Biometrika (2020), **00**, 0, *pp.* 1–8
Printed in Great Britain

doi: 10.1093/biomet/asaa029

Extended stochastic gradient Markov chain Monte Carlo for large-scale Bayesian variable selection

BY QIFAN SONG, YAN SUN, MAO YE AND FAMING LIANG

Department of Statistics, Purdue University, 250 N. University St., West Lafayette, Indiana 47906, U.S.A.
qfsong@purdue.edu sun748@purdue.edu ye207@purdue.edu fmliang@purdue.edu

SUMMARY

Stochastic gradient Markov chain Monte Carlo algorithms have received much attention in Bayesian computing for big data problems, but they are only applicable to a small class of problems for which the parameter space has a fixed dimension and the log-posterior density is differentiable with respect to the parameters. This paper proposes an extended stochastic gradient Markov chain Monte Carlo algorithm which, by introducing appropriate latent variables, can be applied to more general large-scale Bayesian computing problems, such as those involving dimension jumping and missing data. Numerical studies show that the proposed algorithm is highly scalable and much more efficient than traditional Markov chain Monte Carlo algorithms.

Some key words: Dimension jumping; Missing data; Stochastic gradient Langevin dynamics; Subsampling.

1. INTRODUCTION

After six decades of continual development, Markov chain Monte Carlo, MCMC, has proven to be a powerful computational tool for analysing data of complex structures. However, for large datasets, its computational cost can be prohibitive as it requires all of the data to be processed at each iteration. To tackle this difficulty, a variety of scalable algorithms have been proposed in the recent literature. These algorithms can be grouped into a few categories according to the strategies they employed, including stochastic gradient MCMC (Welling & Teh, 2011; Ahn et al., 2012; Chen et al., 2014; Ding et al., 2014; Betancourt, 2015; Ma et al., 2015; Nemeth & Fearnhead, 2019), split-and-merge (Scott et al., 2016; Li et al., 2017; Srivastava et al., 2018; Xue & Liang, 2019), mini-batch Metropolis–Hastings algorithms (Bardenet et al., 2014, 2017; Korattikara et al., 2014; Maclaurin & Adams, 2014; Seita et al., 2017), nonreversible Markov process-based algorithms (Bouchard Coté et al., 2018; Bierkens et al., 2019), Bayesian bootstrapping (Fong et al., 2019), and some discrete sampling algorithms based on the multi-armed bandit (Chen & Ghahramani, 2016).

Although scalable algorithms have been developed for both continuous and discrete sampling problems, they are hard to apply to dimension-jumping problems. Dimension jumping is characterized by variable selection where the number of parameters changes from iteration to iteration in MCMC simulations. Under their current settings, stochastic gradient MCMC and nonreversible Markov process-based algorithms are only applicable to problems for which the parameter space has a fixed dimension, and the log-posterior density is differentiable with respect to the parameters. For split-and-merge algorithms, it is unclear how to aggregate samples of different dimensions drawn from the posterior distributions based on different subsets of data. The multi-armed bandit algorithms are only applicable to problems with a small discrete domain and can be extremely inefficient for high-dimensional variable selection problems. The mini-batch Metropolis–Hastings algorithms are in principle applicable to dimension-jumping problems; however, they are generally difficult to use. For example, the algorithms by Bardenet et al. (2014), Korattikara et al. (2014) and Seita et al. (2017) perform approximate acceptance tests using subsets of data. The amount of data consumed for each test varies significantly from one iteration to another, which compromise their scalability. The algorithms by Maclaurin & Adams (2014) and Bardenet et al. (2017) perform exact tests, but require a lower bound on the parameter distribution across its domain. Unfortunately, the lower bound is usually difficult to obtain.

This paper extends the stochastic gradient Langevin dynamics algorithm to more general large-scale Bayesian computing problems such as variable selection and missing data, by introducing appropriate latent variables. The extended stochastic gradient Langevin dynamics algorithm is highly scalable and much more efficient than traditional MCMC algorithms. Compared to the mini-batch Metropolis–Hastings algorithms, the proposed algorithm is much easier to use, involves only a fixed amount of data at each iteration and does not require any lower bound on the parameter distribution.

2. A BRIEF REVIEW OF STOCHASTIC GRADIENT LANGEVIN DYNAMICS

Let $X_N = (x^1, x^2, \ldots, x^N)$ denote a set of N independent and identically distributed samples drawn from the distribution $f(x \mid \theta)$, where N is the sample size and θ is the parameter. Let $p(X_N \mid \theta) = \prod_{i=1}^N f(x^i \mid \theta)$ denote the likelihood function. Let $\pi(\theta)$ and $\pi(\theta \mid X_N) \propto p(X_N \mid \theta)\pi(\theta)$ denote the prior and posterior distributions of θ, respectively. If θ has a fixed dimension and $\log \pi(\theta \mid X_N)$ is differentiable with respect to θ, then the stochastic gradient Langevin dynamics algorithm (Welling & Teh, 2011) can be applied to simulate from the posterior. This uses iterations of the form

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\epsilon_{t+1}}{2} \widehat{\nabla}_\theta \log \pi(\theta^{(t)} \mid X_N) + \surd(\epsilon_{t+1}\tau)\eta_{t+1}, \quad \eta_{t+1} \sim N(0, I_d),$$

where d is the dimension of θ, I_d is a $d \times d$ identity matrix, ϵ_{t+1} is the step size, also known as the learning rate, τ is the temperature, and $\widehat{\nabla}_\theta \log \pi(\theta^{(t)} \mid X_N)$ denotes an estimate of $\nabla_\theta \log \pi(\theta^{(t)} \mid X_N)$ based on a mini-batch of data. The learning rate can be decreasing or kept as a constant. For the former, the convergence of the algorithm was studied in Teh et al. (2016). For the latter, the convergence of the algorithm was studied in Sato & Nakagawa (2014) and Dalalyan & Karagulyan (2018). Refer to Nemeth & Fearnhead (2019) for more discussions on the theory, implementation and variants of this algorithm.

3. AN EXTENDED STOCHASTIC GRADIENT LANGEVIN DYNAMICS ALGORITHM

3.1. *Monte Carlo estimator*

To extend the applications of the stochastic gradient Langevin dynamics algorithm to varying-dimensional problems such as variable selection and missing data, we first establish an identity for evaluating $\nabla_\theta \log \pi(\theta \mid X_N)$ in the presence of latent variables. As illustrated below, the latent variables can be the model indicator in the variable selection problems or missing values in the missing data problems.

LEMMA 1. *For any latent variable ϑ,*

$$\nabla_\theta \log \pi(\theta \mid X_N) = \int \nabla_\theta \log \pi(\theta \mid \vartheta, X_N)\pi(\vartheta \mid \theta, X_N)\, \mathrm{d}\vartheta, \tag{1}$$

where $\pi(\vartheta \mid \theta, X_N)$ and $\pi(\theta \mid \vartheta, X_N)$ denote the conditional distribution of ϑ and θ, respectively.

Lemma 1 provides a Monte Carlo estimator for $\nabla_\theta \log \pi(\theta \mid X_N)$ by averaging over the samples drawn from the conditional distribution $\pi(\vartheta \mid \theta, X_N)$. The identity (1) is similar to Fisher's identity. The latter has been used in evaluating the gradient of the loglikelihood function in the presence of latent variables; see, e.g., Cappé et al. (2005). When N is large, the computation can be accelerated by subsampling. Let X_n denote a subsample, where n denotes the subsample size. Without loss of generality, we assume that N is a multiple of n, i.e., N/n is an integer. Let $X_{n,N} = \{X_n, \ldots, X_n\}$ denote a duplicated dataset with the subsample, whose total sample size is also N. Following from (1), we have

$$\nabla_\theta \log \pi(\theta \mid X_{n,N}) = \int \nabla_\theta \log \pi(\theta \mid \vartheta, X_{n,N})\pi(\vartheta \mid \theta, X_{n,N})\, \mathrm{d}\vartheta. \tag{2}$$

Since $\nabla_\theta \log \pi(\theta \mid X_{n,N}) = \nabla_\theta \log p(X_{n,N} \mid \theta) + \nabla_\theta \log \pi(\theta)$ and $\log p(X_{n,N} \mid \theta)$ is unbiased for

$\log p(X_N \mid \theta)$, $\nabla_\theta \log \pi(\theta \mid X_{n,N})$ forms an unbiased estimator of $\nabla_\theta \log \pi(\theta \mid X_N)$. Sampling from $\pi(\gamma_S \mid \theta, X_{n,N})$ can be much faster than sampling from $\pi(\gamma_S \mid \theta, X_N)$, as for the former the likelihood only needs to be evaluated on a mini-batch of data.

3.2. *Bayesian variable selection*

As an illustrative example, we consider the problem of variable selection for the linear regression

$$Y = Z^{\mathrm{T}}\beta + \varepsilon, \tag{3}$$

where ε is a zero-mean Gaussian random error with variance σ^2, $\beta \in \mathbb{R}^p$ is the vector of regression coefficients, and $Z = (z_1, z_2, \dots, z_p)$ is the vector of explanatory variables. Let $\gamma_S = (\gamma_S^1, \dots, \gamma_S^p)$ be a binary vector indicating the variables included in model S, and let β_S be the vector of regression coefficients associated with the model S. We are interested in estimating the posterior probability $\pi(\gamma_S \mid X_N)$ for each model $S \in \mathcal{S}$ and the posterior mean $\pi(\rho) = \int \rho(\beta)\pi(\beta \mid X_N)$ for some integrable function $\rho(\cdot)$, where $\mathcal{S}$ comprises 2^p models. Both quantities can be estimated using the reversible jump Metropolis–Hastings algorithm (Green, 1995) by sampling from the posterior distribution $\pi(\gamma_S, \beta_S \mid X_N)$. However, when N is large, the algorithm can be extremely slow due to repeated scans of the full dataset in simulations.

The existing stochastic gradient MCMC algorithms cannot be directly applied to simulate from $\pi(\gamma_S, \beta_S \mid X_N)$ due to the dimension jumping issue involved in model transition. To address this issue, we introduce an auxiliary variable $\theta = (\theta^1, \theta^2, \dots, \theta^p)$, which links γ_S and β_S through

$$\beta_S = \theta * \gamma_S = (\theta^1\gamma_S^1, \theta^2\gamma_S^2, \dots, \theta^p\gamma_S^p), \tag{4}$$

where $*$ denotes elementwise multiplication. Let $\theta_{[S]} = \{\theta^i : \gamma_S^i = 1, i = 1, 2, \dots, p\}$ and $\theta_{[-S]} = \{\theta^i : \gamma_S^i = 0, i = 1, 2, \dots, p\}$ be subvectors of θ corresponding to the nonzero and zero elements of γ_S, respectively. Note that β_S is sparse, with all elements in $\theta_{[-S]}$ being zero, while θ can be dense. Based on the relation (4), we suggest simulating from $\pi(\theta \mid X_N)$ using the stochastic gradient Langevin dynamic algorithm, for which the gradient $\nabla_\theta \log \pi(\theta \mid X_N)$ can be evaluated using Lemma 1 by treating γ_S as the latent variable. Let $\pi(\theta)$ denote the prior of θ. To simplify the computation of $\nabla_\theta \log \pi(\theta \mid \gamma_S, X_N)$, we further assume the a priori independence that $\pi(\theta \mid \gamma_S) = \pi(\theta_{[S]} \mid \gamma_S)\pi(\theta_{[-S]} \mid \gamma_S)$. Then, it is easy to derive

$$\nabla_\theta \log \pi(\theta \mid \gamma_S, X_N) = \begin{cases} \nabla_{\theta_{[S]}} \log p(X_N \mid \theta_{[S]}, \gamma_S) + \nabla_{\theta_{[S]}} \pi(\theta_{[S]} \mid \gamma_S), & \text{for component } \theta_{[S]}, \\ \nabla_{\theta_{[-S]}} \log \pi(\theta_{[-S]} \mid \gamma_S), & \text{for component } \theta_{[-S]}, \end{cases}$$

which can be used in evaluating $\nabla \log \pi(\theta \mid X_N)$ by Lemma 1. If a mini-batch of data is used, the gradient can be evaluated based on (2). This leads to an extended stochastic gradient Langevin dynamics algorithm.

Algorithm 1. Extended stochastic gradient Langevin dynamics for Bayesian variable selection.

Step 1 (Subsampling). Draw a subsample of size n, with or without replacement, from the full dataset X_N at random, and denote the subsample by $X_n^{(t)}$, where t indexes the iteration.

Step 2 (Simulating models). Simulate models $\gamma^{(t)}_{S_1,n}, \dots, \gamma^{(t)}_{S_m,n}$ from the conditional posterior $\pi(\gamma_S \mid \theta^{(t)}, X^{(t)}_{n,N})$ by running a short Markov chain, where $X^{(t)}_{n,N} = \{X^{(t)}_n, \dots, X^{(t)}_n\}$ and $\theta^{(t)}$ is the sample of θ at iteration t.

Step 3 (Updating θ). Update $\theta^{(t)}$ by setting $\theta^{(t+1)} = \theta^{(t)} + (2m)^{-1}\epsilon_{t+1}\sum_{k=1}^{m} \nabla_\theta \log \pi(\theta^{(t)} \mid \gamma^{(t)}_{S_k,n}, X^{(t)}_{n,N}) + \sqrt{(\epsilon_{t+1}\tau)}\eta_{t+1}$, where ϵ_{t+1} is the learning rate, $\eta_{t+1} \sim N(0, I_p)$, τ is the temperature and p is the dimension of θ.

Theorem 1 justifies the validity of this algorithm; the proof is given in the Appendix.

Theorem 1. *Assume that Conditions* (A1)–(A3) *in the Appendix hold, that m, p, n are increasing with N such that $N \geqslant n \succ p$, $m \succ p^{1/2}$, and that a constant learning rate $\epsilon \prec 1/N$ is used. Then, as $N \to \infty$;*

(i) $W_2(\pi_t, \pi_*) \to 0$ *as* $t \to \infty$, *where* π_t *denotes the distribution of* $\theta^{(t)}$, $\pi_* = \pi(\theta \mid X_N)$, *and* $W_2(\cdot, \cdot)$ *denotes the second-order Wasserstein distance between two distributions.*

(ii) If $\rho(\theta)$ *is* α*-Lipschitz for some constant* $\alpha > 0$, *then* $\sum_{t=1}^{T} \rho(\theta^{(t)})/T \xrightarrow{p} \pi_*(\rho)$ *as* $T \to \infty$, *where* $\xrightarrow{p}$ *denotes convergence in probability and* $\pi_*(\rho) = \int_\Theta \rho(\theta)\pi(\theta \mid X_N)\,\mathrm{d}\theta$.

(iii) If Condition (A4) further holds, $\sum_{t=1}^{T} \sum_{i=1}^{m} I(\gamma_{S_i,n}^{(t)} = \gamma_S)/(mT) - \pi(\gamma_S \mid X_N) \xrightarrow{p} 0$ *as* $T \to \infty$.

Part (i) establishes the weak convergence of θ_t; that is, if the total sample size N and the iteration number t are sufficiently large, and the subsample size n and the number of models m simulated at each iteration are reasonably large, then $\pi(\theta^{(t)} \mid X_N)$ will converge to the true posterior $\pi(\theta \mid X_N)$ in 2-Wasserstein distance. Refer to Gibbs & Su (2002) for discussions on the relation between Wasserstein distance and other probability metrics. Parts (ii) and (iii) address our general interests on how to estimate the posterior mean and posterior probability, respectively, based on the samples simulated by Algorithm 1. For parts (i), (ii) and (iii), the explicit convergence rates are given in equations (A1), (A2) and (A6), respectively.

For the choice of $m \succ p^{1/2}$, p can be approximately treated as the maximum size of the models under consideration, which is of the same order as the true model. Therefore, m can be pretty small under the model sparsity assumption. Theorem 1 is established with a constant learning rate. In practice, one may use a decaying learning rate, see, e.g., Teh et al. (2016), where it is suggested to set $\epsilon_t = O(1/t^{\kappa})$ for some $0 < \kappa \leqslant 1$. For the decaying learning rate, Teh et al. (2016) recommended some weighted averaging estimators for $\pi_*(\rho)$. Theorem 2 shows that the unweighted averaging estimators used above still work if the learning rate slowly decays at a rate of $\epsilon_t = O(1/t^{\kappa})$ for $0 < \kappa < 1$. However, if $\kappa = 1$, the weighted averaging estimators are still needed. The proof of Theorem 2 is given in the Supplementary Material.

THEOREM 2. *Assume that the conditions of Theorem 1 hold. If a decaying learning rate* $\epsilon_t = O(1/t^{\kappa})$ *is used for some* $0 < \kappa < 1$, *then parts (i), (ii) and (iii) of Theorem 1 are still valid.*

3.3. *Missing data*

Missing data are ubiquitous over all fields from science to technology. However, under the big data scenario, how to conduct Bayesian analysis in the presence of missing data is still unclear. The existing data augmentation algorithm (Tanner & Wong, 1987) is full-data based and thus can be extremely slow. In this context, we let X_N denote the incomplete data and let θ denote the model parameters. If we treat the missing values as latent variables, then Lemma 1 can be used for evaluating the gradient $\nabla_\theta \log \pi(\theta \mid X_N)$. However, Algorithm 1 cannot be directly applied to missing data problems, since the imputation of the missing data might depend on the subsample only. To address this issue we propose Algorithm S1, given in the Supplementary Material, where the missing values ϑ are imputed from $\pi(\vartheta \mid \theta, X_n)$ at each iteration. Theorems 1 and 2 are still applicable to this algorithm.

4. AN ILLUSTRATIVE EXAMPLE

We now illustrate the performance of Algorithm 1 using a simulated example. More numerical examples are presented in the Supplementary Material. Ten synthetic datasets were generated from the model (3) with $N = 50\,000$, $p = 2001$, $\sigma^2 = 1$, $\beta_1 = \cdots = \beta_5 = 1$, $\beta_6 = \beta_7 = \beta_8 = -1$ and $\beta_0 = \beta_9 = \cdots = \beta_p = 0$, where σ^2 is assumed to be known, and the explanatory variables are normally distributed with a mutual correlation coefficient of 0.5. A hierarchical prior was assumed for the model and parameters; see the Supplementary Material for details. For each dataset, Algorithm 1 was run for 5000 iterations with $n = 200$, $m = 10$ and the learning rate $\epsilon_t \equiv 10^{-6}$, where the first 2000 iterations were discarded for the burn-in process and the samples generated from the remaining iterations were used for inference. At each iteration, the reversible jump Metropolis–Hastings algorithm (Green, 1995) was used for simulating the models $\gamma_{S_i,n}^{(t)}$, $i = 1, 2, \ldots, m$, with the detail given in the Supplementary Material.

Table 1 summarizes the performance of the algorithm. The variables were selected according to the median posterior probability rule (Barbieri & Berger, 2004), which selects only the variables with a marginal inclusion probability greater than 0.5. The Bayesian estimates of parameters were obtained by

Table 1. *Bayesian variable selection with the extended stochastic gradient Langevin dynamics, reversible jump Metropolis–Hastings, split-and-merge and Bayesian lasso algorithms, where the false selection rate, negative selection rate, mean squared errors for false predictors and mean squared errors for true predictors are reported in averages over* 10 *datasets with standard deviations given in the parentheses, and the CPU time was recorded for one dataset on a Linux machine with a* 3.4 *GHz Intel® Core™i*7-3770 *CPU*

Algorithm	FSR	NSR	MSE$_1$	MSE$_0$	CPU (m)
eSGLD	0 (0)	0 (0)	2.91×10^{-3} (1.90×10^{-3})	1.26×10^{-7} (1.18×10^{-8})	3.3
RJMH	0.50 (0.10)	0.16 (0.042)	1.60×10^{-1} (3.89×10^{-2})	2.64×10^{-5} (8.75×10^{-6})	180.1
SaM	0.05 (0.05)	0.013 (0.013)	1.29×10^{-2} (1.27×10^{-2})	1.01×10^{-6} (1.00×10^{-6})	150.4
B-Lasso	0 (0)	0 (0)	2.32×10^{-4} (3.58×10^{-5})	1.40×10^{-7} (5.08×10^{-9})	32.8

eSGLD, extended stochastic gradient Langevin dynamics; RJMH, reversible jump Metropolis–Hastings; SaM, split-and-merge; B-Lasso, Bayesian lasso; FSR, false selection rate; NSR, negative selection rate; MSE$_0$, mean squared errors for false predictors; MSE$_1$, mean squared errors for true predictors, defined in the Supplementary Material.

averaging over a set of thinned, by a factor of 10, posterior samples. For comparison, some existing algorithms were applied to this example, with the results also given in Table 1.

5. Discussion

To the best of our knowledge, this paper provides the first Bayesian method and theory for high-dimensional discrete parameter estimation with mini-batch samples, while the existing methods work for continuous parameters or very low-dimensional discrete problems only. The proposed algorithm can be used for generalized linear models, but also in other applications in data science, e.g., sparse deep learning and accelerating computation for statistical models/problems where latent variables are involved, such as hidden Markov models, random coefficient models and model-based clustering problems.

Algorithm 1 can be further extended by updating θ using a variant of stochastic gradient Langevin dynamics, such as stochastic gradient Hamiltonian Monte Carlo (Chen et al., 2014), stochastic gradient thermostats (Ding et al., 2014), stochastic gradient Fisher scoring (Ahn et al., 2012) or preconditioned stochastic gradient Langevin dynamics (Li et al., 2016). We expect that the advantages of these variants over stochastic gradient Langevin dynamics can be carried over to the extension.

Acknowledgement

This work was partially supported by the U.S. National Science Foundation and the National Institutes of Health. The authors thank the editor, associate editor and referees for their insightful comments.

Supplementary material

Supplementary material available at *Biometrika* online contains the proof of Lemma 1, the proof of Theorem 2, and further numerical examples.

Appendix

Proof of Theorem 1

Let $\pi_* = \pi(\theta \mid X_N)$ denote the posterior density function of θ, and let $\pi_t = \pi(\theta^{(t)} \mid X_N)$ denote the density of $\theta^{(t)}$ generated by Algorithm 1 at iteration t. We are interested in studying the discrepancy between π_* and π_t in the second-order Wasserstein distance. The following conditions are assumed:

Condition A1. the posterior π_* is strongly log-concave and gradient-Lipschitz: $f(\theta) - f(\theta') - \nabla f(\theta')^{\mathrm{T}}(\theta - \theta') \geqslant \frac{q_N}{2}\|\theta - \theta'\|_2^2\ \forall\theta,\theta' \in \Theta$, $\|\nabla f(\theta) - \nabla f(\theta')\|_2 \leqslant Q_N\|\theta - \theta'\|_2\ \forall\theta,\theta' \in \Theta$, where $f(\theta) = -\log\pi(\theta \mid X_N)$, and $c_0'N \leqslant q_N \leqslant Q_N \leqslant c_0N$ for some positive constants c_0 and c_0';

Condition A2. the posterior π_* has bounded second moments: $\int_\Theta \theta^{\mathrm{T}}\theta\pi_*(\theta)\,\mathrm{d}\theta = O(p)$.

Condition A3. $\max_{S\in\mathcal{S}} E_{X_N}[\|\nabla_\theta \log\pi(\theta \mid \gamma_S, X_N)\|^2 \mid \theta] = O\{N^2(\|\theta\|^2 + p)\}$, where E_{X_N} denotes expectation with respect to the distribution of X_N, and $\mathcal{S}$ denotes the set of all possible models;

Condition A4. let $L_N(\gamma_S, \theta) = \log p(X_N \mid \gamma_S, \theta)/N$, and let $\{L_N^{(i)}(\theta) : i = 1, 2, \ldots, |\mathcal{S}|\}$ be the descending order statistics of $\{L_N(\gamma_S, \theta) : S \in \mathcal{S}\}$. Assume that there exists a constant $\delta > 0$ such that $\inf_{\theta\in\Theta}\{L_N^{(1)}(\theta) - L_N^{(2)}(\theta)\} \geqslant \delta$.

Proof. Part (i). In Algorithm 1, the gradient $\nabla \log\pi(\theta^{(t)} \mid X_N)$ is estimated by running a short Markov chain with a mini-batch of data. Since the initial distribution of the Markov chain might not coincide with its equilibrium distribution, the resulting gradient estimate can be biased. Let $\zeta^{(t)} = \frac{1}{m}\sum_{k=1}^m \nabla_\theta \log\pi(\theta^{(t)} \mid \gamma_{S_k,n}^{(t)}, X_{n,N}^{(t)}) - \nabla\log\pi(\theta^{(t)} \mid X_N)$. Following from Condition A3, we have

$$\|E(\zeta^{(t)} \mid \theta^{(t)})\|^2 = O\left\{\frac{N^2(\|\theta^{(t)}\|^2 + p)}{m^2}\right\}, \quad E\|\zeta^{(t)} - E(\zeta^{(t)} \mid \theta^{(t)})\|^2 = O\left\{\frac{N^2(\|\theta^{(t)}\|^2 + p)}{mn}\right\}.$$

Following from Lemma S2 in the Supplementary Material, if $m \succ p^{1/2}$, $\epsilon \prec 1/N \prec (mn)/(Np)$, and $V = O(p)$ holds, then

$$W_2(\pi_t, \pi_*) = (1-\omega)^t W_2(\pi_0, \pi_*) + O\left(\frac{p^{1/2}}{m}\right) + O\{(\epsilon p)^{1/2}\} + O\left\{\left(\frac{\epsilon Np}{mn}\right)^{1/2}\right\} \to 0 \quad \text{as } t \to \infty \tag{A1}$$

for some $\omega > 0$, since $q_N \asymp N$ and $Q_N \asymp N$ hold by Conditions A1 and A2.

Part (ii). Since $\rho(\theta)$ is α-Lipschitz, we have $|\rho(\theta)| \leqslant \alpha\|\theta\| + C'$ for some constant C'. Further, π_* is strongly log-concave, so $\pi_*(|\rho|) < \infty$, i.e., ρ is π_*-integrable. On the other hand,

$$\begin{aligned}\left\|\int \rho(\theta)\,\mathrm{d}\pi_*(\theta) - \int \rho(\tilde\theta)\,\mathrm{d}\pi_t(\tilde\theta)\right\| &= \|E\rho(\theta) - E\rho(\tilde\theta)\| \leqslant E\|\rho(\theta) - \rho(\tilde\theta)\| \\ \leqslant \alpha E\|\theta - \tilde\theta\|_2 \leqslant \alpha\{E\|\theta - \tilde\theta\|_2^2\}^{1/2} &= \alpha W_2(\pi_*, \pi_t) = o(1)\end{aligned}$$

where θ and $\tilde\theta$ are two random variables whose marginal distributions follow π_* and π_t, respectively, $E(\cdot)$ denotes expectation with respect to the joint distribution of θ and $\tilde\theta$, and $(E\|\theta - \theta_t\|_2^2)^{1/2} = W_2(\pi_*, \pi_t)$. This implies that ρ is also π_t-integrable and $\int \rho(\tilde\theta)\,\mathrm{d}\pi_t(\tilde\theta) \to \int \rho(\theta)\,\mathrm{d}\pi_*(\theta)$ as $t \to \infty$.

Further, by the property of Markov chains, the weak law of large numbers applies and thus $\sum_{t=1}^T \rho(\theta^{(t)})/T - \sum_{t=1}^T \int \rho(\tilde\theta)\,\mathrm{d}\pi_t(\tilde\theta)/T = O_p(T^{-1/2})$. Combining this with the above result leads to

$$\sum_{t=1}^T \rho(\theta^{(t)})/T - \pi_*(\rho) = O_p(T^{-1/2}) + \alpha\sum_{t=1}^T W_2(\pi_*, \pi_t)/T \to 0. \tag{A2}$$

Part (iii). To establish the convergence of $\hat\pi(\gamma_S \mid X_N)$, we define $L_N(\gamma_S, \theta^{(t)}) = \log p(X_N \mid \gamma_S, \theta^{(t)})/N$, $L_n(\gamma_S, \theta^{(t)}) = \log p(X_n^{(t)} \mid \gamma_S, \theta^{(t)})/n$ and $\xi_{n,\,S}^{(t)} = L_n(\gamma_S, \theta^{(t)}) - L_N(\gamma_S, \theta^{(t)})$ for any $S \in \mathcal{S}$. For each S, $\xi_{n,\,S}^{(t)}$ is approximately Gaussian with $E(\xi_{n,\,S}^{(t)}) = 0$ and $\mathrm{var}(\xi_{n,\,S}^{(t)}) = O(1/n)$. Therefore, for any positive ν, with probability $1 - |\mathcal{S}|^{-\nu}$, $\max_S |\xi_{n,\,S}|$ is bounded by $\delta_n := \{(2\nu + 2)\log|\mathcal{S}|/n\}^{1/2} = O[\{(\nu + 1)p/n\}^{1/2}]$ according to the tail probability of the Gaussian. This implies, with high probability, that if S is the most likely model, i.e., $L_N^{(t)}(\gamma_S) = L_N^{(1)}(\theta^{(t)})$, then

$$\begin{aligned}
&|\pi(\gamma_S \mid X_{n,\,N}^{(t)}, \theta^{(t)}) - \pi(\gamma_S \mid X_N, \theta^{(t)})| \\
&\quad = \left| \frac{1}{1 + \sum_{S' \neq S} \mathrm{e}^{N\{L_N(\gamma_{S'}, \theta^{(t)}) - L_N(\gamma_S, \theta^{(t)}) + \xi_{n,S'} - \xi_{n,S}\}}} - \frac{1}{1 + \sum_{S' \neq S} \mathrm{e}^{N\{L_N(X_N \mid \gamma_{S'}, \theta^{(t)}) - L_N(X_N \mid \gamma_S, \theta^{(t)})\}}} \right| \\
&\quad = \frac{\sum_{S' \neq S} \mathrm{e}^{N\{L_N(X_N \mid \gamma_{S'}, \theta^{(t)}) - L_N(X_N \mid \gamma_S, \theta^{(t)}) + b_{S'}\}}}{[1 + \sum_{S' \neq S} \mathrm{e}^{N\{L_N(X_N \mid \gamma_{S'}, \theta^{(t)}) - L_N(X_N \mid \gamma_S, \theta^{(t)}) + b_{S'}\}}]^2} N|\xi_{n,S'} - \xi_{n,S}| \\
&\quad \leqslant (2^p - 1)\, \mathrm{e}^{-N(\delta - 2\delta_n)} N 2\delta_n \leqslant \mathrm{e}^{-N\delta/2} \to 0
\end{aligned}$$

if $\nu p \prec n$, i.e., $\delta_n \prec \delta$, and $N \succ p$, where the second equality follows from the mean value theorem by viewing the $N\{L_N(X_N \mid \gamma_{S'}, \theta^{(t)}) - L_N(X_N \mid \gamma_S, \theta^{(t)})\}$ as the arguments of $\pi(\gamma_S \mid X_N, \theta^{(t)})$, and $b_{S'}$ denotes a value between 0 and $(\xi_{n,S'} - \xi_{n,S})$. Similarly, if S is not the most likely model, then we denote S^* as the most likely model and, by the mean value theorem,

$$\begin{aligned}
&|\pi(\gamma_S \mid X_{n,N}^{(t)}, \theta^{(t)}) - \pi(\gamma_S \mid X_N, \theta^{(t)})| \\
&\quad = \left| \frac{\mathrm{e}^{N\{L_N(\gamma_S, \theta^{(t)}) - L_N(\gamma_{S^*}, \theta^{(t)}) + \xi_{n,S} - \xi_{n,S^*}\}}}{1 + \sum_{S' \neq S} \mathrm{e}^{N\{L_N(\gamma_{S'}, \theta^{(t)}) - L_N(\gamma_{S^*}, \theta^{(t)}) + \xi_{n,S'} - \xi_{n,S^*}\}}} - \frac{\mathrm{e}^{N\{L_N(\gamma_S, \theta^{(t)}) - L_N(\gamma_{S^*}, \theta^{(t)})\}}}{1 + \sum_{S' \neq S} \mathrm{e}^{N\{L_N(\gamma_{S'}, \theta^{(t)}) - L_N(\gamma_{S^*}, \theta^{(t)})\}}} \right| \\
&\quad \leqslant [1 + (2^p - 1)\, \mathrm{e}^{-N(\delta - 2\delta_n)} + \mathrm{e}^{2N\delta_n}]\, \mathrm{e}^{-N(\delta - 2\delta_n)} N 2\delta_n \leqslant \mathrm{e}^{-N\delta/2} \to 0.
\end{aligned}$$

In conclusion, with probability $1 - 1/|\mathcal{S}|^\nu$, $|\pi(\gamma_S \mid X_{n,N}^{(t)}, \theta^{(t)}) - \pi(\gamma_S \mid X_N, \theta^{(t)})| < \exp(-N\delta/2)$ for all S, any iteration t and any $\theta^{(t)} \in \Theta$. Then, one could choose some $\nu = (n/p)^{1/2} \to \infty$ such that $\pi(\gamma_S \mid X_{n,N}^{(t)}, \theta^{(t)}) - \pi(\gamma_S \mid X_N, \theta^{(t)})$ is bounded by $\max_S E|\pi(\gamma_S \mid X_{n,N}^{(t)}, \theta^{(t)}) - \pi(\gamma_S \mid X_N, \theta^{(t)})| \leqslant \exp(-N\delta/2) + 1/|\mathcal{S}|^\nu \to 0$ for any iteration t. Conditioned on $\{\theta^{(t)} : t = 1, 2, \ldots\}$, the $[\pi(\gamma_S \mid X_{n,N}, \theta^{(t)}) - \pi(\gamma_S \mid X_N, \theta^{(t)})]$ are independent and each is bounded by 1, so the law of large numbers applies. Therefore, for any $S \in \mathcal{S}$,

$$\frac{1}{T}\sum_{t=1}^{T} \pi(\gamma_S \mid X_{n,N}^{(t)}, \theta^{(t)}) - \frac{1}{T}\sum_{t=1}^{T} \pi(\gamma_S \mid X_N, \theta^{(t)}) = O_p(T^{-1/2}) + \exp(-N\delta/2) + 1/|\mathcal{S}|^\nu \to 0, \tag{A3}$$

provided $p \prec n \leqslant N$. Since $\{\theta^{(t)} : t = 1, 2, \ldots\}$ forms a time-homogeneous Markov chain whose convergence is measured by (A1), and the function $\pi(\gamma_S \mid X_N, \theta)$ is bounded and continuous in θ,

$$\frac{1}{T}\sum_{t=1}^{T} \pi(\gamma_S \mid X_N, \theta^{(t)}) - \pi(\gamma_S \mid X_N) = O_p(T^{-1/2}) \tag{A4}$$

holds for any $S \in \mathcal{S}$. Combining (A4) with (A3) leads to

$$\frac{1}{T}\sum_{t=1}^{T} \pi(\gamma_S \mid X_{n,N}^{(t)}, \theta^{(t)}) - \pi(\gamma_S \mid X_N) = O_p(T^{-1/2}) + \exp(-N\delta/2) + 1/|\mathcal{S}|^\nu \to 0. \tag{A5}$$

Conditioned on $X_{n,N}^{(t)}$ and $\theta^{(t)}$, by the standard theory of MCMC, $m^{-1}\sum_{i=1}^{m} I(\gamma_S^{(t,i)} = \gamma_S)$ forms a consistent estimator of $\pi(\gamma_S \mid X_{n,N}^{(t)}, \theta^{(t)})$ with an asymptotic bias of $O(1/m)$. Since m is increasing with p and N, the estimator is asymptotically unbiased. Combining this result with (A5) leads to

$$\frac{1}{mT}\sum_{t=1}^{T}\sum_{i=1}^{m} I(\gamma_{S_i,n}^{(t)} = \gamma_S) - \pi(\gamma_S \mid X_N) = O_p(T^{-1/2}) + \exp(-N\delta/2) + 1/|\mathcal{S}|^\nu + O_p(m^{-1/2}), \tag{A6}$$

which converges to 0 as $T \to \infty$ and $N \to \infty$. □

References

Ahn, S., Balan, Korattikara, A. & Welling, M. (2012). Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proc. 29th Int. Conf. Mach. Learn.*, pp. 1771–8

Barbieri, M. & Berger, J. (2004). Optimal predictive model selection. *Ann. Statist.* **32**, 870–97.

Bardenet, R., Doucet, A. & Holmes, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. *Proc. Mach. Learn. Res.* **32**, 405–13.

Bardenet, R., Doucet, A. & Holmes, C. C. (2017). On Markov chain Monte Carlo methods for tall data. *J. Mach. Learn. Res.* **18**, 47:1–47:43.

Betancourt, M. (2015). The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling. *Proc. Mach. Learn. Res.* **37**, 533–40.

Bierkens, J., Fearnhead, P. & Roberts, G. (2019). The zig-zag process and super-efficient Monte Carlo for Bayesian analysis of big data. *Ann. Statist.* **47**, 1288–320.

Bouchard Coté, A., Vollmer, S. & Doucet, A. (2018). The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *J. Am. Statist. Assoc.* **113**, 855–67.

Cappé, O., Moulines, E. & Ryden, T. (2005). *Inference in Hidden Markov Models*. New York: Springer.

Chen, T., Fox, E. B. & Guestrin, C. (2014). Stochastic gradient Hamiltonian Monte Carlo. In *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, pp. II-1683–91.

Chen, Y. & Ghahramani, Z. (2016). Scalable discrete sampling as a multi-armed bandit problem. In *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, pp. 2492–501.

Dalalyan, A. S. & Karagulyan, A. G. (2018). User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *arXiv:*1710.00095.

Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R. D. & Neven, H. (2014). Bayesian sampling using stochastic gradient thermostats. In *Proc. 27th Int. Conf. Neural Inf. Proc. Syst.*, vol. 2, pp. 3203–11.

Fong, E., Lyddon, S. & Holmes, C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *Proc. 36th Int. Conf. Mach. Learn.*, pp. 1952–62

Gibbs, A. & Su, F. (2002). On choosing and bounding probability metrics. *Int. Statist. Rev.* **70**, 419–35.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.

Korattikara, A., Chen, Y. & Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis–Hastings budget. In *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, pp. I-181–9.

Li, C., Chen, C., Carlson, D. E. & Carin, L. (2016). Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Proc. 30th AAAI Conf. Artif. Intel.*, pp. 1788–94.

Li, C., Srivastava, S. & Dunson, D. (2017). Simple, scalable and accurate posterior interval estimation. *Biometrika* **104**, 665–80.

Ma, Y.-A., Chen, T. & Fox, E. B. (2015). A complete recipe for stochastic gradient MCMC. In *Proc. 28th Int. Conf. Neural Inf. Proc. Syst.*, vol. 2, pp. 2917–25.

Maclaurin, D. & Adams, R. P. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. In *Proc. 24th Int. Joint Conf. Artif. Intel.*, pp. 4289–95.

Nemeth, C. & Fearnhead, P. (2019). Stochastic gradient Markov chain Monte Carlo. *arXiv:*1907.06986.

Sato, I. & Nakagawa, H. (2014). Approximation analysis of stochastic gradient Langevin dynamics by using Fokker–Planck equation and Ito process. *PMLR* **32**, 982–90.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. & McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *Int. J. Manag. Sci. Eng. Manag.* **11**, 78–88.

Seita, D., Pan, X., Chen, H. & Canny, J. (2017). An efficient minibatch acceptance test for Metropolis–Hastings. *arXiv:*1610.06848v3.

Srivastava, S., Li, C. & Dunson, D. B. (2018). Scalable Bayes via barycenter in Wasserstein space. *J. Mach. Learn. Res.* **19**, 1–35.

Tanner, M. & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussions). *J. Am. Statist. Assoc.* **82**, 528–40.

Teh, W., Thiery, A. & Vollmer, S. (2016). Consistency and fluctuations for stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.* **17**, 1–33.

Welling, M. & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proc. 28th Int. Conf. Mach. Learn.*, pp. 681–8.

Xue, J. & Liang, F. (2019). Double-parallel Monte Carlo for Bayesian analysis of big data. *Statist. Comp.* **29**, 23–32.

[*Received on* 7 *January* 2019. *Editorial decision on* 3 *December* 2019]