Supplementary Material of Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks

Chunyuan Li[1], **Changyou Chen**[1], **David Carlson**[2] **and Lawrence Carin**[1]
[1]Department of Electrical and Computer Engineering, Duke University
[2]Department of Statistics and Grossman Center, Columbia University
chunyuan.li@duke.edu, cchangyou@gmail.com, david.edwin.carlson@gmail.com, lcarin@duke.edu

A. The proof for main theorem

In (Chen, Ding, and Carin 2015), the authors provide the convergence property for general SG-MCMC, here we follow their assumptions and proof techniques, with specific treatment on the 1st-order numerical integrator, and the case of preconditioner.

Details on the assumption

Before the proof, we detail the assumptions needed for Theorem 1. For pSGLD, its associated Stochastic Differential Equation (SDE) has an invariant measure $\rho(\boldsymbol{\theta})$, the posterior average is defined as: $\bar{\phi} \triangleq \int_{\mathcal{X}} \phi(\boldsymbol{\theta})\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ for some test function $\phi(\boldsymbol{\theta})$ of interest. Given samples $(\boldsymbol{\theta}_t)_{t=1}^T$ from pSGLD, we use the *sample average* $\hat{\phi}$ to approximate $\bar{\phi}$. In the analysis, we define a functional ψ that solves the following *Poisson Equation*:

$$\mathcal{L}\psi(\boldsymbol{\theta}_t) = \phi(\boldsymbol{\theta}_t) - \bar{\phi}\,. \tag{1}$$

The solution functional $\psi(\boldsymbol{\theta}_t)$ characterizes the difference between $\phi(\boldsymbol{\theta}_t)$ and the posterior average $\bar{\phi}$ for every $\boldsymbol{\theta}_t$, thus would typically possess a unique solution, which is at least as smooth as ϕ under the elliptic or hypoelliptic settings (Mattingly, Stuart, and Tretyakov 2010). In the unbounded domain of $\boldsymbol{\theta}_t$, to make the presentation simple, we follow (Chen, Ding, and Carin 2015) and make certain assumptions on the solution functional, ψ, of the Poisson equation (1), which are used in the detailed proofs.

The mild assumptions of smoothness and boundedness made in the main paper are detailed as follows.

Assumption 1 *ψ and its up to 3rd-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 $k = (0,1,2,3)$, $C_k, p_k > 0$. Furthermore, the expectation of $\mathcal{V}$ on $\{\boldsymbol{\theta}_t\}$ is bounded: $\sup_t \mathbb{E}\mathcal{V}^p(\boldsymbol{\theta}_t) < \infty$, and $\mathcal{V}$ is smooth such that $\sup_{s\in(0,1)} \mathcal{V}^p\left(s\boldsymbol{\theta} + (1-s)\,Y\right) \leq C\left(\mathcal{V}^p\left(\boldsymbol{\theta}\right) + \mathcal{V}^p\left(Y\right)\right)$, $\forall \boldsymbol{\theta}, Y, p \leq \max\{2p_k\}$ for some $C > 0$.*

Proof of Theorem 1

Based on Assumption 1, we prove the main theorem.

Proof First let us denote

$$\tilde{\mathcal{L}}_t = \Bigg(G(\boldsymbol{\theta}_t)\Big(\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{\theta}_t) + \frac{N}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{d}_{t_i}|\boldsymbol{\theta}_t)\Big) + \Gamma(\boldsymbol{\theta}_t)\Bigg)\cdot\nabla_{\boldsymbol{\theta}} + \frac{1}{2}G(\boldsymbol{\theta})\left(G(\boldsymbol{\theta})^T\right) : \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^T\,, \tag{2}$$

the local generator of our proposed pSGLD with stochastic gradients, where $\boldsymbol{a}\cdot\boldsymbol{b} \triangleq \boldsymbol{a}^\top\boldsymbol{b}$ is the vector inner product, $\mathbf{A} : \mathbf{B} \triangleq \mathrm{tr}\{\mathbf{A}^\top\mathbf{B}\}$ is the matrix double dot product. Furthermore, let $\mathcal{L}$ be the true generator of the Langevin dynamic corresponding to the pSGLD, *e.g.*, replacing the stochastic gradient in $\tilde{\mathcal{L}}_t$ with the true gradient. As a result, we have the relation:

$$\tilde{\mathcal{L}}_t = \mathcal{L} + \Delta V_t\,, \tag{3}$$

where $\Delta V_t \triangleq (N\bar{g}(\boldsymbol{\theta}_t;\mathcal{D}^t) - g(\boldsymbol{\theta}_t;\mathcal{D}^t))^\top G(\boldsymbol{\theta}_t)\nabla_{\boldsymbol{\theta}}$, $g(\boldsymbol{\theta}_t;\mathcal{D}^t)$ is the full gradient, $\bar{g}(\boldsymbol{\theta}_t;\mathcal{D}^t)$ is the stochatic gradient calculated from the t-th minibatch.

In pSGLD, we use the Euler integrator, which is a first order integrator. As a result, according to (Chen, Ding, and Carin 2015), for a test function ϕ, we can decompose it as:

$$\begin{aligned}\mathbb{E}[\psi(\boldsymbol{\theta}_t)] &= e^{\epsilon_t\tilde{\mathcal{L}}_t}\psi(\boldsymbol{\theta}_{(t-1)}) + O(\epsilon_t^2)\\ &= \left(\mathbb{I} + \epsilon_t\tilde{\mathcal{L}}_t\right)\psi(\boldsymbol{\theta}_{(t-1)}) + O(\epsilon_t^2)\,,\end{aligned} \tag{4}$$

where $\mathbb{I}$ is the identity map, *i.e.,* $\mathbb{I}f(x) = f(x)$.

According to the assumptions, there exists a functional ψ that solves the following Poisson Equation:

$$\mathcal{L}\psi(\boldsymbol{\theta}_t) = \phi(\boldsymbol{\theta}_t) - \bar{\phi}\,, \tag{5}$$

where $\bar{\phi}$ is defined in the main text.

Sum over $t = 1, \cdots, T$ in the above equation, take expectation on both sides, and use the Poisson Equation (5) and the relation $\tilde{\mathcal{T}}_t = \mathcal{L} + \Delta V_t$ to expand the first order term. We obtain

$$\sum_{t=1}^{T}\mathbb{E}\left(\psi(\boldsymbol{\theta}_t)\right) = \sum_{t=1}^{T}\psi(\boldsymbol{\theta}_{(t-1)}) + \sum_{t=1}^{T}\epsilon_t\mathcal{T}_t\psi(\boldsymbol{\theta}_{(t-1)}) + \sum_{t=1}^{T}\epsilon_t\Delta V_l\psi(\boldsymbol{\theta}_{(t-1)}) + C\sum_{t=1}^{T}\epsilon_t^2\,. \tag{6}$$

Divide both sides by S_T, we have

$$\hat{\phi} - \bar{\phi} = \frac{\mathbb{E}\psi(\boldsymbol{\theta}_t) - \psi(\boldsymbol{\theta}_0)}{S_T} + \frac{1}{S_T}\sum_{l=1}^{T-1}\left(\mathbb{E}\psi(\boldsymbol{\theta}_{(t-1)}) + \psi(\boldsymbol{\theta}_{(t-1)})\right) + \sum_{t=1}^{T}\frac{\epsilon_t}{S_T}\Delta V_t\psi(\boldsymbol{\theta}_{(t-1)}) + C\frac{\sum_{t=1}^{T}\epsilon_t^2}{S_T}\,. \tag{7}$$

As a result, there exists some positive constant C, such that:

$$\left(\hat{\phi} - \bar{\phi}\right)^2 \leq C\left(\underbrace{\frac{1}{S_T^2}\left(\psi(\boldsymbol{\theta}_0) - \mathbb{E}\psi(\boldsymbol{\theta}_T)\right)^2}_{A_1} + \underbrace{\frac{1}{S_T^2}\sum_{t=1}^{T}\left(\mathbb{E}\psi(\boldsymbol{\theta}_{(t-1)}) - \psi(\boldsymbol{\theta}_{(t-1)})\right)^2}_{A_2} + \sum_{t=1}^{T}\frac{\epsilon_t^2}{S_T^2}\left\|\Delta V_t\right\|^2 + \left(\frac{\sum_{t=1}^{T}\epsilon_t^2}{S_T}\right)^2\right) \tag{8}$$

A_1 can be bounded by assumptions, and A_2 can be easily shown to be bounded by $O(\sqrt{\epsilon_t})$ due to the Gaussian noise. It turns out that the resulting terms have order higher than those from the other terms, thus can be ignored in the expression below. After some simplifications, (8) is bounded by:

$$\mathbb{E}\left(\hat{\phi} - \bar{\phi}\right)^2 \lesssim \sum_t \frac{\epsilon_t^2}{S_T^2}\mathbb{E}\left\|\Delta V_t\right\|^2 + \frac{1}{S_T} + \frac{1}{S_T^2} + \left(\frac{\sum_{t=1}^{L}\epsilon_t^2}{S_T}\right)^2 = C\left(\sum_t \frac{\epsilon_t^2}{S_T^2}\mathbb{E}\left\|\Delta V_t\right\|^2 + \frac{1}{S_T} + \frac{(\sum_{t=1}^{T}\epsilon_t^2)^2}{S_T^2}\right) \tag{9}$$

for some $C > 0$. It is easy to show under the assumptions, all the terms in the above bound approach zero. This completes the first part of the theorem. ■

B. The proof for Corollary 2

To prove Corollary 2, we first show the following results.

Lemma 1 *Assume that the 1st-order and 2nd-order gradient are bounded, then there exists some constant M, for k-th component of $\Gamma(\boldsymbol{\theta}_t)$, we have*

$$\left|\sum_{t=1}^{T}\Gamma_k(\boldsymbol{\theta}_t)\right| \leq MT\frac{(1-\alpha)}{\alpha^{\frac{3}{2}}}\,. \tag{10}$$

Proof Since $\Gamma(\boldsymbol{\theta})$ is a diagnal matrix, we focus on one of its elements thus omit the index k in the following.

First, the iterative form of exponential moving average can be written as a function of the gradients at all the previous timesteps:

$$V(\theta_t) = \alpha V(\theta_{t-1}) + (1-\alpha)\bar{g}^2(\theta_t) \tag{11}$$

$$= (1-\alpha)\sum_{i=1}^{t}\alpha^{t-i}\bar{g}^2(\theta_i) \tag{12}$$

Based on this, for each component of $\Gamma(\boldsymbol{\theta}_t)$, we have

$$\left|\sum_{t=1}^{T}\Gamma(\theta_t)\right| = \left|\sum_{t=1}^{T}(1-\alpha)V^{-\frac{3}{2}}(\theta_t)\bar{g}(\theta_t)\frac{\partial\bar{g}(\theta_t)}{\partial\theta_t}\right| \tag{13}$$

$$= \left|\sum_{t=1}^{T}(1-\alpha)\frac{\bar{g}(\theta_t)}{\left(\alpha V(\theta_{t-1}) + (1-\alpha)\bar{g}^2(\theta_t)\right)^{\frac{3}{2}}}\frac{\partial\bar{g}(\theta_t)}{\partial\theta_t}\right| \tag{14}$$

$$\ll \left|\sum_{t=1}^{T}\frac{(1-\alpha)}{\alpha^{\frac{3}{2}}V^{\frac{3}{2}}(\theta_{t-1})}\frac{\partial\bar{g}(\theta_t)}{\partial\theta_t}\right| \tag{15}$$

With the assumption that the 1st-order and 2nd-order gradient are bounded, we have $\left|V^{-\frac{3}{2}}(\theta_{t-1})\frac{\partial\bar{g}(\theta_t)}{\partial\theta_t}\right| \leq M$, where M is a constant independent of $\{\epsilon_t\}$. Therefore, $\left|\sum_{t=1}^{T}\epsilon_t\Gamma(\theta_t)\right| \ll MT(1-\alpha)/\alpha^{\frac{3}{2}}$. ■

Based on Lemma 1, we now proceed to the proof of Corollary 2.

Proof By dropping the $\Gamma(\boldsymbol{\theta}_t)$ terms, we get a modified version of the local generator corresponding to the SDE of the pSGLD, defined as

$$\tilde{\mathcal{L}}_t = \mathcal{L} + \Delta\tilde{V}_t\,,$$

where $\Delta\tilde{V}_t = \Delta V_t + \Gamma(\boldsymbol{\theta}_t)\cdot\nabla_{\boldsymbol{\theta}}$ with ΔV_t defined in the proof of Theorem 1.

Following the proof of Theorem 1, we can derive the bound for $(\hat{\phi} - \bar{\phi})^2$, which is no more than (8) with an extra term as:

$$\left(\hat{\phi} - \bar{\phi}\right)^2 \leq C\left(\underbrace{\frac{1}{S_T^2}\left(\psi(\boldsymbol{\theta}_0) - \mathbb{E}\psi(\boldsymbol{\theta}_T)\right)^2}_{A_1} + \underbrace{\frac{1}{S_T^2}\sum_{t=1}^{T}\left(\mathbb{E}\psi(\boldsymbol{\theta}_{(t-1)}) - \psi(\boldsymbol{\theta}_{(t-1)})\right)^2}_{A_2} + \sum_{t=1}^{T}\frac{\epsilon_t^2}{S_T^2}\left\|\Delta V_t\right\|^2 + \underbrace{\left\|\sum_{t=1}^{T}\frac{\epsilon_t}{S_T}\Gamma(\boldsymbol{\theta}_t)\right\|^2}_{A_3} + \left(\frac{\sum_{t=1}^{T}\epsilon_t^2}{S_T}\right)^2\right) \tag{16}$$

We can further relax A_3 above as:

$$A_3 \leq \left(\sum_k\left|\sum_{t=1}^{T}\frac{\epsilon_t}{S_T}\Gamma_k(\boldsymbol{\theta}_t)\right|\right)^2 \leq \left(\sum_k\frac{\epsilon_1}{T\epsilon_T}\left|\sum_{t=1}^{T}\Gamma_k(\boldsymbol{\theta}_t)\right|\right)^2 \leq O\left(\frac{(1-\alpha)^2}{\alpha^3}\right)\,, \tag{17}$$

where the last inequality follows by using the bound from Lemma 1. Taking expectation on both sides, we arrive at the MSE:

$$\mathbb{E}\left(\hat{\phi}-\bar{\phi}\right)^2 \leq$$
$$C\left(\sum_t \frac{\epsilon_t^2}{S_T^2}\mathbb{E}\left\|\Delta V_t\right\|^2+\frac{1}{S_T}+\frac{(\sum_{t=1}^T \epsilon_t^2)^2}{S_T^2}+\mathbb{E}\left\|\sum_{t=1}^T \frac{\epsilon_t}{S_T}\Gamma(\boldsymbol{\theta}_t)\right\|^2\right)$$
$$\leq \mathcal{B}_{\text{mse}}+O\left(\frac{(1-\alpha)^2}{\alpha^3}\right). \tag{18}$$

for some $C>0$. ■

C. The proof for Corollary 3

Proof By thinning samples from the pSGLD, we obtain a sequence of subsamples $\{\boldsymbol{\theta}_{t_1},\cdots,\boldsymbol{\theta}_{t_m}\}$ from the original samples $\{\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_n\}$ where $m\leq n$ and $(t_1,\cdots,t_m)$ is a subsequence of $(1,2,\cdots,n)$. Since we use the 1st-order Euler integrator, based on the definition in (Chen, Ding, and Carin 2015), we have for the original samples:

$$\tilde{P}_l f(\boldsymbol{\theta}_l)\triangleq \mathbb{E}f(\boldsymbol{\theta}_l)=e^{\epsilon_l\tilde{\mathcal{L}}_l}f(\boldsymbol{\theta}_l)+O(\epsilon_l^2)\,, \tag{19}$$

where $\tilde{P}_l$ denotes the Kolmogorov operator. Now for samples between t_i and t_j, *i.e.*, $\{\boldsymbol{\theta}_{t_i},\cdots,\boldsymbol{\theta}_{t_j}\}$, we have

$$\tilde{P}_{t_j}f(\boldsymbol{\theta}_i)=\tilde{P}_{t_j}\circ\cdots\circ\tilde{P}_{t_i}f(\boldsymbol{\theta}_i)\,, \tag{20}$$

where $A\circ B$ denotes the composition of the two operators A and B, *i.e.*, A is evaluated on the output of B. Now substitute (19) into (20), and use the Baker-Campbell-Hausdorff formula (Bakhturin 2001) for commutators, we have

$$\begin{aligned}\tilde{P}_{t_j}f(\boldsymbol{\theta}_i)&=e^{\sum_{l=i}^j\epsilon_l\tilde{\mathcal{L}}_l}f(\boldsymbol{\theta}_i)+O(\sum_{l=i}^j\epsilon_l^2)\\&\leq e^{S_{ij}\tilde{\mathcal{L}}_{ij}}f(\boldsymbol{\theta}_i)+O(S_{ij}^2)\,,\end{aligned} \tag{21}$$

where $S_{ij}\triangleq\sum_{l=i}^j\epsilon_l$, $\tilde{\mathcal{L}}_{ij}\triangleq\sum_{l=i}^j\frac{\epsilon_l}{S_{ij}}\tilde{\mathcal{L}}_l$. This means by thinning the samples, going from $\boldsymbol{\theta}_i$ to $\boldsymbol{\theta}_j$ corresponds to a 1st-order local integrator with stepsize S_{ij} and a modified generator of the corresponding SDE as $\tilde{\mathcal{L}}_{ij}$, which is in the same form as the original generator $\mathcal{L}$.

By performing the same derivation with the new generator $\tilde{\mathcal{L}}_{ij}$, we obtain the same MSE as in Theorem 1 in the main text. ■

D. The proof for bias-variance tradeoff

Bias-variance decomposition

$$\text{Risk}\quad:\quad R=\mathbb{E}[(\bar{\phi}-\hat{\phi})^2]=B^2+V \tag{22}$$

Proof

$$\begin{aligned}R&=\mathbb{E}[(\bar{\phi}-\hat{\phi})^2]\\&=\mathbb{E}[(\bar{\phi}-\bar{\phi}_\eta+\bar{\phi}_\eta-\hat{\phi})^2]\\&=\mathbb{E}[(\bar{\phi}-\bar{\phi}_\eta)^2+(\bar{\phi}_\eta-\hat{\phi})^2+2(\bar{\phi}-\bar{\phi}_\eta)(\bar{\phi}_\eta-\hat{\phi})]\\&=\mathbb{E}[(\bar{\phi}-\bar{\phi}_\eta)^2]+\mathbb{E}[(\bar{\phi}_\eta-\hat{\phi})^2]\\&\quad+2\mathbb{E}[(\bar{\phi}-\bar{\phi}_\eta)(\bar{\phi}_\eta-\hat{\phi})]\\&=(\bar{\phi}-\bar{\phi}_\eta)^2+\mathbb{E}[(\bar{\phi}_\eta-\hat{\phi})^2]\\&=B^2+V\end{aligned}$$

where

$$\text{Bias}\quad:\quad B=\bar{\phi}_\eta-\bar{\phi} \tag{23}$$
$$\text{Variance}\quad:\quad V=\mathbb{E}[(\bar{\phi}_\eta-\hat{\phi})^2] \tag{24}$$

■

Variance term in risk of estimator

$$\text{Variance}\quad:\quad V=\mathbb{E}[(\bar{\phi}-\hat{\phi})^2]\approx\frac{A(0)}{M} \tag{25}$$

Proof

$$\begin{aligned}V&=\mathbb{E}[(\bar{\phi}_\eta-\hat{\phi})^2]\\&=\mathbb{E}\Big[\Big(\bar{\phi}_\eta-\frac{1}{T}\sum_{i=1}^T\phi(\boldsymbol{\theta}_i)\Big)^2\Big] &(26)\\&=\frac{1}{T^2}\mathbb{E}\Big[\sum_{i=1}^T\sum_{j=1}^T\Big(\bar{\phi}_\eta-\phi(\boldsymbol{\theta}_i)\Big)\Big(\bar{\phi}_\eta-\phi(\boldsymbol{\theta}_j)\Big)\Big]\\&=\frac{1}{T^2}\sum_{i=1}^T\sum_{j=1}^T A(|i-j|) &(27)\\&=\frac{1}{T^2}\sum_{i=1}^T\left(\sum_{t=-\infty}^{\infty}A(|t|)-\sum_{|t|>2T}A(|t|)\right) &(28)\\&\approx\frac{1}{T^2}\sum_{i=1}^T\sum_{t=-\infty}^{\infty}A(|t|) &(29)\\&=\frac{1}{T}\Big(A(0)+2\sum_{t=1}^{\infty}A(t)\Big) &(30)\\&=\frac{A(0)}{T}\Big(1+2\sum_{t=1}^{\infty}\frac{A(t)}{A(0)}\Big) &(31)\end{aligned}$$

where the term $\sum_{|t|>2T}A(|t|)$ is omitted from (28) to (29), which is usually small according to the property of autocovariance function.

We repeat some defintions from the main paper (Gamerman and Lopes 2006).

$$A(t)=\mathbb{E}[(\bar{\phi}_\eta-\phi(\boldsymbol{\theta}_0))(\bar{\phi}_\eta-\phi(\boldsymbol{\theta}_t))] \tag{32}$$

is the *autocovariance function*, manifesting how strong two samples with a time lag t are correlated. Its normalized version

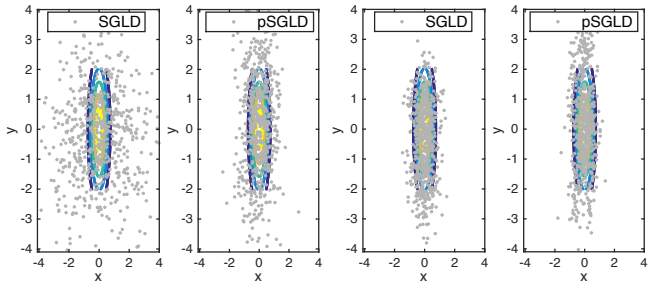
$$\text{ACF}\quad:\quad\gamma(t)=\frac{A(t)}{A(0)} \tag{33}$$

is called the *autocorrelation function* (ACF).

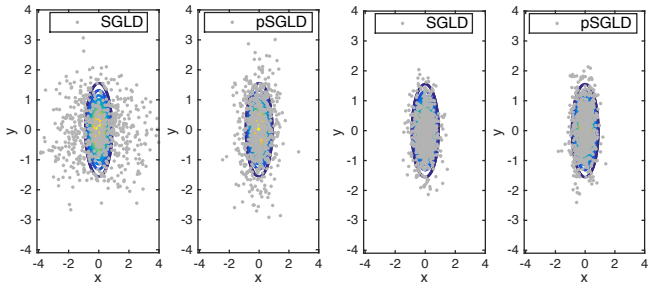
$$\text{ACT}\quad:\quad\tau=\frac{1}{2}+\sum_{t=1}^{\infty}\gamma(t) \tag{34}$$

is the integrated *autocorrelation time* (ACT), which measures the interval between independent samples.

Note that *effective sample size* (ESS) is defined as



(a) $a = 2, \epsilon = 0.3$ (b) $a = 2, \epsilon = 0.1$



(c) $a = 0.5, \epsilon = 0.3$ (d) $a = 0.5, \epsilon = 0.1$

Figure 1: Simulation.

$$\text{ESS} \quad : \quad M = \frac{T}{1 + 2\sum_{t=1}^{\infty} \frac{A(t)}{A(0)}} \tag{35}$$

Plugin the definition into the derivation for variance, we have

$$V \approx \frac{A(0)}{T}\Big(1 + 2\sum_{t=1}^{\infty} \frac{A(t)}{A(0)}\Big) = \frac{A(0)}{M} \tag{36}$$

■

E. More results on simulation

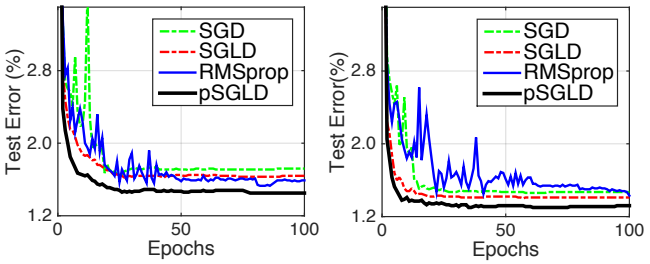
We demonstrate our pSGLD on a simple 2D gaussian example, $\mathcal{N}\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.16 & 0 \\ 0 & a \end{bmatrix}\Big)$. The first 600 samples of both methods for different a and ϵ are shown in Fig. 1.

Comparing the results for different stepsize ϵ at the same a, it can be seen that pSGLD can adapt stepsizes acorrding to the manifold geometry of different dimensions.

When a is rescaled from 0.5 to 2, stepsize $\epsilon = 0.1$ is appropriate for SGLD at $a = 0.5$, but not a good choice at $a = 2$, because the space is not fully explored. This also implies that even if the covariance matrix of a target distribution is mildly rescaled, we do not have to choose a new stepsize for pSGLD. Whilst, the stepsize of the standard SGLD needs to be fine-tuned in order to obtain decent samples.

F. More results on Feedforward Neural Networks

Learning curves for network sizes of 400-400 and 800-800 on MNIST are provided in Fig. 2 (a) and (b), respectively. Similar with results of network size 1200-1200 in the main paper, stochastic sampling methods take less iterations to converge, and the results are more stable than their optimization counterparts. Moreover, it can be seen that pSGLD consistently converges faster and better than SGLD and others.



(a) 400-400 (b) 800-800

Figure 2: Learning curves of FNN at different network sizes.

G. More results on Convolutional Neural Networks

We use another fairly standard network configuration containing 2 convolutional layers on MNIST dataset. It is followed by a single fully-connected layer (Chen-Yu et al. 2015), containing **500** hidden nodes that uses ReLU. Both convolutional layers use 5×5 filter size with 32 and 64 channels, respectively, 2×2 max pooling are used after each convolutional layer. 100 epochs are used, and L is set to 20. The stepsizes for pSGLD and RSMprop are set to $\epsilon = \{1, 2\} \times 10^{-3}$ via grid search. For SGLD and SGD, this is $\epsilon = \{1, 2\} \times 10^{-1}$.

A comparison of test errors is shown in Table 1, with the corresponding learning curves in Fig. 3. Again, under the same network architecture, CNN trained with traditional SGD gives an error of 0.81%, while pSGLD has a significant improvement, with an error of 0.56%.

Method	Test error
pSGLD	**0.56%**
SGLD	0.76%
RMSprop	0.64%
SGD	0.81%

Table 1: Results of CNN.

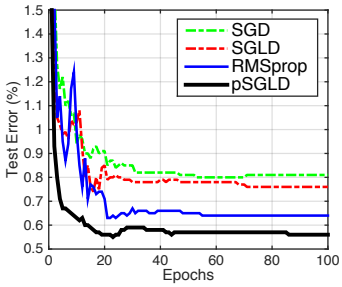


Figure 3: Learning curves.

We also tested a similar 3-layer CNN with 32-32-64 channels on Cifar-10 RGB image dataset (Krizhevsky and Hinton 2009), which consists of $50,000$ samples for training and $10,000$ samples for testing. No data augmentation is employed for the dataset. We keep the same setting for pSGLD and SGLD from MNIST, and show the comparison on Cifar-10 in Fig. 4. pSGLD converges faster and reach a lower error.

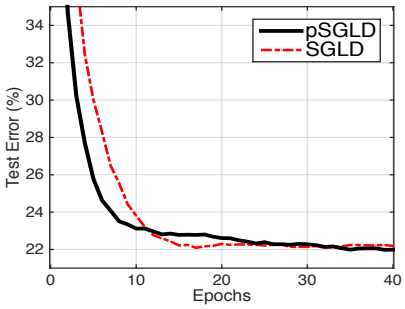


Figure 4: Test learning curves of CNN on Cifar-10 dataset.

References

Bakhturin, Y. A. 2001. *Campbell–Hausdorff formula*. Encyclopedia of Mathematics, Springer.

Chen, C.; Ding, N.; and Carin, L. 2015. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *NIPS*.

Chen-Yu, L.; Saining, X.; Patrick, G.; Zhengyou, Z.; and Zhuowen, T. 2015. Deeply-supervised nets. *AISTATS*.

Gamerman, D., and Lopes, H. F. 2006. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images.

Mattingly, J. C.; Stuart, A. M.; and Tretyakov, M. V. 2010. Construction of numerical time-average and stationary measures via Poisson equations. *SIAM J. NUMER. ANAL.* 48(2):552–577.