_{c}+\log\left(\left|Z+\frac{\sqrt{\frac{\gamma}{m}}\mu_{c}+\lambda}{\lambda\tilde{\sigma}_{c}}\right|\right) \\ &= \log\left(\frac{\sigma_{c}}{\lambda}\sqrt{\frac{\gamma}{m}}\right) + \log\left(\left|Z+\frac{\mu_{c}+\sqrt{\frac{m}{\gamma}}\lambda}{\sigma_{c}}\right|\right) \\ &= \log\left(\frac{\sigma_{c}}{\lambda}\sqrt{\frac{\gamma}{m}}\right) + \frac{1}{2}\log\left(\left(Z+\frac{\mu_{c}+\sqrt{\frac{m}{\gamma}}\lambda}{\sigma_{c}}\right)^{2}\right) \\ &\stackrel{d}{=} \log\left(\frac{\sigma_{c}}{\lambda}\sqrt{\frac{\gamma}{m}}\right) + \frac{1}{2}\log\left(W^{(h,l)}\right), \\ &\text{where } W^{(h,l)} \sim \chi^{2}\left(1,\frac{(\mu_{c}+\sqrt{\frac{m}{\gamma}}\lambda)^{2}}{\sigma_{c}^{2}}\right). \end{split}$$

where $\chi^2\left(k,\lambda\right)$ denotes the non-central χ^2 distribution with k degrees of freedom and non-centrality parameter λ . So $\log\left(\left|\frac{\sqrt{\frac{\gamma}{m}}d_m^{(h,l)}}{\lambda}+1\right|\right)$ is a mixture of log of non-central χ^2 variables with component means and variances given by Lemma S10

$$\eta_c := \operatorname{E} \log \left(\left| \frac{\sqrt{\frac{\gamma}{m}} \bar{d}_m^{(h,l)}}{\lambda} + 1 \right| |I^{(k,l)} = c \right) = \log \left(\frac{\sigma_c}{\lambda} \sqrt{\frac{\gamma}{m}} \right) + \frac{1}{2} \left(\log 2 + \operatorname{E}_{J_c} \left(\psi^{(0)} (1/2 + J_c) \right) \right)$$

and

$$\nu_c^2 := V \log \left(\left| \frac{\sqrt{\frac{\gamma}{m}} \bar{d}_m^{(h,l)}}{\lambda} + 1 \right| | I^{(k,l)} = c \right) = \frac{1}{4} \left[E_{J_c} \left(\psi^{(1)} (1/2 + J_c) \right) + V_{J_c} \left(\psi^{(0)} (1/2 + J_c) \right) \right]$$

where the J_c are independent and $J_c \sim \text{Pois}\left(\frac{(\mu_c + \sqrt{\frac{m}{\gamma}}\lambda)^2}{2\sigma_c^2}\right)$. By the mean and variance of a finite mixture we then have

$$\eta := \operatorname{E} \log \left(\left| \frac{\sqrt{\frac{\gamma}{m}} \bar{d}_m^{(h,l)}}{\lambda} + 1 \right| \right) = \sum_{c=1}^C \omega_c \eta_c$$

$$\operatorname{V} \log \left(\left| \frac{\sqrt{\frac{\gamma}{m}} \bar{d}_m^{(h,l)}}{\lambda} + 1 \right| \right) = \sum_{c=1}^C \omega_c (\nu_c^2 + (\eta_c - \eta)^2).$$

Finally,

$$V\left(\log\left|\widehat{L}_{B}\right|\right) = E_{\mathcal{X}_{1:\lambda}}V\left(\log\left|\widehat{L}_{B}\right||\mathcal{X}_{1:\lambda}\right) + V_{\mathcal{X}_{1:\lambda}}E\left(\log\left|\widehat{L}_{B}\right||\mathcal{X}_{1:\lambda}\right)$$

$$= E_{\mathcal{X}_{1:\lambda}}\left[\left(\sum_{l=1}^{\lambda} \mathcal{X}_{l}\right)\sum_{c=1}^{C} \omega_{c}(\nu_{c}^{2} + (\eta_{c} - \eta)^{2})\right] + V_{\mathcal{X}_{1:\lambda}}\left(\left(\sum_{l=1}^{\lambda} \mathcal{X}_{l}\right)\eta\right)$$

$$= \lambda \sum_{c=1}^{C} \omega_{c}(\nu_{c}^{2} + (\eta_{c} - \eta)^{2}) + \lambda \eta^{2}.$$

 $V\left(\log\left|\widehat{L}_{B}\right|\right)$ is finite since $\psi^{(1)}(1/2+J)\leq\pi^{2}/2$ for all $J\geq0$ and for all $J\geq0$

$$\left(\psi^{(0)}(1/2+J)\right)^2 = \left(\psi^{(0)}(1) - 2\log 2 + 2\sum_{k=1}^J \frac{1}{2k-1}\right)^2 < \left(\psi^{(0)}(1) - 2\log 2 + 2J\right)^2$$

and the Poisson has finite first and second moments.

S4.4. Proofs for Appendix C.

Proof of Lemma S6. The result follows from numerically optimizing the expression

$$CT(\sigma^2, \rho) := \frac{IF(\sigma^2, \rho)}{\sigma^2},$$

with IF in Part (iii) of Eq. (S3).