

A Convergence Analysis for A Class of Practical Variance-Reduction Stochastic Gradient MCMC

Changyou Chen^{1*}, Wenlin Wang², Yizhe Zhang³, Qinliang Su⁴ & Lawrence Carin²

¹*SUNY at Buffalo, Buffalo, NY 14260, USA;*

²*Duke University, Durham, NC 27708, USA;*

³*Microsoft Research, Redmond, WA 98052, USA;*

⁴*Sun Yat-sen University, Guangzhou, 510275, P. R. China*

Appendix A Basic Setup for Stochastic Gradient MCMC

Given data $\mathbf{D} = \{\mathbf{d}_1, \dots, \mathbf{d}_N\}$, a generative model

$$p(\mathbf{D} | \boldsymbol{\theta}) = \prod_{i=1}^N p(\mathbf{d}_i | \boldsymbol{\theta}),$$

with model parameter $\boldsymbol{\theta} \in \mathbb{R}^r$, and prior $p(\boldsymbol{\theta})$, we want to compute the posterior distribution:

$$\rho(\boldsymbol{\theta}) \triangleq p(\boldsymbol{\theta} | \mathbf{D}) \propto p(\mathbf{D} | \boldsymbol{\theta})p(\boldsymbol{\theta}) \triangleq e^{-U(\boldsymbol{\theta})},$$

where

$$\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) - \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} \log p(\mathbf{d}_i | \boldsymbol{\theta}). \quad (\text{A1})$$

Consider the SDE:

$$d\mathbf{x}_t = F(\mathbf{x}_t)dt + g(\mathbf{x}_t)d\mathbf{w}_t, \quad (\text{A2})$$

where $\mathbf{x} \in \mathbb{R}^d$ is the state variable, typically $\mathbf{x} \supseteq \boldsymbol{\theta}$ is an augmentation of the model parameter, thus $r \leq d$; t is the time index, $\mathbf{w}_t \in \mathbb{R}^d$ is d -dimensional Brownian motion; functions $F: \mathbb{R}^r \rightarrow \mathbb{R}^d$ and $g: \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ are assumed to satisfy the usual Lipschitz continuity condition [1]. In Langevin dynamics, we have $\mathbf{x} = \boldsymbol{\theta}$ and

$$\begin{aligned} F(\boldsymbol{\theta}_t) &= -\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}_t) \\ g(\boldsymbol{\theta}_t) &= \sqrt{2}. \end{aligned}$$

For the SDE in (A2), the generator \mathcal{L} is defined as:

$$\mathcal{L}\psi \triangleq \frac{1}{2} \nabla \psi \cdot F + \frac{1}{2} g(\boldsymbol{\theta})g(\boldsymbol{\theta})^* : D^2 \psi, \quad (\text{A3})$$

where ψ is a measurable function, $D^k \psi$ means the k -derivative of ψ , $*$ means transpose. $\mathbf{a} \cdot \mathbf{b} \triangleq \mathbf{a}^T \mathbf{b}$ for two vectors \mathbf{a} and \mathbf{b} , $\mathbf{A} : \mathbf{B} \triangleq \text{trace}(\mathbf{A}^T \mathbf{B})$ for two matrices \mathbf{A} and \mathbf{B} . Under certain assumptions, we have that there exists a function ϕ on \mathbb{R}^d such that the following Poisson equation is satisfied [2]:

$$\mathcal{L}\psi = \phi - \bar{\phi}, \quad (\text{A4})$$

where $\bar{\phi} \triangleq \int \phi(\boldsymbol{\theta})\rho(d\boldsymbol{\theta})$ denotes the model average, with ρ being the equilibrium distribution for the SDE (A2).

* Corresponding author (email: cchangyou@gmail.com)

In stochastic gradient Langevin dynamics (SGLD), we update the parameter θ at step l , denoted as θ_l^D , using the following descretized method:

$$\theta_{l+1} = \theta_l - \nabla_{\theta} \tilde{U}_l(\theta_l) h_{l+1} + \sqrt{2h_{l+1}} \zeta_{l+1},$$

where h_{l+1} is the step size, ζ_l a Gaussian random variable with mean 0 and variance 1, $\nabla_{\theta} \tilde{U}_l$ is an unbiased estimate of $\nabla_{\theta} U$ in (A1) with a random minibatch of size n , e.g.,

$$\nabla_{\theta} \tilde{U}_l(\theta_l) = \nabla_{\theta} \log p(\theta_l) + \frac{N}{n} \sum_{i=1}^n \nabla_{\theta} \log p(\mathbf{x}_{\pi_i} | \theta_l), \quad (\text{A5})$$

where $\{\pi_1, \dots, \pi_n\}$ is a subset of a random permutation of $\{1, \dots, N\}$.

In our analysis, we are interested in the *mean square error* (MSE) at iteration L , defined as

$$\text{MSE}_L \triangleq \mathbb{E} \left(\hat{\phi}_L - \bar{\phi} \right)^2,$$

where $\hat{\phi}_L \triangleq \frac{1}{L} \sum_{l=1}^L \phi(\theta_l)$ denotes the sample average, $\bar{\phi}$ is the true posterior average defined in (A4).

In this paper, for the function $f: \mathbb{R}^m \rightarrow \mathbb{R}$ in an \mathcal{L}^p space, i.e., a space of functions for which the p -th power of the absolute value is Lebesgue integrable, we consider the standard norm $\|f\|_p$ defined as ($\|f\|_{\infty}$ is simplified as $\|f\|$):

$$\|f\|_p \triangleq \left(\int_{\mathbb{R}^m} |f(\mathbf{x})|^p d\mathbf{x} \right)^{1/p} < \infty.$$

In order to guarantee well-behaved SDEs and the corresponding numerical integrators, following existing literatures such as [3, 4], we impose the following assumptions.

Assumption 1. The SDE (A2) is ergodic. Furthermore, the solution of (A4) exists, and the solution functional ψ of the Poisson equation (A4) satisfies the following properties:

- ψ and its up to 3th-order derivatives $\mathcal{D}^k \psi$, are bounded by a function \mathcal{V} , i.e., $\|\mathcal{D}^k \psi\| \leq C_k \mathcal{V}^{p_k}$ for $k = (0, 1, 2, 3, 4)$, $C_k, p_k > 0$.
- The expectation of \mathcal{V} on $\{\mathbf{x}_l\}$ is bounded: $\sup_l \mathbb{E} \mathcal{V}^{p_l}(\mathbf{x}_l) < \infty$.
- \mathcal{V} is smooth such that $\sup_{s \in (0,1)} \mathcal{V}^p(s\mathbf{x} + (1-s)\mathbf{y}) \leq C(\mathcal{V}^p(\mathbf{x}) + \mathcal{V}^p(\mathbf{y}))$, $\forall \mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^m, p \leq \max\{2p_k\}$ for some $C > 0$.

Appendix B Proofs of Extended Results for Standard SG-MCMC

First, let us make a clarification for ΔV_l , which will be used through out the paper. According to the definition of ΔV_l , we note that $\Delta V_l \psi = (\nabla_{\theta} U_l(\theta) - \nabla_{\theta} \tilde{U}_l(\theta)) \cdot \nabla \psi$ for the solution functional ψ of the Poisson equation (A4). Since $\|\Delta V_l \psi\| \leq \|\nabla_{\theta} U_l(\theta) - \nabla_{\theta} \tilde{U}_l(\theta)\| \|\nabla \psi\|$, and $\|\nabla \psi\|$ is assumed to be bounded for a test function ψ , we omit the operator ∇ in our following analysis (which only contributes to a constant), manifesting a slight abuse of notation for conciseness.

The proofs of Lemma 2 and Theorem 1 are closely related. We will first prove Theorem 1, the proof for Lemma 2 then directly follows.

Proof. [Proof of Theorem 1]

Let $\alpha_{li} = \nabla_{\theta} \log p(\mathbf{d}_i | \theta_l)$, and

$$z_i = \begin{cases} 1 & \text{if data } i \text{ is selected} \\ 0 & \text{otherwise} \end{cases},$$

then we have

$$\Delta V_l = \sum_{i=1}^N \mathbb{E} \alpha_{li} \left(1 - \frac{N}{n} z_i \right) \rightarrow \mathbb{E} |\Delta V_l|^2 = \sum_{i=1}^N \sum_{j=1}^N \mathbb{E} \alpha_{li} \mathbb{E} \alpha_{lj} \left(1 - \frac{N}{n} z_i \right) \left(1 - \frac{N}{n} z_j \right)$$

Since

$$\begin{aligned} \mathbb{E} z_i &= \frac{1}{N} + \frac{N-1}{N} \frac{1}{N-1} + \dots + \frac{N-1}{N} \frac{N-2}{N-1} \dots \frac{N-m+1}{N-m+2} \frac{1}{N-m+1} \\ &= \frac{n}{N}, \end{aligned}$$

we have $\mathbb{E} \Delta V_l = 0$, i.e., $\nabla \tilde{U}_l(\theta)$ is an unbiased estimate of $\nabla U(\theta)$.

In addition, we have

$$\begin{aligned} & \mathbb{E} \left(1 - \frac{N}{n} z_i \right) \left(1 - \frac{N}{n} z_j \right) \\ &= \mathbb{E} \left[1 - \frac{N}{n} z_i - \frac{N}{n} z_j + \frac{N^2}{n^2} z_i z_j \right] \\ &= 1 - 2 \frac{N}{n} \frac{n}{N} + \frac{N^2}{n^2} \mathbb{E} z_i z_j \end{aligned}$$

1) Strictly speaking, θ should be indexed by ‘‘time’’ instead of ‘‘step’’, i.e., $\theta_{\sum_{l=1}^t h_l}$, instead of θ_l . We adopt the later for notation simplicity in the following. This applies for the general case of \mathbf{x} .

$$= \frac{N^2}{n^2} \mathbb{E} z_i z_j - 1.$$

When $i = j$,

$$\begin{aligned} \mathbb{E} \left(1 - \frac{N}{n} z_i \right) \left(1 - \frac{N}{n} z_j \right) &= \frac{N^2}{n^2} \mathbb{E} z_i^2 - 1 \\ &= \frac{N^2}{n^2} \mathbb{E} z_i - 1 = \frac{N}{n} - 1. \end{aligned}$$

When $i \neq j$, because

$$\mathbb{E} z_i z_j = p(i \text{ selected})p(j \text{ selected} | i \text{ selected}) = \frac{n}{N} \frac{n-1}{N-1}.$$

We have

$$\mathbb{E} \left(1 - \frac{N}{n} z_i \right) \left(1 - \frac{N}{n} z_j \right) = \frac{N^2}{n^2} \mathbb{E} z_i z_j - 1 = \frac{N}{n} \frac{n-1}{N-1} - 1.$$

As a result,

$$\begin{aligned} &\mathbb{E} |\Delta V_l|^2 \\ &= \left(\sum_{i=1}^N \mathbb{E} \alpha_{li}^2 \right) \left(\frac{N}{n} - 1 \right) + 2 \sum_{i < j} \mathbb{E} \alpha_{li} \alpha_{lj} \left(\frac{N}{n} \frac{n-1}{N-1} - 1 \right) \\ &= \left(\frac{N}{n} - 1 \right) \sum_{i,j} \mathbb{E} \alpha_{li} \alpha_{lj} + 2 \sum_{i < j} \mathbb{E} \alpha_{li} \alpha_{lj} \left(\frac{N}{n} \frac{n-1}{N-1} - \frac{N}{n} \right) \\ &= \left(\frac{N}{n} - 1 \right) \sum_{i,j} \mathbb{E} \alpha_{li} \alpha_{lj} - 2 \sum_{i < j} \mathbb{E} \alpha_{li} \alpha_{lj} \frac{N}{n} \frac{N-n}{N-1} \\ &= \frac{(N-n)N^2}{n} \left(\frac{1}{N^2} \sum_{i,j} \mathbb{E} \alpha_{li} \alpha_{lj} - \frac{2}{N(N-1)} \sum_{i < j} \mathbb{E} \alpha_{li} \alpha_{lj} \right) \\ &\triangleq \frac{(N-n)N^2}{n} \Gamma_l. \end{aligned} \tag{B1}$$

Because we assume using a 1st-order numerical integrator, according to Lemma 1, and combining (B1) from above, we have the bound for the MSE $\mathbb{E} \left(\hat{\phi}_L - \bar{\phi} \right)^2$ as:

$$\mathbb{E} \left(\hat{\phi}_L - \bar{\phi} \right)^2 \leq C \left(\frac{(N-n)N^2 \Gamma_M}{nL} + \frac{1}{Lh} + h^2 \right).$$

Proof. [Proof of Lemma 2] The lemma follows directly from (B1) and the fact that

$$\mathbb{E} |\Delta V_l|^2 \geq 0.$$

Proof. [Proof of the optimal MSE bound of Theorem 1]

From the assumption, we have

$$T \propto nL. \tag{B2}$$

The MSE bounded is obtained by directly substituting (B2) into the MSE bound in Lemma 2, resulting in

$$\text{MSE: } \mathbb{E} \left(\hat{\phi}_L - \bar{\phi} \right)^2 \leq C \left(\frac{(N-n)N^2 \Gamma_M}{T} + \frac{n}{Th} + h^2 \right)$$

After optimizing the above bound over h , we have

$$\mathbb{E} \left(\hat{\phi}_L - \bar{\phi} \right)^2 \leq C \left(\frac{(N-n)N^2 \Gamma_M}{T} + \frac{n^{2/3}}{T^{2/3}} \right). \tag{B3}$$

Proof. [Proof of Corollary 1]

To examine the property of the MSE bound (B3) w.r.t. n , we first note that the derivative can be written as:

$$f \triangleq \frac{\partial}{\partial n} \mathbb{E} \left(\hat{\phi}_L - \bar{\phi} \right)^2 = O \left(\frac{2}{3T^{2/3}n^{1/3}} - \frac{\Gamma_M N^2}{T} \right).$$

As a result, we have the following three cases:

1) When $f < 0$, i.e., the bound is decreasing when n increasing, we have $T < \frac{27}{8} \Gamma_M^3 N^6 n$. Because n is in the range of $[1, N]$, and we require $f < 0$ for all n 's, the minimum value of $\frac{27}{8} \Gamma_M^3 N^6 n$ is obtained when taking $n = 1$. Consequently, we have that when $T < \frac{27}{8} \Gamma_M^3 N^6$, the optimal MSE bound (B3) is decreasing w.r.t. n . The minimum MSE bound is thus achieved at $n = N$. This case corresponds to the limited-computation-budget case.

2) When $f > 0$, *i.e.*, the bound is increasing when n increasing, we have $T > \frac{27}{8}\Gamma_M^3 N^6 n$. Because n is in the range of $[1, N]$, and we require $f > 0$ for all n 's, the maximum value of $\frac{27}{8}\Gamma_M^3 N^6 n$ is obtained when taking $n = N$. Consequently, we have that when $T > \frac{27}{8}\Gamma_M^3 N^7$, the optimal MSE bound (B3) is increasing w.r.t. n . The minimum MSE bound is thus achieved at $n = 1$. This case corresponds to the long-run case (computational budget is large enough).

3) When the computational budget is in between the above two cases, the optimal MSE bound (B3) first increases then decreases w.r.t. n in range $[1, N]$. The optimal MSE bound is thus obtained either at $n = 1$ or at $n = N$, depending on (N, T, Γ_M) .

Appendix C Proofs of theorems for vrSG-MCMC

Proof. [Proof of Lemma 3] From the definitions, we know that α_{li} is the same as β_{li} except evaluating on different model parameters, denoted as θ_l and $\tilde{\theta}_l$, respectively. Note that $\tilde{\theta}_l$ is an *outdated* version of θ_l , with difference at most m . The proof of Lemma 3 is then an application of a lemma from [5], which is stated in Lemma 1 below.

Lemma 1 (Lemma 8 in [5]). Let θ_l and $\tilde{\theta}_l$ be two parameters where $\tilde{\theta}_l$ is τ -step older than θ_l , then we have

$$\left\| \mathbb{E} \left(\nabla_{\theta} \log p(\mathbf{d} | \theta_l) - \nabla_{\theta} \log p(\mathbf{d} | \tilde{\theta}_l) \right) \right\| = O(\tau h).$$

Based on the definitions in Algorithm 1, we can consider β_{li} as an outdated version of α_{li} , with time difference m . As a result, Lemma 3 follows by replacing τ with m in Lemma 1.

The following is a formal proof of the unbiasedness of $\nabla_{\theta} \tilde{U}(\theta_l)$, stated in the ‘‘Convergence rate’’ section in the main text.

Proof. [Proof of the unbiasedness of $\nabla_{\theta} \tilde{U}(\theta_l)$]

First note that in variance reduction, the following stochastic gradient is used:

$$\begin{aligned} \nabla_{\theta} \tilde{U}(\theta_l) &= \frac{N}{n_2} \sum_{i=1}^N \left(\nabla_{\theta} \log p(\mathbf{x}_i | \theta_l) - \nabla_{\theta} \log p(\mathbf{x}_i | \tilde{\theta}_l) \right) z_i \\ &\quad + \frac{N}{n_1} \sum_{i=1}^N \sum_{i=1}^N \nabla_{\theta} \log p(\mathbf{x}_i | \tilde{\theta}_l) b_i. \end{aligned} \quad (\text{C1})$$

$$\Delta V_l = \sum_{i=1}^N \alpha_{li} \left(1 - \frac{N}{n_2} z_i \right) + \sum_{i=1}^N \beta_{li} \left(\frac{N}{n_2} z_i - \frac{N}{n_1} b_i \right). \quad (\text{C2})$$

Because $\mathbb{E} z_i = \frac{n_2}{N}$, $\mathbb{E} b_i = \frac{n_1}{N}$, it is easy to verify that $\mathbb{E} \Delta V_l = 0$. As a result, the unbiasedness holds.

Appendix D Proof of Theorem 2

Before proving Theorem 2, let us first simplify $\mathbb{E} \|\Delta V_l\|^2$ in the MSE bound. In the following, we decompose it into several terms which can be simplified separately. Our goal is to show that the proposed vrSG-MCMC algorithm induces a smaller $\mathbb{E} \|\Delta V_l\|^2$ term, thus leading to a faster convergence rate. Note we can rewrite ΔV_l in terms of $\{\alpha_{li}, \beta_{li}, z_i, b_i\}$ as:

$$\Delta V_l = \sum_{i=1}^N \alpha_{li} \left(1 - \frac{N}{n_2} z_i \right) + \sum_{i=1}^N \beta_{li} \left(\frac{N}{n_2} z_i - \frac{N}{n_1} b_i \right).$$

Consequently, we have

$$\begin{aligned} \mathbb{E} \|\Delta V_l\|^2 &= \underbrace{\sum_{i,j} \mathbb{E} \alpha_{li}^T \alpha_{lj} \left(1 - \frac{N}{n_2} z_i \right) \left(1 - \frac{N}{n_2} z_j \right)}_{A_l} \\ &\quad + \underbrace{\sum_{i,j} \mathbb{E} \beta_{li}^T \beta_{lj} \left(\frac{N}{n_2} z_i - \frac{N}{n_1} b_i \right) \left(\frac{N}{n_2} z_j - \frac{N}{n_1} b_j \right)}_{B_l} \\ &\quad + 2 \underbrace{\sum_{i,j} \mathbb{E} \alpha_{li}^T \beta_{lj} \left(1 - \frac{N}{n_2} z_i \right) \left(\frac{N}{n_2} z_j - \frac{N}{n_1} b_j \right)}_{C_l}. \end{aligned} \quad (\text{D1})$$

Now (D1) can be further simplified by summing over all the binary random variables $\{z_i\}$ and $\{b_i\}$. After summing out the binary random variables $\{z_i, b_i\}$, we arrive formula summarized in the following proposition:

Proposition 1. The terms A_l , B_l and C_l in (D1) can be simplified as:

$$\begin{aligned} A_l &= \left(\frac{N}{n_2} - 1 \right) \sum_{ij} \mathbb{E} \alpha_{li}^T \alpha_{lj} - 2 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} \mathbb{E} \alpha_{li}^T \alpha_{lj} \\ B_l &= \left(\frac{N}{n_2} + \frac{N}{n_1} - 2 \right) \sum_{ij} \mathbb{E} \beta_{li}^T \beta_{lj} \end{aligned}$$

$$\begin{aligned}
& -2 \left(\frac{N(N-n_2)}{n_2(N-1)} + \frac{N(N-n_1)}{n_1(N-1)} \right) \sum_{i < j} \mathbb{E} \beta_{li}^T \beta_{lj} \\
C_l & = 2 \left(1 - \frac{N}{n_2} \right) \sum_{ij} \mathbb{E} \alpha_{li}^T \beta_{lj} + 4 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} \mathbb{E} \alpha_{li}^T \beta_{lj} .
\end{aligned}$$

Proof. [Proof of Proposition 1]

First, for the A_l term, from the proof of Theorem 1, we know that

$$A_l = \left(\frac{N}{n_2} - 1 \right) \sum_{ij} \mathbb{E} \alpha_{li} \alpha_{lj} - 2 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} \mathbb{E} \alpha_{li} \alpha_{lj} ,$$

which is the value of $\mathbb{E} \|\Delta V_l\|^2$ for standard SG-MCMC.

The derivations for B_l and C_l go as follows. For B_l , we have

$$\begin{aligned}
& \mathbb{E} \left(\frac{N}{n_2} z_i - \frac{N}{n_1} b_i \right) \left(\frac{N}{n_2} z_j - \frac{N}{n_1} b_j \right) \\
& = \mathbb{E} \left(\frac{N^2}{n_2^2} z_i z_j + \frac{N^2}{n_1^2} b_i b_j - \frac{N^2}{n_1 n_2} z_i b_j - \frac{N^2}{n_1 n_2} b_i z_j \right) \\
& = \mathbb{E} \left(\frac{N^2}{n_2^2} z_i z_j + \frac{N^2}{n_1^2} b_i b_j - 2 \right) .
\end{aligned}$$

If $i = j$,

$$\mathbb{E} \left(\frac{N^2}{n_2^2} z_i z_j + \frac{N^2}{n_1^2} b_i b_j - 2 \right) = \mathbb{E} \left(\frac{N^2}{n_2^2} z_i + \frac{N^2}{n_1^2} b_i - 2 \right) = \frac{N}{n_2} + \frac{N}{n_1} - 2 .$$

If $i \neq j$,

$$\begin{aligned}
& \mathbb{E} \left(\frac{N^2}{n_2^2} z_i z_j + \frac{N^2}{n_1^2} b_i b_j - 2 \right) \\
& = \frac{N^2}{n_2^2} \frac{n_2}{N} \frac{n_2-1}{N-1} + \frac{N^2}{n_1^2} \frac{n_1}{N} \frac{n_1-1}{N-1} - 2 \\
& = \frac{N}{n_2} \frac{n_2-1}{N-1} + \frac{N}{n_1} \frac{n_1-1}{N-1} - 2 .
\end{aligned}$$

$$\begin{aligned}
B_l & = \left(\frac{N}{n_2} + \frac{N}{n_1} - 2 \right) \left(\sum_i \mathbb{E} \beta_i^2 \right) + 2 \left(\frac{N}{n_2} \frac{n_2-1}{N-1} + \frac{N}{n_1} \frac{n_1-1}{N-1} - 2 \right) \left(\sum_{i < j} \mathbb{E} \beta_{li} \beta_{lj} \right) \\
& = \left(\frac{N}{n_2} + \frac{N}{n_1} - 2 \right) \sum_{ij} \mathbb{E} \beta_{li} \beta_{lj} + 2 \left(\frac{N}{n_2} \frac{n_2-1}{N-1} + \frac{N}{n_1} \frac{n_1-1}{N-1} - \frac{N}{n_2} - \frac{N}{n_1} \right) \sum_{i < j} \mathbb{E} \beta_{li} \beta_{lj} \\
& = \left(\frac{N}{n_2} + \frac{N}{n_1} - 2 \right) \sum_{ij} \mathbb{E} \beta_{li} \beta_{lj} + 2 \left(\frac{N(N-n_2)}{n_2(N-1)} + \frac{N(N-n_1)}{n_1(N-1)} \right) \sum_{i < j} \mathbb{E} \beta_{li} \beta_{lj}
\end{aligned}$$

Similarly, for C_l , we have

$$\begin{aligned}
C_l & = 2 \mathbb{E} \sum_i \sum_j \alpha_{li} \beta_{lj} \left(\frac{N}{n_2} z_j - \frac{N^2}{n_2^2} z_i z_j - \frac{N}{n_1} b_j + \frac{N^2}{n_1 n_2} z_i b_j \right) \\
& = 2 \mathbb{E} \sum_i \alpha_{li} \beta_{lj} \left(1 - \frac{N}{n_2} \right) + 2 \mathbb{E} \sum_{i \neq j} \alpha_{li} \beta_{lj} \left(1 - \frac{N}{n_2} \frac{n_2-1}{N-1} \right) \\
& = 2 \sum_{ij} \mathbb{E} \alpha_{li} \beta_{lj} \left(1 - \frac{N}{n_2} \right) + 4 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} \mathbb{E} \alpha_{li} \beta_{lj} .
\end{aligned}$$

This completes the proof.

The following derivations verify an intuition: with larger minibatch size n_1 , we can get smaller MSEs. This is not directly relevant to the proof of Theorem 2. Readers can choose to skip this part without affecting the flow of the proof.

To show that, let's first look at the term $B_l + C_l$ defined above. We have that

$$\begin{aligned}
B_l + C_l & = \left(\frac{N}{n_1} + \frac{N}{n_2} - 2 \right) \sum_{ij} \mathbb{E} \beta_{li} \beta_{lj} \\
& - 2 \sum_{i < j} \mathbb{E} \beta_{li} \beta_{lj} \left(\frac{N(N-n_2)}{n_2(N-1)} + \frac{N(N-n_1)}{n_1(N-1)} \right) + 2 \left(1 - \frac{N}{n_2} \right) \sum_{ij} \mathbb{E} \alpha_{li} \beta_{lj} + 4 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} \mathbb{E} \alpha_{li} \beta_{lj}
\end{aligned}$$

When $n_1 = N$, the case of using the whole data to calculate *old* gradient \tilde{g} [6], we have

$$B_l + C_l$$

$$\begin{aligned}
&= \left(\frac{N}{n_2} - 1\right) \sum_{ij} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} - 2 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} + 2 \left(1 - \frac{N}{n_2}\right) \sum_{ij} \mathbb{E} \boldsymbol{\alpha}_{li} \boldsymbol{\beta}_{lj} + 4 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\alpha}_{li} \boldsymbol{\beta}_{lj} \\
&= \left(\frac{N}{n_2} - 1\right) \sum_{ij} (\mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} - 2 \mathbb{E} \boldsymbol{\alpha}_{li} \boldsymbol{\beta}_{lj}) + 2 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} (2 \mathbb{E} \boldsymbol{\alpha}_{li} \boldsymbol{\beta}_{lj} - \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj}) \\
&= \frac{(N-n_2)N^2}{n_2} \left[\frac{1}{N^2} \sum_{ij} (\mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} - 2 \mathbb{E} \boldsymbol{\alpha}_{li} \boldsymbol{\beta}_{lj}) + \frac{2}{N(N-1)} \sum_{i < j} (2 \mathbb{E} \boldsymbol{\alpha}_{li} \boldsymbol{\beta}_{lj} - \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj}) \right] \triangleq M_{BC}
\end{aligned}$$

When $n_1 \neq N$, we have

$$\begin{aligned}
&B_l + C_l \\
&= M_{BC} + \frac{N-n_1}{n_1} \sum_{ij} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} - 2 \frac{N(N-n_1)}{n_1(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} \\
&= M_{BC} + \frac{(N-n_1)N^2}{n_1} \left[\frac{1}{N^2} \sum_{ij} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} - \frac{2}{N(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} \right].
\end{aligned}$$

According to Lemma 2, we have that $\left[\frac{1}{N^2} \sum_{ij} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} - \frac{2}{N(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj}\right] \geq 0$. As a result, the value of $B_l + C_l$ in the case of $n_1 \neq N$ is larger than that in the case of $n_1 = N$, resulting in a larger MSE bound.

Now it is ready to prove Theorem 2.

Proof. [Proof of Theorem 2]

Note that term A_l corresponds to the $\mathbb{E} \Delta V_l$ term in standard SG-MCMC, where no variance reduction is performed. As a result, in order to prove that vrSG-MCMC induces a lower MSE bound, what remains to be shown is to prove $B_l + C_l \leq 0$.

First, let us simplify term C_l , which results in:

$$\begin{aligned}
C_l &= 2 \left(1 - \frac{N}{n_2}\right) \sum_{ij} \mathbb{E} \boldsymbol{\alpha}_{li} \boldsymbol{\beta}_{lj} + 4 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\alpha}_{li} \boldsymbol{\beta}_{lj} \\
&= 2 \left(1 - \frac{N}{n_2}\right) \sum_{ij} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} + 4 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} + 2 \left(1 - \frac{N}{n_2}\right) N \cdot O(mh) + 4 \frac{N(N-n_2)}{n_2(N-1)} \frac{N(N-1)}{2} \cdot O(mh) \\
&= 2 \left(1 - \frac{N}{n_2}\right) \sum_{ij} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} + 4 \frac{N(N-n_2)}{n_2(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} + O(mh),
\end{aligned}$$

where the second equality is obtained by applying the independence property of $\boldsymbol{\alpha}_{li}$ and $\boldsymbol{\beta}_{lj}$, as well as the result from Lemma 3. Consequently, $B_l + C_l$ can be simplified as

$$B_l + C_l = \left(\frac{N}{n_1} - \frac{N}{n_2}\right) \sum_{ij} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} - 2 \left(\frac{N(N-n_2)}{n_2(N-1)} - \frac{N(N-n_1)}{n_1(N-1)}\right) \sum_{i < j} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} + O(mh)$$

By substituting the above formula in to the MSE bound in Lemma 1, we have that:

$$\mathbb{E} (\hat{\phi}_L - \bar{\phi})^2 = O\left(\frac{A_M}{L} + \frac{1}{Lh} + h^{2K} + \frac{mh}{L} - \frac{\lambda_M}{L}\right).$$

To further simplify the $B_l + C_l$ term, we have

$$\begin{aligned}
B_l + C_l &= O(mh) + \frac{N^3(n_2-n_1)}{n_1 n_2} \frac{1}{N^2} \sum_{ij} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} - 2 \frac{N^2(N-n_2)n_1 - N^2(N-n_1)n_2}{n_1 n_2} \frac{1}{N(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} \\
&= \frac{N^3(n_2-n_1)}{n_1 n_2} \left(\frac{1}{N^2} \sum_{ij} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} - 2 \frac{1}{N(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} \right) + O(mh)
\end{aligned}$$

According to Lemma 2, $\left[\frac{1}{N^2} \sum_{ij} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj} - 2 \frac{1}{N(N-1)} \sum_{i < j} \mathbb{E} \boldsymbol{\beta}_{li} \boldsymbol{\beta}_{lj}\right] \geq 0$. Consequently, we have $B_l + C_l \leq 0$ up to an order of $O(mh)$. This completes the proof of $\lambda_M \geq 0$.

Appendix E Discussion of the Theoretical Results of Dubey et al. 2016

[6] proved the following MSE bound for SVRG-LD, by extending results of the standard SG-MCMC [4]:

$$\mathbb{E} (\hat{\phi}_L - \bar{\phi})^2 = O\left(\frac{N^2 \min\{2\sigma^2, m^2(D^2 h^2 \sigma^2 + hd)\}}{nL} + \frac{1}{Lh} + h^2\right), \quad (\text{E1})$$

where (d, D, σ) are constants related to the data and the true posterior. Using similar techniques (shown in the paper), the MSE bound for SGLD is given by

$$\mathbb{E} \left(\hat{\phi}_L - \bar{\phi} \right)^2 = O \left(\frac{N^2 \sigma^2}{nL} + \frac{1}{Lh} + h^2 \right). \quad (\text{E2})$$

From the proof of their theorem (eq. 13 in their appendix), we note that the constant “2” inside the “min” in (E1) is not negligible when comparing to the bound for SGLD. As a result, the bound associated with this term is strictly larger than the bound for SGLD. This means that compared with SGLD, the MSE bound for SVRG-LD should be written in the form of

$$\mathbb{E} \left(\hat{\phi}_L - \bar{\phi} \right)^2 = O \left(\frac{N^2 m^2 (D^2 h^2 \sigma^2 + hd)}{nL} + \frac{1}{Lh} + h^2 \right). \quad (\text{E3})$$

As a result, the comparison between (E3) and (E2) becomes more complicated, because it now depends on other parameters such as the stepsize. It is thus not clear if SVRG-LD would improve the MSE bound of SGLD.

In contrast, our theoretical results (Theorem 2) guarantee an improvement of vrSG-MCMC over the correspond SG-MCMC, which is a stronger result than that in [6].

Appendix F Additional Experimental Results

Appendix F.1 Supplemental results on logistic regression and deep learning

We plot the corresponding results in terms of number of passes through data versus training error/loss in Figure F1, F2, F3 and F4.

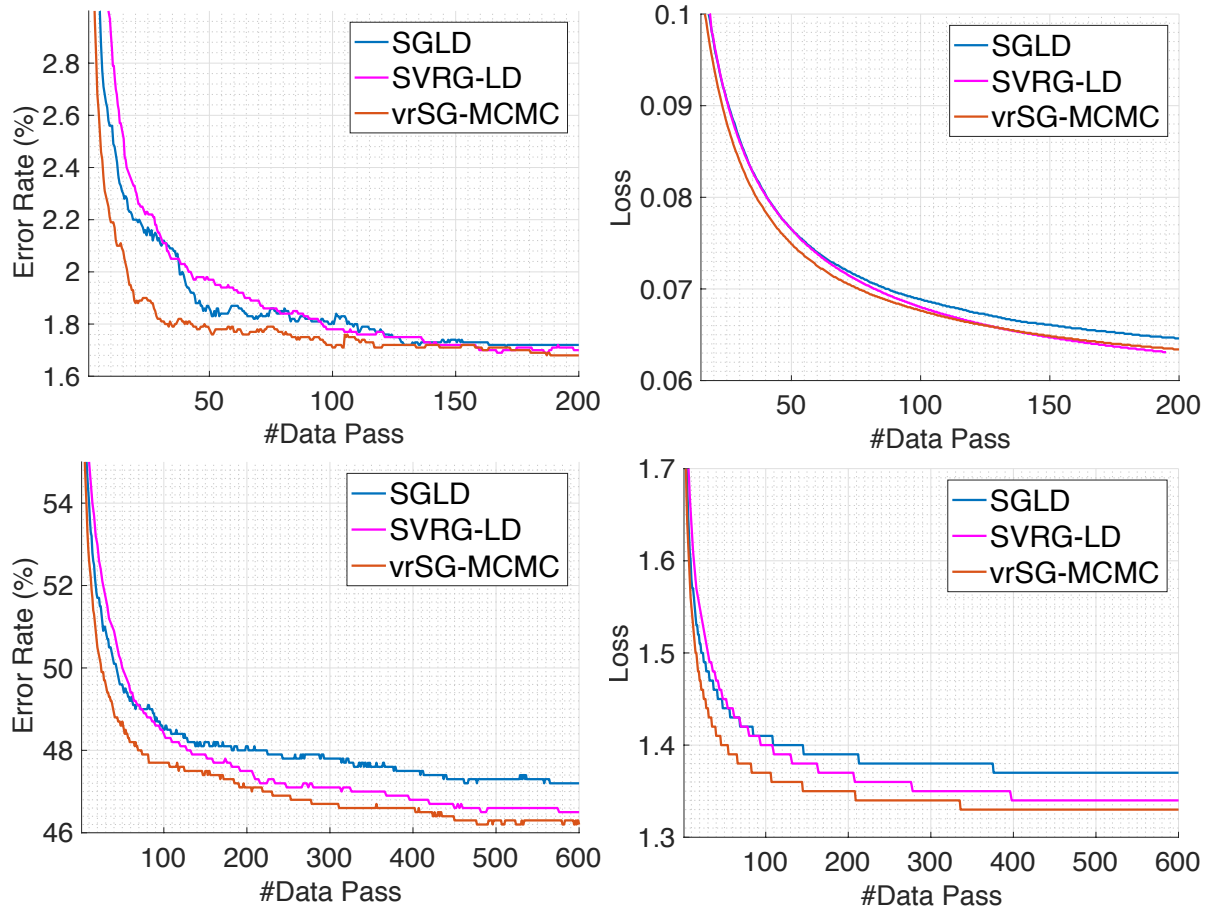


Figure F1 Number of passes through data vs. testing error (left) / loss (right) on MNIST (top) and CIFAR-10 (bottom) datasets.

References

- 1 A. P. Ghosh. *Backward and Forward Equations for Diffusion Processes*. Wiley Encyclopedia of Operations Research and Management Science, 2011.

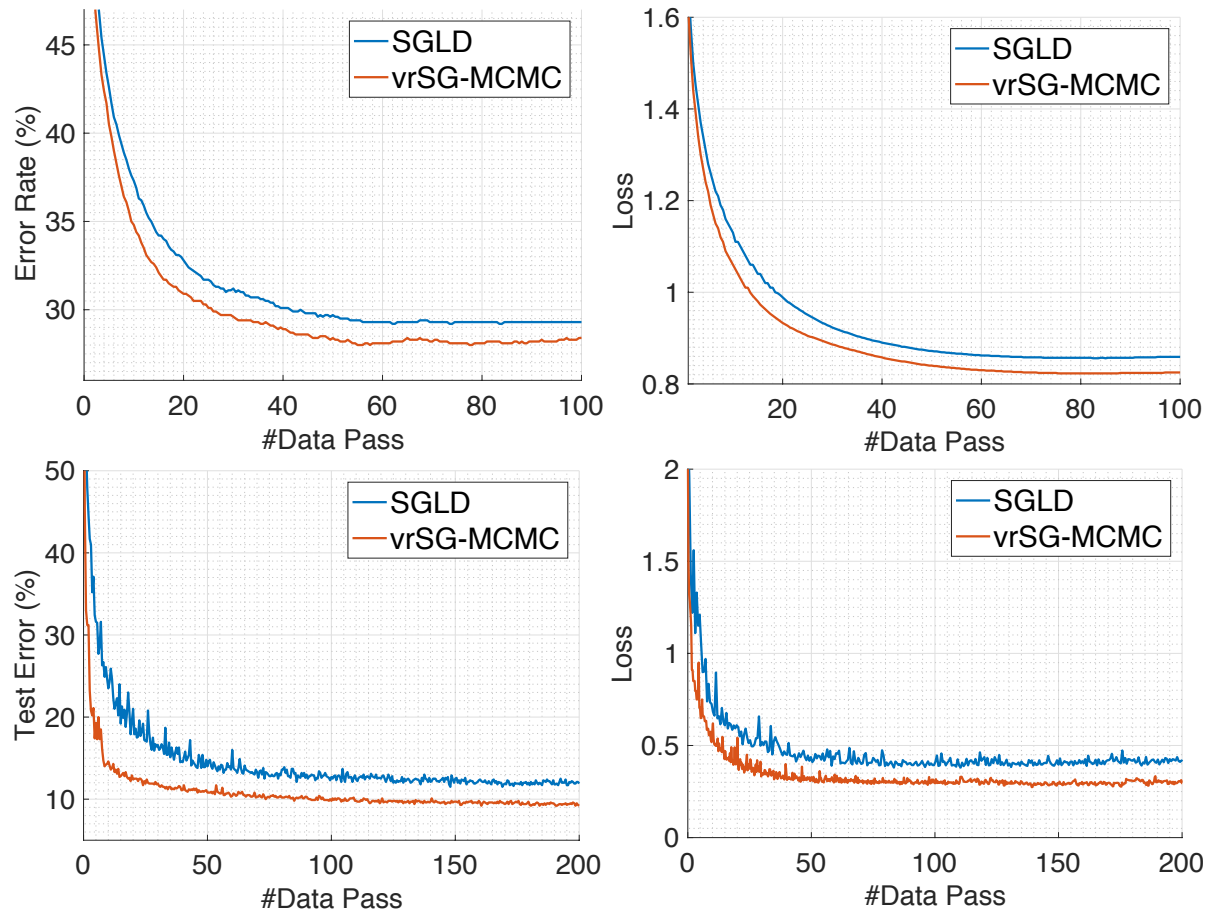


Figure F2 Number of passes through data vs. testing error (left) / loss (right) with CNN-4 (top) and ResNet (bottom) on CIFAR-10.

- 2 J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov. Construction of numerical time-average and stationary measures via Poisson equations. *SIAM J. NUMER. ANAL.*, 48(2):552–577, 2010.
- 3 S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. Exploration of the (Non-)Asymptotic bias and variance of stochastic gradient Langevin dynamics. *JMLR*, 2016.
- 4 C. Chen, N. Ding, and L. Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *NIPS*, 2015.
- 5 C. Chen, N. Ding, C. Li, Y. Zhang, and L. Carin. Stochastic gradient MCMC with stale gradients. In *NIPS*, 2016.
- 6 A. Dubey, S. J. Reddi, B. Póczos, A. J. Smola, and E. P. Xing. Variance reduction in stochastic gradient Langevin dynamics. In *NIPS*, 2016.

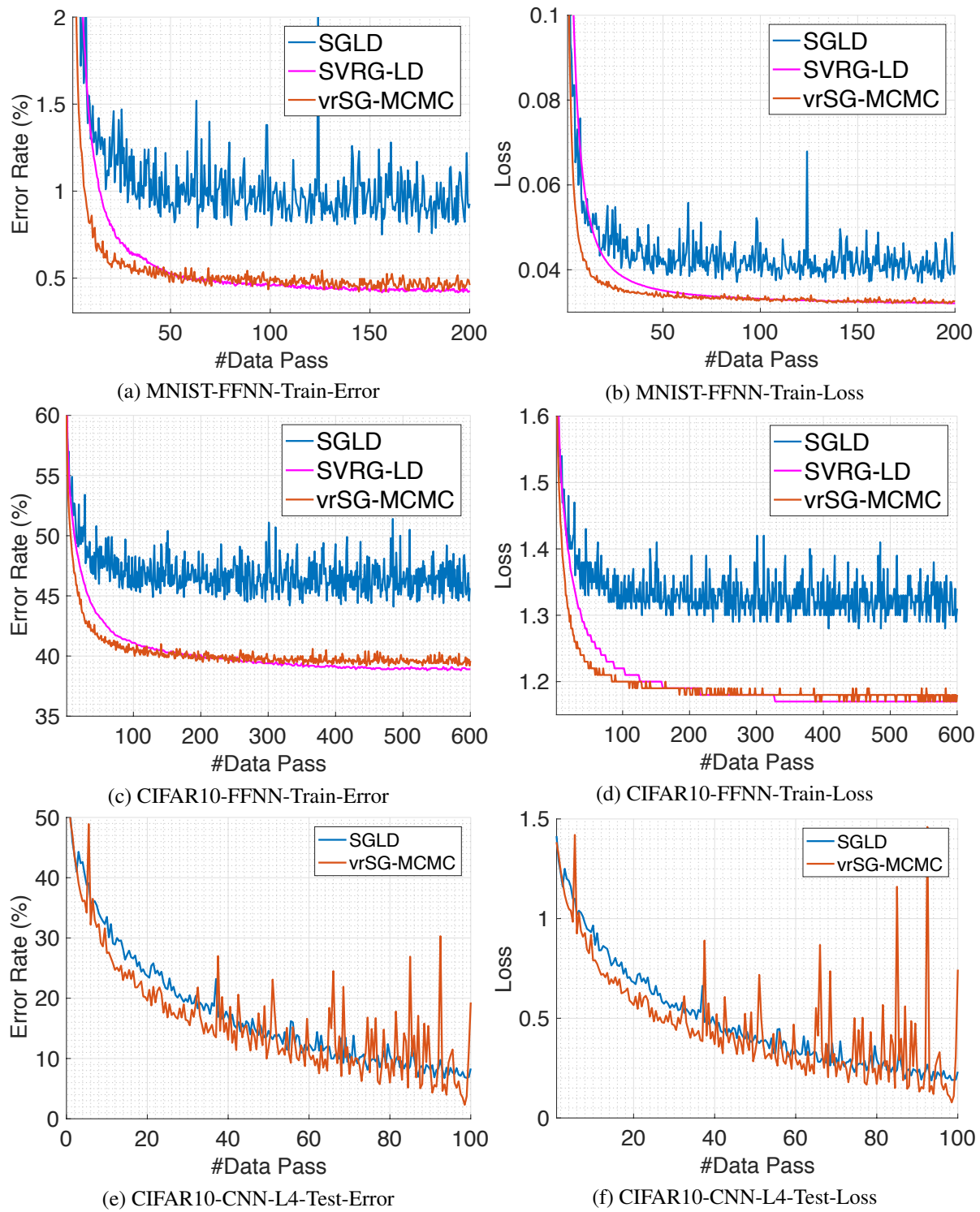


Figure F3 Number of passes through data vs. training error / loss on MNIST and Cifar-10 datasets.

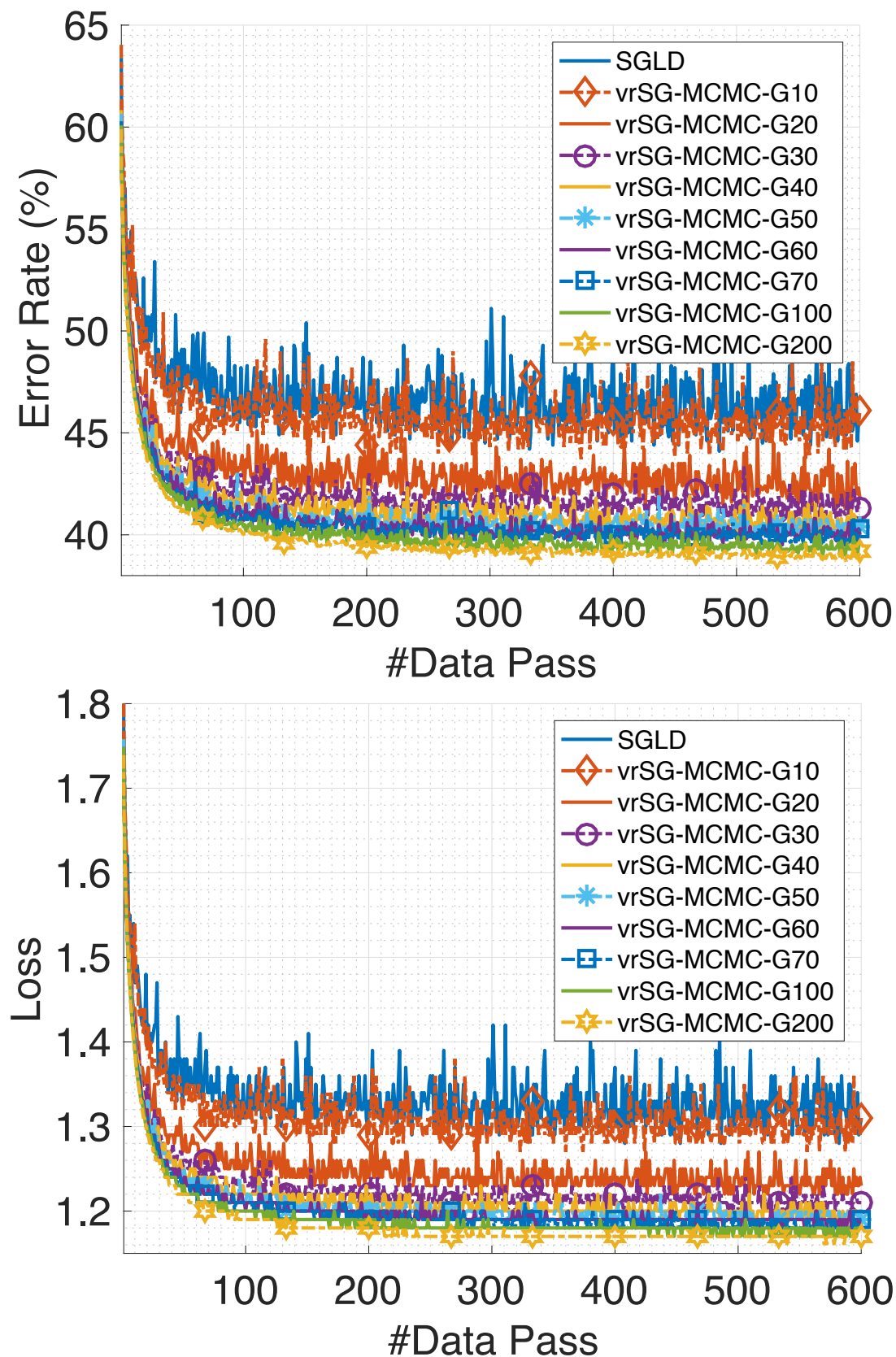


Figure F4 Number of passes through data vs. training errors (left) / loss (right) on the CIFAR-10 dataset. All are with varying n_1 values.