Communication Efficient Stochastic Gradient MCMC for Neural Networks Supplementary Material

Chunyuan Li[1], **Changyou Chen**[2], **Yunchen Pu**[3], **Ricardo Henao**[4], **and Lawrence Carin**[4]
[1]Microsoft Research, Redmond [2]University at Buffalo, SUNY [3]Facebook [4]Duke University

A. Algorithm of Downpour pSGLD

Algorithm 1 Downpour pSGLD

Input: ϵ_t, β_1, λ, $\pi \in \mathbb{N}$
Output: $\{\boldsymbol{\theta}_l\}_{l=1}^{L}$
Initialize: $t^{(p)}=0$, $l=0$, $\boldsymbol{\nu}^{(p)}=0$
$\tilde{\boldsymbol{\theta}} \sim \mathcal{N}(0,\mathbf{I})$, $\boldsymbol{\theta}^{(p)}=\tilde{\boldsymbol{\theta}}$
while maximum time not reached **do**
  % `Estimate gradient from` $\mathcal{D}_{\mathtt{M}}$
  $\tilde{\boldsymbol{f}}_t^{(p)} = \nabla \tilde{U}(\boldsymbol{\theta}_t^{(p)})$
  % `Preconditioning`
  $\boldsymbol{v}_t^{(p)} \leftarrow \beta_1 \boldsymbol{v}_{t-1} + (1-\beta_1)\tilde{\boldsymbol{f}}_t \odot \tilde{\boldsymbol{f}}_t$
  $\mathbf{G}_t^{(p)} \leftarrow \text{diag}\left(\mathbf{1} \oslash \left(\lambda \mathbf{1} + \sqrt{\boldsymbol{v}_t^{(p)}}\right)\right)$
  % `Parameter update with pSGLD`
  $\boldsymbol{\xi}_t^{(p)} \sim \mathcal{N}(0,\mathbf{I})$
  $\boldsymbol{\theta}_{t+1}^{(p)} \leftarrow \boldsymbol{\theta}_t^{(p)} - \epsilon_t \mathbf{G}_t^{(p)} \tilde{\boldsymbol{f}}_t^{(p)} + \sqrt{\epsilon_t \mathbf{G}_t^{(p)}}\boldsymbol{\xi}_t^{(p)}$
  $\boldsymbol{\nu}_{t+1}^{(p)} \leftarrow \boldsymbol{\nu}_t^{(p)} - \epsilon_t \mathbf{G}_t^{(p)} \tilde{\boldsymbol{f}}_t^{(p)} + \sqrt{\epsilon_t \mathbf{G}_t^{(p)}}\boldsymbol{\xi}_t^{(p)}$
  $t^{(p)} \leftarrow t^{(p)}+1$
  % $\pi-$`Downpour Communication`
  if $t^{(p)}$ divide π **then**
    $\boldsymbol{\theta}_{l+1} \leftarrow \boldsymbol{\theta}_l + \boldsymbol{\nu}_{t+1}^{(p)}$
    $\boldsymbol{\theta}_{t+1}^{(p)} \leftarrow \boldsymbol{\theta}_{l+1}$ $\boldsymbol{\nu}_{t+1}^{(p)} \leftarrow 0$
    $l = l+1$
  end if
end while

Preconditioned SGLD (pSGLD) (Li *et al.* 2016) specifies the general dynamics of SG-MCMC as $\boldsymbol{z}=\boldsymbol{\theta}$, $H(\boldsymbol{\theta}) = U(\boldsymbol{\theta})$, $W(\boldsymbol{\theta}) = \mathbf{G} = G(\boldsymbol{\theta})$, $Q(\boldsymbol{\theta}) = \mathbf{0}$, and $\Gamma(\boldsymbol{\theta})$ is a vector with its i-th element as $\sum_j \frac{\partial}{\partial \boldsymbol{\theta}_j}(\mathbf{G}_{ij})$. pSGLD utilizes magnitudes of recent gradients to construct a diagonal preconditioner $\mathbf{G}$. The downpour pSGLD is shown in Algorithm 1. There are two tuning parameters in pSGLD: λ controls the extremes of the curvature in the preconditioner (default λ=10^{-5}), and β_1 balances the weights of historical and current gradients. We use a default value of β_1=0.99, to construct an exponentially decaying sequence.

B. MSE bounds for Communication Protocols

We extends the proofs in (Chen *et al.* 2015) to study the impact of the proposed periodic communication protocols on the MSE bounds for accelerating SG-MCMC algorithms. We follow the same assumptions of (Chen *et al.* 2015; Teh *et al.* 2016; Vollmer *et al.* 2015).

The MSE bound for standard SG-MCMC is:

$$\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \leq \mathcal{B}_0 \triangleq C(\mathcal{E}_1 + \mathcal{E}_2), \text{ with} \quad (1)$$

$$\mathcal{E}_1 = \sum_l \frac{\epsilon_l^2}{S_L^2}\mathbb{E}\left\|\Delta V_l\right\|^2 \text{ and } \mathcal{E}_2 = \frac{1}{S_L} + \frac{(\sum_l \epsilon_l^2)^2}{S_L^2}, \quad (2)$$

B2. Proof for Downpour Protocol

Proof To simplify the proof, we note that the master collects a sample every π iterations for each worker, this corresponds to thinning the worker chains, with thinning interval π. According to Corollary 2 of (Li *et al.* 2016), this is equivalent to collecting samples from the chains with stepsize $\epsilon_l^{(p)} = \sum_{t=t^{(p)}-\pi}^{t^{(p)}} \epsilon_t$. As a result, the algorithm is transformed to a simpler algorithm, in which the chains communicate with the master instantly.

Compared with standard SG-MCMC, there is an additional error associated with the gradient approximation, *i.e.*, $\boldsymbol{\theta}_{l+1}$ is obtained with a stochastic gradient $\tilde{\boldsymbol{f}}_{l-\pi_l}$ evaluated on "old" parameters $\boldsymbol{\theta}_{l-\pi_l}$ for some integer π_l, instead of $\boldsymbol{\theta}_l$. In other words, $\tilde{\boldsymbol{f}}_{l-\pi_l}$ has been used to approximate $\tilde{\boldsymbol{f}}_l$, resulting in additional approximation error.

Only the term $\mathcal{E}_1$ relies on the gradient, in which ΔV_l needs to be replaced with

$$\begin{aligned}\Delta \tilde{V}_l &\triangleq \tilde{\boldsymbol{f}}_{l-\pi_l} - \boldsymbol{f}_l \\ &= (\tilde{\boldsymbol{f}}_{l-\pi_l} - \tilde{\boldsymbol{f}}_l) + (\tilde{\boldsymbol{f}}_l - \boldsymbol{f}_l) \\ &= (\tilde{\boldsymbol{f}}_{l-\pi_l} - \tilde{\boldsymbol{f}}_l) + \Delta V_l \,,\end{aligned}$$

where the "old" parameter is at most $\pi(P-1)$ steps away from the latest parameter, as there are totally P workers:

$$\pi_l \leq \pi(P-1)\,.$$

As a result, we need to bound the term $\sum_l \frac{\epsilon_l^2}{S_L^2}\mathbb{E}\left\|\Delta \tilde{V}_l\right\|^2$,

which gives

$$\sum_l \frac{\epsilon_l^2}{S_L^2}\mathbb{E}\left\|\Delta\tilde{V}_l\right\|^2$$
$$\leq \sum_l \frac{\epsilon_l^2}{S_L^2}\mathbb{E}\left\|\Delta V_l\right\|^2 + \sum_l \frac{\epsilon_l^2}{S_L^2}\mathbb{E}\left\|\tilde{\boldsymbol{f}}_{l-\pi_l} - \tilde{\boldsymbol{f}}_l\right\|^2$$
$$\leq \mathcal{E}_1 + D_1 \sum_l \frac{\epsilon_l^2}{S_L^2}\mathbb{E}\left\|\boldsymbol{\theta}_{l-\pi_l} - \boldsymbol{\theta}_l\right\|^2 ,$$

where the last inequality follows from the Lipschitz property. Now, using the integrator property as in (Chen *et al.* 2015), we have $\mathbb{E}\|\boldsymbol{\theta}_l - \boldsymbol{\theta}_{l+\delta}\| = O(\delta\epsilon_l^K)$ for a K-th-order integrator. Because we are using a 1st-order integrator ($K = 1$), we have

$$\mathbb{E}\left\|\boldsymbol{\theta}_{l-\pi_l} - \boldsymbol{\theta}_l\right\|^2 = \left(\mathbb{E}\left\|\boldsymbol{\theta}_{l-\pi_l} - \boldsymbol{\theta}_l\right\|\right)^2 + \mathrm{Var}\left(\boldsymbol{\theta}_{l-\pi_l} - \boldsymbol{\theta}_l\right)$$
$$\leq \left(\mathbb{E}\left\|\boldsymbol{\theta}_{l-\pi_l} - \boldsymbol{\theta}_l\right\|\right)^2 \leq D_2\pi_l^2 ,$$

for some constant D_2.

As a result, we have

$$\sum_l \frac{\epsilon_l^2}{S_L^2}\mathbb{E}\left\|\Delta\tilde{V}_l\right\|^2 \leq \mathcal{E}_1 + D_3\sum_l \frac{\epsilon_l^2}{S_L^2}\pi_l^2$$
$$\leq \mathcal{E}_1 + D_3\sum_l \frac{\epsilon_l^2}{S_L^2}\pi^2(P-1)^2 .$$

Substitute the above inequality into the MSE bound, we get the MSE bound for downpour protocol as

$$\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \leq C_1(\mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3), \text{ with}$$
$$\mathcal{E}_3 = \pi^2(P-1)^2\frac{(\sum_l \epsilon_l^2)^2}{S_L^2} ,$$

The proof for bias and variance follow similar derivation. ■

B2. Proof for Elastic Protocol with $\alpha = 1$

Proof

$$\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 = \mathbb{E}\left(\sum_{p=1}^P \frac{S_L^{(p)}}{\sum_p S_L^{(p)}}\left(\hat{\phi}_L^{(p)} - \bar{\phi}\right)\right)^2$$
$$= \sum_{p=1}^P \frac{{S_L^{(p)}}^2}{(\sum_p S_L^{(p)})^2}\mathbb{E}\left(\hat{\phi}_L^{(p)} - \bar{\phi}\right)^2$$
$$+ \sum_{i\neq j}\frac{S_L^{(i)}S_L^{(j)}}{(\sum_p S_L^{(p)})^2}\mathbb{E}\left[\hat{\phi}_L^{(i)} - \bar{\phi}\right]\mathbb{E}\left[\hat{\phi}_L^{(j)} - \bar{\phi}\right]$$
$$\leq \sum_{p=1}^P \frac{{S_L^{(p)}}^2}{(\sum_p S_L^{(p)})^2}\mathbb{E}\left(\hat{\phi}_L^{(p)} - \bar{\phi}\right)^2$$
$$+ \sum_{i\neq j}\frac{S_L^{(i)}S_L^{(j)}}{(\sum_p S_L^{(p)})^2}\left|\mathbb{E}\hat{\phi}_L^{(i)} - \bar{\phi}\right|\left|\mathbb{E}\hat{\phi}_L^{(j)} - \bar{\phi}\right| .$$

From (Chen *et al.* 2015), the bias and MSE bounds above are formulated as

$$\mathbb{E}\left(\hat{\phi}_L^{(p)} - \bar{\phi}\right)^2 = O\left(\mathcal{E}_1^{(p)} + \frac{1}{S_L^{(p)}} + \frac{(\sum_l {\epsilon_l^{(p)}}^2)^2}{{S_L^{(p)}}^2}\right)$$
$$\left|\mathbb{E}\hat{\phi}_L^{(p)} - \bar{\phi}\right| = O\left(\frac{1}{S_L^{(p)}} + \frac{\sum_l {\epsilon_l^{(p)}}^2}{S_L^{(p)}}\right)$$

Substituting the bounds into the above MSE for elastic protocal, we have

$$\mathbb{E}\left(\hat{\phi}_L - \bar{\phi}\right)^2 \leq \sum_{p=1}^P \frac{{S_L^{(p)}}^2}{(\sum_p S_L^{(p)})^2}\left(\mathcal{E}_1^{(p)} + \frac{1}{S_L^{(p)}} + \frac{(\sum_l {\epsilon_l^{(p)}}^2)^2}{{S_L^{(p)}}^2}\right)$$
$$+ \sum_{i\neq j}\frac{S_L^{(i)}S_L^{(j)}}{(\sum_p S_L^{(p)})^2}\left(\frac{1}{S_L^{(i)}} + \frac{\sum_l {\epsilon_l^{(i)}}^2}{S_L^{(i)}}\right)\left(\frac{1}{S_L^{(j)}} + \frac{\sum_l {\epsilon_l^{(j)}}^2}{S_L^{(j)}}\right)$$
$$\leq D_4\left(\sum_{p=1}^P \frac{{S_L^{(p)}}^2}{(\sum_p S_L^{(p)})^2}\mathcal{E}_1^{(p)} + \frac{1}{\sum_p S_L^{(p)}} + \sum_{p=1}^P \frac{(\sum_l {\epsilon_l^{(p)}}^2)^2}{(\sum_p S_L^{(p)})^2}\right.$$
$$\left. + \frac{P^2 - P}{(\sum_p S_L^{(p)})^2} + \sum_{i\neq j}^P \frac{({\epsilon_l^{(i)}}^2)({\epsilon_l^{(j)}}^2)}{(\sum_p S_L^{(p)})^2}\right) ,$$

for some constant D_4.

$$\leq D_5\Big(\sum_{p=1}^P \frac{{S_L^{(p)}}^2}{(\sum_p S_L^{(p)})^2}\mathcal{E}_1^{(p)}$$
$$+ \frac{1}{\sum_p S_L^{(p)}} + \sum_{i,j=1}^P \frac{(\sum_l {\epsilon_l^{(i)}}^2 + \sum_l {\epsilon_l^{(j)}}^2)^2}{(\sum_p S_L^{(p)})^2}\Big)$$

The proof for bias and variance follow similar derivation, where the variance (Chen *et al.* 2016) is

$$\mathcal{V}_2 = V_2\left(\frac{1}{\sum_p S_L^{(p)}} + \sum_{p=1}^P \frac{\sum_l (\epsilon_l^{(p)})^2}{(\sum_p S_L^{(p)})^4}\Big(\sum_l \epsilon_l^{(p)}\Big)^2\right)$$
$$\leq V_2\left(\frac{1}{\sum_p S_L^{(p)}} + \sum_{i,j=1}^P \frac{(\sum_l {\epsilon_l^{(i)}}^2 + \sum_l {\epsilon_l^{(j)}}^2)^2}{(\sum_p S_L^{(p)})^2}\right)$$

■

C. Additional Experimental Results

Details on settings We have focused on the impact of π and P in the main text. In SG-MCMC algorithms, the hyperparameters for our methods also include mini-batch size, thinning interval, number of burn-in's, step-size and variances of the Gaussian priors. We discuss and specify them in the following.

- **Mini-batch size M:** Following conventional settings on these datasets, we choose $M = 100$ throughout the experiments.
- **Burn-in:** To obtain a good initialization for parameter samples from regions of higher probability, "burn-in" is usually used in MCMC to dispose of samples at the beginning of an MCMC run. We choose zero burn-in for MNIST/SVHN, a small burn-in period for WP in our parallel SG-MCMC methods. Because we aim to use multiple workers to explore the space.

- **Thinning interval:** While the proposed periodic protocol can play the role of thinning, we also manually thin the master chain: we collect samples from the master chain every R seconds in wall-clock time for testing. This is because (i) the same number of model samples is kept in testing with standard SG-MCMC, for fair comparison, and (ii) it can reduce the computational burden during testing. Specifically, $R = 5$ for MNIST, $R = 15$ for SVHN, and $R = 30$ for WP.
- **Step-size:** The default step-size for SGHMC and SGLD is 1×10^{-2}, and 1×10^{-3} for pSGLD.
- **Gaussian priors:** The variance of the Gaussian distribution encapsulates the prior belief of users on how strongly these weights should concentrate around 0. A larger variance in the prior encourages a wider range of weight choices, thus higher uncertainty in learning the weights. We fix $\sigma^2 = 1$ to focus on the study of communication protocols, and refer readers to (Li *et al.* 2016) for details about the impact of variance in modeling weight uncertainty.

More Results on FNN Figure 1 shows the test classification errors for various parallel algorithms.

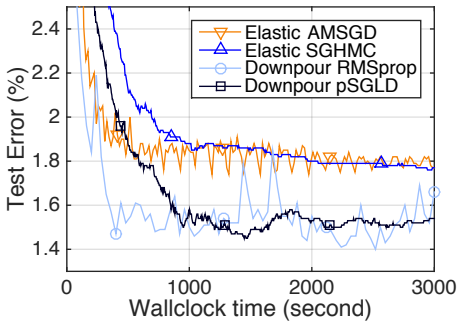

Figure 1: Comparison of Classification errors on FNN.

More Results on CNN Figure 2 shows the test classification errors of parallel SG-MCMC algorithms with 1 and 4 workers. Lower classification errors can be achieved when using 4 workers compared to 1 worker, *i.e.*, standard SG-MCMC.

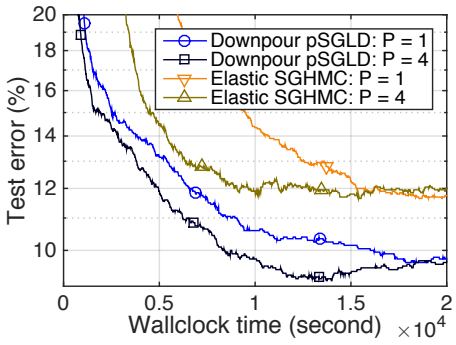

Figure 2: Classification error on CNN.

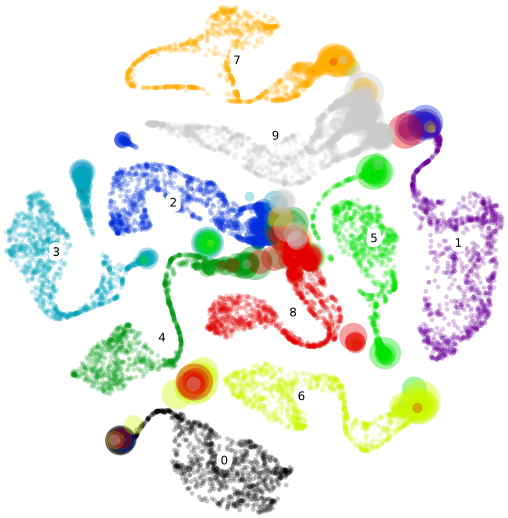

Figure 3: t-SNE visualization of MNIST.

Table 1: NLL on WP dataset.

Gates		LSTM		GRU	
#Layers		1-layer	2-layer	1-layer	2-layer
Downpour pSGLD	P=1	1.778	1.408	1.705	1.578
	P=4	**1.582**	**1.266**	**1.548**	**1.400**
Elastic SGHMC	P=1	1.513	1.429	1.608	1.351
	P=4	**1.403**	**1.315**	**1.423**	**1.282**

More Results on RNN Table 1 shows the final NLL for various setup of RNNs and algorithms. Lower NLLs can be achieved when using 4 workers compared to 1 worker, *i.e.*, standard SG-MCMC.

More Results t-SNE Visualization Following the same procedure of the t-SNE visualization for SVHN, we also embed the 10,000 test digits of MNIST in Fig. 3, using 20 model samples of the 400-400 FNN, obtained by downpour pSGLD algorithm. Though many methods can achieve high classification accuracy on MNIST dataset, there are some digits that are intrinsically hard to classify. In practice, it may be beneficial to leverage the uncertainty to further boost performance.

References

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

C. Chen, N. Ding, C. Li, Y. Zhang, and L. Carin. Stochastic gradient MCMC with stale gradients. In *NIPS*, 2016.

C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *AAAI*, 2016.

Y. W. Teh, A. H. Thiéry, and S. J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *JMLR*, 2016.

S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. (Non-)asymptotic properties of stochastic gradient Langevin dynamics. Technical report, 2015.