Supplementary Material: Distributed Stochastic Gradient MCMC

Sungjin Ahn SUNGJIA@ICS.UCI.EDU
Department of Computer Science, University of California, Irvine

Babak Shahbaba BABAKS@UCI.EDU
Department of Statistics, University of California, Irvine

Max Welling M.WELLING@UVA.NL
Machine Learning Group, University of Amsterdam

1. Valid SGLD Estimators

Definition 1. An estimator $f(\theta, Z; X)$, where Z is a set of auxiliary random variables associated with the estimator, is said to be a *valid SGLD estimator* if $\mathbb{E}_Z[f(\theta, Z; X)] = \bar{g}(\theta; X)$, where $\mathbb{E}_Z$ denotes expectation w.r.t. the distribution $p(Z; X)$ and it has finite variance $\mathbb{V}_Z[f(\theta, Z; X)] < \infty$.

Proposition 1.1. *For each shard $s = 1, \ldots, S$, given shard size, N_s, and the normalized shard selection frequency, q_s, such that $N_s > 0$, $\sum_{s=1}^S N_s = N$, $q_s \in (0, 1)$, and $\sum_{s=1}^S q_s = 1$, the following estimator is a valid SGLD estimator,*

$$\bar{g}_d(\theta; X_s^n) \stackrel{def}{=} \frac{N_s}{N q_s} \bar{g}(\theta_t; X_s^n) \tag{1}$$

where shard s is sampled by a scheduler $h(\mathcal{Q})$ with frequencies $\mathcal{Q} = \{q_1, \ldots, q_S\}$.

Proof. We first decompose the expectation of the estimator $\mathbb{E}[\bar{g}_d(\theta; X_s^n)|X]$ w.r.t. (1) the shard s and (2) the minibatch X_s^n conditioned on the shard s, as follows

$$\mathbb{E}[\bar{g}_d(\theta; X_s^n)|X] = \mathbb{E}_s[\mathbb{E}_{X_s^n}[\bar{g}_d(\theta; X_s^n)|s]|X]. \tag{2}$$

Then, plugging Eqn. (1) in Eqn. (2) and rearranging, we obtain

$$\begin{aligned} &= \mathbb{E}_s\left[\mathbb{E}_{X_s^n}\left[\frac{N_s}{nNq_s}\sum_{x\in X_s^n} g(\theta; x)\middle| s\right] \middle| X\right] \\ &= \mathbb{E}_s\left[\frac{N_s}{Nq_s}\mathbb{E}_{X_s^n}\left[\frac{1}{n}\sum_{x\in X_s^n} g(\theta; x)\middle| s\right] \middle| X\right]. \end{aligned} \tag{3}$$

Note here that given X, the inner expectation w.r.t. the minibatches of shard s, X_s^n, is equal to the mean score over the shard X_s. That is,

$$\mathbb{E}_{X_s^n}\left[\frac{1}{n}\sum_{x\in X_s^n} g(\theta; x)\middle| s, X\right] = \frac{1}{N_s}\sum_{x\in X_s} g(\theta; x). \tag{4}$$

Substituting this for the inner expectation, in Eqn. (3), we have

$$\mathbb{E}_s\left[\frac{N_s}{Nq_s}\frac{1}{N_s}\sum_{x\in X_s} g(\theta; x)\right] \tag{5}$$

$$= \frac{1}{N}\mathbb{E}_s\left[\frac{1}{q_s}\sum_{x\in X_s} g(\theta; x)\right] \tag{6}$$

$$= \frac{1}{N}\sum_{s=1}^S p(s)\frac{1}{q_s}\sum_{x\in X_s} g(\theta; x). \tag{7}$$

Because we choose a shard s by $h(\mathcal{Q})$, $p(s)$ is equal to q_s. Thus, by plugging $p(s) = q_s$ in Eqn. (7) and rearranging, we obtain

$$\begin{aligned} &= \frac{1}{N}\sum_{s=1}^S q_s \frac{1}{q_s}\sum_{x\in X_s} g(\theta; x) \\ &= \frac{1}{N}\sum_{s=1}^S \sum_{x\in X_s} g(\theta; x) \\ &= \frac{1}{N}\sum_{x\in X} g(\theta; x) \\ &= \bar{g}(\theta; X). \end{aligned} \tag{8}$$

which completes the proof for the validity of the estimator $\bar{g}_d$,

$$\mathbb{E}[\bar{g}_d(\theta; X_s^n)|X] = \bar{g}(\theta; X). \tag{9}$$

□

Corollary 1.2. *A trajectory sampler with a finite* $\tau \geq 1$*, obtained by redefining the worker (shard) selection process* $h(\mathcal{Q})$ *in Proposition 1.1 by the process* $h(\mathcal{Q}, \tau)$ *below, is a valid SGLD sampler.* $h(\mathcal{Q}, \tau)$: *for chain* c *at iteration* t*, choose the next worker* s^c_{t+1} *by*

$$s^c_{t+1} = \begin{cases} \tilde{h}(\mathcal{Q}), & \text{if } t = k\tau \text{ for } k = 0, 1, 2, \ldots \\ s^c_t, & \text{otherwise,} \end{cases} \tag{10}$$

where $\tilde{h}(\mathcal{Q})$ *is an arbitrary scheduler with selection probabilities* $\mathcal{Q}$.

Proof. Because the trajectory lengths are all equal to τ for all workers $s = 1, \ldots, S$ and $\tilde{h}(\mathcal{Q})$ conforms to the frequencies $\mathcal{Q}$, the worker (shard) selection frequencies of the trajectory sampling process $h(\mathcal{Q}, \tau)$ also satisfies $\mathcal{Q}$. As a result, in the proof of Proposition 1.1, the probability $p(s) = q_s$ is retained even if we replace $h(\mathcal{Q})$ in Proposition 1.1 by $h(\mathcal{Q}, \tau)$. Because changing the worker selection process only affects $p(s)$ in the proof of Proposition 1.1, the proof directly applies to the corollary. □

Corollary 1.3. *Given* τ_s*, where* $1 \leq \tau_s < \infty$ *for* $s = 1, \ldots, S$*, the adaptive trajectory sampler, obtained by redefining the worker (shard) selection process* $h(\mathcal{Q})$ *in Proposition 1.1 by the process* $h(\mathcal{Q}, \{\tau_s\})$ *below, is a valid SGLD sampler.* $h(\mathcal{Q}, \{\tau_s\})$: *for chain* c *at iteration* t*, choose the next worker* s^c_{t+1} *by*

$$s^c_{t+1} = \begin{cases} \tilde{h}(1/S), & \text{if } t = k\tau_{s^c_t} \text{ for } k = 0, 1, 2, \ldots \\ s^c_t, & \text{otherwise,} \end{cases} \tag{11}$$

where $\tilde{h}(1/S)$ *is a scheduler with uniform selection probabilities.*

Proof. Because we select the worker uniformly by $\tilde{h}(1/S)$, only the trajectory lengths $\{\tau_{s^1}, \ldots, \tau_{s^C}\}$ affect the shard selection frequency of the process $h(\mathcal{Q}, \{\tau_s\})$. Since the trajectory length τ_s is proportional to q_s ($\tau_s \stackrel{\text{def}}{=} \bar{\tau} S q_s$), taking τ_s consecutive updates for uniformly selected random worker s satisfies the frequency $\mathcal{Q}$. Therefore, the proof of Proposition 1.1 also directly applies to the corollary. □