Federated Stochastic Gradient Langevin Dynamics

Khaoula el Mekkaoui[*†1] **Diego Mesquita**[*1] **Paul Blomstedt**[2] **Samuel Kaski**[1,3]

[1]Helsinki Institute for Information Technology,Department of Computer Science,Aalto University, Finland
[2]F-Secure, Finland
[3]Department of Computer Science, University of Manchester, UK

A BACKGROUND ON CONVERGENCE ANALYSIS FOR SGLD

Regularity conditions. Let ψ be the functional that solves the Poisson equation $\mathcal{L}\psi = \phi - \hat{\phi}$. Assume ψ is bounded up to its third order derivative by a function Γ, such that $\|\mathcal{D}^k\psi\| \leq C_k\Gamma^{p_k}$ with $C_k, p_k > 0\,\forall k \in \{0, \dots, 3\}$ with $\mathcal{D}^k$ denoting the kth order derivative. Assume as well that the expectation of Γ w.r.t. θ_t is bounded ($\sup_t \mathbb{E}\Gamma^p[\theta_t] \leq \infty$) and that Γ is smooth such that $\sup_{s\in(0,1)} \Gamma^p(s\theta + (1-s)\theta') \leq C(\Gamma^p(\theta) + \Gamma^p(\theta'))$, $\forall \theta, \theta'$, $p \leq \max_k 2p_k$, for some $C > 0$.

Under these regularity conditions, Chen et al. [2015] showed the following result.

Theorem 1 (See Chen et al. [2015]). *Let U_t be an unbiased estimate of U, the unnormalized negative log posterior, and $h_t = h$ for all $t \in \{1, \dots, T\}$. Let $\Delta V_t = (\nabla U_t - \nabla U)\cdot\nabla$. Under the assumptions above, for a smooth test function ϕ, the MSE OF SGLD at time $K = hT$ is bounded for some $C > 0$ independent of (T, h) as:*

$$\mathbb{E}[(\overline{\phi} - \hat{\phi})^2] \leq C\left(\frac{\frac{1}{T}\sum_t \mathbb{E}[\|\Delta V_t\|^2]}{T} + \frac{1}{Th} + h^2\right) \tag{1a}$$

Equation (1a) can also be written as:

$$\mathbb{E}[(\overline{\phi} - \hat{\phi})^2] \leq C\left(\frac{\frac{1}{T}\sum_t \mathbb{E}[\|\Delta V_t\psi(\theta_t)\|^2]}{T} + \frac{1}{Th} + h^2\right). \tag{2a}$$

For further analysis we add the assumption that $(\Delta V_t\psi(\theta))^2 \leq C'\|\nabla U_t(\theta) - \nabla U(\theta)\|^2$ for some $C' > 0$.

B PROOF OF THEOREM 1: CONVERGENCE OF DSGLD

Here, we follow the footprints of Chen et al. [2015] later adopted by Dubey et al. [2016]. Thus, we focus on bounding $\frac{1}{T}\sum_t \mathbb{E}[(\Delta V_t\psi(\theta_t))^2]$, when $U_t(\theta_t) = v_{s_t}(\theta_t)$. For some $C' > 0$, we have:

$$\frac{1}{C'T}\sum_t \mathbb{E}[(\Delta V_t\psi(\theta_t))^2] \leq \frac{1}{T}\sum_t \mathbb{E}[\|\nabla U_t(\theta_t) - \nabla U(\theta_t)\|^2] \tag{3a}$$

$$= \frac{1}{T}\sum_t \mathbb{E}\left[\left\|\frac{1}{f_s}\frac{N_s}{m}\nabla \log p(\mathbf{x}^{(m)}_{s_t}|\theta_t) - \nabla \log p(\mathbf{x}|\theta_t)\right\|^2\right] \tag{3b}$$

$$= \frac{1}{Tm^2}\sum_t \mathbb{E}\left[\left\|\textstyle\sum_{x_i\in\mathbf{x}^{(m)}_{s_t}} \displaystyle\frac{1}{f_s}N_s\nabla \log p(x_i|\theta_t) - \nabla \log p(\mathbf{x}|\theta_t)\right\|^2\right] \tag{3c}$$

[*]Equal contribution.

[†]Corresponding author: khaoula.elmekkaoui@aalto.fi

$$= \frac{1}{m^2}\mathbb{E}_s\mathbb{E}_{\mathbf{x}_{s_t}^{(m)}|s_t}\left[\sum\nolimits_{x_i\in\mathbf{x}_{s_t}^{(m)}}\left\|\frac{1}{f_s}N_s\nabla\log p(x_i|\theta_t) - \nabla\log p(\mathbf{x}|\theta_t)\right\|^2\right] \tag{3d}$$

$$\leq \frac{1}{m^2}\mathbb{E}_s\left[\mathbb{E}_{\mathbf{x}_{s_t}^{(m)}|s_t}\left[\sum\nolimits_{x_i\in\mathbf{x}_{s_t}^{(m)}}\left\|\frac{1}{f_s}N_s\nabla\log p(x_i|\theta_t)\right\|^2\right]\right] \tag{3e}$$

$$= \frac{1}{m^2}\mathbb{E}_s\left[m\frac{N_s^2}{f_s^2}\mathbb{E}_{x_i|s_t}\left[\left\|\nabla\log p(x_i|\theta_t)\right\|^2\right]\right] \tag{3f}$$

$$\leq \frac{1}{m}\mathbb{E}_s\left[\frac{N_s^2}{f_s^2}\gamma_s^2\right] = \frac{1}{m}\sum_s f_s\frac{N_s^2}{f_s^2}\gamma_s^2 = \frac{1}{m}\sum_s\frac{N_s^2}{f_s}\gamma_s^2 \tag{3g}$$

Here, $\mathbb{E}_{\mathbf{x}_{s_t}^{(m)}|s_t}$ denotes that the expectation is taken w.r.t. a mini-batch of size m with elements drawn with replacement and equal probability from shard s_t. Expectations without explicit subscripts are taken w.r.t. all random variables. To advance from Equation (3c) to (3d), we use law of iterated expectations and the fact that $\mathbb{E}[\|\sum_i r_i\|^2] = \sum_i\mathbb{E}[\|r_i\|^2]$ for zero-mean independent r_i's. To advance from Equation (3d) to (3e), we use $\mathbb{E}[\|r - \mathbb{E}[r]\|^2] \leq \mathbb{E}[\|r\|^2]$. Substituting Equation (3g) in Equation (2a) yields the desired result.

C PROOF OF LEMMA 1: UNBIASEDNESS AND FINITE VARIANCE

Recall that the for the DSGLD update [Ahn et al., 2014] we have

$$\mathbb{E}_{s,\mathbf{x}_{s_t}^{(m)}}\left[\frac{1}{f_s}\frac{N_s}{m}\nabla\log p(\mathbf{x}_{s_t}^{(m)}|\theta_t)\right] = \nabla\log p(\mathbf{x}|\theta_t).$$

Furthermore, for *conducive gradients* we have that:

$$\mathbb{E}_s\left[\nabla q(\theta_t) - \frac{1}{f_s}\nabla q_s(\theta_t)\right] = q(\theta_t) - \sum_s f_s\frac{1}{f_s}q_s(\theta_t) = 0.$$

Since the FSGLD estimator is the sum of the DSGLD estimator and the *conducive gradient*, it is unbiased.

The sufficient condition for the DSGLD estimator to have finite variance is that the unnormalized log posterior is Lipschitz continuous. Similarly, since $q_1,\ldots,q_S$ are also Lipschitz continuous, their first derivatives are bounded, so the conducive gradient is a convex combination of bounded functions and has finite variance. Thus, their sum, the FSGLD estimator has finite variance.

D PROOF OF THEOREM 2: CONVERGENCE OF FSGLD

We now bound $\frac{1}{T}\sum_t\mathbb{E}[(\Delta V_t\psi(\theta_t))^2]$ for the FSGLD update equation, when $U_t(\theta_t) = v_{s_t}(\theta_t) + g_{s_t}(\theta_t)$.

$$\frac{1}{C'T}\sum_t\mathbb{E}[(\Delta V_t\psi(\theta_t))^2]$$

$$\leq \frac{1}{T}\sum_t\mathbb{E}[\|\nabla U_t(\theta_t) - \nabla U(\theta_t)\|^2] \tag{4a}$$

$$= \frac{1}{Tm^2}\sum_t\mathbb{E}\left[\left\|\sum_{x_i\in\mathbf{x}_{s_t}^{(m)}}\frac{1}{f_s}\left(N_s\nabla\log p(x_i|\theta_t) - \nabla\log q_s(\theta_t)\right) + \nabla\log q(\theta_t) - \nabla\log p(\mathbf{x}|\theta_t)\right\|^2\right] \tag{4b}$$

$$= \frac{1}{m^2}\mathbb{E}_s\mathbb{E}_{\mathbf{x}_{s_t}^{(m)}|s_t}\left[\sum_{x_i\in\mathbf{x}_{s_t}^{(m)}}\left\|\frac{1}{f_s}\left(N_s\nabla\log p(x_i|\theta_t) - \nabla\log q_s(\theta_t)\right) + \nabla\log q(\theta_t) - \nabla\log p(\mathbf{x}|\theta_t)\right\|^2\right] \tag{4c}$$

$$\leq \frac{1}{m^2}\mathbb{E}_s\left[\mathbb{E}_{\mathbf{x}_{s_t}^{(m)}|s_t}\left[\sum_{x_i\in\mathbf{x}_{s_t}^{(m)}}\left\|\frac{1}{f_s}\left(N_s\nabla\log p(x_i|\theta_t) - \nabla\log q_s(\theta_t)\right)\right\|^2\right]\right] \tag{4d}$$

$$= \frac{1}{m^2}\mathbb{E}_s\left[m\frac{N_s^2}{f_s^2}\mathbb{E}_{x_i|s_t}\left[\left\|\nabla \log p(x_i|\theta_t) - N_s^{-1}\nabla \log q_s(\theta_t)\right\|^2\right]\right] \tag{4e}$$

$$= \frac{1}{m^2}\sum_s f_s m\frac{N_s^2}{f_s^2}\mathbb{E}_{x_i|s_t}\left[\left\|\nabla \log p(x_i|\theta_t) - N_s^{-1}\nabla \log q_s(\theta_t)\right\|^2\right] \tag{4f}$$

$$\leq \frac{1}{m}\sum_s \frac{N_s^2}{f_s}\epsilon_s^2 \tag{4g}$$

We proceed from Equation (4b) to (4c) using the law of iterated expectations and the fact that $\mathbb{E}[\|\sum_i r_i\|^2] = \sum_i \mathbb{E}[\|r_i\|^2]$ for zero-mean independent r_i's. To transition from Equation (4c) to (4d), we use $\mathbb{E}[\|r - \mathbb{E}[r]\|^2] \leq \mathbb{E}[\|r\|^2]$. The last line is obtained using *Lemma* 2. We use this bound and Equation 2a to get the desired result.

E MORE DETAILS ON REMARK 3

Following *Lemma* 2 and assuming that ϵ_s^2 is a tight bound, i.e.

$$\epsilon_s^2 := \frac{1}{N_s}\sum_{x_i\in\mathbf{x}_s}\left\|\nabla \log p(x_i|\theta) - \frac{1}{N_s}\nabla \log q_s(\theta)\right\|^2, \tag{5a}$$

choosing q_s that minimizes ϵ_s^2 is equivalent to finding

$$\min_{q_s}\max_{\theta}\frac{1}{N_s}\sum_{x_i\in\mathbf{x}_s}\left\|\nabla \log p(x_i|\theta) - \frac{1}{N_s}\nabla \log q_s(\theta)\right\|^2, \tag{5b}$$

which is equal to

$$\frac{1}{N_s}\min_{q_s}\max_{\theta}\sum_{x_i\in\mathbf{x}_s}\left[\|\nabla \log p(x_i|\theta)\|^2 + \frac{1}{N_s^2}\|\nabla \log q_s(\theta)\|^2 - \frac{2}{N_s}\left(\nabla \log q_s(\theta)\right)^\intercal \nabla \log p(x_i|\theta)\right], \tag{5c}$$

and can be further developed into

$$\frac{1}{N_s}\min_{q_s}\max_{\theta}\left[\sum_{x_i\in\mathbf{x}_s}\|\nabla \log p(x_i|\theta)\|^2\right] + \frac{1}{N_s}\|\nabla \log q_s(\theta)\|^2 - \frac{2}{N_s}\left(\nabla \log q_s(\theta)\right)^\intercal \nabla \log p(\mathbf{x}_s|\theta). \tag{5d}$$

Completing the squares and using $\max a + b \leq \max a + \max b$, we get the following upper-bound for Equation 5b:

$$\frac{1}{N_s^2}\min_{q_s}\max_{\theta}\left[\|\nabla \log q_s(\theta) - \nabla \log p(\mathbf{x}_s|\theta)\|^2\right] + \frac{1}{N_s}\max_{\theta}\left[\frac{1}{N_s}\|\nabla \log p(\mathbf{x}_s|\theta)\|^2 + \sum_{x_i\in\mathbf{x}_s}\|\nabla \log p(x_i|\theta)\|^2\right], \tag{5e}$$

in which only the first term depends on q_s.

F ADDITIONAL EXPERIMENTS

In this section, we provide additional results for Bayesian linear regression. Since we can leverage the simple likelihood function and compute surrogates analytically, this setting is especially useful to understand the behavior of our method.

F.1 LINEAR REGRESSION

In this set of experiments, we apply FSGLD to Bayesian linear regression and analyze its performance, which we measure in terms of MSE averaged over posterior samples.

Model The inputs of our model are $\boldsymbol{Z} = \{x_i, y_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. The likelihood of the ith output $y_i \in \{0, 1\}$, given the input vector x_i, is $p\left(y_i|x_i\right) = \mathcal{N}\left(y_i|\beta^\top x_i, \sigma_e\right)$, and we place the prior $p(\beta) = \mathcal{N}\left(\beta|\mathbf{0}, \lambda^{-1}I\right)$.

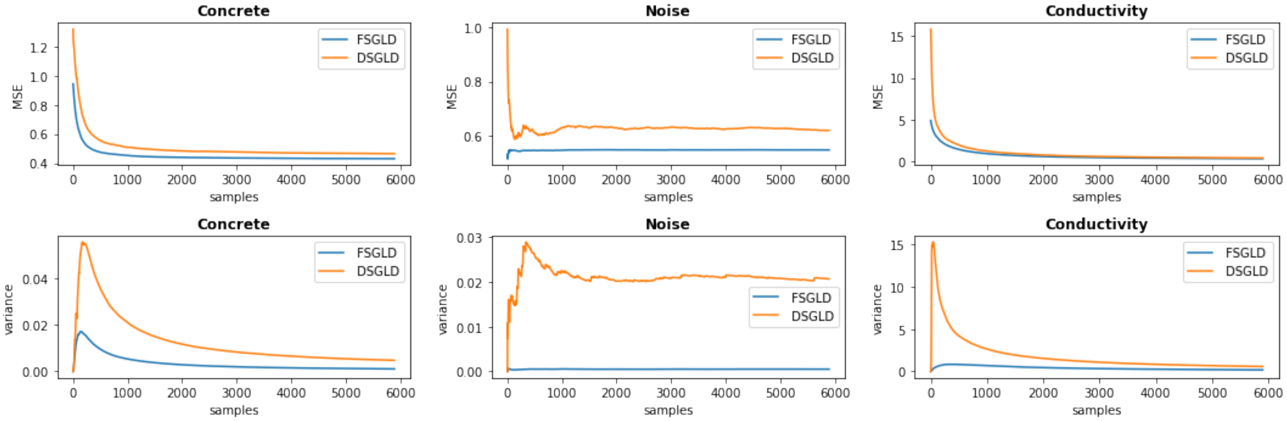

Figure 1: Average MSE and variance along time computed for DSGLD and FSGLD as a function of the number of samples. Overall, FSGLD converges to better performance than DSGLD. Additionally, FSGLD shows lower variance for all datasets.

Setting We run experiments on three different datasets[1] from the UCI repository: Concrete (1030 samples, 9 features); Noise (1503 samples, 6 features); Conductivity (17389 samples, 81 features)We normalize and partition our datasets into (80%) training and (20%) test sets. In all our experiments, both DSGLD and FSGLD have the same hyper-parameters. We sample $S = 10$ disjoint data subsets for $r = 1000$ rounds each having 600 iteration per round, with fixed step-size $h_t = 10^{-5}$ and mini-batch size $m = 10$. All shards are chosen with same probability $f_1 = \cdots = f_S = 1/S$. We also burn-in the first ten thousand samples and thin the remaining by a hundred. We set $q_s(\theta) = \mathcal{N}(\theta|\mu, \Sigma)$, with $\Sigma = (\mathbf{x}^\top \mathbf{x})^{-1}$ and $\mu = (\sum y_i x_i)\Sigma^{-1}$, for each $s = 1 \ldots S$. We repeated the same experiment for 10 different random seeds. We report the average test MSE and its variance as a function of the number of posterior samples.

Results Figure 1 shows the cumulative MSE and its variance. Overall, FSGLD converges faster than DSGLD in MSE, with the notable exception of the Conductivity dataset, for which both methods converge virtually at the same time. In the case of the Noise dataset, FSGLD additionally converges to a much lower MSE. Notably, our method also presents clearly lower variance for all datasets.

F.2 METRIC LEARNING

While we employed MCMC-based surrogates for $q_1, \ldots, q_S$, we can also use FSGLD with coarser approximations. As an example, we also run FSGLD on the metric learning posterior of subsection 5.2 using Laplace approximations. Notably, Figure 2 shows that both options lead to similar average results, but Laplace approximations result in higher variance.

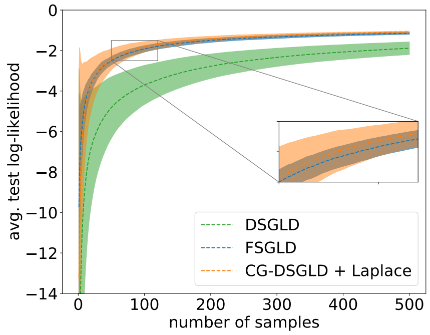

Figure 2: In avgerage, CG-DSGLD with Laplace or MCMC-based q_s's perform on par. Notably, MCMC yields smaller variance in the metric learning experiment.

[1] Datasets can be downloaded from `https://archive.ics.uci.edu/ml/index.html`

References

Sungjin Ahn, Babak Shahbaba, and Max Welling. Distributed stochastic gradient MCMC. In *International Conference on Machine Learning (ICML)*, 2014.

Jack Baker, Paul Fearnhead, Emily B. Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29(3):599–615, May 2019.

Changyou Chen, Nan Ding, and Lawrence Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Tianqi Chen, Emily B. Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning (ICML)*, 2014.

S. De and T. Goldstein. Efficient distributed sgd with variance reduction. In *International Conference on Data Mining (ICDM)*, 2016.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014a.

Aaron Defazio, Justin Domke, and Caetano. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning (ICML)*, 2014b.

Kumar Avinava Dubey, Sashank J. Reddi, Sinead A. Williamson, Barnabás Póczos, Alexander J. Smola, and Eric P. Xing. Variance reduction in stochastic gradient langevin dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Jakub Konecný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *ArXiv e-print*, 2016.

Boyue Li, Shicong Cen, Yuxin Chen, and Yuejie Chi. Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. In *Artificial Intelligence and Statistics (AISTATS)*, 2020.

Chunyuan Li, Changyou Chen, Yunchen Pu, Ricardo Henao, and Lawrence Carin. Communication-efficient stochastic gradient MCMC for neural networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Diego Mesquita, Paul Blomstedt, and Samuel Kaski. Embarrassingly parallel MCMC using deep invertible transformations. In *Uncertainty in Artificial Intelligence (UAI)*, 2019.

Tigran Nagapetyan, Andrew B. Duncan, Leonard Hasenclever, Sebastian J. Vollmer, Lukasz Szpruch, and Konstantinos Zygalakis. The true cost of stochastic gradient Langevin dynamics. *ArXiv e-print*, 2017.

Radford Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, New York, 2011.

Willie Neiswanger, Chong Wang, and Eric P. Xing. Asymptotically exact, embarrassingly parallel MCMC. In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

C. Nemeth and C. Sherlock. Merging MCMC subposteriors through Gaussian-process approximations. *Bayesian Analysis*, 13(2):507–530, 2018.

Brian D. Ripley. *Stochastic Simulation*. John Wiley & Sons, Inc., USA, 1987.

Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(1):193–225, 2016.

Alexander Terenin, Daniel Simpson, and David Draper. Asynchronous gibbs sampling. In *Artificial Intelligence and Statistics (AISTATS)*, 2020.

Sebastian J Vollmer, Konstantinos C Zygalakis, and Yee Whye Teh. Exploration of the (non-) asymptotic bias and variance of stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(1):5504–5548, 2016.

Xiangyu Wang, Fangjian Guo, Katherine A. Heller, and David B. Dunson. Parallelizing MCMC with random partition trees. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning (ICML)*, 2011.

Liu Yang, Rong Jin, and Rahul Sukthankar. Bayesian active distance metric learning. In *Uncertainty in Artificial Intelligence (UAI)*, 2007.

Yuan Yang, Jianfei Chen, and Jun Zhu. Distributing the stochastic gradient sampler for large-scale LDA. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.