Asynchronous Stochastic Quasi-Newton MCMC for Non-Convex Optimization

SUPPLEMENTARY DOCUMENT

Umut Şimşekli [1] **Çağatay Yıldız** [2] **Thanh Huy Nguyen** [1] **Gaël Richard** [1] **A. Taylan Cemgil** [3]

1. The Approximate Euler-Maruyama Scheme

1.1. Connection with gradient descent with momentum

The standard Euler-Maruyama scheme for the SDE (8) can be developed as follows:

$$\theta_{n+1} = \theta_n + hH_n(\theta_n)p_n, \tag{S1}$$

$$p_{n+1} = p_n - hH_n(\theta_n)\nabla_\theta U(\theta_n) - h\gamma p_n + \frac{h}{\beta}\Gamma_n(\theta_n) + \sqrt{\frac{2h\gamma}{\beta}}Z_{n+1} \tag{S2}$$

$$= (1-h\gamma)p_n - hH_n(\theta_n)\nabla_\theta U(\theta_n) + \frac{h}{\beta}\Gamma_n(\theta_n) + \sqrt{\frac{2h\gamma}{\beta}}Z_{n+1} \tag{S3}$$

where h is the step-size and $\{Z_n\}_{n=1}^N$ is a collection of standard Gaussian random variables.

We can obtain simplified update rules if we define $u_n \triangleq hp_n$ and use it in (S3). The modified update rules are given as follows:

$$hp_{n+1} = hp_n - h^2 H_n(\theta_n)\nabla_\theta U(\theta_n) - h^2\gamma p_n + \frac{h^2}{\beta}\Gamma_n(\theta_n) + \sqrt{\frac{2h^3\gamma}{\beta}}Z_{n+1} \tag{S4}$$

$$u_{n+1} = u_n - h^2 H_n(\theta_n)\nabla_\theta U(\theta_n) - h\gamma u_n + \frac{h^2}{\beta}\Gamma_n(\theta_n) + \sqrt{\frac{2h^3\gamma}{\beta}}Z_{n+1} \tag{S5}$$

$$= \underbrace{(1-h\gamma)}_{\gamma'}u_n - \underbrace{h^2}_{h'}H_n(\theta_n)\nabla_\theta U(\theta_n) + \frac{h^2}{\beta}\Gamma_n(\theta_n) + \sqrt{\frac{2h^3\gamma}{\beta}}Z_{n+1} \tag{S6}$$

$$= \gamma' u_n - h' H_n(\theta_n)\nabla_\theta U(\theta_n) + \frac{h'}{\beta}\Gamma_n(\theta_n) + \sqrt{\frac{2h'(1-\gamma')}{\beta}}Z_{n+1}, \tag{S7}$$

where $\gamma' \in (0,1)$. If we use the modified Euler scheme as described in (Neal, 2010) and replace p_n with p_{n+1} in (S1), we obtain the following update equation:

$$\theta_{n+1} = \theta_n + hH_n(\theta_n)p_{n+1} \tag{S8}$$

$$= \theta_n + H_n(\theta_n)u_{n+1}. \tag{S9}$$

[1]LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France [2]Department of Computer Science, Aalto University, Espoo, 02150, Finland [3]Department of Computer Engineering, Boğaziçi University, 34342, Bebek, Istanbul, Turkey. Correspondence to: Umut Şimşekli <umut.simsekli@telecom-paristech.fr>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Note that, when $\beta \to \infty$ we have the following update rules:

$$u_{n+1} = \gamma' u_n - h' H_n(\theta_n) \nabla_\theta U(\theta_n) \tag{S10}$$
$$\theta_{n+1} = \theta_n + H_n(\theta_n) u_{n+1}, \tag{S11}$$

which coincides with Gradient descent with momentum when $H_n(\theta) = I$ for all n.

1.2. Numerical integration with stale stochastic gradients

We now focus on (S1) and (S2). We first drop the term Γ_n, replace the gradients ∇U with the stochastic gradients, and then modify the update rules by using stale parameters θ_{n-l_n} and p_{n-l_n}. The resulting scheme is given as follows:

$$\theta_{n+1} = \theta_n + h H_n(\theta_{n-l_n}) p_{n-l_n}, \tag{S12}$$
$$p_{n+1} = p_n - h H_n(\theta_{n-l_n}) \nabla_\theta \tilde{U}_n(\theta_{n-l_n}) - h\gamma p_{n-l_n} + \sqrt{\frac{2h\gamma}{\beta}} Z_{n+1}. \tag{S13}$$

By using a similar argument to the one used in Section 1.1, we define $u_n \triangleq h p_n$, $h' = h^2$, $\gamma' = h\gamma$, and obtain the following update equations:

$$\theta_{n+1} = \theta_n + H_n(\theta_{n-l_n}) u_{n-l_n}, \tag{S14}$$
$$u_{n+1} = u_n - h' H_n(\theta_{n-l_n}) \nabla_\theta \tilde{U}_n(\theta_{n-l_n}) - \gamma' u_{n-l_n} + \sqrt{\frac{2h'\gamma'}{\beta}} Z_{n+1}. \tag{S15}$$

2. Proof of Proposition 1

Proof. We start by rewriting the SDE given in (8) as follows:

$$dX_t = \left\{ - \left(\underbrace{\begin{bmatrix} 0 & 0 \\ 0 & \frac{\gamma}{\beta} I \end{bmatrix}}_{\mathbf{D}} + \underbrace{\begin{bmatrix} 0 & -\frac{H_t(\theta_t)}{\beta} \\ \frac{H_t(\theta_t)}{\beta} & 0 \end{bmatrix}}_{\mathbf{Q}_t(X)} \right) \underbrace{\begin{bmatrix} \beta \nabla_\theta U(\theta_t) \\ \beta p_t \end{bmatrix}}_{\nabla_X \mathcal{E}(X_t)} + \underbrace{\begin{bmatrix} 0 \\ \frac{1}{\beta} \Gamma_t(\theta_t) \end{bmatrix}}_{\mathbf{\Gamma}_t(X_t)} \right\} dt + \sqrt{2\mathbf{D}} dW_t. \tag{S16}$$

Here, we observe that $\mathbf{D}$ is positive semi-definite, $\mathbf{Q}$ is anti-symmetric. Furthermore, for all $i \in \{1, 2, \ldots, 2d\}$ we observe that

$$\Big[\mathbf{\Gamma}_t(X)\Big]_i = \sum_{j=1}^{2d} \frac{\partial [\mathbf{D} + \mathbf{Q}_t(X)]_{ij}}{\partial_{X_j}}. \tag{S17}$$

The assumptions **H**1 and 2 directly imply that the function $-(\mathbf{D}+\mathbf{Q}_t(X))\nabla_X \mathcal{E}(X) + \mathbf{\Gamma}_t(X)$ is locally Lipschitz continuous in X for all t. Then, the desired result is obtained by applying Theorem 1 of (Ma et al., 2015) and Proposition 4.2.2 of (Kunze, 2012). □

3. Proof of Lemma 1

3.1. Preliminaries

In the rest of this document, if there is no specification, the notation $\mathbb{E}\big[F\big]$ will denote the expectation taken over *all the random sources* contained in F.

Before providing the proof of Lemma 1, let us consider the following Itô diffusion:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \tag{S18}$$

where $X_t \in \mathbb{R}^{2d}$, $b : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$, $\sigma : \mathbb{R}^{2d} \to \mathbb{R}^{2d\times 2d}$, and W_t is Brownian motion in $\mathbb{R}^{2d}$. The generator $\mathcal{L}$ for (S18) is formally defined as follows:

$$\mathcal{L}f(X_t) \triangleq \lim_{h\to 0^+} \frac{\mathbb{E}[f(X_{t+h})|X_t] - f(X_t)}{h} = \Big(b(X_t)\cdot\nabla + \frac{1}{2}\big(\sigma(X_t)\sigma(X_t)^\top\big) : \nabla\nabla^\top\Big) f(X_t), \tag{S19}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is any twice differentiable function whose support is compact. Here, $a \cdot b$ denotes the inner product between vectors a and b, $A : B$ by definition is equal to $tr\{A^\top B\}$ for some matrices A and B. In our study, the generator for the diffusion (S16) is then defined as follows: (define $n = t/h$ and use (S19))

$$\mathcal{L}_n \triangleq \Big(H_n p_n \cdot \nabla_\theta - \big(H_n \nabla_\theta U(\theta_n) + \gamma p_n - \frac{1}{\beta}\Gamma_n(\theta_{n-l_n})\big)\cdot\nabla_p\Big) + \mathbf{D} : \nabla_X \nabla_X^\top. \tag{S20}$$

We also define the following operator for the approximate Euler-Maruyama scheme with delayed updates:

$$\tilde{\mathcal{L}}_n \triangleq \Big(H_n p_n \cdot \nabla_\theta - \big(H_n \nabla_\theta \tilde{U}(\theta_{n-l_n}) + \gamma p_n\big)\cdot\nabla_p\Big) + \mathbf{D} : \nabla_X \nabla_X^\top. \tag{S21}$$

By using the definitions $\mathcal{L}_n$ and $\tilde{\mathcal{L}}_n$, we obtain the following identity:

$$\tilde{\mathcal{L}}_n = \mathcal{L}_n - \Delta V_n, \tag{S22}$$

where $\Delta V_n \triangleq \Big(H_n(\theta_{n-l_n})(\nabla_\theta \tilde{U}_n(\theta_{n-l_n}) - \nabla_\theta U(\theta_n)) + \frac{1}{\beta}\Gamma(\theta_{n-l_n})\Big)\cdot\nabla_p$. This operator essentially captures all the errors induced by the approximate integrator.

We now proceed to the proof of Lemma 1. The proof uses several technical lemmas that are given in Section 6.

3.2. Proof of Lemma 1

Proof. We first consider the Euler-Maruyama integrator of the SDE (S16), to combine (S1) and (S3) into a single equation, given as follows:

$$X_{n+1} = X_n - h(\mathbf{D} + \mathbf{Q}_n(X_n))\nabla\mathcal{E}(X_n) + h\mathbf{\Gamma}_{n+1}(X_n) + \sqrt{2h\mathbf{D}}Z'_{n+1}$$

where Z'_n is a standard Gaussian random variable in $\mathbb{R}^{2d}$, h is the step-size, $\mathbf{D}$, $\mathbf{Q}$, and $\mathbf{\Gamma}$ are defined in (S16). We then modify this scheme such that we replace the gradient $\nabla\mathcal{E}$ with the stale stochastic gradients and we discard the term $\mathbf{\Gamma}$. The resulting numerical integrator is given as follows:

$$X_{n+1} = X_n - h(\mathbf{D} + \mathbf{Q}_n(X_{n-l_n}))\nabla\tilde{\mathcal{E}}_n(X_{n-l_n}) + \sqrt{2h\mathbf{D}}Z'_{n+1}. \tag{S23}$$

Note that (S23) coincides with the proposed algorithm, given in (5).

In the sequel, we follow a similar strategy to (Chen et al., 2016b). However, we have additional difficulties caused by the usage of L-BFGS matrices, which are reflected in the operator ΔV_n. Since we are using the Euler-Maruyama integrator, we have the following inequality (Chen et al., 2015):

$$\mathbb{E}[\psi(X_n)|X_{n-1}] = (\mathbb{I} + h\tilde{\mathcal{L}}_n)\psi(X_{n-1}) + \mathcal{O}(h^2). \tag{S24}$$

By summing both sides of (S24) over n, taking the expectation, and using (S22), we obtain the following:

$$\sum_{n=1}^{N}\mathbb{E}[\psi(X_n)] = \psi(X_0) + \sum_{n=1}^{N-1}\mathbb{E}[\psi(X_n)] - h\sum_{n=1}^{N}\mathbb{E}[\Delta V_n\psi(X_{n-1})] + h\sum_{n=1}^{N}\mathbb{E}[\mathcal{L}_n\psi(X_{n-1})] + \mathcal{O}(Nh^2). \tag{S25}$$

By rearranging the terms and dividing all the terms by Nh, we obtain:

$$\frac{\mathbb{E}\psi(X_N) - \psi(X_0)}{Nh} = \frac{-\sum_{n=1}^{N}\mathbb{E}[\Delta V_n\psi(X_{n-1})] + \sum_{n=1}^{N}\mathbb{E}[\mathcal{L}_n\psi(X_{n-1})]}{N} + \mathcal{O}(h). \tag{S26}$$

By using the Poisson equation given in (12) for each $\mathcal{L}_n\psi(X_{n-1})$ and rearranging the terms, we obtain:

$$\mathbb{E}[\frac{1}{N}\sum_n(U(\theta_n)-\bar{U}_\beta)] = \frac{\mathbb{E}[\psi(X_N)]-\psi(X_0)}{Nh} + \frac{\sum_{n=1}^N \mathbb{E}[\Delta V_n \psi(X_{n-1})]}{N} + \mathcal{O}(h). \tag{S27}$$

By assumption **H**3, the term $\mathbb{E}[\psi(X_N)] - \psi(X_0)$ is uniformly bounded. Then, by Assumption **H**3 and Lemma S3, we obtain the following bound:

$$\mathbb{E}[\frac{1}{N}\sum_n(U(\theta_n)-\bar{U}_\beta)] = \mathcal{O}\Big(\frac{1}{Nh} + \max(l_{\max},1)h + \frac{1}{\beta}\Big), \tag{S28}$$

as desired. □

Remark 1. *Theorem 1 significantly differentiates from other recent results. First of all, none of the references we are aware of provides an analysis for an asynchronous stochastic L-BFGS algorithm. Aside from this fact, when compared to (Chen et al., 2016a), our bound handles the case of delayed updates and provides an explicit dependence on β. When compared to (Chen et al., 2016b), our analysis considers the tempered case and handles the additional difficulties brought by the L-BFGS matrices and their derivatives. On the other hand, our analysis is also significantly different than the ones presented in (Raginsky et al., 2017) and (Xu et al., 2017), as it considers the asynchrony and L-BFGS matrices, and provides a bound for the ergodic error.*

4. Proof of Lemma 2

Proof. We use the same proof technique given in (Raginsky et al., 2017)[Proposition 11]. We assume that π_θ admits a density with respect to the Lebesgue measure, denoted as $\rho(\theta) \triangleq \frac{1}{Z_\beta}\exp(-\beta U(\theta))$, where Z_β is the normalization constant: $Z_\beta \triangleq \int_{\mathbb{R}^d}\exp(-\beta U(\theta))d\theta$. We start by using the definition of $\bar{U}_\beta$, as follows:

$$\bar{U}_\beta = \int_{\mathbb{R}^d} U(\theta)\pi_\theta(d\theta) = \frac{1}{\beta}(\mathcal{H}(\rho) - \log Z_\beta), \tag{S29}$$

where $\mathcal{H}(\rho)$ is the *differential entropy*, defined as follows:

$$\mathcal{H}(\rho) \triangleq -\int_{\mathbb{R}^d}\rho(\theta)\log\rho(\theta)d\theta. \tag{S30}$$

We now aim at upper-bounding $\mathcal{H}(\rho)$ and lower-bounding $\log Z_\beta$. By Assumption **H**6, the distribution π_θ has a finite second order moment, therefore its differential entropy is upper-bounded by the differential entropy of a Gaussian distribution that has the same second order moment. Then, we obtain

$$\mathcal{H}(\rho) \le \frac{1}{2}\log[(2\pi e)^d\det(\Sigma)] \tag{S31}$$

$$\le \frac{1}{2}\log[(2\pi e)^d\Big(\frac{\mathrm{tr}(\Sigma)}{d}\Big)^d] \tag{S32}$$

$$\le \frac{d}{2}\log\Big(2\pi e\frac{C_\beta}{\beta d}\Big), \tag{S33}$$

where Σ denotes the covariance matrix of the Gaussian distribution. In (S32) we used the relation between the arithmetic and geometric means, and in (S33) we used Assumption **H**6.

We now lower-bound $\log Z_\beta$. By definition, we have

$$\log Z_\beta = \log\int_{\mathbb{R}^d}\exp(-\beta U(\theta))d\theta \tag{S34}$$

$$= -\beta U^\star + \log\int_{\mathbb{R}^d}\exp(\beta(U^\star - U(\theta)))d\theta \tag{S35}$$

$$\ge -\beta U^\star + \log\int_{\mathbb{R}^d}\exp(-\frac{\beta L\|\theta-\theta^\star\|^2}{2})d\theta \tag{S36}$$

$$= -\beta U^\star + \frac{d}{2}\log(\frac{2\pi}{L\beta}). \tag{S37}$$

Here, in (S36) we used Assumption **H**1 and (Nesterov, 2013)(Lemma 1.2.3).

Finally, by combining (S29), (S33), and (S37), we obtain

$$\bar{U}_\beta - U^\star = \frac{1}{\beta}(\mathcal{H}(\rho) - \log Z_\beta) - U^\star \tag{S38}$$

$$\leq \frac{1}{\beta}\Big(\frac{d}{2}\log\big(2\pi e\frac{C_\beta}{\beta d}\big) + \beta U^\star - \frac{d}{2}\log(\frac{2\pi}{L\beta})\Big) - U^\star \tag{S39}$$

$$= \frac{1}{\beta}\frac{d}{2}\log\Big(\frac{eC_\beta L}{d}\Big) \tag{S40}$$

$$= \mathcal{O}\Big(\frac{1}{\beta}\Big). \tag{S41}$$

This finalizes the proof. □

5. Proof of Theorem 1

Proof. We decompose the error, as follows:

$$\Big|\mathbb{E}\hat{U}_N - U^\star\Big| = \Big|\mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}(U(\theta_n) - U^\star)]\Big| \tag{S42}$$

$$= \Big|\mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}\big(U(\theta_n) - \bar{U}_\beta]\big) + \big(\bar{U}_\beta - U^\star\big)\Big| \tag{S43}$$

$$\leq \underbrace{\Big|\mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}\big(U(\theta_n) - \bar{U}_\beta]\big)\Big|}_{\mathcal{A}_1} + \underbrace{\Big(\bar{U}_\beta - U^\star\Big)}_{\mathcal{A}_2}, \tag{S44}$$

where the term $\mathcal{A}_1$ is bounded by Lemma 1 and the term $\mathcal{A}_2$ is bounded by Lemma 2. This finalizes the proof. □

6. Technical Lemmas

For convenience, let us introduce the following notations: $\bar{X}_k \triangleq (X_0, \ldots, X_k)$. Let us also denote Ω_n the (uniform) random subsample, which is chosen independently of $(\bar{X}_n)$, used for iteration n.

Lemma S1. *Let $f_k(X) \triangleq \|X - X_{k-1}\|$. Under the assumptions **H**2-5, the following bound holds:*

$$\mathbb{E}_{\bar{X}_n}\big[\|\nabla_\theta U(\theta_{n-l_n}) - \nabla_\theta U(\theta_n)\|\big] = \mathcal{O}\Big(l_{max}h\max_{i\in[\![n-l_{max}+1,n]\!]}\mathbb{E}\big[\mathcal{L}_i f_i(X_{i-1})\big] + h^2\Big) \tag{S45}$$

where $\mathbb{E}_{\bar{X}_k}$ denotes the expectation taken over the random variables $X_0, \ldots, X_k$.

Proof. The proof is similar to [(Chen et al., 2016b), Lemma 8], we provide the proof for completeness. We first consider the following estimate which uses the Lipschitz property of $\nabla_\theta U(\theta)$:

$$\begin{aligned}
\mathbb{E}_{\bar{X}_n}\big[\|\nabla_\theta U(\theta_{n-l_n}) - \nabla_\theta U(\theta_n)\|\big] &\leq L\mathbb{E}_{\bar{X}_n}\big[\|\theta_{n-l_n} - \theta_n\|\big] \\
&\leq L\mathbb{E}_{\bar{X}_n}\Big[\Big\|\sum_{i=n-l_n}^{n-1}(\theta_i - \theta_{i+1})\Big\|\Big] \\
&\leq L\sum_{i=n-l_n}^{n-1}\mathbb{E}_{\bar{X}_n}\Big[\Big\|\theta_i - \theta_{i+1}\Big\|\Big] \\
&\leq L\sum_{i=n-l_n}^{n-1}\mathbb{E}_{\bar{X}_n}\Big[\Big\|X_i - X_{i+1}\Big\|\Big] \\
&= L\sum_{i=n-l_n}^{n-1}\mathbb{E}_{\bar{X}_n}\big[f_{i+1}(X_{i+1})\big].
\end{aligned} \tag{S46}$$

Using law of total expectation, we have

$$\begin{aligned}\mathbb{E}_{\bar{X}_n}\big[f_{i+1}(X_{i+1})\big] &= \mathbb{E}\big[f_{i+1}(X_{i+1})\big]\\ &= \mathbb{E}\big[\mathbb{E}\big[f_{i+1}(X_{i+1})|X_i\big]\big]\\ &= \mathbb{E}\big[e^{h\mathcal{L}_{i+1}}f_{i+1}(X_i)+\mathcal{O}(h^2)\big]\\ &= \mathbb{E}\big[f_{i+1}(X_i)+h\mathcal{L}_{i+1}f_{i+1}(X_i)+\mathcal{O}(h^2)\big]\\ &\leq h\mathbb{E}\big[\mathcal{L}_{i+1}f_{i+1}(X_i)\big]+\mathcal{O}(h^2). \end{aligned} \tag{S47}$$

The third equality is due to the fact that Euler integrator is a first order integrator. Then we applied Assumption **H**5 and $f_{i+1}(X_i)=0$ to obtain the last two lines. Finally, by combining (S46) and (S47), we obtain:

$$\begin{aligned}\mathbb{E}_{\bar{X}_n}\big[\|\nabla_\theta U(\theta_{n-l_n})-\nabla_\theta U(\theta_n)\|\big] &\leq L\sum_{i=n-l_n}^{n-1}\big(h\mathbb{E}\big[\mathcal{L}_{i+1}f_{i+1}(X_i)\big]+\mathcal{O}(h^2)\big)\\ &\leq L\sum_{i=n-l_{\max}}^{n-1}\big(h\mathbb{E}\big[\mathcal{L}_{i+1}f_{i+1}(X_i)\big]+\mathcal{O}(h^2)\big)\\ &\leq Ll_{\max}h\max_{i\in[\![n-l_{\max}+1,n]\!]}\mathbb{E}\big[\mathcal{L}_i f_i(X_{i-1})\big]+\mathcal{O}(h^2).\end{aligned}$$

This completes the proof. □

Lemma S2. *If Assumption **H**2 holds then the following bound holds:*

$$\|\mathbf{\Gamma}_n\| = \mathcal{O}(\frac{1}{\beta}), \tag{S48}$$

where $\mathbf{\Gamma}_n$ is defined in (S16)*.*

Proof. If $l_n > 0$ then $\|\Gamma_n(\theta_n)\| = 0$ since H_n will not depend on θ_n (see (9) for the definition of Γ_n). For $l_n = 0$, by the Lipschitz continuity of H_n, the first order partial derivatives of H_n are all bounded by L_H. Then, $\|\mathbf{\Gamma}_n\| = \frac{1}{\beta}\|\Gamma_n\|$ is therefore bounded by a quantity that is proportional to β^{-1}. □

Lemma S3. *Let $f_k(X) \triangleq \|X - X_{k-1}\|$. Under the assumptions **H**2-5, the following bound holds:*

$$\mathbb{E}\big[\Delta V_n\psi(X_{n-1})\big] = \mathcal{O}\Big(l_{max}h\max_{i\in[\![n-l_{max}+1,n]\!]}\mathbb{E}\big[\mathcal{L}_i f_i(X_{i-1})\big]+h^2+\beta^{-1}\Big). \tag{S49}$$

Proof. First, by using the triangular inequality we have:

$$\begin{aligned}\|\mathbb{E}\big[\Delta V_n\psi(X_{n-1})\big]\| =& \|\mathbb{E}\big[(H_n(\theta_{n-l_n})(\nabla_\theta\tilde{U}_{n-l_n}(\theta_{n-l_n})-\nabla_\theta U(\theta_n))+\frac{1}{\beta}\Gamma_n(\theta_{n-l_n}))\cdot\nabla_p\psi(X_{n-1})\big]\|\\ \leq& \|\mathbb{E}\big[H_n(\theta_{n-l_n})(\nabla_\theta\tilde{U}_{n-l_n}(\theta_{n-l_n})-\nabla_\theta U(\theta_n))\cdot\nabla_p\psi(X_{n-1})\big]\|\\ &+\|\mathbb{E}\big[\frac{1}{\beta}\Gamma_n(\theta_{n-l_n})\cdot\nabla_p\psi(X_{n-1})\big]\|.\end{aligned} \tag{S50}$$

Applying Assumption **H**3 and Lemma S2, we obtain the bound for the second term in the above sum:

$$A_1 \triangleq \|\mathbb{E}\big[\frac{1}{\beta}\Gamma_n(\theta_{n-l_n})\cdot\nabla_p\psi(X_{n-1})\big]\| = \mathcal{O}(\beta^{-1}). \tag{S51}$$

We note that the expectation is taken over $(\bar{X}_n, \Omega_n)$, where $\bar{X}_n$ and Ω_n are independent. Hence, the first term in (S50) can be rewritten as follows:

$$\begin{aligned}A_2 \triangleq& \|\mathbb{E}\big[H_n(\theta_{n-l_n})(\nabla_\theta\tilde{U}_{n-l_n}(\theta_{n-l_n})-\nabla_\theta U(\theta_n))\cdot\nabla_p\psi(X_{n-1})\big]\|\\ =& \|\mathbb{E}_{\bar{X}_n}\big[\mathbb{E}_{\Omega_n}\big[H_n(\theta_{n-l_n})(\nabla_\theta\tilde{U}_{n-l_n}(\theta_{n-l_n})-\nabla_\theta U(\theta_n))\cdot\nabla_p\psi(X_{n-1})\big]\big]\|\\ =& \Big\|\mathbb{E}_{\bar{X}_n}\Big[\mathbb{E}_{\Omega_n}\big[H_n(\theta_{n-l_n})(\nabla_\theta\tilde{U}_{n-l_n}(\theta_{n-l_n})-\nabla_\theta\tilde{U}_{n-l_n}(\theta_n))\cdot\nabla_p\psi(X_{n-1})\big]+\mathbb{E}_{\Omega_n}\big[H_n(\theta_{n-l_n})(\nabla_\theta\tilde{U}_{n-l_n}(\theta_n)\\ &-\nabla_\theta U_n(\theta_n))\cdot\nabla_p\psi(X_{n-1})\big]\Big]\Big\|.\end{aligned}$$

As H_n and $\nabla_p \psi(X_{n-1})$ are independent of the random subsample Ω_n, we have

$$\begin{aligned}\mathbb{E}_{\Omega_n}\left[H_n(\theta_{n-l_n})(\nabla_\theta \tilde{U}_{n-l_n}(\theta_n) - \nabla_\theta U_n(\theta_n)) \cdot \nabla_p \psi(X_{n-1})\right] &= H_n(\theta_{n-l_n})\mathbb{E}_{\Omega_n}\left[\nabla_\theta \tilde{U}_{n-l_n}(\theta_n) - \nabla_\theta U_n(\theta_n)\right] \cdot \nabla_p \psi(X_{n-1}) \\ &= 0.\end{aligned}$$

As a result,

$$\begin{aligned}A_2 &= \|\mathbb{E}_{\bar{X}_n}\left[\mathbb{E}_{\Omega_n}\left[H_n(\theta_{n-l_n})(\nabla_\theta \tilde{U}_{n-l_n}(\theta_{n-l_n}) - \nabla_\theta \tilde{U}_{n-l_n}(\theta_n)) \cdot \nabla_p \psi(X_{n-1})\right]\right]\| \\ &= \|\mathbb{E}_{\bar{X}_n}\left[H_n(\theta_{n-l_n})(\nabla_\theta U(\theta_{n-l_n}) - \nabla_\theta U(\theta_n)) \cdot \nabla_p \psi(X_{n-1})\right]\| \\ &\leq C\mathbb{E}_{\bar{X}_n}\left[\|\nabla_\theta U(\theta_{n-l_n}) - \nabla_\theta U(\theta_n)\|\right] \\ &= \mathcal{O}\Big(l_{\max} h \max_{i \in [\![n-l_{\max}+1, n]\!]} \mathbb{E}\left[\mathcal{L}_i f_i(X_{i-1})\right] + h^2\Big). \end{aligned} \tag{S52}$$

The inequality in (S52) is deduced from the fact that H_n is bounded by (Berahas et al., 2016)[Lemma3.3] and $\nabla_p \psi(X_{n-1})$ is bounded by assumptions, and the last equality is due to Lemma S1. Finally, by combining (S50), (S51), and (S52), we obtain (S49), which concludes the proof. □

7. Additional Experimental Results

In this section, we provide the result where we illustrate the iteration speedup of as-L-BFGS on the ML-1M dataset.

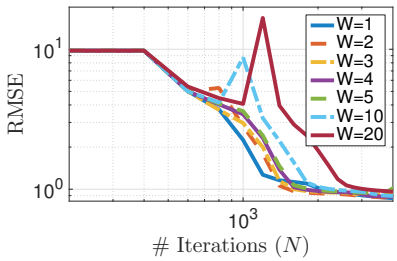

Figure S1. The convergence behavior of as-L-BFGS on the ML-1M dataset for increasing number of workers.

8. Algorithm Parameters Used in the Experiments

8.1. Linear Gaussian model

Table 1 lists the algorithm parameters for the synthetic data experiments. We fixed the L-BFGS memory sizes for mb-L-BFGS and as-L-BFGS to $M = 3$. The remaining parameters are the step sizes (h, h'), timeout duration of mb-L-BFGS server (T_{mb}), the friction parameter (γ'), and the inverse temperature (β) of as-L-BFGS.

Table 1. The list of algorithm parameters that are used in the experiments on the linear Gaussian model.

a-SGD	mb-L-BFGS		as-L-BFGS		
h	h	T_{mb} (base units)	h'	γ'	β
1×10^{-3}	5×10^{-2}	10	4×10^{-4}	3×10^{-2}	5×10^{2}

Table 2 lists the parameters of the simulator. The parameters are (i) μ_m: the average computational time spent by the master node at each iteration, (ii) μ_w: the average computational time spent by a single worker at each iteration, and (iii) τ: the time spent for communication per iteration. In all cases we set $\tau = 10$, $N_\Omega = N_Y/100$, $N_O = N_\Omega/3$.

8.2. Large-scale matrix factorization

Table 3 lists the algorithm parameters for different data sets. We fixed the L-BFGS memory sizes for mb-L-BFGS and as-L-BFGS to $M = 3$. In all experiments we set $\rho = 3$, $N_\Omega = N_Y/100$, $N_O = N_\Omega/3$.

Table 2. The list of simulator parameters that are used in the experiments on the linear Gaussian model.

a-SGD		mb-L-BFGS		as-L-BFGS	
μ_m	μ_w	μ_m	μ_w	μ_m	μ_w
0	$1000 \times \frac{N_\Omega}{N}$	30	$1000 \times \frac{N_\Omega}{N_Y}$	0	$1000 \times \frac{N_\Omega}{N_Y} + 60$

Table 3. The list of algorithm parameters that are used in the experiments on the large scale matrix factorization.

	a-SGD	mb-L-BFGS		as-L-BFGS		
	h	h	T_{mb} (m. sec.)	h'	γ'	β
ML-1M	1×10^{-6}	5×10^{-7}	400	2×10^{-8}	1×10^{-1}	1×10^{3}
ML-10M	2×10^{-7}	1×10^{-8}	3400	1×10^{-9}	3×10^{-2}	1×10^{3}
ML-20M	1×10^{-7}	1×10^{-8}	4500	1×10^{-9}	1×10^{-3}	1×10^{3}

References

Berahas, Albert S, Nocedal, Jorge, and Takác, Martin. A multi-batch L-BFGS method for machine learning. In *Advances in Neural Information Processing Systems*, pp. 1055–1063, 2016.

Chen, C., Ding, N., and Carin, L. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Advances in Neural Information Processing Systems*, pp. 2269–2277, 2015.

Chen, C., Carlson, D., Gan, Z., Li, C., and Carin, L. Bridging the gap between stochastic gradient MCMC and stochastic optimization. In *AISTATS*, 2016a.

Chen, C., Ding, N., Li, C., Zhang, Y., and Carin, L. Stochastic gradient MCMC with stale gradients. In *Advances In Neural Information Processing Systems*, pp. 2937–2945, 2016b.

Kunze, M. Stochastic differential equations. Lecture notes, University of Ulm, 2012.

Ma, Y. A., Chen, T., and Fox, E. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pp. 2899–2907, 2015.

Neal, R. M. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54, 2010.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pp. 1674–1703, 2017.

Xu, P., Chen, J., and Gu, Q. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *arXiv preprint arXiv:1707.06618*, 2017.