Stochastic Quasi-Newton Langevin Monte Carlo

SUPPLEMENTARY DOCUMENT

Umut Şimşekli UMUT.SIMSEKLI@TELECOM-PARISTECH.FR
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France

Roland Badeau ROLAND.BADEAU@TELECOM-PARISTECH.FR
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France

A. Taylan Cemgil TAYLAN.CEMGIL@BOUN.EDU.TR
Dept. of Computer Engineering, Boğaziçi University, 34342, Bebek, İstanbul, Turkey

Gaël Richard GAEL.RICHARD@TELECOM-PARISTECH.FR
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France

1. The L-BFGS Algorithm

In this section, we illustrate the two-loop recursion of the L-BFGS algorithm in Algorithm 1. The pseudo-code is based on (Nocedal & Wright, 2006).

Algorithm 1: L-BFGS two loop recursion.

```
input : $g_t = \nabla\tilde{U}_{\Omega_t}(\theta_t)$, $M$, $H_t^1 = \gamma I$
output: $\xi = H_t g_t$
$\xi_t \leftarrow g_t$
for $i = t-1, \cdots, t-M+1$ do
    $\rho_i \leftarrow \frac{1}{y_i^\top s_i}$
    $\alpha_i \leftarrow \rho_i s_i^\top \xi$
    $\xi \leftarrow \xi - \alpha_i y_i$
$\xi \leftarrow \gamma\xi$
for $i = t-M+1, \cdots, t-1$ do
    $\beta \leftarrow \rho_i y_i^\top \xi$
    $\xi \leftarrow \xi + s_i(\alpha_i - \beta)$
return $\xi$
```

2. Proof of the Main Theorem

In this section, we provide the proof of the main theorem. For completeness, let us first introduce certain definitions and theorems that will be used in our proof. Note that we have translated the following theorems to our notation and framework.

2.1. Preliminaries

Definition 1 (Hölder Continuity, (Duan, 2015)(p. 22)).
A real valued function f on $\mathbb{R}^D$ is called uniformly Hölder continuous, if there exist $K, \alpha > 0$ such that

$$|f(x) - f(y)| \leq K|x - y|^\alpha \tag{1}$$

for $x, y \in \mathcal{B} = \mathbf{dom}(f)$. We use $\mathcal{C}^{\alpha}(\mathcal{B})$ for denoting the space of functions that are locally Hölder continuous in $\mathcal{B}$ with exponent α. We also use $\mathcal{C}^{k,\alpha}(\mathcal{B})$ for denoting the space of continuous functions in $\mathcal{B}$ whose kth-order derivatives are locally Hölder continuous in $\mathcal{B}$ with exponent α.

Theorem 1 (Fokker-Planck Equation, (Duan, 2015)(p. 105))**.**
Let us consider an Itō diffusion that is described by the following stochastic differential equation (SDE):

$$d\theta_t = b(\theta_t)dt + \sigma(\theta_t)dW_t \tag{2}$$

where $\theta \in \mathbb{R}^D$, $b(\cdot)$ is a D dimensional vector function, $\sigma(\cdot)$ is a $D \times D$ dimensional matrix function, and W_t is a D dimensional Brownian motion. Assume that θ_t has the conditional probability density $p(\theta, t)$. Then, the time evolution of $p(\theta, t)$ is described by the following partial differential equation:

$$\partial_t p(\theta, t) = A^* p(\theta, t) \tag{3}$$

where A^* is the adjoint of the generator of the SDE given in Equation 2 and known as the Fokker-Planck operator, defined as follows:

$$A^* h = -\nabla \cdot (bh) + \frac{1}{2} tr[\nabla^2(\sigma\sigma^\top h)]. \tag{4}$$

Here $x \cdot y$ denotes $x^\top y$ and ∇^2 denotes the Hessian matrix.

Theorem 2 (Existence and Uniqueness for Fokker-Planck Equation, (Duan, 2015) Theorem 5.8)**.**
Consider the SDE in Equation 2 and the following system:

$$\partial_t p(\theta, t) = A^* p(\theta, t), \qquad p(\theta, 0) = p_0(\theta) \tag{5}$$

where A^* is the Fokker-Planck operator defined in Equation 4, $\mathcal{B} \subset \mathbb{R}^D$. Assume that A^* is uniformly elliptic on $\mathcal{B}$, i.e. there is a positive constant κ such that

$$\sum_{i,j=1}^{D} [\sigma(\theta)\sigma^\top(\theta)]_{ij}\xi_i\xi_j \geq \kappa|\xi|^2 \tag{6}$$

for $\theta \in \mathcal{B}$ and all $\xi \in \mathbb{R}^D$. If $b(\cdot)$ and $\sigma(\cdot)$, the first-order derivatives of $b(\cdot)$, the second-order derivatives of $\sigma(\cdot)$, and $p_0(\cdot)$ are all uniformly Hölder continuous with exponent α in $\mathcal{B}$ and are all bounded in $\mathcal{C}^{\alpha}(\mathcal{B})$, then the unique solution $p(\cdot)$ to Equation 5 exists and is in $\mathcal{C}^{2,\alpha}(\mathcal{B})$.

Theorem 3 (Stationary Distribution of SDEs, (Ma et al., 2015) Theorem 1)**.**
Assume $\theta \in \mathbb{R}^D$ is a random variable and $x \equiv \{x_n\}_{n=1}^N$ denotes observed data points where $x_n \in \mathbb{R}^P$. We have $p(\theta|x) \propto \exp\big(-U(\theta)\big)$, where $U(\theta) = -[\log p(x|\theta) + \log p(\theta)]$. Let us consider the SDE given in Equation 2 with the following form:

$$b(\theta) = -[H(\theta) + Q(\theta)]\nabla U(\theta) + \Gamma(\theta), \qquad \sigma(\theta) = \sqrt{2H(\theta)} \tag{7}$$

where $\Gamma_i(\theta) = \sum_{j=1}^{D} \partial_{\theta_j}[H_{ij}(\theta) + Q_{ij}(\theta)]$, $H(\theta)$ is a positive semi-definite matrix and $Q(\theta)$ is a skew-symmetric matrix. Then $p(\theta|x)$ is a stationary distribution of the dynamics given in Equations 2 and 7. If $H(\theta)$ is positive definite, then the stationary distribution is unique.

Remark 1. *Positive definiteness of $H(\theta)$ implies uniform ellipticity of the Fokker-Planck operator A^*. However, we also need Hölder continuity for uniqueness as described in Theorem 2.*

Theorem 4 (Bias and MSE of SG-MCMC, (Chen et al., 2015) Theorem 5)**.**
Consider the SDE given in Equation 2 with the following form:

$$b(\theta_t) = -H_t\nabla U(\theta_t), \qquad \sigma(\theta_t) = \sqrt{2H_t} \tag{8}$$

where H_t is a positive definite matrix that does not depend on θ_t. Let $\{\theta_t\}_{t=1}^T$ be a sequence that is obtained by discretizing this SDE via the Euler-Maruyama integrator:

$$\theta_t = \theta_{t-1} - \epsilon_t H_t \nabla\tilde{U}(\theta_{t-1}) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, 2\epsilon_t H_t) \tag{9}$$

where the full gradients are replaced with stochastic gradients, that are defined as follows:

$$\nabla\tilde{U}(\theta_t) = -[\nabla \log p(\theta_t) + \frac{N}{N_\Omega} \sum_{n\in\Omega} \nabla \log p(x_n|\theta_t)]. \tag{10}$$

Here, $\Omega \subset \{1, 2, \dots, N\}$, and $N_\Omega = |\Omega| \geq 1$. Also consider an ergodic Itō diffusion with invariant measure π and a smooth test function $h(\theta)$. The posterior expectation is defined as

$$\bar{h} = \int h(\theta)\pi(\theta) d\theta. \tag{11}$$

Assume that $\{\theta_t\}_t$ is generated by using Equation 9, where the step-sizes satisfy the following conditions: (i) $\{\epsilon_t\}_t$ is decreasing, (ii) $\sum_{t=1}^\infty \epsilon_t = \infty$, and (iii) $\lim_{T\to\infty} \frac{\sum_{t=1}^T \epsilon_T^2}{\sum_{t=1}^T \epsilon_t} = 0$. Then, the sample average is defined as:

$$\hat{h} = \frac{1}{W_T} \sum_{t=1}^T \epsilon_t h(\theta_t), \tag{12}$$

where $W_T = \sum_{t=1}^T \epsilon_t$. Let us define the function ψ that solves the following Poisson equation:

$$\mathcal{L}\psi(\theta_t) = h(\theta_t) - \bar{h}, \tag{13}$$

where $\mathcal{L}$ is the generator of the SDE given in Equations 2 and 8. Assume that ψ and its up to third-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, $\sup_t \mathbb{E}\mathcal{V}^p(\theta_t) < \infty$ and $\sup_{s\in(0,1)} \mathcal{V}^p\big(sx + (1-s)y\big) \leq C\big(\mathcal{V}^p(x) + \mathcal{V}^p(y)\big), \forall x, y, p \leq 2\max_k p_k$ for some $C > 0$. Then, the bias and the mean squared-error (MSE) of an SG-MCMC algorithm can be bounded as follows:

$$\Big|\mathbb{E}[\hat{h}] - \bar{h}\Big| = \mathcal{O}\Big(\frac{1}{W_T} + \frac{\sum_{t=1}^T \epsilon_t^2}{W_T}\Big) \qquad \textit{(Bias)} \tag{14}$$

$$\mathbb{E}[(\hat{h} - \bar{h})^2] = \mathcal{O}\Big(\sum_{t=1}^T \frac{\epsilon_t^2}{W_T^2}\|\Delta V_t\|^2 + \frac{1}{W_T} + \frac{(\sum_{t=1}^T \epsilon_t^2)^2}{W_T^2}\Big) \qquad \textit{(MSE)} \tag{15}$$

where ΔV_t is defined as $\Delta V_t \triangleq (\nabla\tilde{U}(\theta_t) - \nabla U(\theta_t))^\top H_t \nabla$, and $\|\Delta V_t\|$ denotes the operator norm.

Remark 2. *The original version of Theorem 4 is more general in the sense that it also covers Hamiltonian dynamics and other integrators besides the Euler-Maruyama scheme. Moreover, in the original theorem, H_t is taken as the identity matrix. However, it is straightforward to extend this theorem to positive definite H_t, as exemplified in (Li et al., 2016).*

2.2. The Main Result

In this section we provide a formal proof for our theorem. Before proceeding to the theorem, let us first state our assumptions.

Condition 1.
The step-sizes satisfy the following properties: $\sum_{t=1}^\infty \epsilon_t = \infty$, $\sum_{t=1}^\infty \epsilon_t^2 < \infty$ and $\epsilon_t = \epsilon_{t'}$ for t and t' such that $\lfloor \frac{t-1}{M} \rfloor = \lfloor \frac{t'-1}{M} \rfloor$.

Condition 2.
The trust region parameter is chosen such that $\lambda > \max\{0, -1/\lambda_t^{min}\}$ for all t, where λ_t^{min} is the smallest eigenvalue of the L-BFGS approximation to $(\nabla^2 U(\theta_t))^{-1}$.

Condition 3.
$\nabla U(\theta)$ is Lipschitz continuous, so that $|\nabla U(\theta) - \nabla U(\theta')| \leq L|\theta - \theta'|, \quad \forall \theta, \theta' \in \mathbb{R}^D$.

Condition 4.
Let $\mathcal{L}$ be an operator defined as follows: $\mathcal{L}f(\theta_t) \triangleq -[H_t \nabla U(\theta_t)]^\top \nabla f(\theta_t) + \mathrm{tr}[H_t^\top \nabla^2 f(\theta_t)]$. Consider the functional ψ that solves the following Poisson equation: $\mathcal{L}\psi(\theta_t) = h(\theta_t) - \bar{h}$. Assume that ψ and its up to third-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 $C_k, p_k > 0$ and $k = 0, 1, 2, 3$ where $\mathcal{D}^k$ denotes the derivative of order k. Also assume $\sup_t \mathbb{E}\mathcal{V}^p(\theta_t) < \infty$ and $\sup_{s \in (0,1)} \mathcal{V}^p\big(sx + (1-s)y\big) \leq C\big(\mathcal{V}^p(x) + \mathcal{V}^p(y)\big), \forall x, y, p \leq 2\max_k p_k$ for some $C > 0$.

Theorem 5 (Convergence of HAMCMC)**.**
Assume the number of iterations is chosen as $T = KM$ where $K \in \mathbb{N}_+$, M is the memory size, and $\{\theta_t\}_{t=1}^T$ are obtained by HAMCMC that has the following update equation:

$$\theta_t = \theta_{t-M} - \epsilon_t H_t(\theta_{t-2M+1:t-1}^{\neg(t-M)}) \nabla \tilde{U}(\theta_{t-M}) + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, 2\epsilon_t H_t(\theta_{t-2M+1:t-1}^{\neg(t-M)})), \tag{16}$$

where $\theta_{t-2M+1:t-1}^{\neg(t-M)} \equiv \{\theta_{t-2M+1}, \ldots, \theta_{t-M-1}, \theta_{t-M+1}, \ldots, \theta_{t-1}\}$ and $H_t(\cdot)$ is computed via stochastic L-BFGS. Then, under Conditions 1– 4, the following holds:

(a) $\big|\mathbb{E}[\hat{h}] - \bar{h}\big| = \mathcal{O}\big(\frac{1}{L_K} + \frac{Y_K}{L_K}\big)$

(b) $\mathbb{E}\big[(\hat{h} - \bar{h})^2\big] = \mathcal{O}\Big(\sum_{k=1}^{K} \frac{\epsilon_{kM}^2}{L_K^2} \mathbb{E}\|\Delta V_{k^*}\|^2 + \frac{1}{L_K} + \frac{Y_K^2}{L_K^2}\Big)$

where $\bar{h}$, and $\hat{h}$ are defined in Equations 11, and 12, respectively, for a smooth function $h(\cdot)$. Furthermore, we define $L_K \triangleq \sum_{k=1}^K \epsilon_{kM}$, $Y_K \triangleq \sum_{k=1}^K \epsilon_{kM}^2$, and the operator $\Delta V_{k^\star} = \Delta V_{m_k^* + kM}$, where $m_k^* = \arg\max_{1 \leq m \leq M} \mathbb{E}\|\Delta V_{m+kM}\|^2$, $\Delta V_t \triangleq (\nabla \tilde{U}(\theta_t) - \nabla U(\theta_t))^\top H_t \nabla$, and $\|\Delta V_t\|$ denotes the operator norm.

Proof. The usual way of analyzing Langevin and Hamiltonian MC algorithms is to first analyze the underlying continuous dynamics, then consider the algorithm as a discrete-time approximation. This approach is not directly applicable in our case. Therefore, we follow a different approach by exploiting the conditional independence structure of HAMCMC given in Figure 1.

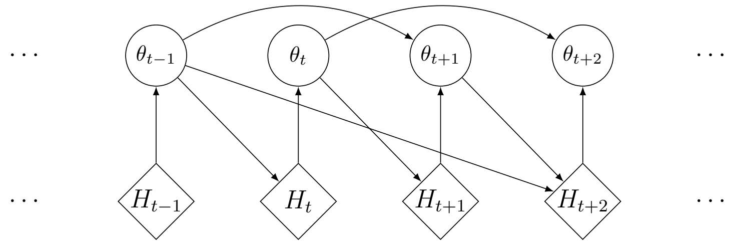

Figure 1. Illustration of HAMCMC with $M = 2$. The circle nodes represent the generated samples and the diamond nodes represent a deterministic function of the particular samples. The arrows illustrates the conditional independence structure.

An important property that is revealed by Figure 1 is that H_t is conditionally independent of θ_{t-M}, given $\theta_{t-2M+1:t-1}^{\neg(t-M)}$. This property will be useful in the analysis. By using this structure, we can convert Equation 16 to a first-order Markov chain on a product space, such that

$$\Theta_t \equiv \{\theta_{t-2M+2}, \ldots, \theta_t\} \tag{17}$$

where $\Theta_t \in \mathbb{R}^{D(2M-1)}$. With this construction, we can see that Equation 16 is a Markov chain that uses one of M different transition kernels at each iteration, where the kernels have the following structure:

$$\mathcal{T}_m(\Theta_t, \Theta') = \Big(\prod_{m'=1}^{M-1} \delta(\tilde{\Theta}_{t,m'}, \Theta'_{m'})\Big) \mathcal{K}(\Theta_{t,m+M-1}, \Theta'_{t,m+M-1} | \Theta_{t,-(m+M-1)}) \prod_{\substack{m' \in \{M \ldots, 2M-1\} \\ m' \neq m}} \delta(\Theta_{t,m'}, \Theta'_{m'}) \tag{18}$$

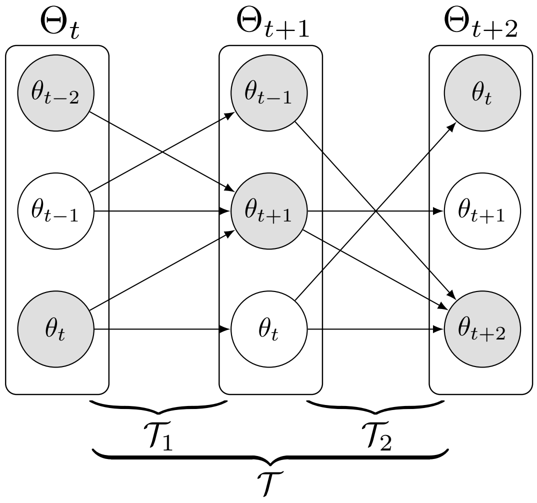

Figure 2. First-order Markov chain representation of HAMCMC. The memory variable is selected as $M = 2$. The shaded nodes represent the variables that are used in L-BFGS computations.

for $m = 1, \dots, M$. Here, $\Theta_{t,m}$ denotes the mth element of Θ_t, $\Theta_{t,-m}$ denotes $\Theta_t \setminus \Theta_{t,m}$, and $\tilde{\Theta}_{t,m}$ denotes the $(m+1)$th oldest element of Θ_t (i.e. $\tilde{\Theta}_{t,m} \equiv \theta_{t-2M+2+m}$). At each iteration, the kernel $\mathcal{T}_m$ modifies $\Theta_{t,m+M-1}$ by using a base kernel $\mathcal{K}$ that is conditioned on $\Theta_{t,-(m+M-1)}$, and rearranges the samples that will be used in the L-BFGS computations. By using the intermediate kernels $\mathcal{T}_m$, we can represent the kernel of the whole HAMCMC algorithm as the composition of these kernels, given as follows:

$$\mathcal{T}(\Theta_t, \Theta') = [\mathcal{T}_1 \circ \cdots \circ \mathcal{T}_M](\Theta_t, \Theta'). \tag{19}$$

Therefore, if $\mathcal{K}$ targets the correct distribution, so does $\mathcal{T}$. We illustrate this construction in Figure 2.

Now, let us investigate the base kernel $\mathcal{K}$, that has the following form:

$$\mathcal{K}(\Theta_{t,m}, \Theta'_{t,m} | \Theta_{t,-m}) = \mathcal{N}(\Theta'_{t,m}; \mu, \Sigma), \tag{20}$$

where $\mu = \epsilon H(\Theta_{t,-m}) \nabla \tilde{U}(\Theta_{t,m})$, and $\Sigma = 2\epsilon H(\Theta_{t,-m})$. The key property of $\mathcal{K}$ is that, since the L-BFGS computations only involve $\Theta_{t,-m}$, $H(\cdot)$ is independent of $\Theta_{t,m}$. Therefore, perhaps not surprisingly, $\mathcal{K}$ appears to be the transition kernel of the SGLD algorithm with a preconditioning matrix $H(\Theta_{t,-m})$. Henceforth, we can analyze $\mathcal{K}$ as an approximation to the continuous-time diffusion, with the following form:

$$d\theta_t = -H_t \nabla U(\theta_t) dt + \sqrt{2H} dW_t. \tag{21}$$

Firstly, we need to show that there exists a unique stationary distribution for this SDE and this distribution is the posterior distribution that we are interested in. By Condition 2, we know that the lowest eigenvalue of H_t is strictly greater than 0, hence, H_t is positive definite. Therefore, the Fokker-Planck operator of this SDE is uniformly elliptic. Then, by Condition 3, we know that ∇U is Hölder continuous with exponent $\alpha = 1$. Under these conditions, Theorem 2 states that the unique stationary distribution for this SDE exists.

Next, we investigate the stationary distribution of this SDE. Since H_t is independent of θ_t, we have $\frac{\partial}{\partial \theta_i} H_t = 0$ for $i = 1, \dots, D$, therefore the correction term $\Gamma(\theta) = 0$ (see Equation 7). After choosing $Q(\theta) = 0$, we observe that Equation 21 is in the same form as Equation 7. Therefore, by Theorem 3, the unique stationary distribution of Equation 21 is the Bayesian posterior.

The rest of the proof is built on the fact that HAMCMC can be decomposed into M different but related SGLD algorithms whose (state independent) preconditioning matrices are computed via stochastic L-BFGS. More formally, we can group the samples $\{\theta_t\}_{t=1}^T$ into M different series as $\{\theta_{m+kM}\}_{k=0}^{K-1}$ for $m = 1, \dots, M$, where each $\{\theta_{m+kM}\}_{k=0}^{K-1}$ can be considered as being obtained from a different SGLD algorithm.

Part (a):

By Condition 1, we can rewrite the sample average as the average of M different sample averages that are obtained via different SGLD algorithms, given as follows:

$$\hat{h} = \frac{1}{W_T} \sum_{t=1}^{T} \epsilon_t h(\theta_t) = \frac{1}{M} \sum_{m=1}^{M} \hat{h}_m \tag{22}$$

where

$$\hat{h}_m \triangleq \frac{1}{L_K} \sum_{k=0}^{K-1} \epsilon_{m+kM} h(\theta_{m+kM}). \tag{23}$$

Then, we can bound the bias as follows:

$$\begin{aligned}
\left|\mathbb{E}[\hat{h}] - \bar{h}\right| &= \Big|\big(\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[\hat{h}_m]\big) - \bar{h}\Big| &(24)\\
&= \frac{1}{M} \Big| \sum_{m=1}^{M} \Big(\mathbb{E}[\hat{h}_m] - \bar{h}\Big)\Big| &(25)\\
&\leq \frac{1}{M} \sum_{m=1}^{M} \Big|\mathbb{E}[\hat{h}_m] - \bar{h}\Big| &(26)\\
&\leq \frac{1}{M} \sum_{m=1}^{M} C_m \Big(\frac{1}{L_K} + \frac{Y_K}{L_K}\Big) &(27)
\end{aligned}$$

for some $C_m > 0$ for all m. Equation 27 uses Conditions 1, 4, and the first part of Theorem 4. Define $C^* \triangleq \max_m C_m$, then we obtain the desired bound as:

$$\left|\mathbb{E}[\hat{h}] - \bar{h}\right| \leq C^* \Big(\frac{1}{L_K} + \frac{Y_K}{L_K}\Big) \implies \left|\mathbb{E}[\hat{h}] - \bar{h}\right| = \mathcal{O}\Big(\frac{1}{L_K} + \frac{Y_K}{L_K}\Big). \tag{28}$$

This completes part (a).

Part (b):

For bounding the MSE, we follow a similar strategy to the one presented for part (b). We start by bounding $(\hat{h} - \bar{h})^2$, as follows:

$$\begin{aligned}
(\hat{h} - \bar{h})^2 &= \Big(\big(\frac{1}{M} \sum_{m=1}^{M} \hat{h}_m\big) - \bar{h}\Big)^2 &(29)\\
&= \frac{1}{M^2} \Big(\sum_{m=1}^{M} (\hat{h}_m - \bar{h})\Big)^2 &(30)\\
&\leq \frac{1}{M} \sum_{m=1}^{M} (\hat{h}_m - \bar{h})^2 &(31)
\end{aligned}$$

Taking the expectation of both sides results in:

$$\begin{aligned}
\mathbb{E}\big[(\hat{h} - \bar{h})^2\big] &\leq \frac{1}{M} \mathbb{E}\big[\sum_{m=1}^{M} (\hat{h}_m - \bar{h})^2\big] &(32)\\
&= \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\big[(\hat{h}_m - \bar{h})^2\big] &(33)
\end{aligned}$$

Then, by using Conditions 1, 4, and the second part of Theorem 4 we obtain

$$\mathbb{E}\big[(\hat{h} - \bar{h})^2\big] \le \frac{1}{M}\sum_{m=1}^{M} C_m \Big(\sum_{k=1}^{K} \frac{\epsilon_{kM}^2}{L_K^2}\mathbb{E}\|\Delta V_{m+kM}\|^2 + \frac{1}{L_K} + \frac{Y_K^2}{L_K^2}\Big) \tag{34}$$

for some $C_m > 0, \forall m$. Define $C^* = \max_m C_m$, $m_k^* = \arg\max_m \mathbb{E}\|\Delta V_{m+kM}\|^2$, and $\Delta V_{k^\star} = \Delta V_{m_k^* + kM}$, and we obtain the desired bound as:

$$\mathbb{E}\big[(\hat{h} - \bar{h})^2\big] \le C^* \Big(\sum_{k=1}^{K} \frac{\epsilon_{kM}^2}{L_K^2}\mathbb{E}\|\Delta V_{k^\star}\|^2 + \frac{1}{L_K} + \frac{Y_K^2}{L_K^2}\Big) \implies \mathbb{E}\big[(\hat{h} - \bar{h})^2\big] = \mathcal{O}\Big(\sum_{k=1}^{K} \frac{\epsilon_{kM}^2}{L_K^2}\mathbb{E}\|\Delta V_{k^\star}\|^2 + \frac{1}{L_K} + \frac{Y_K^2}{L_K^2}\Big)$$

This completes part (b) and concludes the proof.

□

Remark 3. *Condition 4 is a special case of the assumption given in Theorem 4. Here, we have customized this condition within HAMCMC framework.*

Remark 4. *Note that a similar construction that uses Markov chains on product spaces has been presented in (Zhang & Sutton, 2011), where the authors consider a different context in which they aim to obtain separable Hamiltonians, that would avoid costly numerical integration steps. Also note that PSGLD cannot be redefined by using Markov chains on product spaces, since the volatility in PSGLD depends on the full history of the samples.*

3. Algorithm Parameters Used in the Experiments

Linear Gaussian Model:

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

	SGLD	PSGLD			SGRLD	HAMCMC		
	a_ϵ	a_ϵ	α	λ	a_ϵ	a_ϵ	γ	λ
$D = 2$	1×10^{-5}	5×10^{-2}	0.9	1×10^{-3}	5×10^{-2}	1×10^{-6}	50	1
$D = 10$	1×10^{-6}	1×10^{-1}	0.9	1×10^{-3}	5×10^{-1}	1×10^{-8}	50	1
$D = 100$	8×10^{-6}	1×10^{-1}	0.9	1×10^{-3}	5×10^{-1}	1×10^{-8}	50	1

Alpha-Stable Matrix Factorization:

Table 2. The list of algorithm parameters that are used in the experiments on αMF.

SGLD	PSGLD			HAMCMC		
a_ϵ	a_ϵ	α	λ	a_ϵ	γ	λ
1×10^{-6}	2×10^{-5}	0.9	1×10^{-3}	1×10^{-8}	0.01	1000

Distributed Matrix Factorization:

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

SGLD	PSGLD			HAMCMC		
ϵ	ϵ	α	λ	ϵ	γ	λ
1×10^{-4}	2×10^{-4}	0.99	2	5×10^{-4}	0.1	0.5

References

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.

Duan, J. *An Introduction to Stochastic Dynamics*. Cambridge University Press, New York, 2015.

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

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.

Nocedal, J. and Wright, S. J. *Numerical optimization*. Springer, Berlin, 2006.

Zhang, Y. and Sutton, C. A. Quasi-Newton methods for Markov Chain Monte Carlo. In *Advances in Neural Information Processing Systems*, pp. 2393–2401, 2011.