Scalable Metropolis–Hastings for Exact Bayesian Inference with Large Datasets

Rob Cornish [1] **Paul Vanetti** [1] **Alexandre Bouchard-Côté** [2] **George Deligiannidis** [1 3] **Arnaud Doucet** [1 3]

Abstract

Bayesian inference via standard Markov Chain Monte Carlo (MCMC) methods is too computationally intensive to handle large datasets, since the cost per step usually scales like $\Theta(n)$ in the number of data points n. We propose the *Scalable Metropolis–Hastings* (SMH) kernel that exploits Gaussian concentration of the posterior to require processing on average only $O(1)$ or even $O(1/\sqrt{n})$ data points per step. This scheme is based on a combination of factorized acceptance probabilities, procedures for fast simulation of Bernoulli processes, and control variate ideas. Contrary to many MCMC subsampling schemes such as fixed step-size Stochastic Gradient Langevin Dynamics, our approach is exact insofar as the invariant distribution is the true posterior and not an approximation to it. We characterise the performance of our algorithm theoretically, and give realistic and verifiable conditions under which it is geometrically ergodic. This theory is borne out by empirical results that demonstrate overall performance benefits over standard Metropolis–Hastings and various subsampling algorithms.

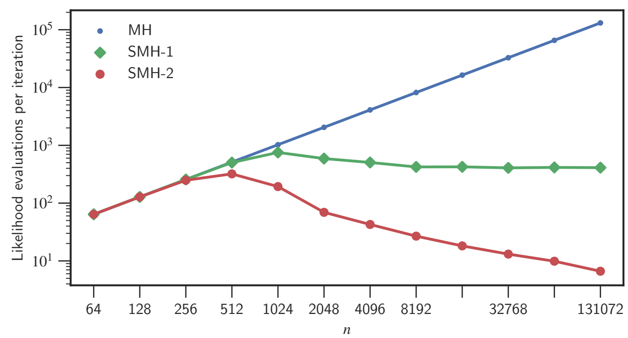

Figure 1. Average number of likelihood evaluations per iteration required by SMH for a 10-dimensional logistic regression posterior as the number of data points n increases. SMH-1 uses a first-order approximation to the target and SMH-2 a second-order one.

1. Introduction

Bayesian inference is concerned with the posterior distribution $p(\theta|y_{1:n})$, where $\theta \in \Theta = \mathbb{R}^d$ denotes parameters of interest and $y_{1:n} = (y_1, \cdots, y_n) \in \mathcal{Y}^n$ are observed data. We assume the prior admits a Lebesgue density $p(\theta)$ and that the data are conditionally independent given θ with likelihoods $p(y_i|\theta)$, which means

$$p(\theta|y_{1:n}) \propto p(\theta) \prod_{i=1}^{n} p(y_i|\theta).$$

In most cases of interest, $p(\theta|y_{1:n})$ does not admit a closed-form expression and so we must resort to a Markov Chain Monte Carlo (MCMC) approach. However, standard MCMC schemes can become very computationally expensive for large datasets. For example, the Metropolis–Hastings (MH) algorithm requires computing a likelihood ratio $p(y_{1:n}|\theta')/p(y_{1:n}|\theta)$ at each iteration. A direct implementation of this algorithm thus requires computational cost $\Theta(n)$ per step, which is prohibitive for large n.

Many ideas for mitigating this cost have been suggested; see Bardenet et al. (2017) for a recent review. Broadly speaking these approaches are distinguished by whether they exactly preserve the true posterior as the invariant distribution of the Markov chain produced. Approximate methods that have been proposed include *divide-and-conquer* schemes, which run parallel MCMC chains on a partition of the data (Neiswanger et al., 2013; Scott et al., 2016). Other approaches replace the likelihood ratio in MH with an approximation computed from a subsample of observations. The error introduced can be controlled heuristically using central limit theorem approximations (Korattikara et al., 2014) or rigorously via concentration inequalities (Bardenet et al., 2014; 2017; Quiroz et al., 2018a). Another popular class of schemes is based on Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011; Dubey et al., 2016; Baker et al., 2018; Brosse et al., 2018; Chatterji et al., 2018),

[1]University of Oxford, Oxford, United Kingdom [2]University of British Columbia, Vancouver, Canada [3]The Alan Turing Institute, London, United Kingdom. Correspondence to: Rob Cornish <rcornish@robots.ox.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

which is a time-discretized Langevin dynamics where the gradient of the log-likelihood is approximated by subsampling. SGLD is usually implemented using a fixed step-size discretization, which does not exactly preserve the posterior distribution. Finally, Quiroz et al. (2016) and Dang et al. (2017) propose schemes that do not preserve the posterior exactly, but yield consistent estimates of posterior expectations after an importance sampling correction.

In addition to these approximate methods, several MCMC methods exist that do preserve the target as invariant distribution while only requiring access to a subset of the data at each iteration. However, various restrictions of these approaches have so far limited their widespread use. Firefly Monte Carlo (Maclaurin & Adams, 2014) considers an extended target that can be evaluated using a subset of the data at each iteration, but requires the user specify global lower bounds to the likelihood factors that can be difficult to derive. It is as yet also unclear what the convergence properties of this scheme are. Delayed acceptance schemes have been proposed based on a factorized version of the MH acceptance probability (Banterle et al., 2015) and on a random subsample of the data (Payne & Mallick, 2018). These methods allow rejecting a proposal without computing every likelihood term, but still require evaluating each term in order to accept. Quiroz et al. (2018b) combine the latter with the approximate subsampling approach of Quiroz et al. (2018a) to mitigate this problem. Finally, various non-reversible continuous-time MCMC schemes based on Piecewise Deterministic Markov Processes have been proposed which, when applied to large-scale datasets (Bouchard-Côté et al., 2018; Bierkens et al., 2019), only require evaluating the gradient of the log-likelihood for a subset of the data. However, these schemes can be difficult to understand theoretically, falling outside the scope of existing geometric ergodicity results, and can be challenging to implement.

In this paper we present a novel MH-type subsampling scheme that exactly preserves the posterior as the invariant distribution while still enjoying attractive theoretical properties and being straightforward to implement and tune. We make use of a combination of a factorized MH acceptance probability (Ceperley, 1995; Christen & Fox, 2005; Banterle et al., 2015; Michel et al., 2019; Vanetti et al., 2018) and fast methods for sampling non-homogeneous Bernoulli processes (Shanthikumar, 1985; Devroye, 1986; Fukui & Todo, 2009; Michel et al., 2019; Vanetti et al., 2018) to allow iterating without computing every likelihood factor. The combination of these ideas has proven useful for some physics models (Michel et al., 2019), but a naïve application is not efficient for large-scale Bayesian inference. Our contribution here is an MH-style MCMC kernel that realises the potential computational benefits of this method in the Bayesian setting. We refer to this kernel as *Scalable Metropolis-Hastings* (SMH) and, in addition to empirical results, provide a rigorous theoretical analysis of its behaviour under realistic and verifiable assumptions. In particular, we show SMH requires on average only $O(1)$ or even $O(1/\sqrt{n})$ cost per step as illustrated in Figure 1, has a non-vanishing average acceptance probability in the stationary regime, and is geometrically ergodic under mild conditions.

Key to our approach is the use of *control variate* ideas, which allow us to exploit the concentration around the mode frequently observed for posterior distributions with large datasets. Control variate ideas based on posterior concentration have been used successfully for large-scale Bayesian analysis in numerous recent contributions (Dubey et al., 2016; Bardenet et al., 2017; Baker et al., 2018; Brosse et al., 2018; Bierkens et al., 2019; Chatterji et al., 2018; Quiroz et al., 2018a). In our setting, this may be understood as making use of a computationally cheap approximation of the posterior.

The Supplement contains all our proofs as well as a guide to our notation in Section A.

2. Factorised Metropolis-Hastings

We first review the use of a factorised acceptance probability inside an MH-style algorithm. For now we assume a generic target $\pi(\theta)$ before specialising to the Bayesian setting below.

2.1. Transition Kernel

Assume our target $\pi(\theta)$ and proposal $q(\theta, \theta')$ factorise like

$$\pi(\theta) \propto \prod_{i=1}^{m} \pi_i(\theta) \qquad q(\theta, \theta') \propto \prod_{i=1}^{m} q_i(\theta, \theta')$$

for some $m \geq 1$ and some choice of non-negative functions π_i and q_i. These factors are not themselves required to be integrable; for instance, we may take any $\pi_i, q_i \equiv 1$. Define the *Factorised Metropolis-Hastings* (FMH) kernel

$$P_{\mathrm{FMH}}(\theta, A) := \left(1 - \int q(\theta, \theta')\alpha_{\mathrm{FMH}}(\theta, \theta')d\theta'\right)\mathbb{I}_A(\theta) + \int_A q(\theta, \theta')\alpha_{\mathrm{FMH}}(\theta, \theta')d\theta', \quad (1)$$

where $\theta \in \Theta$, $A \subseteq \Theta$ is measurable, and the FMH acceptance probability is defined

$$\alpha_{\mathrm{FMH}}(\theta, \theta') := \prod_{i=1}^{m} \underbrace{1 \wedge \frac{\pi_i(\theta')q_i(\theta', \theta)}{\pi_i(\theta)q_i(\theta, \theta')}}_{=:\alpha_{\mathrm{FMH}\,i}(\theta, \theta')}. \quad (2)$$

It is straightforward and well-known that P_{FMH} is π-reversible; see Section B.1 in the Supplement for a proof. Factorised acceptance probabilities have appeared numerous times in the literature and date back at least to (Ceperley, 1995). The MH acceptance probability α_{MH} and kernel P_{MH} correspond to α_{FMH} and P_{FMH} when $m = 1$.

2.2. Poisson Subsampling Implementation

The acceptance step of P_{FMH} can be implemented by sampling directly m independent Bernoulli trials with success probability $1 - \alpha_{\mathrm{FMH}i}$, and returning θ' if every trial is a failure. Since we can reject θ' as soon as a single success occurs, this allows us potentially to reject θ' without computing each factor at each iteration (Christen & Fox, 2005; Banterle et al., 2015).

However, although this can lead to efficiency gains in some contexts, it remains of limited applicability for Bayesian inference with large datasets since we are still forced to compute every factor whenever we accept a proposal. It was realized independently by Michel et al. (2019) and Vanetti et al. (2018) that if one has access to lower bounds on $\alpha_{\mathrm{FMH}i}(\theta,\theta')$, hence to an upper bound on $1-\alpha_{\mathrm{FMH}i}(\theta,\theta')$, then techniques for fast simulation of Bernoulli random variables can be used that potentially avoid this problem. One such technique is given by the discrete-time thinning algorithms introduced in (Shanthikumar, 1985); see also (Devroye, 1986, Chapter VI Sections 3.3-3.4). This is used in (Michel et al., 2019).

We use here an original variation of a scheme developed in (Fukui & Todo, 2009). Denote

$$\lambda_i(\theta,\theta') := -\log \alpha_{\mathrm{FMH}i}(\theta,\theta'),$$

and assume we have the bounds

$$\lambda_i(\theta,\theta') \leq \varphi(\theta,\theta')\psi_i := \overline{\lambda}_i(\theta,\theta') \tag{3}$$

for nonnegative φ, ψ_i. This condition holds for a variety of statistical models: for instance, if π_i is log-Lipschitz and q is symmetric with (say) $q_i = q^{1/m}$, then

$$\lambda_i(\theta,\theta') \leq K_i\|\theta-\theta'\|. \tag{4}$$

This case illustrates that (3) is usually a *local* constraint on the target and therefore not as strenuous as the global lower-bounds required by Firefly (Maclaurin & Adams, 2014). We exploit this to provide a methodology for producing φ and ψ mechanically when we consider Bayesian targets in Section 3. Letting $\overline{\lambda}(\theta,\theta') := \sum_{i=1}^m \overline{\lambda}_i(\theta,\theta')$, it follows that if

$$N \sim \mathrm{Poisson}\left(\overline{\lambda}(\theta,\theta')\right)$$

$$X_1,\cdots,X_N \overset{\text{iid}}{\sim} \mathrm{Categorical}((\overline{\lambda}_i(\theta,\theta')/\overline{\lambda}(\theta,\theta'))_{1\leq i\leq m})$$

$B_j \sim \mathrm{Bernoulli}(\lambda_{X_j}(\theta,\theta')/\overline{\lambda}_{X_j}(\theta,\theta'))$ independently for $1 \leq j \leq N$

then $\mathbb{P}(B=0) = \alpha_{\mathrm{FMH}}(\theta,\theta')$ where $B = \sum_{j=1}^N B_j$ (and $B = 0$ if $N = 0$). See Proposition C.1 in the Supplement for a proof. These steps may be interpreted as sampling a discrete Poisson point process with intensity $\lambda_i(\theta,\theta')$ on $i \in \{1,\cdots,m\}$ via thinning (Devroye, 1986). Thus, to perform the FMH acceptance step, we can simulate these B_j and check whether each is 0.

We may exploit (3) to sample each X_j and B_j in $O(1)$ time per MCMC step as $m \to \infty$ after paying some once-off setup costs. Note that

$$\overline{\lambda}(\theta,\theta') = \varphi(\theta,\theta')\sum_{i=1}^m \psi_i, \tag{5}$$

so that we may compute $\overline{\lambda}(\theta,\theta')$ in $O(1)$ time per iteration by simply evaluating $\varphi(\theta,\theta')$ if we pre-compute $\sum_{i=1}^m \psi_i$ ahead of our run. This incurs a one-time cost of $\Theta(m)$, but assuming our run is long enough this will be negligible overall. Similarly, note that

$$\frac{\overline{\lambda}_i(\theta,\theta')}{\overline{\lambda}(\theta,\theta')} = \frac{\psi_i}{\sum_{j=1}^m \psi_j},$$

so that $\mathrm{Categorical}((\overline{\lambda}_i(\theta,\theta')/\overline{\lambda}(\theta,\theta'))_{1\leq i\leq m})$ does not depend on θ,θ'. Thus, we can sample each X_i in $O(1)$ time using Walker's alias method (Walker, 1977; Kronmal & Peterson, 1979) having paid another once-off $\Theta(m)$ cost.

Algorithm 1 shows how to implement P_{FMH} using this approach. Observe that if $N < m$ we are guaranteed not to evaluate every target factor even if we accept the proposal θ'. Of course, since N is random, in general it is not obvious that $N \ll m$ will necessarily hold on average, and indeed this will not be so for a naïve factorisation. We show in Section 3 how to use Algorithm 1 as the basis of an efficient subsampling method for Bayesian inference.

In many cases we will not have bounds of the form (3) for every factor. However, Algorithm 1 can still be useful provided the computational cost for computing these extra factors is $O(1)$. In this case we can directly simulate a Bernoulli trial for each additional factor, which by assumption does not change the asymptotic complexity of this method.

2.3. Geometric Ergodicity

We consider now the theoretical implications of using P_{FMH} rather than P_{MH}. We refer the reader to Section B.2 in the Supplement for a review of the relevant definitions and theory of Markov chains. It is straightforward to show and well-known that the following holds.

Proposition 2.1. *For all* $\theta,\theta' \in \Theta$, $\alpha_{\mathrm{FMH}}(\theta,\theta') \leq \alpha_{\mathrm{MH}}(\theta,\theta')$.

See Section B in the Supplement for a proof. As such, we do not expect FMH to enjoy better convergence properties than MH. Indeed, Proposition 2.1 immediately entails that FMH produces ergodic averages of higher asymptotic variance

Algorithm 1 Efficient implementation of the FMH kernel. Setup() is called once prior to starting the MCMC run.

```
function Setup()
    Ψ ← Σ_{i=1}^m ψ_i
    τ ← AliasTable((ψ_i/Ψ)_{1≤i≤m})
end function

function FmhKernel(θ)
    θ' ~ q(θ, ·)
    N ~ Poisson(φ(θ, θ')Ψ)
    for j ∈ 1, ..., N do
        X_j ~ τ
        B_j ~ Bernoulli(λ_{X_j}(θ, θ')/λ̄_{X_j}(θ, θ'))
        if B_j = 1 then
            return θ
        end if
    end for
    return θ'
end function
```

than standard MH (Peskun, 1973; Tierney, 1998). Moreover P_{FMH} can fail to be geometrically ergodic even when P_{MH} is, as noticed by Banterle et al. (2015). Geometric ergodicity is a desirable property of MCMC algorithms because it ensures the central limit theorem holds for some ergodic averages (Roberts & Rosenthal, 1997, Corollary 2.1). The central limit theorem in turn is the foundation of principled stopping criteria based on Monte Carlo standard errors (Jones & Hobert, 2001).

To address the fact that P_{FMH} might not be geometrically ergodic, we introduce the *Truncated FMH* (TFMH) kernel P_{TFMH} which is obtained by simply replacing in (1) the term $\alpha_{\mathrm{FMH}}(\theta, \theta')$ with the acceptance probability

$$\alpha_{\mathrm{TFMH}}(\theta, \theta') := \begin{cases} \alpha_{\mathrm{FMH}}(\theta, \theta'), & \overline{\lambda}(\theta, \theta') < R \\ \alpha_{\mathrm{MH}}(\theta, \theta'), & \text{otherwise}, \end{cases} \quad (6)$$

for some choice of $R \in [0, \infty]$. Observe that FMH is a special case of TFMH with $R = \infty$. When $\overline{\lambda}(\theta, \theta')$ is symmetric in θ and θ', Proposition B.3 in the Supplement shows that P_{TFMH} is still π-reversible. The following theorem shows that under mild conditions TFMH inherits the desirable convergence properties of MH.

Theorem 2.1. *If P_{MH} is φ-irreducible, aperiodic, and geometrically ergodic, then P_{TFMH} is too if*

$$\delta := \inf_{\overline{\lambda}(\theta, \theta') < R} \alpha_{\mathrm{FMH}}(\theta, \theta') \vee \alpha_{\mathrm{FMH}}(\theta', \theta) > 0. \quad (7)$$

In this case, $\mathrm{Gap}(P_{\mathrm{FMH}}) \geq \delta \mathrm{Gap}(P_{\mathrm{MH}})$*, and for* $f \in L^2(\pi)$

$$\mathrm{var}(f, P_{\mathrm{TFMH}}) \leq (\delta^{-1} - 1)\mathrm{var}(f, \pi) + \delta^{-1}\mathrm{var}(f, P_{\mathrm{MH}}).$$

Here $\mathrm{Gap}(P)$ denotes the spectral gap and $\mathrm{var}(f, P)$ the asymptotic variance of the ergodic averages of f. See Section B.2 in the Supplement for full definitions and a proof. Proposition B.1 in the Supplement shows that $\alpha_{\mathrm{FMH}}(\theta, \theta') \vee \alpha_{\mathrm{FMH}}(\theta', \theta) = \alpha_{\mathrm{FMH}}(\theta, \theta')/\alpha_{\mathrm{MH}}(\theta, \theta')$, and hence (7) quantifies the worst-case cost we pay for using the FMH acceptance probability rather than the MH one. The condition (7) is easily seen to hold in the common case that each π_i is bounded away from 0 and ∞ on $\{\theta, \theta' \in \Theta \mid \overline{\lambda}(\theta, \theta') < R\}$, which is a fairly weak requirement when $R < \infty$.

Recall from the previous section that P_{FMH} requires computing $N \sim \mathrm{Poisson}(\overline{\lambda}(\theta, \theta'))$ factors for a given θ, θ'. In this way, TFMH yields the additional benefit of controlling the maximum expected number of factors we will need to compute via the choice of R. An obvious choice is to take $R = m$, which ensures we will not compute more factors for FMH than for MH on average. Thus, overall, TFMH yields the computational benefits of α_{FMH} when our bounds (3) are tight (usually near the mode), and otherwise falls back to MH as a default (usually in the tails).

3. FMH for Bayesian Big Data

We now consider the specific application of FMH to the problem of Bayesian inference for large datasets, where $\pi(\theta) \propto p(\theta) \prod_{i=1}^{n} p(y_i|\theta)$. It is frequently observed that such targets concentrate at a rate $1/\sqrt{n}$ around the mode as $n \to \infty$, in what is sometimes referred to as the Bernstein-von Mises phenomenon. We describe here how to leverage this phenomenon to devise an effective subsampling algorithm based on Algorithm 1. Our approach is based on control variate ideas similar to Dubey et al. (2016); Bardenet et al. (2017); Bierkens et al. (2019); Baker et al. (2018); Chatterji et al. (2018); Quiroz et al. (2018a). We emphasise that all these techniques also rely on a posterior concentration assumption but none of them only requires processing $O(1/\sqrt{n})$ data points per iteration as we do.

To see why this approach is needed, observe that the most natural factorisations of the posterior have $m \asymp n$. This introduces a major pitfall: each new factor introduced can only lower the value of $\alpha_{\mathrm{FMH}}(\theta, \theta')$, which in the aggregate can therefore mean $\alpha_{\mathrm{FMH}}(\theta, \theta') \to 0$ as $n \to \infty$.

Consider heuristically a naïve application of Algorithm 1 to π. Assuming a flat prior for simplicity, the obvious factorisation takes $m = n$ and each $\pi_i(\theta) = p(y_i|\theta)$. Suppose the likelihoods are log-Lipschitz and that we use the bounds (4) derived above. For smooth likelihoods, if the Lipschitz constants K_i are chosen minimally, these bounds will be tight in the limit as $\|\theta - \theta'\| \to 0$. Consequently, if we scale $\|\theta - \theta'\|$ as $1/\sqrt{n}$ to match the concentration of the target, then $\alpha_{\text{FMH}}(\theta, \theta') \asymp \exp(-\overline{\lambda}(\theta, \theta')) \to 0$ since

$$\overline{\lambda}(\theta, \theta') = \underbrace{\|\theta - \theta'\|}_{=\Theta(1/\sqrt{n})} \underbrace{\sum_{i=1}^{n} K_i}_{=\Theta(n)} = \Theta(\sqrt{n}).$$

Recall that Algorithm 1 requires the computation of at most $N \sim \text{Poisson}(\overline{\lambda}(\theta, \theta'))$ factors, and hence in this case we do obtain a reduced expected cost per iteration of $\Theta(\sqrt{n})$ as opposed to $\Theta(n)$. Nevertheless, we found empirically that the increased asymptotic variance produced by the decaying acceptance probability entails an overall loss of performance compared with standard MH. We could consider using a smaller stepsize such as $\|\theta - \theta'\| = O(1/n)$ which would give a stable acceptance probability, but then our proposal would not match the $1/\sqrt{n}$ concentration of the posterior. We again found this increases the asymptotic variance to the extent that it negates the benefits of subsampling overall.

3.1. Scalable Metropolis–Hastings

Our approach is based on controlling $\overline{\lambda}(\theta, \theta')$, which ensures both a low computational cost and a large acceptance probability. We assume an initial factorisation

$$\pi(\theta) \propto p(\theta) \prod_{i=1}^{n} p(y_i|\theta) \propto \prod_{i=1}^{m} \widetilde{\pi}_i(\theta) \tag{8}$$

for some m (not necessarily equal to n) and $\widetilde{\pi}_i$ (e.g. using directly the factorisation of prior and likelihoods). Let

$$U_i(\theta) := -\log \widetilde{\pi}_i(\theta) \qquad U(\theta) := \sum_{i=1}^{m} U_i(\theta).$$

We choose some fixed $\widehat{\theta} \in \Theta$ not depending on i that is near the mode of π like Dubey et al. (2016); Bardenet et al. (2017); Bierkens et al. (2019); Baker et al. (2018); Chatterji et al. (2018); Quiroz et al. (2018a). Assuming sufficient differentiability, we then approximate U_i with a k-th order Taylor expansion around $\widehat{\theta}$, which we denote by

$$\widehat{U}_{k,i}(\theta) \approx U_i(\theta).$$

We also define

$$\widehat{\pi}_{k,i}(\theta) := \exp(-\widehat{U}_{k,i}(\theta)) \approx \widetilde{\pi}_i(\theta).$$

In practice we are exclusively interested in the cases $k = 1$ and $k = 2$, which correspond to first and second-order approximations respectively. Explicitly, in these cases

$$\begin{aligned} \widehat{U}_{1,i}(\theta) &= U(\widehat{\theta}) + \nabla U_i(\widehat{\theta})^\top (\theta - \widehat{\theta}), \\ \widehat{U}_{2,i}(\theta) &= \widehat{U}_{1,i}(\theta) + \frac{1}{2}(\theta - \widehat{\theta})^\top \nabla^2 U_i(\widehat{\theta})(\theta - \widehat{\theta}), \end{aligned}$$

where ∇ denotes the gradient and ∇^2 the Hessian. Letting

$$\widehat{U}_k(\theta) := \sum_{i=1}^{m} \widehat{U}_{k,i}(\theta) \qquad \widehat{\pi}_k(\theta) := \exp(-\widehat{U}_k(\theta)),$$

additivity of the Taylor expansion further yields

$$\begin{aligned} \widehat{U}_1(\theta) &= U(\widehat{\theta}) + \nabla U(\widehat{\theta})^\top (\theta - \widehat{\theta}) \\ \widehat{U}_2(\theta) &= \widehat{U}_1(\theta) + \frac{1}{2}(\theta - \widehat{\theta})^\top \nabla^2 U(\widehat{\theta})(\theta - \widehat{\theta}). \end{aligned} \tag{9}$$

Thus when $\nabla^2 U(\widehat{\theta}) \succ 0$ (i.e. symmetric positive-definite), $\widehat{\pi}_2(\theta)$ is seen to be a Gaussian approximation to π around the (approximate) mode $\widehat{\theta}$.

We use the $\widehat{\pi}_{k,i}$ to define the *Scalable Metropolis-Hastings* (SMH or SMH-k) acceptance probability

$$\alpha_{\text{SMH-}k}(\theta, \theta') := \left(1 \wedge \frac{\widehat{\pi}_k(\theta')q(\theta', \theta)}{\widehat{\pi}_k(\theta)q(\theta, \theta')}\right) \prod_{i=1}^{m} 1 \wedge \frac{\widetilde{\pi}_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\widehat{\pi}_{k,i}(\theta')\widetilde{\pi}_i(\theta)}. \tag{10}$$

Note that SMH-k is a special case of FMH with $m + 1$ factors given by

$$\pi = \underbrace{\widehat{\pi}_k}_{=\pi_{m+1}} \prod_{i=1}^{m} \underbrace{\frac{\widetilde{\pi}_i}{\widehat{\pi}_{k,i}}}_{=\pi_i} \qquad q = \underbrace{q}_{=q_{m+1}} \prod_{i=1}^{m} \underbrace{1}_{=q_i} \tag{11}$$

and hence defines a valid acceptance probability. (Note that $\widehat{\pi}_1$ is not integrable, but recall this is not required of FMH factors.) We could consider any factorisation of q, but we will not make use of this generality.

$\widehat{\pi}_k(\theta)$ can be computed in constant time after precomputing the relevant partial derivatives at $\widehat{\theta}$ before our MCMC run. This allows us to deal with $1 \wedge \widehat{\pi}_k(\theta')q(\theta', \theta)/\widehat{\pi}_k(\theta)q(\theta, \theta')$ by directly simulating a Bernoulli trial with this value as its success probability. For the remaining factors we have

$$\lambda_i(\theta, \theta') = -\log\left(1 \wedge \frac{\widetilde{\pi}_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\widetilde{\pi}_i(\theta)\widehat{\pi}_{k,i}(\theta')}\right).$$

We can obtain a bound of the form (3) provided U_i is $(k+1)$-times continuously differentiable. In this case, if we can find constants

$$\overline{U}_{k+1,i} \geq \sup_{\substack{\theta \in \Theta \\ |\beta| = k+1}} |\partial^\beta U_i(\theta)|, \tag{12}$$

(here β is multi-index notation; see Section A of the Supplement) it follows that

$$\overline{\lambda}(\theta, \theta') := \underbrace{(\|\theta - \widehat{\theta}\|_1^{k+1} + \|\theta' - \widehat{\theta}\|_1^{k+1})}_{=\varphi(\theta, \theta')} \sum_{i=1}^{m} \underbrace{\frac{\overline{U}_{k+1,i}}{(k+1)!}}_{=\psi_i} \tag{13}$$

defines an upper bound of the required form (5). See Proposition D.1 in the Supplement for a derivation. Observe this is symmetric in θ and θ' and therefore can be used to define a truncated version of SMH as described in Section 2.3.

Although we concentrate on Taylor expansions here, other choices of $\widehat{\pi}_i$ may be useful. For instance, it may be possible to make $\widetilde{\pi}_i/\widehat{\pi}_i$ log-Lipschitz or log-concave and obtain better bounds. However, Taylor expansions have the advantage of generality and (13) is sufficiently tight for us.

Heuristically, if the posterior concentrates like $1/\sqrt{n}$, if we scale our proposal like $1/\sqrt{n}$, and if $\widehat{\theta}$ is not too far (specifically $O(1/\sqrt{n})$) from the mode, then both $\|\theta-\widehat{\theta}\|$ and $\|\theta'-\widehat{\theta}\|$ will be $O(1/\sqrt{n})$, and $\varphi(\theta,\theta')$ will be $O(n^{-(k+1)/2})$. If moreover $m \asymp n$, then the summation will be $O(n)$ and hence overall $\overline{\lambda}(\theta,\theta') = O(n^{(1-k)/2})$. When $k=1$ this is $O(1)$ and when $k=2$ this is $O(1/\sqrt{n})$, which entails a substantial improvement over the naïve approach. In particular, we expect stable acceptance probabilities in both cases, constant expected cost in n for $k=1$, and indeed $O(1/\sqrt{n})$ *decreasing* cost for $k=2$. We make this argument rigorous in Theorem 3.1 below.

Beyond what is already needed for MH, $\overline{U}_{k+1,i}$ and $\widehat{\theta}$ are all the user must provide for our method. In practice neither of these seems problematic in typical settings. We have found deriving $\overline{U}_{k+1,i}$ to be a fairly mechanical procedure, and give examples for two models in Section 4. Likewise, while computing $\widehat{\theta}$ does entail some cost, we have found that standard gradient descent finds an adequate result in time negligible compared with the full MCMC run.

3.2. Choice of Proposal

We now consider the choice of proposal q and its implications for the acceptance probability. As mentioned, it is necessary to ensure that, roughly speaking, $\|\theta-\theta'\| = O(n^{-1/2})$ to match the concentration of the target. In this section we describe heuristically how to ensure this. Theorem 3.1 below and Section F.1.2 in the Supplement give precise statements of what is required.

Two main classes of q are of interest to us. When q is symmetric, (10) simplifies to

$$\alpha_{\text{SMH-}k}(\theta,\theta') = \left(1 \wedge \frac{\widehat{\pi}_k(\theta')}{\widehat{\pi}_k(\theta)}\right) \prod_{i=1}^{m} 1 \wedge \frac{\widetilde{\pi}_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\widetilde{\pi}_i(\theta)\widehat{\pi}_{k,i}(\theta')}. \tag{14}$$

We can realise this with the correct scaling with for example

$$q(\theta,\theta') = \mathrm{Normal}(\theta' \mid \theta, \frac{\sigma^2}{n} I_d), \tag{15}$$

where $\sigma > 0$ is fixed in n. Alternatively, we can more closely match the covariance of our proposal to the covariance of our target with

$$q(\theta,\theta') = \mathrm{Normal}(\theta' \mid \theta, \sigma^2[\nabla^2 U(\widehat{\theta})]^{-1}). \tag{16}$$

Under usual circumstances $[\nabla^2 U(\widehat{\theta})]^{-1}$ is approximately (since in general this will include a non-flat prior term) proportional to the inverse observed information matrix, and hence the correct $O(n^{-1/2})$ scaling is achieved automatically. See Section F.1.2 in the Supplement for more details.

We can improve somewhat on a symmetric proposal if we choose q to be $\widehat{\pi}_k$*-reversible* in the sense that

$$\widehat{\pi}_k(\theta) q(\theta,\theta') = \widehat{\pi}_k(\theta') q(\theta',\theta)$$

for all θ,θ'; see, e.g., (Tierney, 1994; Neal, 1999; Kamatani, 2018). In this case we obtain

$$\alpha_{\text{SMH-}k}(\theta,\theta') = \prod_{i=1}^{m} 1 \wedge \frac{\widetilde{\pi}_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\widetilde{\pi}_i(\theta)\widehat{\pi}_{k,i}(\theta')}.$$

Note that using a $\widehat{\pi}_k$-reversible proposal allows us to drop the first term in (14), and hence obtain a higher acceptance probability for the same θ,θ'. Moreover, when $k=2$, we see from (9) that a $\widehat{\pi}_k$-reversible proposal corresponds to an MCMC kernel that targets a Gaussian approximation to π, and may therefore be more suited to the geometry of π than a symmetric one.

We now consider how to produce $\widehat{\pi}_k$-reversible proposals. For q of the form

$$q(\theta,\theta') = \mathrm{Normal}(\theta' \mid A\theta + b, C)$$

where $A, C \in \mathbb{R}^{d\times d}$ with $C \succ 0$ and $b \in \mathbb{R}^d$, Theorem E.1 in the Supplement gives necessary and sufficient conditions for $\widehat{\pi}_1$ and $\widehat{\pi}_2$-reversibility. Specific useful choices that satisfy these conditions and ensure the correct scaling are then as follows. For $\widehat{\pi}_1$ we can use for example

$$A = I_d \qquad b = -\frac{\sigma}{2n}\nabla U(\widehat{\theta}) \qquad C = \frac{\sigma}{n} I_d \tag{17}$$

for some $\sigma > 0$, where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix. For $\widehat{\pi}_2$, assuming $\nabla^2 U(\widehat{\theta}) \succ 0$ (which will hold if $\widehat{\theta}$ is sufficiently close to the mode), we can use a variation of the *preconditioned-Crank Nicholson* proposal (pCN) (Neal, 1999) defined by taking

$$A = \sqrt{\rho} I_d \qquad C = (1-\rho)[\nabla^2 U(\widehat{\theta})]^{-1}$$
$$b = (1-\sqrt{\rho})(\widehat{\theta} - [\nabla^2 U(\widehat{\theta})]^{-1} \nabla U(\widehat{\theta}))$$

where $\rho \in [0,1)$. When $\rho = 0$ this corresponds to an independent Gaussian proposal: $\theta' \sim \widehat{\pi}_2$. Note that this can be re-interpreted as the exact discretization of an Hamiltonian dynamics for the Gaussian $\widehat{\pi}_2$.

3.3. Performance

We now show rigorously that SMH addresses the issues of a naive approach and entails an overall performance benefit. In our setup we assume some unknown data-generating distribution P_0, with data $Y_i \overset{\text{iid}}{\sim} P_0$. We denote the (random) targets by $\pi^{(n)}(\theta) := p(\theta|Y_{1:n})$, for which we assume a factorisation (8) involving $m^{(n)}$ terms. We denote the mode of $\pi^{(n)}$ by $\theta^{(n)}_{\text{MAP}}$, and our estimate of the mode by $\widehat{\theta}^{(n)}$. Observe that $\theta^{(n)}_{\text{MAP}} \equiv \theta^{(n)}_{\text{MAP}}(Y_{1:n})$ is a deterministic function of the data, and we assume this holds for $\widehat{\theta}^{(n)} \equiv \widehat{\theta}^{(n)}(Y_{1:n})$ also. In general $\widehat{\theta}^{(n)}$ may depend on additional randomness, say $W_{1:n}$, if for instance it is the output of a stochastic gradient descent algorithm. In that case, our statements should involve conditioning on $W_{1:n}$ but are otherwise unchanged.

Given n data points we denote the proposal by $q^{(n)}$, and model the behaviour of our chain at stationarity by considering $\theta^{(n)} \sim \pi^{(n)}$ and $\theta'^{(n)} \sim q^{(n)}(\theta^{(n)}, \cdot)$ sampled independently of all other randomness given $Y_{1:n}$. The following theorem allows us to show that both the computational cost and the acceptance probability of SMH remain stable as $n \to \infty$. See Section F in the Supplement for a proof.

Theorem 3.1. *Suppose each U_i is $(k+1)$-times continuously differentiable, each $\overline{U}_{k+1,i} \in L^{k+2}$, and $\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{k+1,i}|Y_{1:n}] = O_{P_0}(n)$. Likewise, assume each of $\|\theta^{(n)} - \theta^{(n)}_{\text{MAP}}\|$, $\|\theta^{(n)} - \theta'^{(n)}\|$, and $\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\text{MAP}}\|$ is in L^{k+2}, and each of $\mathbb{E}[\|\theta^{(n)} - \theta^{(n)}_{\text{MAP}}\|^{k+1}|Y_{1:n}]$, $\mathbb{E}[\|\theta^{(n)} - \theta'^{(n)}\|^{k+1}|Y_{1:n}]$, and $\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\text{MAP}}\|^{k+1}$ is $O_{P_0}(n^{-(k+1)/2})$ as $n \to \infty$. Then $\overline{\lambda}$ defined by (13) satisfies*

$$\mathbb{E}[\overline{\lambda}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] = O_{P_0}(n^{(1-k)/2}).$$

For given $\theta^{(n)}$ and $\theta'^{(n)}$, recall that the method described in Section 2.2 requires the computation of at most $N^{(n)} \sim \text{Poisson}(\overline{\lambda}(\theta^{(n)}, \theta'^{(n)}))$ factors. Under the conditions of Theorem 3.1, we therefore have

$$\begin{aligned}
\mathbb{E}[N^{(n)}|Y_{1:n}] &= \mathbb{E}[\mathbb{E}[N^{(n)}|\theta^{(n)}, \theta'^{(n)}, Y_{1:n}]|Y_{1:n}] \\
&= \mathbb{E}[\overline{\lambda}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] \\
&= O_{P_0}(n^{(1-k)/2}).
\end{aligned}$$

In other words, with arbitrarily high probability with respect to the data-generating distribution, SMH requires processing on average only $O(1)$ data points per step for a first-order approximation, and $O(1/\sqrt{n})$ for a second-order one.

This result also ensures that the acceptance probability for SMH does not vanish as $n \to \infty$. Denoting by $\widehat{\pi}^{(n)}_k$ our approximation in the case of n data points, observe that

$$\begin{aligned}
0 &\leq \mathbb{E}[-\log \alpha_{\text{FMH}}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] \\
&\leq \mathbb{E}[-\log(1 \wedge \frac{\widehat{\pi}^{(n)}_k(\theta'^{(n)}) q^{(n)}(\theta'^{(n)}, \theta^{(n)})}{\widehat{\pi}^{(n)}_k(\theta^{(n)}) q^{(n)}(\theta^{(n)}, \theta'^{(n)})})|Y_{1:n}] \\
&\qquad + \mathbb{E}[\overline{\lambda}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}]
\end{aligned}$$

Here the second right-hand side term is $O_{P_0}(n^{(1-k)/2})$ by Theorem 3.1. For a $\widehat{\pi}_k$-reversible proposal the first term is simply 0, while for a symmetric proposal Theorem F.2 in the Supplement shows it is $O_{P_0}(1)$. In either case, we see that the acceptance probability is stable in the limit of large n. In the case of a $\widehat{\pi}_2$-reversible proposal, we in fact have $\mathbb{E}[\alpha_{\text{FMH}}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] \overset{P_0}{\to} 1$.

Note that both these implications also apply if we use a truncated version of SMH as per Section 2.3. This holds since in general TFMH ensures both that the expected number of factor evaluations is not greater than for FMH, and that the acceptance probability is not less than for FMH.

The conditions of Theorem 3.1 hold in realistic scenarios. The integrability assumptions are mild and mainly technical. We will see in Section 4 that in practice $\overline{U}_{k+1,i} \equiv \overline{U}_{k+1}(Y_i)$ is usually a function of Y_i, in which case

$$\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{k,i}|Y_{1:n}] = \sum_{i=1}^{n} \overline{U}_{k+1}(Y_i) = O_{P_0}(n)$$

by the law of large numbers. In general, we might also have one $\overline{U}_{k,i}$ for the prior also, but the addition of this term still gives the same asymptotic behaviour.

The condition $\mathbb{E}[\|\theta^{(n)} - \theta^{(n)}_{\text{MAP}}\|^{k+1}|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2})$ essentially states that the posterior must concentrate at rate $O(1/\sqrt{n})$ around the mode. This is a consequence of standard, widely-applicable assumptions that are used to prove Bernstein-von Mises. See Section F.1.1 of the Supplement for more details. Note in particular that we do not require our model to be well-specified (i.e. we do not need $P_0 = p(y|\theta_0)$ for some $\theta_0 \in \Theta$). The remaining two O_{P_0} conditions correspond to the heuristic conditions given in Section 3.1. In particular, the proposal should scale like $1/\sqrt{n}$. We show Section F.1.2 of the Supplement that this condition holds for the proposals described in Section 3.2. Likewise, $\widehat{\theta}$ should be distance $O(1/\sqrt{n})$ from the mode. When the posterior is log-concave it can be shown this holds for instance for stochastic gradient descent after performing a single pass through the data (Baker et al., 2018, Section 3.4). In practice, we interpret this condition to mean that $\widehat{\theta}$ should be as close as possible to θ_{MAP}, but that some small margin for error is acceptable.

4. Experimental Results

In this section we apply SMH to Bayesian logistic regression. A full description of the model and upper bounds (12) we used is given in Section G.2 of the Supplement. We also provide there an additional application our method to robust linear regression. We chose these models due to the availability of lower bounds on the likelihoods required by Firefly.

In our experiments we took $d = 10$. For both SMH-1 and SMH-2 we used truncation as described in Section 2.3, with $R = n$. Our estimate of the mode $\widehat{\theta}$ was computed using stochastic gradient descent. We compare our algorithms to standard MH, Firefly, and Zig-Zag (Bierkens et al., 2019), which all have the exact posterior as the invariant distribution. We used the MAP-tuned variant of Firefly (which also makes use of $\widehat{\theta}$) with implicit sampling (this uses an algorithmic parameter $q_{d\to b} = 10^{-3}$; the optimal choice of $q_{d\to b}$ is an open question) and the lower bounds specified in Section 3.1 of Maclaurin & Adams (2014).

Figure 1 (in Section 1) shows the average number of likelihood evaluations per step and confirms the predictions of Theorem 3.1. Figure 2 displays the effective sample sizes (ESS) for the posterior mean estimate of one regression coefficient, rescaled by execution time. For large n, SMH-2 significantly outperforms competing techniques. For all methods except Zig-Zag we used the proposal (16) with $\sigma = 1$, which automatically scales according to the concentration of the target.

We also separately considered the performance of the pCN proposal. Figure 3 shows the effect of varying ρ. As the target concentrates, the Gaussian approximation of the target improves and an independent proposal ($\rho = 0$) becomes optimal. Finally, we also illustrate the average acceptance rate when varying ρ in Figure 4.

Since SMH-2 makes use of a Gaussian approximation to the posterior $\widehat{\pi}_2$, we finally consider the benefit that our method yields over simply using $\widehat{\pi}_2$ directly. Observe in Figure 4 that the acceptance probability of MH with the independent proposal differs non-negligibly from 1 for reasonably large values of n, which indicates that our method yields a non-trivial increase in accuracy. For very large n, the discrepancy vanishes as expected and SMH and other subsampling methods based on control variates become less useful in practice. See Section G of the Supplement for further results along these lines. We believe however that our approach could form the basis of subsampling methods in more general and interesting settings such as random effect models and leave this as an important piece of future work.

Code to reproduce our experiments is available at `github.com/pjcv/smh`.

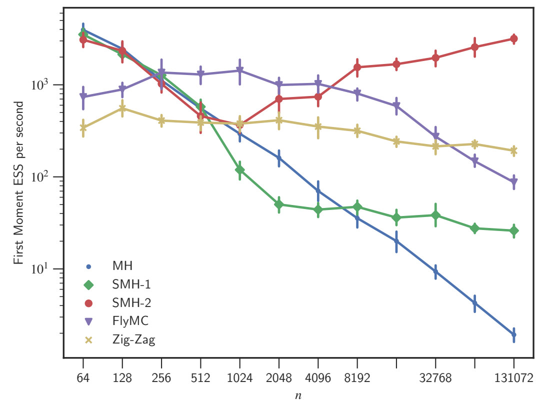

Figure 2. ESS for first regression coefficient, scaled by execution time (higher is better).

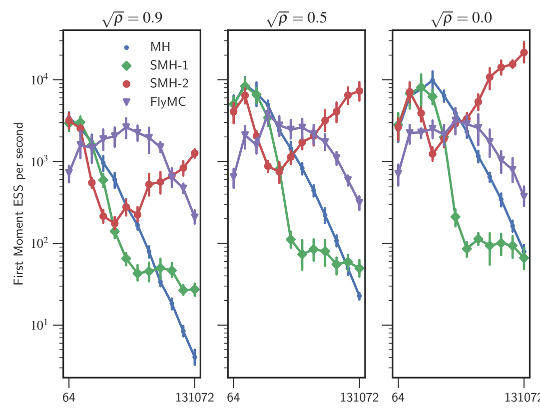

Figure 3. Effect of ρ on ESS. ESS for first regression coefficient, scaled by execution time (higher is better).

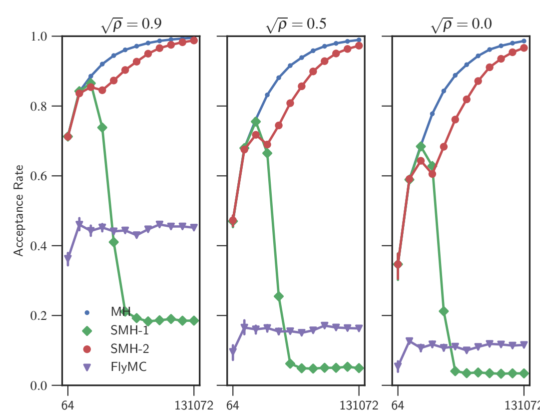

Figure 4. Acceptance rates for pCN proposals.

Acknowledgements

Rob Cornish is supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems (EP/L015897/1) and NVIDIA. Arnaud Doucet is partially supported by the U.S. Army Research Laboratory, the U. S. Army Research Office, and by the U.K. Ministry of Defence (MoD) and the U.K. EPSRC under grant numbers EP/R013616/1 and EP/R034710/1.

References

Baker, J., Fearnhead, P., Fox, E. B., and Nemeth, C. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 2018. URL https://doi.org/10.1007/s11222-018-9826-2. to appear.

Banterle, M., Grazian, C., Lee, A., and Robert, C. P. Accelerating Metropolis–Hastings algorithms by delayed acceptance. *arXiv preprint arXiv:1503.00996*, 2015.

Bardenet, R., Doucet, A., and Holmes, C. Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. I–405–I–413. JMLR.org, 2014. URL http://dl.acm.org/citation.cfm?id=3044805.3044852.

Bardenet, R., Doucet, A., and Holmes, C. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017. URL http://jmlr.org/papers/v18/15-205.html.

Bierkens, J., Fearnhead, P., and Roberts, G. The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. *The Annals of Statistics*, 47:1288–1320, 2019.

Bouchard-Côté, A., Vollmer, S. J., and Doucet, A. The Bouncy Particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association*, 113:855–867, 2018.

Brosse, N., Durmus, A., and Moulines, E. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pp. 8268–8278, 2018.

Ceperley, D. M. Path integrals in the theory of condensed helium. *Reviews of Modern Physics*, 67(2):279, 1995.

Chatterji, N. S., Flammarion, N., Ma, Y.-A., Bartlett, P. L., and Jordan, M. I. On the theory of variance reduction for stochastic gradient Monte Carlo. *arXiv preprint arXiv:1802.05431*, 2018.

Christen, J. A. and Fox, C. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810, 2005. ISSN 10618600. URL http://www.jstor.org/stable/27594150.

Dang, K.-D., Quiroz, M., Kohn, R., Tran, M.-N., and Villani, M. Hamiltonian Monte Carlo with energy conserving subsampling. *arXiv preprint arXiv:1708.00955*, 2017.

Devroye, L. *Non-Uniform Random Variate Generation*. Springer, 1986.

Dubey, K. A., Reddi, S. J., Williamson, S. A., Poczos, B., Smola, A. J., and Xing, E. P. Variance reduction in stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems*, pp. 1154–1162, 2016.

Fukui, K. and Todo, S. Order-n cluster Monte Carlo method for spin systems with long-range interactions. *Journal of Computational Physics*, 228(7):2629–2642, 2009.

Jones, G. L. and Hobert, J. P. Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo. *Statistical Science*, 16(4):312–334, 2001. ISSN 0883-4237.

Kamatani, K. Efficient strategy for the Markov chain Monte Carlo in high-dimension with heavy-tailed target probability distribution. *Bernoulli*, 24(4B):3711–3750, 2018.

Korattikara, A., Chen, Y., and Welling, M. Austerity in MCMC land: Cutting the Metropolis–Hastings budget. In *International Conference on Machine Learning*, pp. 181–189, 2014.

Kronmal, R. A. and Peterson, A. V. J. On the alias method for generating random variables from a discrete distribution. *The American Statistician*, 33(4):214–218, 1979.

Maclaurin, D. and Adams, R. P. Firefly Monte Carlo: Exact MCMC with subsets of data. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pp. 543–552, Arlington, Virginia, United States, 2014. AUAI Press. ISBN 978-0-9749039-1-0. URL http://dl.acm.org/citation.cfm?id=3020751.3020808.

Michel, M., Tan, X., and Deng, Y. Clock Monte Carlo methods. *Physical Review E*, 99(1):010105, 2019.

Neal, R. M. Regression and classification using Gaussian process priors. In *Bayesian Statistics 6*, pp. 475–501. Oxford Univ. Press, 1999.

Neiswanger, W., Wang, C., and Xing, E. Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*, 2013.

Payne, R. D. and Mallick, B. K. Two-stage Metropolis–Hastings for tall data. *Journal of Classification*, 35 (1):29–51, 2018. ISSN 1432-1343. doi: 10.1007/s00357-018-9248-z. URL https://doi.org/10.1007/s00357-018-9248-z.

Peskun, P. H. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973. ISSN 00063444. URL http://www.jstor.org/stable/2335011.

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., and Dang, K.-D. The block-Poisson estimator for optimally tuned exact subsampling MCMC. *arXiv e-prints*, art. arXiv:1603.08232, 2016.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 2018a. to appear.

Quiroz, M., Tran, M.-N., Villani, M., and Kohn, R. Speeding up MCMC by delayed acceptance and data subsampling. *Journal of Computational and Graphical Statistics*, 27(1):12–22, 2018b. doi: 10.1080/10618600.2017.1307117. URL https://doi.org/10.1080/10618600.2017.1307117.

Roberts, G. and Rosenthal, J. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997. doi: 10.1214/ECP.v2-981. URL https://doi.org/10.1214/ECP.v2-981.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11:78–88, 2016. URL http://www.tandfonline.com/doi/full/10.1080/17509653.2016.1142191.

Shanthikumar, J. Discrete random variate generation using uniformization. *European Journal of Operational Research*, 21(3):387–398, 1985.

Tierney, L. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

Tierney, L. A note on Metropolis–Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8(1):1–9, 1998.

Vanetti, P., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. Piecewise deterministic Markov chain Monte Carlo. *arXiv preprint arXiv:1707.05296v2*, 2018.

Walker, A. J. An efficient method for generating discrete random variables with general distributions. *ACM Trans. Math. Softw.*, 3(3):253–256, 1977. ISSN 0098-3500. doi: 10.1145/355744.355749. URL http://doi.acm.org/10.1145/355744.355749.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 681–688, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5. URL http://dl.acm.org/citation.cfm?id=3104482.3104568.