Appendix for Stochastic Gradient Monomial Gamma Sampler

A. The Main Theorem

We provide the following theorem to characterize the stationary distribution of the stochastic process with SDEs in (12).

Theorem 3. *The stochastic process generated from SDEs* (12) *converges to a stationary distribution* $p(\Gamma) \propto \exp(-H(\Gamma))$*, where* $H(\Gamma)$ *is defined as in* (9).

Proof. We first show that the Fokker-Planck equation holds for the proposed SDE and probability density $p(\Gamma)$,

$$\nabla_\Gamma \cdot p(\Gamma)V(\Gamma) = \nabla_\Gamma \nabla_\Gamma^T : [p(\Gamma)D(\Gamma)]$$

Here the $\nabla \triangleq (\partial/\partial\theta, \partial/\partial p, \partial/\partial\xi)$. $\cdot$ represents a vector inner product and $:$ denotes the matrix double dot product, *i.e.*, $X : Y = \mathrm{Tr}(X^T Y)$. In order to show FP equation holds, we look at both side of the equation.

The left hand side can be written as

$$\begin{aligned}
& \nabla_\Gamma \cdot p(\Gamma)V(\Gamma) \\
= \; & [\frac{\partial V(\Gamma)}{\partial \Gamma} - \frac{\partial H(\Gamma)}{\partial \Gamma} V(\Gamma)]p(\Gamma) \\
= \; & \{\sigma_\theta[(\nabla U(\theta))^2 - \nabla^2 U(\theta)] \\
& + \sigma_p[(\nabla K(p))^2 - \nabla^2 K(p)] \\
& + \sigma_\xi[(\nabla F(\xi))^2 - \nabla^2 F(\xi)]\}p(\Gamma)
\end{aligned}$$

For the right hand side,

$$\begin{aligned}
& \nabla_\Gamma \nabla_\Gamma^T : [p(\Gamma)D(\Gamma)] \\
= \; & \sigma_\theta \nabla\nabla^T : p(\Gamma) + \sigma_p \nabla\nabla^T : p(\Gamma) + \sigma_\xi \nabla\nabla^T : p(\Gamma) \\
= \; & \{\sigma_\theta[(\nabla U(\theta))^2 - \nabla^2 U(\theta)] \\
& + \sigma_p[(\nabla K(p))^2 - \nabla^2 K(p)] \\
& + \sigma_\xi[(\nabla F(\xi))^2 - \nabla^2 F(\xi)]\}p(\Gamma)
\end{aligned}$$

For stationary distribution,

$$\frac{\partial p(\Gamma, t)}{\partial t} = 0$$

As a result, the equality in (13) holds. The stochastic process defined by (12) is preserved by the dynamic. Alternatively, one can leverage the recipe from (Ma et al., 2015) to recover the same conclusion, by setting semi-definite matrix $D = \mathrm{Diag}([\sigma_\theta, \sigma_p, \sigma_\xi])$ and skew-symmetric Q to be

$$\begin{pmatrix} 0 & -I & 0 \\ I & 0 & \gamma\nabla K(p) \\ 0 & -\gamma\nabla K(p) & 0 \end{pmatrix}$$

Note that under the softened kinetics, the $K_c(p)$ is twice differentiable, and $\nabla K_c(p)$ is Lipschitz continuous. Thus the Fokker-Planck equation holds, leading to a stationary distribution invariant to target distribution. Another remark is that the resampling process for p and ξ will still lead to the same invariante distribution $p(\Gamma)$, since the resampling process is directly drawing sample from the marginal distribution. Finally, it can be proved that the corresponding Itô diffusion of our algorithm in (12) is non-reversible. This speed up the convergence speed to equilibrium, because it is known that a reversible process convergences slower than its non-reversible counter part (Hwang et al., 2005). □

B. Details for softened kinetics

We provide the details for the derivation of softened kinetics. Note that in the SDE (12), only $\nabla K_c(p)$ and $\nabla^2 K_c(p)$ is involved. For $a = 1$, we consider

$$\begin{aligned}
K_c(p) &= -g(p) + 2/c \log(1 + e^{cg(p)}), \\
g(p) &= p/m.
\end{aligned} \tag{13}$$

which gives

$$\begin{aligned}
\nabla K_c(p) &= \tfrac{1}{m}\psi(g(p)), \\
\nabla^2 K_c(p) &= \tfrac{1}{m^2}\psi'(g(p)).
\end{aligned}$$

Where, $\psi(x) = \frac{e^{cx}-1}{e^{cx}+1}$ is the hyperbolic tangent function (tanh) with the softening parameter c, $\psi'(x) = \frac{2ce^{cx}}{(e^{cx}+1)^2}$.

For $a = 2$, we consider

$$\begin{aligned}
K_c(p) &= g(p) + \frac{4}{c(1 + e^{cg(p)})}, \\
g(p) &= |p|^{1/2}/m.
\end{aligned} \tag{14}$$

which gives

$$\begin{aligned}
\nabla K_c(p) &= \tfrac{1}{2m}\mathrm{sign}(p)\psi(g(p))^2 |p|^{-1/2}, \\
\nabla^2 K_c(p) &= \tfrac{1}{2m^2}\psi(g(p))\psi'(g(p))|p|^{-1} \\
& \quad - \tfrac{1}{4m}\psi^2(g(p))|p|^{-3/2}.
\end{aligned}$$

In general, for arbitrary a, we consider setting the

$$\nabla K_c(p) = \frac{a}{m}\psi(g(p))^a |p|^{-1/a},$$

Such specification will yield a differentiable softened kinetics function by computing the integral, which is tractable for positive value of a. However, in practice, as suggested by (Zhang et al., 2016) the optimal a would usually between $[0.5, 2]$. We would suggest consider using $a = 1$ or $a = 2$ for general inference tasks.

C. Synthetic multi-well potential problem

The five-well potential is defined as:

$$U(\theta) \triangleq e^{\frac{3}{4}\theta^2 - \frac{3}{2}\sum_{i=1}^{10} c_i \sin\left(\frac{1}{4}\pi i(\theta+4)\right)},$$

where $c = (-0.47, -0.83, -0.71, -0.02, 0.24, 0.01, 0.27, -0.37, 0.87, -0.37)$ is a vector, c_i is the i-th element of c.

D. Symmetric Splitting Integrators for SGMGT

The first-ordered Euler integration results in high discretization error in Hamiltonian dynamic updating of HMC. In (Chen et al., 2016), a symmetric splitting scheme is leveraged to reduce the numerical error. We applied the softened kinetics $K_c(p)$, and set $F(\xi)$ as $\frac{(\xi-\sigma_p)^2}{2\gamma}$. In this symmetric splitting scheme, the Hamiltonian is split into sub-componenents, and for each sub-componenents an individual SDE is applied on. The resulting discretization is symplectic and second-ordered:

$$
\begin{aligned}
A : d\Gamma &= \begin{pmatrix} -\sigma_\theta \nabla \tilde{U}(\theta) + \nabla K_c(p) \\ 0 \\ f(\Gamma) \end{pmatrix} dt/2 \\
B : d\Gamma &= \begin{pmatrix} 0 \\ -\xi \cdot \nabla K_c(p) \\ 0 \end{pmatrix} dt/2 \\
O : d\Gamma &= \begin{pmatrix} 0 \\ -\nabla \tilde{U}(\theta) \\ 0 \end{pmatrix} dt + D(\Gamma) dW
\end{aligned}
$$

Here we denote $f(\Gamma) \triangleq \gamma[(\nabla K_c(p))^2 - (\nabla^2 K_c(p))] - \frac{\sigma_\xi}{\gamma}(\xi - \sigma_p)$ for clarity. The sub-SDE under sub-SDE B is analytically solvable. Following (Chen et al., 2015), for $a \neq 1/2$, the updating procedure follows an ABOBA scheme, given by

Table 4. Experimental setup for discriminative RBM

Algorithms	σ_p	σ_θ	σ_ξ	γ	c	h
SGNHT	10	-	-	1	-	2e-4
SGNHT-D	10	0.1	0.1	1	-	2e-4
SGMGT-D (a=1)	10	0.1	0.1	1	3	1e-5
SGMGT-D (a=2)	10	0.1	0.1	1	5	5e-5

$$
\begin{aligned}
&A : \theta_{t+1/3} = \theta_t + \nabla K_c(p)h/2, \xi_{t+1/3} = \xi_t + f(\Gamma)h/2 \\
&B : p_{t+1/3} = [p_t^{(2a-1)/a} - \frac{2a-1}{a^2}\xi_{t+1/2}h/2]^{a/(2a-1)} \\
&O : \theta_{t+2/3} = \theta_{t+1/3} + \sqrt{2\sigma_\theta}\epsilon_\theta \\
&\quad p_{t+2/3} = p_{t+1/3} - \nabla \tilde{U}(\theta))h/2 + \sqrt{2\sigma_p}\epsilon_p, \\
&\quad \xi_{t+2/3} = \xi_{t+1/3} + \sqrt{2\sigma_\xi}\epsilon_\xi \\
&B : p_{t+1} = [p_{t+2/3}^{(2a-1)/a} - \frac{2a-1}{a^2}\xi_{t+2/3}h/2]^{a/(2a-1)} \\
&A : \theta_{t+1} = \theta_{t+2/3} + \nabla K_c(p)h/2, \xi_{t+1} = \xi_{t+2/3} + f(\Gamma)h/2
\end{aligned}
$$

When $a = 1/2$, it follows the splitting scheme with standard SGNHT (Chen et al., 2015).

E. Experimental setups for DRBM

The hyper-parameter setups for the DRBM experiments are provided as below. We select the hyperparameters based on the performance on validation dataset. The algorithm will be early stopped if the validation error start to increase. The selection is based on a grid search. For σ_p, σ_ξ and σ_θ we select from $\{0.001, 0.01, 0.1, 1, 10\}$. For the softening parameter c we select from $\{3, 5, 8\}$. We fixed the $m = 1$ and $\gamma = 1$. The stepsize is chosen from $\{1e-5, 2e-5, 5e-5, 1e-4, 2e-4, 5e-4\}$. The T_p and T_ξ are set as 100 and 100, respectively.

For SGLD, we use a stepsize of $1e-5$

F. Experimental setups for RNNs

The hyper-parameter setups for the RNNs experiments are similar to the DRBM experiments. For σ_p, σ_ξ and σ_θ we select from $\{0.01, 0.1, 1, 10\}$. For the softening parameter c we select from $\{3, 5, 8\}$. We fixed the $m = 1$ and $\gamma = 1$. The stepsize of SGMGT-D/SGMGT is chosen from $\{1e-3, 1.5e-3, 2e-3, 2.5e-3, 3e-3\}$. The T_p and T_ξ are set as 100 and 100, respectively. We also incorporate a decay scheme for stepsize, *i.e.* the stepsize is divided by a decaying factor $\alpha = 1.1$ for each scan of dataset (*i.e.* each epoch). The gradient estimated on a subset of data is clipped to have a maximum value of 5 as in (Chen et al., 2016) for each dimension to prevent updates from a large

Table 5. Experimental setup for discriminative RNNs

Algorithms	m	σ_p	σ_θ	σ_ξ	γ	c
SGNHT	1	10	-	-	1	-
SGNHT-D	1	10	0.01	0.01	1	-
SGMGT/SGMGT-D (a=1)	1	10	0.1	0.01	1	5
SGMGT/SGMGT-D (a=2)	1	10	0.1	0.01	1	3

gradient value to blow up the objective loss. For JSB we use a stepsize of $2e-3$ for SGMGT, for other three datasets (Piano, Muse, Nott) we use a stepsize of $3e-3$. For SGLD, we use a stepsize of $1e-3$, for SGNHT the stepsize is set as $5e-5$. The other hyperparameters are provided in 5

G. Additional figure for RNNs experiment

We provide the traceplot of one parameter in RNN experiment of JSB dataset. We choose this parameter at random. Generally, the SGMGT with $a = 2$ seems to demonstrate more random walk behavior than SGMGT with $a = 1$

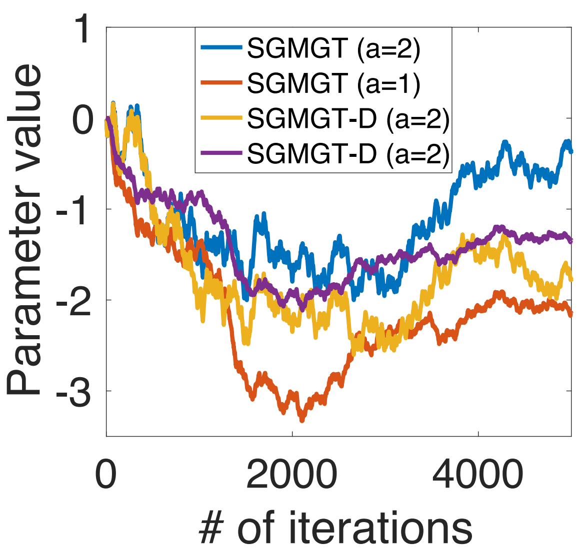

Figure 6. Traceplot for RNN experiments

H. Additional results for RNN experiments

Here we provide the results of several optimization methods, the results are taken from Chen et al. (2016).

Table 6. Test negative log-likelihood results on polyphonic music datasets using RNN.

Algorithms	Piano.	Nott.	Muse.	JSB.
Adam	8.00	3.70	7.56	8.51
RMSprop	7.70	3.48	7.22	8.52
SGD-M	8.32	3.60	7.69	8.59
SGD	11.13	5.26	10.08	10.81
HF	7.66	3.89	7.19	8.58
SGD-M	8.37	4.46	8.13	8.71

I. Convergence property

Proof. This follows the proof for general SG-MCMC algorithms. Specifically, in SGMGT, the generator of the corresponding SDE is defined as:

$$\mathcal{L}f(x) \triangleq \left(F(x) \cdot \nabla + \frac{1}{2} \left(\Sigma\Sigma^T\right) : \nabla\nabla^T \right) f(x) \, ,$$

where

$$x = (\theta, p, \xi),$$
$$F(x) = \begin{pmatrix} -\sigma_\theta \nabla U(\theta) + \nabla K_c(p) \\ -\nabla\tilde{U}(\theta) - (\sigma_p + \gamma \nabla F(\xi)) \nabla K_c(p) \\ \gamma\left[(\nabla K_c(p))^2 - \nabla^2 K_c(p)\right] - \sigma_\xi \nabla F(\xi) \end{pmatrix},$$
$$\Sigma = \begin{pmatrix} \sqrt{2\sigma_\theta} & 0 & 0 \\ 0 & \sqrt{2\sigma_p} & 0 \\ 0 & 0 & \sqrt{2\sigma_\xi} \end{pmatrix}.$$

After introducing stochastic gradients, in each iteration t, the generator is perturbed by:

$$\Delta V_t = \left(\nabla\tilde{U}(\theta) - \nabla U(\theta)\right) \cdot (\nabla - \sigma_\theta \nabla) \, ,$$

such that $\tilde{\mathcal{L}}_t = \mathcal{L} + \Delta V_t$, where $\tilde{\mathcal{L}}_t$ is the local generator for the SDE in iterator t.

After defining these notation, we follows the proofs of Theorem 2 and Theorem 3 in (Chen et al., 2015).

The proof for the bias: Following Theorem 2 in Chen et al. (2015), in the decreasing step size setting, the split flow can be written as:

$$\mathbb{E}\left(\psi(\mathbf{X}_{lh})\right) = \left(\mathbb{I} + h_l \tilde{\mathcal{L}}_l\right) \psi(\mathbf{X}_{(l-1)h})$$
$$+ \sum_{k=2}^{K} \frac{h_l^k}{k!} \tilde{\mathcal{L}}_l^2 \psi(\mathbf{X}_{(l-1)h}) + O(h_l^{K+1}) \, .$$

Similarly, the expected difference between $\tilde{\phi}$ and $\bar{\phi}$ can be simplified using the step size sequence (h_l) as:

$$\mathbb{E}\left(\tilde{\phi}-\bar{\phi}\right) \tag{15}$$

$$=\frac{1}{S_L}\left(\mathbb{E}\left(\psi(\mathbf{X}_{Lh})\right)-\psi(\mathbf{X}_0)\right) \tag{16}$$

$$-\sum_{k=2}^{K}\sum_{l=1}^{L}\frac{h_l^k}{k!S_L}\tilde{\mathcal{L}}_l^k\psi(\mathbf{X}_{(l-1)h})+O(\frac{\sum_{l=1}^{L}h_l^{K+1}}{S_L}) \tag{17}$$

Similar to the derivation in Chen et al. (2015), we can derive the following bounds $k=(2,\cdots,K)$:

$$\sum_{l=1}^{L}h_l^k\mathbb{E}\tilde{\mathcal{L}}_l^k\psi(\mathbf{X}_{(l-1)h}) \tag{18}$$

$$=O\left(\sum_{l=1}^{L}\left((h_l^{k-1}-h_{l-1}^{k-1})\tilde{\mathcal{L}}_l^{k-1}\psi(\mathbf{X}_{(l-1)h})+h_l^{K+1}\right)\right)$$

$$=O\left(1+\sum_{l=1}^{L}h_l^{K+1}\right). \tag{19}$$

Substitute (18) into (15) and collect low order terms, we have:

$$\mathbb{E}\left(\tilde{\phi}-\bar{\phi}\right) \tag{20}$$

$$=\frac{1}{S_L}\left(\mathbb{E}\left(\psi(\mathbf{X}_{Lh})\right)-\psi(\mathbf{X}_0)\right)+O(\frac{\sum_{l=1}^{L}h_l^{K+1}}{S_L}). \tag{21}$$

As a result, the bias can be expressed as:

$$\begin{aligned}\left|\mathbb{E}\tilde{\phi}-\bar{\phi}\right| &\leq\left|\frac{1}{S_L}\left(\mathbb{E}\left[\psi(\mathbf{X}_{Lh})\right]-\psi(\mathbf{X}_0)\right)+O(\frac{\sum_{l=1}^{L}h_l^{K+1}}{S_L})\right|\\ &\lesssim\left|\frac{1}{S_L}\right|+\left|\frac{\sum_{l=1}^{L}h_l^{K+1}}{S_L})\right|\\ &=O\left(\frac{1}{S_L}+\frac{\sum_{l=1}^{L}h_l^{K+1}}{S_L}\right).\end{aligned}$$

Taking $L\to\infty$, both terms go to zero by assumption.

The proof for the MSE: Following similar derivations as in Theorem 2 in Chen et al. (2015), we have that

$$\begin{aligned}\sum_{l=1}^{L}\mathbb{E}\left(\psi(\mathbf{X}_{lh})\right)=&\sum_{l=1}^{L}\psi(\mathbf{X}_{(l-1)h})+\sum_{l=1}^{L}h_l\mathcal{L}\psi(\mathbf{X}_{(l-1)h})\\ &+\sum_{l=1}^{L}h_l\Delta V_l\psi(\mathbf{X}_{(l-1)h})\\ &+\sum_{k=2}^{K}\sum_{l=1}^{L}\frac{h_l^k}{k!}\tilde{\mathcal{L}}_l^k\psi(\mathbf{X}_{(l-1)h})+C\sum_{l=1}^{L}h_l^{K+1}.\end{aligned}$$

Substitute the Poisson equation into the above equation and divided both sides by S_L, we have

$$\begin{aligned}\hat{\phi}-\bar{\phi}=&\frac{\mathbb{E}\psi(\mathbf{X}_{Lh})-\psi(x_0)}{S_L}\\ &+\frac{1}{S_L}\sum_{l=1}^{L-1}\left(\mathbb{E}\psi(\mathbf{X}_{(l-1)h})+\psi(\mathbf{X}_{(l-1)h})\right)\\ &+\sum_{l=1}^{L}\frac{h_l}{S_L}\Delta V_l\psi(\mathbf{X}_{(l-1)h})\\ &+\sum_{k=2}^{K}\sum_{l=1}^{L}\frac{h_l^k}{k!S_L}\tilde{\mathcal{L}}_l^k\psi(\mathbf{X}_{(l-1)h})+C\frac{\sum_{l=1}^{L}h_l^3}{S_L}.\end{aligned}$$

As a result, there exists some positive constant C, such that:

$$\mathbb{E}\left(\hat{\phi}-\bar{\phi}\right)^2\leq C\mathbb{E}\left(\frac{1}{S_L^2}\underbrace{\left(\psi(\mathbf{X}_0)-\mathbb{E}\psi(\mathbf{X}_{Lh})\right)^2}_{A_1}\right. \tag{22}$$

$$\begin{aligned}&+\underbrace{\frac{1}{S_L^2}\sum_{l=1}^{L}\left(\mathbb{E}\psi(\mathbf{X}_{(l-1)h})-\psi(\mathbf{X}_{(l-1)h})\right)^2}_{A_2}\\ &+\sum_{l=1}^{L}\frac{h_l^2}{S_L^2}\left\|\Delta V_l\right\|^2+\underbrace{\sum_{k=2}^{K}\left(\sum_{l=1}^{L}\frac{h_l^k}{k!S_L}\tilde{\mathcal{L}}_l^k\psi(\mathbf{X}_{(l-1)h})\right)^2}_{A_3}\\ &\left.+\left(\frac{\sum_{l=1}^{L}h_l^3}{S_L}\right)^2\right)\end{aligned} \tag{23}$$

A_1 can be bounded by assumptions, and A_2 is shown to be bounded by using the fact that $\mathbb{E}\psi(\mathbf{X}_{(l-1)h})-\psi(\mathbf{X}_{(l-1)h})=O(\sqrt{h_l})$ from Theorem 2 in Chen et al. (2015). Furthermore, similar to the proof of Theorem 2 in Chen et al. (2015), the expectation of A_3 can also be bounded by using the formula $\mathbb{E}[\mathbf{X}^2]=(\mathbb{E}\,\mathbf{X})^2+\mathbb{E}[(\mathbf{X}-\mathbb{E}\,\mathbf{X})^2]$ and (18). It turns out that the resulting terms have order higher than those from the other terms, thus can be ignored in the expression below. After some simplifications, (22) is bounded by:

$$\mathbb{E}\left(\hat{\phi}-\bar{\phi}\right)^2 \tag{24}$$

$$\begin{aligned}&\lesssim\sum_l\frac{h_l^2}{S_L^2}\mathbb{E}\left\|\Delta V_l\right\|^2+\frac{1}{S_L}+\frac{1}{S_L^2}+\left(\frac{\sum_{l=1}^{L}h_l^{K+1}}{S_L}\right)^2\\ &=C\left(\sum_l\frac{h_l^2}{S_L^2}\mathbb{E}\left\|\Delta V_l\right\|^2+\frac{1}{S_L}+\frac{(\sum_{l=1}^{L}h_l^{K+1})^2}{S_L^2}\right)\end{aligned} \tag{25}$$

for some $C>0$, this completes the first part of the theorem. We can see that according to the assumption, the last two terms in (24) approach to 0 when $L \to \infty$. If we further assume $\frac{\sum_{l=1}^{\infty} h_l^2}{S_L^2} = 0$, then the first term in (24) approaches to 0 because:

$$\sum_l \frac{h_l^2}{S_L^2} \mathbb{E} \left\| \Delta V_l \right\|^2 \le \left(\sup_l \mathbb{E} \left\| \Delta V_l \right\|^2 \right) \frac{\sum_l h_l^2}{S_L^2} \to 0 \, .$$

As a result, we have $\lim_{L\to\infty} \mathbb{E} \left(\hat{\phi} - \bar{\phi} \right)^2 = 0$.

□

J. Proof for Lemma 1

To prove Lemma 1, we first introduce the following lemma from (Geyer, 2005).

Lemma 3 (Geyer (2005))**.** *Suppose μ is a probability distribution and for each z in the domain of domain of μ there is a Markov kernel P_z satisfying $\pi = \pi P_z$, and suppose that the map $(z, x) \rightarrowtail P_z(x, A)$ is jointly measurable for each A. Then*

$$Q(x, A) = \int \mu(\mathrm{d}z) P_z(x, A)$$

defines a kernel Q that is Markov and satisfies $\pi = \pi Q$.

Proof. Detailed proof can be found in Chapter 3 of Geyer (2005). □

Now it is ready to prove Lemma 1.

Proof of Lemma 1. First, we note that the momentum (or other auxiliary variables) is resampled from the stationary distribution of the Itô diffusion. As a result, for each model parameter θ, it corresponds to a Markov kernel P_θ with the stationary Gaussian density. According to Lemma 3, the composition of the numerical integrator in SGMGT and the resampling forms a Markov kernel $Q(\theta, A)$, such that

$$\pi_h = \pi_h Q \, .$$

The above equation means that π_h is also the stationary distribution of the Markov kernel Q, which completes the proof. □

K. Proof for Lemma 2

Proof. First, the optimal bias and MSE bounds in Proposition 2 are given by:

$$\text{Bias: } \left| \mathbb{E}\hat{\phi}_T - \bar{\phi} \right| = O\left(T^{-1/2} \right) \, ,$$
$$\text{MSE: } \mathbb{E}\left(\hat{\phi} - \bar{\phi} \right)^2 = O\left(T^{-2/3} \right) \, .$$

Let the number of samples in each resampling period to be $(T_l)_{l=1}^{L}$, and denote $T \triangleq \sum_{l=1}^{L} T_l$. Further denote the sample average in the l-th resampling period to be:

$$\hat{\phi}_{T_l} \triangleq \frac{1}{T_l} \sum_{l=1}^{T_l} \phi(x_l^{(T_l)}) \, ,$$

where $\{x_l^{(T_l)}\}$ denotes samples in the l-th resampling period. The final sample average is defined as:

$$\hat{\phi}_T \triangleq \sum_{l=1}^{L} \frac{T_l}{\sum_{l'=1}^{T_{l'}}} \hat{\phi}_{T_l} \, .$$

As a result, the bias can be bounded as:

$$\begin{aligned}
&\left| \mathbb{E}\hat{\phi}_T - \bar{\phi} \right| = \left| \mathbb{E} \sum_{l=1}^{L} \frac{T_l}{\sum_{l'=1}^{T_{l'}}} \hat{\phi}_{T_l} - \bar{\phi} \right| \\
=& \frac{1}{\sum_l T_l} \left| \sum_{l=1} T_l \left(\mathbb{E}\hat{\phi}_{T_l} - \bar{\phi} \right) \right| \\
\le& \sum_l \frac{T_l}{\sum_{l'} T_{l'}} \left| \mathbb{E}\hat{\phi}_{T_l} - \bar{\phi} \right| \\
=& \sum_l \frac{T_l}{\sum_{l'}} T_{l'} O\left(\frac{1}{T_l h} + h \right) \\
=& \sum_l \frac{1}{\sum_{l'}} T_{l'} O\left(\frac{1}{h} + T_l h \right)
\end{aligned}$$

Optimizing over h, we have

$$\begin{aligned}
&\left| \mathbb{E}\hat{\phi}_T - \bar{\phi} \right| = \sum_l \frac{1}{\sum_{l'}} T_{l'} O\left(T_l^{1/2} \right) \\
\le& O\left(\frac{(\sum_l T_l)^{1/2}}{\sum_l T_l} \right) = O\left(T^{-1/2} \right) \, ,
\end{aligned}$$

which is the same as the optimal bias bound for SGMGT.

The proof for the optimal MSE bound follows similarly. □

L. Stochastic slice sampling

In this section, we leverage the connection between slice sampling and HMC (Zhang et al., 2016), to investigate the approach to perform slice sampling with subset of data.

Slice sampling (Neal, 2003) augments the density $p(\theta)/C$ (where $C > 0$ is a normalization constant) with slice variables u, such that the joint distribution $p(\theta, u) = 1/C$, s.t. $0 < u < p(\theta)$. To sample from the target distribution, slice sampling is performed in a Gibbs sampling manner, i.e., alternating between uniformly sampling the slice variable (*slice sampling step*) u, and uniformly generating new samples θ (*conditional sampling step*), from a restricted domain such that the (unnormalized) density function values for θ are less than the sampled slice variable u.

Slice sampling allows moves that can adaptively fit the scale of the local density structure, thus yielding rapid mixing. When the dataset is large, however, full-data density evaluations can be very expensive. One recent attempt to use subset data for slice sampling incorporates a hypothesis test sub-procedure when performing the *conditional sampling step* (DuBois et al., 2014). However, the rejection rate could be large if the mini-batch size is small. Furthermore, samples from the algorithm are biased due to the hypothesis test step.

One straightforward approach to perform stochastic slice sampling is by evaluating the likelihood on a subset of data during the *conditional sampling step* when performing standard slice sampling. This approach, detailed in the SM is referred as naïve stochastic slice sampling (Naïve stochastic SS). As shown in Figure 7 in the SM, applying this naïve implementation to a Bayesian linear regression problem would yield over-dispersed samples.

The reason why naïve stochastic slice sampling fails can be explained by following the logic of Zhang et al. (2016) and Chen et al. (2014); Betancourt (2015). In (Zhang et al., 2016), the authors demonstrate the connection between slice sampling and Hamiltonian Monte Carlo, revealed by Hamiltonian-Jacobi Equation. As a result, performing slice sampling can be equivalently realized in an HMC formulation.

We consider mapping naïve stochastic slice sampling to its equivalent HMC space parameterized by model parameter θ and momentum p as in (26) (where the monomial parameter $a = 1$, with notation from Zhang et al. (2016)). This results in an HMC formulation that is equivalent to the naïve stochastic gradient HMC in Chen et al. (2014), but with different kinetic function, as in (3) when $a = 1$. Similar to (Chen et al., 2014), the entropy of the joint distribution of (θ, p) would always increase due to the stochastic noise, explaining the over-dispersion distribution that we observe in Figure 7 in the SM.

Fortunately, one can leverage the connection between slice sampling and HMC from Zhang et al. (2016) to perform an improved stochastic slice sampling. This is done by adopting the SDE of SGHMC in (7) and substituting the Gaussian kinetic with a softened Laplace kinetic (*i.e.* $K_c(p)$ when $a = 1$) as in (11). A friction term $A\nabla K_c(p)$ is incorporate to offset the stochastic noise, resulting in

$$d\theta = \nabla K_c(p)dt, \tag{26}$$

$$dp = -[\nabla\tilde{U}(\theta) + A\nabla K_c(p)]dt + \sqrt{2(AI - \hat{B}(\theta))}dW\,.$$

The resulting stochastic Laplace HMC algorithm (detailed in the SM) from (26) is (asymptotically) invariant to the target distribution, and performs equivalently to a correct stochastic slice sampling in one-dimensional cases, as $c \to \infty$. In Figure 7, the stochastic Laplace HMC sampler can ameliorate the over-dispersion of sampled posterior distribution than naïve stochastic slice sampling.

M. Naive stochastic slice sampling and Stochastic Laplacian HMC

The naïve stochastic slice sampling can be described in Algorithm 1

Algorithm 1 Naïve stochastic SS.

Input: Initial parameter θ_0.
for $t = 1, 2, \ldots$ **do**
  Sampling a mini-batch $\tilde{x}_t$.
  Evaluate stochastic negative log-density $\tilde{U}_{\tilde{x}_t}(\theta_{t-1}) \triangleq \exp[-\log p(\theta_{t-1}) - \frac{N}{N'}\sum_{x'\in\tilde{x}_t}\log p(x'|\theta_{t-1})]$.
  Uniformly sample u_t from $(0, \exp[-\tilde{U}_{\tilde{x}_t}(\theta_{t-1})])$.
  Sample θ_t from $\{\theta : \tilde{U}_{\tilde{x}_t}(\theta) < \log(-u_t)\}$ using doubling and shrinking (Neal, 2003).
end for

According to Zhang et al. (2016), Algorithm 1 has deep connection to Algorithm 2 in HMC formulation, in univariate scenarios.

Algorithm 2 Naïve stochastic SS in HMC space.

Input: Initial parameter θ_0.
for $t = 1, 2, \ldots$ **do**
  Sampling a mini-batch $\tilde{x}_t$.
  Sampling each momentum p independently (for each θ dimension) from a Laplacian distribution $\mathcal{L}(m)$, where $m > 0$ is the mass parameter.
  for $s = 1, 2, \ldots$ **do**
    Evaluate stochastic gradient, $\nabla\tilde{U}(\theta)$, from (5) on mini-batch $\tilde{x}_t$.
    Perform leap-frog updates using (4) by substituting the $\nabla U(\theta)$ with $\nabla\tilde{U}(\theta)$ and substituting $\nabla K(p)$ with $\text{sign}(p)/m$.
  end for
end for

By adding a friction term as in Chen et al. (2014) we provide the corrected SG-MCMC Algorithm 2 that corresponds to stochastic slice sampling.

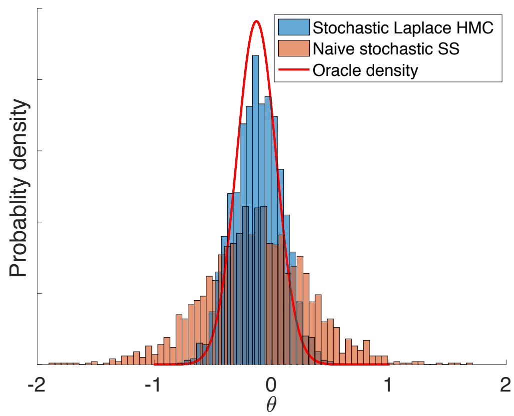

Figure 7. Naïve stochastic slice sampling *vs.* Stochastic Laplacian HMC

Algorithm 3 Stochastic Laplacian HMC

Input: Initial parameter θ_0.
for $t = 1, 2, \ldots$ **do**
  Sampling a momentum p from a distribution $\propto \exp(-K_c(p))$ with $K_c(p)$ defined in (11), where c is the softened parameter.
  for $s = 1, 2, \ldots$ **do**
    Sampling a mini-batch $\tilde{x}_t$.
    Evaluating stochastic gradient $\nabla\tilde{U}(\theta)$ from (5) on mini-batch $\tilde{x}_t$.
    Performing leap-frog updating using SDE in (26).
  end for
end for

We provide the empirical density drawn by naïve stochastic slice sampling and stochastic Laplacian HMC on a synthetic Bayesian linear regression problem with one feature dimension. For each instance i, $y_i \sim \mathcal{N}(betax_i, 1)$. We estimate the posterior of the single parameter β. The synthetic dataset has 100 training samples. We use a minibatch size of 30 for each method, and collect 2,000 Monte Carlo iterations. For stochastic Laplacian HMC we use a stepsize of 0.1 the diffusion parameter A is set to be 7 and the soften parameter is set to be 1. From 7, the stochastic Laplacian HMC can better recover the target distribution.