Bridging the Gap between Stochastic Gradient MCMC and Stochastic Optimization: Supplementary Material

Changyou Chen[†] **David Carlson**[‡] **Zhe Gan**[†] **Chunyuan Li**[†] **Lawrence Carin**[†]
[†]Department of Electrical and Computer Engineering, Duke University
[‡]Department of Statistics and Grossman Center for Statistics of Mind, Columbia University

A Solutions for the sub-SDEs

We provide analytic solutions for the split sub-SDEs in Section 4.1. For stepsize h, the solutions are given in (6).

$$A:\begin{cases}\boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1}+\mathbf{G}_1(\boldsymbol{\theta})\,\mathbf{p}\,h\\ \mathbf{p}_t &= \mathbf{p}_{t-1}\\ \boldsymbol{\Xi}_t &= \boldsymbol{\Xi}_{t-1}+\left(\mathbf{Q}-\frac{1}{\beta}I\right)h\end{cases},\qquad (6)$$

$$B:\begin{cases}\boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1}\\ \mathbf{p}_t &= \exp\left(-\boldsymbol{\Xi}h\right)\mathbf{p}_{t-1}\\ \boldsymbol{\Xi}_t &= \boldsymbol{\Xi}_{t-1}\end{cases},$$

$$O:\begin{cases}\boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1}\\ \mathbf{p}_t &= \mathbf{p}_{t-1}+\Big(-G_1(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta})+\frac{1}{\beta}\nabla_{\boldsymbol{\theta}}G_1(\boldsymbol{\theta})\\ &\qquad +G_1(\boldsymbol{\theta})(\boldsymbol{\Xi}-G_2(\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}G_2(\boldsymbol{\theta})\Big)h\\ &\qquad +(\tfrac{2}{\beta}G_2(\boldsymbol{\theta}))^{\frac{1}{2}}\odot\zeta_t\\ \boldsymbol{\Xi}_t &= \boldsymbol{\Xi}_{t-1}\end{cases}$$

B Proof of Lemma 1

For a general stochastic differential equation of the form

$$\mathrm{d}\,\mathbf{x}=F(\mathbf{x})\mathrm{d}t+\sqrt{2}D^{1/2}(\mathbf{x})\mathrm{d}\,\mathbf{w}\ ,\qquad (7)$$

where $\mathbf{x}\in\mathbf{R}^N$, $F:\mathbf{R}^N\to\mathbf{R}^N$, $D:\mathbf{R}^M\to\mathbf{R}^{N\times P}$ are measurable functions with P, and $\mathbf{w}$ is standard P-dimensional Brownian motion. (1) is a special case of the general form (7) with

$$\mathbf{x}=(\boldsymbol{\theta},\mathbf{p},\boldsymbol{\Xi})\qquad (8)$$

$$F(\mathbf{x})=\begin{pmatrix}\mathbf{G}_1(\boldsymbol{\theta})\,\mathbf{p}\\ -\mathbf{G}_1(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta})-\boldsymbol{\Xi}\,\mathbf{p}+\frac{1}{\beta}\nabla_{\boldsymbol{\theta}}\,\mathbf{G}_1(\boldsymbol{\theta})\\ \quad+\mathbf{G}_1(\boldsymbol{\theta})(\boldsymbol{\Xi}-\mathbf{G}_2(\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}\,\mathbf{G}_2(\boldsymbol{\theta})\\ \mathbf{Q}-\frac{1}{\beta}\mathbf{I}\end{pmatrix}$$

$$D(\mathbf{x})=\begin{pmatrix}\mathbf{0}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\frac{1}{\beta}\,\mathbf{G}_2(\boldsymbol{\theta})&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}\end{pmatrix}$$

We write the joint distribution of $\mathbf{x}$ as

$$\rho(\mathbf{x})=\frac{1}{Z}\exp\left\{-H(\mathbf{x})\right\}\triangleq\frac{1}{Z}\exp\left\{-U(\boldsymbol{\theta})-E(\boldsymbol{\theta},\mathbf{p},\boldsymbol{\Xi})\right\}\ .$$

A reformulation of the main theorem in Ding et al. (2014) gives the following lemma, which is used to prove Lemma 1 in the main text.

Lemma 4. *The stochastic process of $\vec{\theta}$ generated by the stochastic differential equation (7) has the target distribution $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})=\frac{1}{Z}\exp\{-U(\boldsymbol{\theta})\}$ as its stationary distribution, if $\rho(\mathbf{x})$ satisfies the following marginalization condition:*

$$\exp\{-U(\boldsymbol{\theta})\}\propto\int\exp\{-U(\boldsymbol{\theta})-E(\boldsymbol{\theta},\mathbf{p},\boldsymbol{\Xi})\}\mathrm{d}\,\mathbf{p}\,\mathrm{d}\boldsymbol{\Xi}\,,\qquad (9)$$

and if the following condition is also satisfied:

$$\nabla\cdot(\rho F)=\nabla\nabla^\top:(\rho D)\,,\qquad (10)$$

where $\nabla\triangleq(\partial/\partial\boldsymbol{\theta},\partial/\partial\,\mathbf{p},\partial/\boldsymbol{\Xi})$, "$\cdot$" represents the vector inner product operator, ":" represents a matrix double dot product, i.e., $\mathbf{X}:\mathbf{Y}\triangleq tr(\mathbf{X}^\top\mathbf{Y})$.

Proof of Lemma 1. We first have reformulated (1) using the general SDE form of (7), resulting in (8). Lemma 1 states the joint distribution of $(\boldsymbol{\theta},\mathbf{p},\boldsymbol{\Xi})$ is

$$\rho(\mathbf{x})=\frac{1}{Z}\exp\left(-\frac{1}{2}\,\mathbf{p}^\top\,\mathbf{p}-U(\boldsymbol{\theta})\right.$$
$$\left.-\frac{1}{2}\mathrm{tr}\left\{\left(\boldsymbol{\Xi}-\mathbf{G}_2(\boldsymbol{\theta})\right)^\top\left(\boldsymbol{\Xi}-\mathbf{G}_2(\boldsymbol{\theta})\right)\right\}\right)\ ,\qquad (11)$$

with $H(\mathbf{x})=\frac{1}{2}\,\mathbf{p}^\top\,\mathbf{p}+U(\boldsymbol{\theta})+\frac{1}{2}\mathrm{tr}\left\{\left(\boldsymbol{\Xi}-\mathbf{G}_2(\boldsymbol{\theta})\right)^\top\left(\boldsymbol{\Xi}-\mathbf{G}_2(\boldsymbol{\theta})\right)\right\}$. The marginalization condition (9) is trivially satisfied, we are left to verify condition (10). Substituting $\rho(\mathbf{x})$ and F into (10), we have the left-hand side

$$
\begin{aligned}
&\text{LHS} = \sum_i \frac{\partial}{\partial \mathbf{x}_i}\left(\rho F_i\right) \\
=&\sum_i \frac{\partial \rho}{\partial \mathbf{x}_i} F_i + \frac{\partial F_i}{\partial \mathbf{x}_i}\rho \\
=&\sum_i \left(\frac{\partial F_i}{\partial \mathbf{x}_i} - \frac{\partial H}{\partial \mathbf{x}_i} F_i\right)\rho \\
=&\Bigg(\sum_i \nabla_{\boldsymbol{\theta}_i}(\mathbf{G}_1)_{i:}\,\mathbf{p} - \sum_i \text{diag}(\boldsymbol{\Xi}) \\
&- \sum_i \beta\left(\nabla_{\boldsymbol{\theta}_i} U - \sum_j (\boldsymbol{\Xi}_{ij} - (\mathbf{G}_2)_{ij})\nabla_{\boldsymbol{\theta}_i}(\mathbf{G}_2)_{ij}\right)(\mathbf{G}_1\,\mathbf{p})_i \\
&- \beta p^T\left(-\mathbf{G}_1\,\nabla_{\boldsymbol{\theta}} U - \boldsymbol{\Xi}\,\mathbf{p} + \frac{1}{\beta}\nabla_{\boldsymbol{\theta}}\,\mathbf{G}_1 + \mathbf{G}_1(\boldsymbol{\Xi} - \mathbf{G}_2)\nabla_{\boldsymbol{\theta}}\,\mathbf{G}_2\right) \\
&-\beta\sum_i (\boldsymbol{\Xi}_{ii} - (\mathbf{G}_2)_{ii})\left(\mathbf{Q}_{ii} - \frac{1}{\beta}\right)\Bigg)\rho \\
&= \frac{1}{\beta}\text{tr}\left\{\mathbf{G}_2(\mathbf{p}\,\mathbf{p}^T - \mathbf{I})\right\}\rho\,.
\end{aligned}
$$

It is easy to see for the right-hand side

$$
\begin{aligned}
\text{RHS} &= \sum_i\sum_j \frac{1}{\beta}(\mathbf{G}_2)_{ij}\frac{\partial^2}{\partial\,\mathbf{x}_i\,\partial\,\mathbf{x}_j}\rho \\
&= \frac{1}{\beta}\sum_i\sum_j (\mathbf{G}_2)_{ij}\frac{\partial}{\partial\,\mathbf{p}_j}\left(-\frac{\partial H}{\partial\,\mathbf{p}_i}\rho\right) \\
&= \frac{1}{\beta}\sum_i (\mathbf{G}_2)_{ii}\left(\mathbf{p}_i^2 - 1\right)\rho \\
&\equiv \text{LHS}\,.
\end{aligned}
$$

According to Lemma 4, the joint distribution (11) is the equilibrium distribution of (1). □

C Proof of Theorem 2

We start by proving the bias result of Theorem 2.

Proof of the bias. For our 2nd-order integrator, according to the definition, we have:

$$
\begin{aligned}
&\mathbb{E}[\psi(\mathbf{X}_t)] = \tilde{P}_h^l\psi(\mathbf{X}_{t-1}) = e^{h\tilde{\mathcal{L}}_t}\psi(\mathbf{X}_{t-1}) + O(h^3) \\
&= \left(\mathbb{I} + h\tilde{\mathcal{L}}_t\right)\psi(\mathbf{X}_{t-1}) + \frac{h^2}{2}\tilde{\mathcal{L}}_t^2\psi(\mathbf{X}_{t-1}) + O(h^3)\,,
\end{aligned} \tag{12}
$$

where $\mathcal{L}_t$ is the generator of the SDE for the t-th iteration, *i.e.*, using stochastic gradient instead of the full gradient, $\mathbb{I}$ is the identity map. Compared to the prove of Chen et al. (2015), we need to consider the approximation error for $\nabla_{\boldsymbol{\theta}} G_1(\boldsymbol{\theta})$. As a result, (12) needs to be rewritten as:

$$
\begin{aligned}
&\mathbb{E}[\psi(\mathbf{X}_t)] \qquad\qquad (13)\\
&= \left(\mathbb{I} + h(\tilde{\mathcal{L}}_t + \mathcal{B}_t)\right)\psi(\mathbf{X}_{t-1}) + \frac{h^2}{2}\tilde{\mathcal{L}}_t^2\psi(\mathbf{X}_{t-1}) + O(h^3)\,,
\end{aligned}
$$

where $\mathcal{B}_t$ is from (3). Sum over $t = 1, \cdots, L$ in (13), take expectation on both sides, and use the relation $\tilde{\mathcal{L}}_t + \mathcal{B}_t = \mathcal{L}_{\beta_t} + \Delta V_t$ to expand the first order term. We obtain

$$
\begin{aligned}
&\sum_{t=1}^{L}\mathbb{E}[\psi(\mathbf{X}_t)] = \psi(\mathbf{X}_0) + \sum_{t=1}^{L-1}\mathbb{E}[\psi(\mathbf{X}_t)] \\
&+ h\sum_{t=1}^{L}\mathbb{E}[\mathcal{L}_{\beta_t}\psi(\mathbf{X}_{t-1})] + h\sum_{t=1}^{L}\mathbb{E}[\Delta V_t\psi(\mathbf{X}_{t-1})] \\
&+ \frac{h^2}{2}\sum_{t=1}^{L}\mathbb{E}[\tilde{\mathcal{L}}_t^2\psi(\mathbf{X}_{t-1})] + O(Lh^3).
\end{aligned}
$$

We divide both sides by Lh, use the Poisson equation (4), and reorganize terms. We have:

$$
\begin{aligned}
&\mathbb{E}[\frac{1}{L}\sum_t\left(\phi(\mathbf{X}_t) - \bar{\phi}_{\beta_t}\right)] = \frac{1}{L}\sum_{t=1}^{L}\mathbb{E}[\mathcal{L}_{\beta_t}\psi(\mathbf{X}_{t-1})] \\
=&\frac{1}{Lh}\left(\mathbb{E}[\psi(\mathbf{X}_t)] - \psi(\mathbf{X}_0)\right) - \frac{1}{L}\sum_t\mathbb{E}[\Delta V_t\psi(\mathbf{X}_{t-1})] \\
&- \frac{h}{2L}\sum_{t=1}^{L}\mathbb{E}[\tilde{\mathcal{L}}_t^2\psi(\mathbf{X}_{t-1})] + O(h^2)
\end{aligned} \tag{14}
$$

Now we try to bound $\tilde{\mathcal{L}}_t^2$. Based on ideas from Mattingly et al. (2010), we apply the following procedure. First replace ψ with $\tilde{\mathcal{L}}_t\psi$ from (13) to (14), and apply the same logic for $\tilde{\mathcal{L}}_t\psi$ as for ψ in the above derivations, but this time expand in (13) up to the order of $O(h^2)$, instead of the previous order $O(h^3)$. After simplification, we obtain:

$$
\sum_t\mathbb{E}[\tilde{\mathcal{L}}_t^2\psi(\mathbf{X}_{t-1})] = O\left(\frac{1}{h} + Lh\right) \tag{15}
$$

Substituting (15) into (14), after simplification, we have: $\mathbb{E}\left(\frac{1}{L}\sum_t\left(\phi(\mathbf{X}_t) - \bar{\phi}_{\beta_t}\right)\right)$

$$
\begin{aligned}
=&\frac{1}{Lh}\underbrace{\left(\mathbb{E}[\psi(\mathbf{X}_t)] - \psi(\mathbf{X}_0)\right)}_{C_1} - \frac{1}{L}\sum_t\mathbb{E}[\Delta V_t\psi(\mathbf{X}_{t-1})] \\
&- O\left(\frac{h}{Lh} + h^2\right) + C_3h^2\,,
\end{aligned}
$$

for some $C_3 \geq 0$. According to the assumption, the term C_1 is bounded. As a result, collecting low order

terms, the bias can be expressed as:

$$
\begin{aligned}
&\left|\mathbb{E}\hat{\phi} - \bar{\phi}\right| \\
=&\left|\mathbb{E}\left(\frac{1}{L}\sum_t \left(\phi(\mathbf{X}_t) - \bar{\phi}_{\beta_t}\right)\right) + \frac{1}{L}\sum_t \bar{\phi}_{\beta_t} - \bar{\phi}\right| \\
\leq&\left|\mathbb{E}\left(\frac{1}{L}\sum_t \bar{\phi}_{\beta_t} - \bar{\phi}\right)\right| + \left|\mathbb{E}\left(\frac{1}{L}\sum_t \left(\phi(\mathbf{X}_t) - \bar{\phi}_{\beta_t}\right)\right)\right| \\
\leq& C\phi(\boldsymbol{\theta}^*)\left(\frac{1}{L}\sum_{t=1}^{L}\int_{\boldsymbol{\theta}\neq\boldsymbol{\theta}^*} e^{-\beta_t \hat{U}(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}\right) \\
&+\left|\frac{C_1}{Lh} - \frac{\sum_t \mathbb{E}\Delta V_t \psi(\mathbf{X}_{t-1})}{L} + C_3 h^2\right| \\
\leq& C\phi(\boldsymbol{\theta}^*)\left(\frac{1}{L}\sum_{t=1}^{L}\int_{\boldsymbol{\theta}\neq\boldsymbol{\theta}^*} e^{-\beta_t \hat{U}(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}\right) + \left|\frac{C_1}{Lh}\right| \\
&+\left|\frac{\sum_t \mathbb{E}\Delta V_t \psi(\mathbf{X}_{t-1})}{L}\right| + \left|C_3 h^2\right| \\
\leq& C\phi(\boldsymbol{\theta}^*)\left(\frac{1}{L}\sum_{t=1}^{L}\int_{\boldsymbol{\theta}\neq\boldsymbol{\theta}^*} e^{-\beta_t \hat{U}(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}\right) \\
&+ D\left(\frac{1}{Lh} + \frac{\sum_t \|\mathbb{E}\Delta V_t\|}{L} + h^2\right) ,
\end{aligned}
$$

where the last equation follows from the finiteness assumption of ψ, $\|\cdot\|$ denotes the operator norm and is bounded in the space of ψ due to the assumptions. This completes the proof. □

We will now prove the MSE result .

Proof of the MSE bound. Similar to the proof of Theorem 2, for our 2nd–order integrator we have:

$$
\begin{aligned}
\mathbb{E}\left(\psi_{\beta_t}(\mathbf{X}_t)\right) = &\left(\mathbb{I} + h(\mathcal{L}_{\beta_t} + \Delta V_t)\right)\psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) \\
&+ \frac{h^2}{2}\tilde{\mathcal{L}}_t^2 \psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) + O(h^3) .
\end{aligned}
$$

Sum over t from 1 to $L+1$ and simplify, we have:

$$
\begin{aligned}
\sum_{t=1}^{L}\mathbb{E}\left(\psi_{\beta_t}(\mathbf{X}_t)\right) = &\sum_{t=1}^{L}\psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) \\
&+ h\sum_{t=1}^{L}\mathcal{L}_{\beta_t}\psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) + h\sum_{t=1}^{L}\Delta V_t \psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) \\
&+ \frac{h^2}{2}\sum_{t=1}^{L}\tilde{\mathcal{L}}_t^2\psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) + O(Lh^3) .
\end{aligned}
$$

Substitute the Poisson equation (4) into the above equation, divide both sides by Lh and rearrange related terms, we have

$$
\begin{aligned}
\frac{1}{L}\sum_{t=1}^{L}\left(\phi(\mathbf{X}_t) - \bar{\phi}_{\beta_t}\right) = &\frac{1}{Lh}\left(\mathbb{E}\psi_{\beta_L}(\mathbf{X}_{Lh}) - \psi_{\beta_0}(\mathbf{X}_0)\right) \\
&- \frac{1}{Lh}\sum_{t=1}^{L}\left(\mathbb{E}\psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) - \psi_{\beta_{t-1}}(\mathbf{X}_{t-1})\right) \\
&- \frac{1}{L}\sum_{t=1}^{L}\Delta V_t \psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) - \frac{h}{2L}\sum_{t=1}^{L}\tilde{\mathcal{L}}_t^2\psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) + O(h^2)
\end{aligned}
$$

Taking the square of both sides, it is then easy to see there exists some positive constant C, such that

$$
\begin{aligned}
&\left(\frac{1}{L}\sum_{t=1}^{L}\left(\phi(\mathbf{X}_t) - \bar{\phi}_{\beta_t}\right)\right)^2 \qquad (16) \\
\leq& C\left(\underbrace{\frac{\left(\mathbb{E}\psi_{\beta_L}(\mathbf{X}_{Lh}) - \psi_{\beta_0}(\mathbf{X}_0)\right)^2}{L^2h^2}}_{A_1}\right. \\
&+\underbrace{\frac{1}{L^2h^2}\sum_{t=1}^{L}\left(\mathbb{E}\psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) - \psi_{\beta_{t-1}}(\mathbf{X}_{t-1})\right)^2}_{A_2} \\
&+\frac{1}{L^2}\sum_{t=1}^{L}\Delta V_t^2\psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) \\
&\left.+\underbrace{\frac{h^2}{2L^2}\left(\sum_{t=1}^{L}\tilde{\mathcal{L}}_t^2\psi_{\beta_{t-1}}(\mathbf{X}_{t-1})\right)^2}_{A_3} + h^4\right)
\end{aligned}
$$

A_1 is easily bounded by the assumption that $\|\psi\| \leq V^{r_0} < \infty$. A_2 is bounded because it can be shown that $\mathbb{E}\left(\psi_{\beta_t}(\mathbf{X}_t)\right) - \psi_{\beta_t}(\mathbf{X}_t) \leq C_1\sqrt{h} + O(h)$ for $C_1 \geq 0$. Intuitively this is true because the only difference between $\mathbb{E}\left(\psi_{\beta_t}(\mathbf{X}_t)\right)$ and $\psi_{\beta_t}(\mathbf{X}_t)$ lies in the additional Gaussian noise with variance h. A formal proof is given in Chen et al. (2015). Furthermore, A_3 is

bounded by the following arguments:

$$
\begin{aligned}
A_3 &= \underbrace{\frac{h^2}{2L^2}\left(\sum_{t=1}^{L}\mathbb{E}\left[\tilde{\mathcal{L}}_t^2\psi_{\beta_{t-1}}(\mathbf{X}_{t-1})\right]\right)^2}_{B_1} \\
&+ \underbrace{\frac{h^2}{2L^2}\mathbb{E}\left(\sum_{t=1}^{L}\left(\tilde{\mathcal{L}}_t^2\psi_{\beta_{t-1}}(\mathbf{X}_{t-1}) - \mathbb{E}\tilde{\mathcal{L}}_t^2\psi_{\beta_{t-1}}(\mathbf{X}_{t-1})\right)\right)^2}_{B_2} \\
&\lesssim B_1 + \left(\frac{h^2}{Lh}\sum_{t=1}^{L}\tilde{\mathcal{L}}_t^2\psi_{\beta_{t-1}}(\mathbf{X}_{t-1})\right)^2 \\
&+ \left(\frac{h^2}{Lh}\sum_{t=1}^{L}\left(\mathbb{E}\tilde{\mathcal{L}}_t^2\psi_{\beta_{t-1}}(\mathbf{X}_{t-1})\right)\right)^2 \\
&\leq O\left(\frac{1}{2L^2} + L^2h^2\right) + \frac{1}{Lh}\left(\frac{h^2}{L}\sum_{t=1}^{L}(\tilde{\mathcal{L}}_t^2\psi(\mathbf{X}_{t-1}))^2\right) \\
&+ O\left(\frac{1}{L^2h^2} + h^4\right) \\
&= O\left(\frac{1}{Lh} + L^4\right)
\end{aligned}
$$

Collecting low order terms we have:

$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{L}\sum_{t=1}^{L}\left(\phi(\mathbf{X}_t) - \bar{\phi}_{\beta_t}\right)\right)^2 \\
=&O\left(\frac{\frac{1}{L}\sum_t \mathbb{E}\left\|\Delta V_t\right\|^2}{L} + \frac{1}{Lh} + h^4\right) . \qquad (17)
\end{aligned}
$$

Finally, we have:

$$
\begin{aligned}
&\mathbb{E}\left(\hat{\phi} - \bar{\phi}\right)^2 < \mathbb{E}\left(\frac{1}{L}\sum_t\left(\phi(\mathbf{X}_t) - \bar{\phi}_{\beta_t}\right)\right)^2 \\
&+ \mathbb{E}\left(\frac{1}{L}\sum_{t=1}^{L}\left(\phi(\mathbf{X}_t) - \bar{\phi}_{\beta_t}\right)\right)^2 \\
\leq& C\phi(\boldsymbol{\theta}^*)^2\left(\frac{1}{L}\sum_{t=1}^{L}\int_{\boldsymbol{\theta}\neq\boldsymbol{\theta}^*} e^{-\beta_t\hat{U}(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}\right)^2 \\
&+ O\left(\frac{\frac{1}{L}\sum_t \mathbb{E}\left\|\Delta V_t\right\|^2}{L} + \frac{1}{Lh} + h^4\right) \\
\leq& C\phi(\boldsymbol{\theta}^*)^2\left(\frac{1}{L}\sum_{t=1}^{L}\int_{\boldsymbol{\theta}\neq\boldsymbol{\theta}^*} e^{-\beta_t\hat{U}(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}\right)^2 \\
&+ D\left(\frac{\frac{1}{L}\sum_t \mathbb{E}\left\|\Delta V_t\right\|^2}{L} + \frac{1}{Lh} + h^4\right) .
\end{aligned}
$$

□

D Proof of Corollary 3

Proof. The *refinement* stage corresponds to $\beta \to \infty$. We can prove that in this case, the integration terms in the bias and MSE in Theorem 2 converge to 0.

To show this, define a sequence of functions $\{g_m\}$ as:

$$
g_m \triangleq -\frac{1}{L}\sum_{l=m}^{L+m-1} e^{-\beta_l\hat{U}(\boldsymbol{\theta})} . \qquad (18)
$$

it is easy to see the sequence $\{g_m\}$ satisfies $g_{m_1} < g_{m_2}$ for $m_1 < m_2$, and $\lim_{m\to\infty} g_m = 0$. According to the monotone convergence theorem, we have

$$
\begin{aligned}
\lim_{m\to\infty}\int g_m &\triangleq \lim_{m\to\infty}\int -\frac{1}{L}\sum_{l=m}^{L+m-1} e^{-\beta_l\hat{U}(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} \\
&= \int \lim_{m\to\infty} g_m = 0 .
\end{aligned}
$$

As a result, the integration terms in the bounds for the bias and MSE vanish, leaving only the terms stated in Corollary 3. This completes the proof. □

E Reformulation of the Santa Algorithm

In this section we give a version of the Santa algorithm that matches better than our actual implementation, shown in Algorithm 3–7.

Algorithm 3: Santa

Input: η_t (learning rate), σ, λ, *burnin*, $\beta = \{\beta_1, \beta_2, \cdots\} \to \infty$, $\{\boldsymbol{\zeta}_t \in \mathbf{R}^p\} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.
Initialize $\boldsymbol{\theta}_0$, $\mathbf{u}_0 = \sqrt{\eta} \times \mathcal{N}(0, I)$, $\boldsymbol{\alpha}_0 = \sqrt{\eta}C$, $\mathbf{v}_0 = 0$;
for $t = 1, 2, \ldots$ **do**
  Evaluate $\tilde{\mathbf{f}}_t = \nabla_{\boldsymbol{\theta}}\tilde{U}_t(\boldsymbol{\theta}_{t-1})$ on the t-th minibatch ;
  $\mathbf{v}_t = \sigma\,\mathbf{v}_{t-1} + \frac{1-\sigma}{m^2}\tilde{\mathbf{f}}_t \odot \tilde{\mathbf{f}}_t$;
  $\mathbf{g}_t = 1 \oslash \sqrt{\lambda + \sqrt{\mathbf{v}_t}}$;
  if $t <$ *burnin* **then**
    /* *exploration* */
    $(\boldsymbol{\theta}_t, \mathbf{u}_t, \boldsymbol{\alpha}_t) = \text{Exploration\_S}(\boldsymbol{\theta}_{t-1}, \mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1})$
    or
    $(\boldsymbol{\theta}_t, \mathbf{u}_t, \boldsymbol{\alpha}_t) = \text{Exploration\_E}(\boldsymbol{\theta}_{t-1}, \mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1})$
  else
    /* *refinement* */
    $(\boldsymbol{\theta}_t, \mathbf{u}_t, \boldsymbol{\alpha}_t) = \text{Refinement\_S}(\boldsymbol{\theta}_{t-1}, \mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1})$
    or
    $(\boldsymbol{\theta}_t, \mathbf{u}_t, \boldsymbol{\alpha}_t) = \text{Refinement\_E}(\boldsymbol{\theta}_{t-1}, \mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1})$
  end
end

Algorithm 4: Exploration_S $(\boldsymbol{\theta}_{t-1}, \mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1})$

$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \mathbf{g}_t \odot \mathbf{u}_{t-1}/2$;
$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + (\mathbf{u}_{t-1} \odot \mathbf{u}_{t-1} - \eta/\beta_t)/2$;
$\mathbf{u}_t = \exp(-\boldsymbol{\alpha}_t/2) \odot \mathbf{u}_{t-1}$;
$\mathbf{u}_t = \mathbf{u}_t - \mathbf{g}_t \odot \tilde{\mathbf{f}}_t \eta + \sqrt{2\,\mathbf{g}_{t-1}\,\eta^{3/2}/\beta_t} \odot \boldsymbol{\zeta}_t$;
$\mathbf{u}_t = \exp(-\boldsymbol{\alpha}_t/2) \odot \mathbf{u}_t$;
$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_t + (\mathbf{u}_t \odot \mathbf{u}_t - \eta/\beta_t)/2$;
$\boldsymbol{\theta}_t = \boldsymbol{\theta}_t + \mathbf{g}_t \odot \mathbf{u}_t/2$;
Return $(\boldsymbol{\theta}_t, \mathbf{u}_t, \boldsymbol{\alpha}_t)$

Algorithm 5: Refinement_S $(\boldsymbol{\theta}_{t-1}, \mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1})$

$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}$;
$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \mathbf{g}_t \odot \mathbf{u}_{t-1}/2$;
$\mathbf{u}_t = \exp(-\boldsymbol{\alpha}_t/2) \odot \mathbf{u}_{t-1}$;
$\mathbf{u}_t = \mathbf{u}_t - \mathbf{g}_t \odot \tilde{\mathbf{f}}_t \eta$;
$\mathbf{u}_t = \exp(-\boldsymbol{\alpha}_t/2) \odot \mathbf{u}_t$;
$\boldsymbol{\theta}_t = \boldsymbol{\theta}_t + \mathbf{g}_t \odot \mathbf{u}_t/2$;
Return $(\boldsymbol{\theta}_t, \mathbf{u}_t, \boldsymbol{\alpha}_t)$

F Relationship of *refinement* Santa to Adam

In the Adam algorithm (see Algorithm 1 of Kingma and Ba (2015)), the key steps are:

$$\begin{aligned}
\tilde{\mathbf{f}}_t &\triangleq \nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta}_{t-1}) \\
\mathbf{v}_t &= \sigma\,\mathbf{v}_{t-1} + (1-\sigma)\tilde{\mathbf{f}}_t \odot \tilde{\mathbf{f}}_t \\
\mathbf{g}_t &= 1 \oslash \sqrt{\lambda + \sqrt{\mathbf{v}_t}} \\
\tilde{\boldsymbol{u}}_t &= (\mathbf{1} - \mathbf{b}_1) \odot \tilde{\boldsymbol{u}}_{t-1} + \mathbf{b}_1 \odot \tilde{\mathbf{f}}_t \\
\boldsymbol{\theta}_t &= \boldsymbol{\theta}_t + \eta(\mathbf{g}_t \odot \mathbf{g}_t) \odot \tilde{\boldsymbol{u}}_t
\end{aligned}$$

Here, we maintain the square root form of $\mathbf{g}_t$, so the square is equivalent to the preconditioner used in Adam. As well, in Adam, the vector $\mathbf{b}_1$ is set to the same constant between 0 and 1 for all entries. An equivalent formulation of this is:

$$\begin{aligned}
\tilde{\mathbf{f}}_t &\triangleq \nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta}_{t-1}) \\
\mathbf{v}_t &= \sigma\,\mathbf{v}_{t-1} + (1-\sigma)\tilde{\mathbf{f}}_t \odot \tilde{\mathbf{f}}_t \\
\mathbf{g}_t &= 1 \oslash \sqrt{\lambda + \sqrt{\mathbf{v}_t}} \\
\boldsymbol{u}_t &= (\mathbf{1} - \mathbf{b}_1) \odot \boldsymbol{u}_{t-1} - \eta(\mathbf{g}_t \odot \mathbf{b}_1 \odot \tilde{\mathbf{f}}_t) \\
\boldsymbol{\theta}_t &= \boldsymbol{\theta}_t - \mathbf{g}_t \odot \boldsymbol{u}_t
\end{aligned}$$

The only differences between these steps and the Euler integrator we present in our Algorithm 1 are that our $\mathbf{b}_1$ has a separate constant for each entry, and the second term in $\boldsymbol{u}$ does not include the $\mathbf{b}_1$ in our formulation. If we modify our algorithm to multiply the gradient by $\mathbf{b}_1$, then our algorithm, under the same assumptions as Adam, will have a similar regret bound of $O(\sqrt{T})$ for a convex problem.

Algorithm 6: Exploration_E $(\boldsymbol{\theta}_{t-1}, \mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1})$

$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + (\mathbf{u}_{t-1} \odot \mathbf{u}_{t-1} - \eta/\beta_t)$;
$\mathbf{u}_t = (1 - \boldsymbol{\alpha}_t) \odot \mathbf{u}_{t-1} - \eta\,\mathbf{g}_t \odot \tilde{\mathbf{f}}_t + \sqrt{2\,\mathbf{g}_{t-1}\,\eta^{3/2}/\beta_t} \odot \boldsymbol{\zeta}_t$;
$\boldsymbol{\theta}_t = \boldsymbol{\theta}_t + \mathbf{g}_t \odot \mathbf{u}_t$;
Return $(\boldsymbol{\theta}_t, \mathbf{u}_t, \boldsymbol{\alpha}_t)$

Algorithm 7: Refinement_E $(\boldsymbol{\theta}_{t-1}, \mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1})$

$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}$;
$\mathbf{u}_t = (1 - \boldsymbol{\alpha}_t) \odot \mathbf{u}_{t-1} - \eta\,\mathbf{g}_t \odot \tilde{\mathbf{f}}_t$;
$\boldsymbol{\theta}_t = \boldsymbol{\theta}_t + \mathbf{g}_t \odot \mathbf{u}_t$;
Return $(\boldsymbol{\theta}_t, \mathbf{u}_t, \boldsymbol{\alpha}_t)$

Because the focus of this paper is not on the regret bound, we only briefly discuss the changes in the theory. We note that Lemma 10.4 from Kingma and Ba (2015) will hold with element-wise $\mathbf{b}_1$.

Lemma 5. *Let* $\gamma_i \triangleq \frac{b_{1,i}^2}{\sqrt{\sigma}}$. *For* $b_{1,i}, \sigma \in [0,1)$ *that satisfy* $\frac{\beta_1^2}{\sqrt{\beta_2}} < 1$ *and bounded* $\tilde{f}_t$, $||\tilde{f}_t||_2 \le G$, $||\tilde{f}_t||_\infty \le G_\infty$, *the following inequality holds*

$$\sum_{t=1}^{T} \frac{u_i^2}{\sqrt{t g_i^2}} \le \frac{2}{1-\gamma_i} ||\tilde{f}_{1:T,i}||_2$$

which contains an element-dependent γ_i *compared to Adam.*

Theorem 10.5 of Kingma and Ba (2015) will hold with the same modifications and assumptions for a $\mathbf{b}$ with distinct entries; the proof in Kingma and Ba (2015) is already element-wise, so it suffices to replace their global parameter γ with distinct $\gamma_i \triangleq \frac{b_{1,i}^2}{\sqrt{\sigma}}$. This will give a regret of $O(\sqrt{T})$, the same as Adam.

G Additional Results

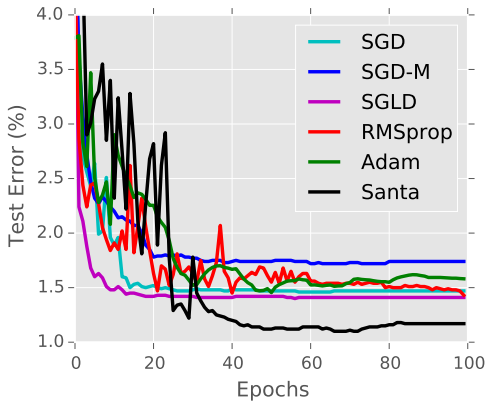

Figure 4: MNIST using FNN with size of 800.

Learning curves of different algorithms on MNIST using FNN with size of 800 are plotted in Figure 4. Learning curves of different algorithms on four polyphonic music datasets using RNN are shown in Figure 6.

We additionally test Santa on the ImageNet dataset. We use the GoogleNet architecture, which is a 22 layer deep model. We use the default setting defined in the Caffe package[8]. We were not able to make other stochastic optimization algorithms except SGD with momentum and the proposed Santa work on this dataset. Figure 5 shows the comparison on this dataset. We did not tune the parameter setting, note the default setting is favourable by SGD with momentum. Nevertheless, Santa still significantly outperforms SGD with momentum in term of convergence speed.

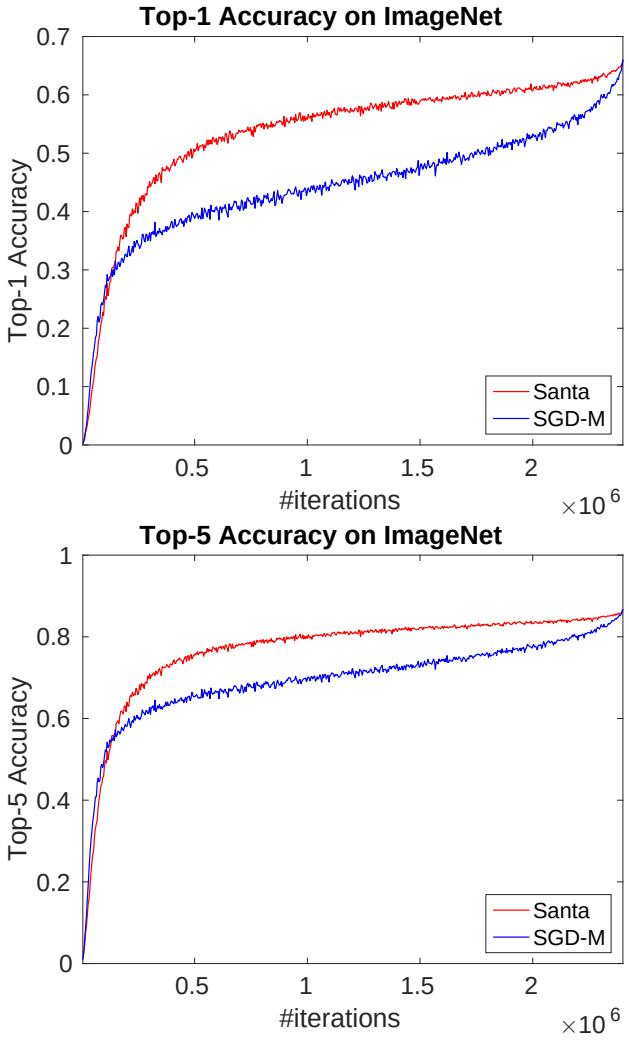

Figure 5: Santa vs. SGD with momentum on ImageNet. We used ImageNet11 for training.

[8] *https://github.com/cchangyou/Santa/tree/master/caffe/models/bvlc_googlenet*

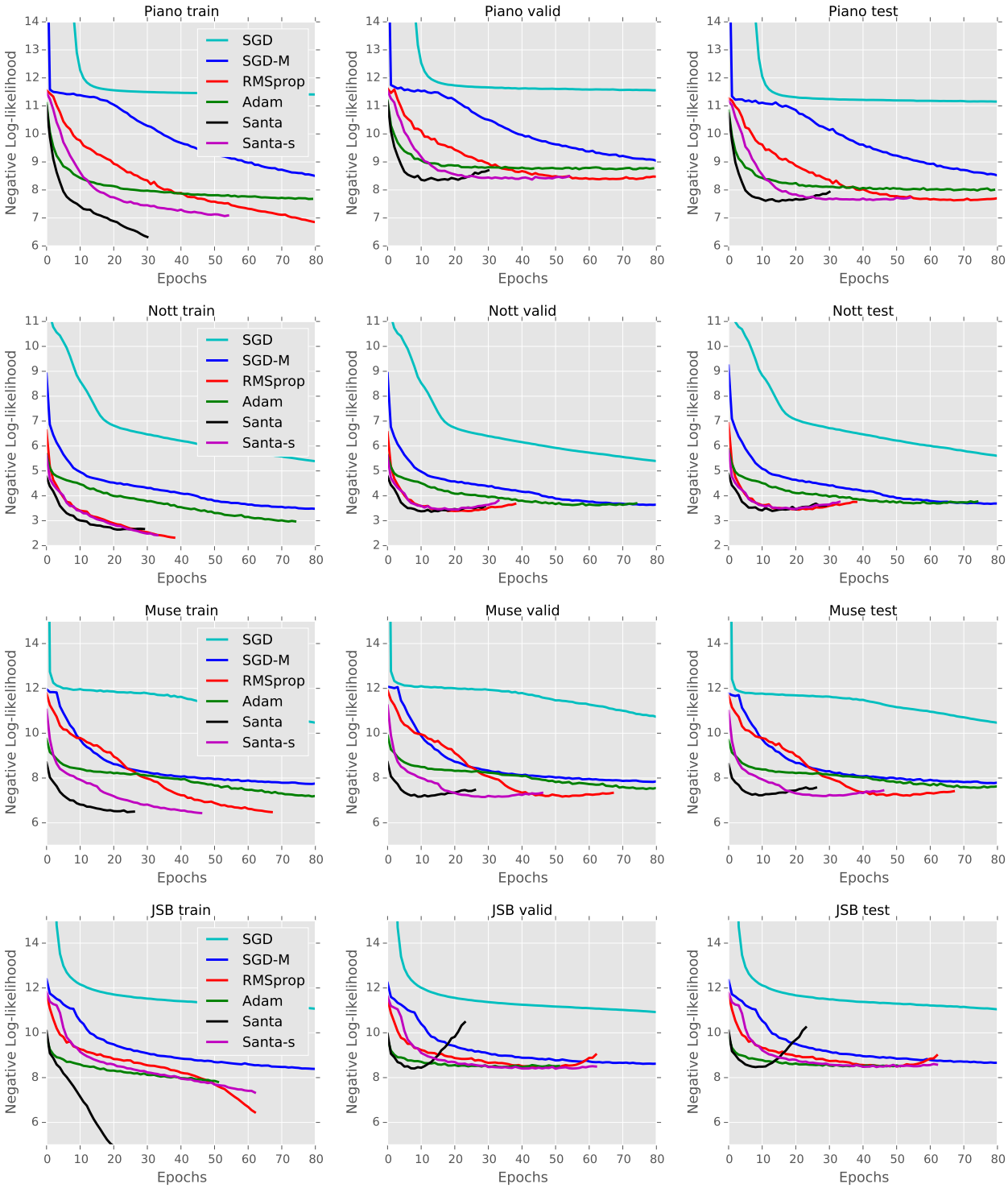

Figure 6: Learning curves of different algorithms on four polyphonic music datasets using RNN.