A PROOF OF THEOREM 1

Proof: Denote the set of all n-dimensional probability vectors by Σ^n, the set of sparse probability vectors by $\mathcal{S}$, and the set of non-sparse (dense) probability vectors by $\mathcal{D} = \Sigma^n \setminus \mathcal{S}$. Denote $B = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n]$, then the optimization problem can be written as

$$\min \sum_{i=1}^{n} p_i \|\boldsymbol{b}_i\|^2$$
$$\text{s.t.} \begin{cases} B\boldsymbol{p} = \boldsymbol{0} \\ \boldsymbol{p}^T \boldsymbol{1}_n = 1 \\ p_i \geq 0, i = 1, 2, \cdots, n \end{cases}$$

Note that the feasible region is always non-empty (take $\boldsymbol{p}$ to be a uniform distribution) and is also closed and bounded, hence this linear programming is always solvable. Denote the set of all minimizers by $\mathcal{M}$. Note that $\mathcal{M}$ depends on $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n$ and is in this sense random.

The Lagrange function is

$$L(\boldsymbol{p}, \boldsymbol{\lambda}, \mu, \boldsymbol{\omega}) = \boldsymbol{p}^T \boldsymbol{s} - \boldsymbol{\lambda}^T B\boldsymbol{p} - \mu(\boldsymbol{p}^T \boldsymbol{1}_n) - \boldsymbol{\omega}^T \boldsymbol{p}$$

where $\boldsymbol{s} = [\|\boldsymbol{b}_1\|^2, \|\boldsymbol{b}_2\|^2, \cdots, \|\boldsymbol{b}_n\|^2]^T$ and $\boldsymbol{\lambda}, \mu, \boldsymbol{\omega}$ are dual variables. The optimality condition reads as

$$\frac{\partial L}{\partial \boldsymbol{p}} = \boldsymbol{s} - B^T \boldsymbol{\lambda} - \mu \boldsymbol{1}_n - \boldsymbol{\omega} = \boldsymbol{0}$$

Dual feasibilty and complementary slackness require

$$\omega_i \leq 0, i = 1, 2, \cdots, n$$
$$\boldsymbol{\omega}^T \boldsymbol{p} = 0$$

Consider the probability of the event {a dense probability vector can solve the above minimization problem}, i.e., $\mathbb{P}(\mathcal{M} \cap \mathcal{D} \neq \emptyset)$. It is upper bounded by

$$\mathbb{P}(\mathcal{M} \cap \mathcal{D} \neq \emptyset) \leq \mathbb{P}(\boldsymbol{p} \in \mathcal{D} \text{ and } \boldsymbol{p} \text{ solves KKT condition})$$

Since $\boldsymbol{p} \in \mathcal{D}$, complementary slackness implies that at least $d + 2$ entries in $\boldsymbol{\omega}$ are zero. Denote the indices of these entries by $\mathcal{J}$. For every $j \in \mathcal{J}$, by optimality condition, we have $s_j - \boldsymbol{\lambda}^T \boldsymbol{b}_j - \mu = 0$, i.e.,

$$\|\boldsymbol{b}_j\|^2 - \boldsymbol{\lambda}^T \boldsymbol{b}_j - \mu = 0$$

Take the first $d + 1$ indices in $\mathcal{J}$, and note a geometric fact that $d + 1$ points in a d-dimensional space must be on the surface of a hypersphere of at most $d - 1$ dimension, which we denote by $\mathcal{S} = S^{q-1} + \boldsymbol{x}$ for some vector $\boldsymbol{x}$ and integer $q \leq d$. Because b_i's distribution is absolutely continuous, we have

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{p} \in \mathcal{D} \text{ and } \boldsymbol{p} \text{ solves KKT condition}) \\
\leq &\mathbb{P}(\boldsymbol{p} \in \mathcal{D} \text{ and } \boldsymbol{b}_j \in S, \forall j \in \mathcal{J}) \\
\leq &\mathbb{P}(\boldsymbol{b}_j \in S, \forall j \in \mathcal{J}) \\
= &\mathbb{P}(\boldsymbol{b}_{j_k} \in S, k = d + 2, \cdots, |\mathcal{J}|) \\
= &\prod_{k=d+2}^{|\mathcal{J}|} \mathbb{P}(\boldsymbol{b}_{j_k} \in S) \qquad \text{(independence)} \\
= &0 \qquad \text{(absolute continuous)}
\end{aligned}$$

Hence $\mathbb{P}(\mathcal{M} \cap \mathcal{D} \neq \emptyset) = 0$ and

$$\begin{aligned}
1 &= \mathbb{P}(\mathcal{M} \neq \emptyset) \\
&= \mathbb{P}((\mathcal{M} \cap \mathcal{S}) \cup (\mathcal{M} \cap \mathcal{D}) \neq \emptyset) \\
&\leq \mathbb{P}(\mathcal{M} \cap \mathcal{S} \neq \emptyset) + \mathbb{P}(\mathcal{M} \cap \mathcal{D} \neq \emptyset) \\
&= \mathbb{P}(\mathcal{M} \cap \mathcal{S} \neq \emptyset)
\end{aligned}$$

Therefore we have

$$\mathbb{P}(\mathcal{M} \cap \mathcal{S} \neq \emptyset) = 1$$

■

B PROOF OF THEOREM 2

Proof: The transition kernel of EM discretization with full gradient can be explicitly written as

$$\begin{aligned}&P^{EM}(\boldsymbol{\theta}_{k+1},\boldsymbol{r}_{k+1}|\boldsymbol{\theta}_k,\boldsymbol{r}_k)\\=&\delta(\boldsymbol{\theta}_{k+1}-(\boldsymbol{\theta}_k+\boldsymbol{r}_k h))\\\times&\Phi\left(\frac{\boldsymbol{r}_{k+1}-\boldsymbol{r}_k+h\gamma\boldsymbol{r}_k+h\nabla V(\boldsymbol{\theta}_k)}{\sigma\sqrt{h}}\right)\frac{1}{\sigma\sqrt{h}}\end{aligned}$$

where $\delta(\cdot)$ is the Dirac delta function and $\Phi(\cdot)$ is the probability density of d-dimensional standard normal distribution.

Denote the unnormalized probability measure of index I_k by

$$\tilde{p}_i=\exp\left\{-\frac{\|\boldsymbol{x}+\sum_{j=1}^n\boldsymbol{a}_j\|^2}{2}+\frac{\|\boldsymbol{x}+n\boldsymbol{a}_i\|^2}{2}\right\}$$

and the normalization constant by

$$\hat{Z}=\sum_{i=1}^n\int\tilde{p}_i d\boldsymbol{r}_{k+1}.$$

Then the transition kernel of EWSG can be written as

$$\begin{aligned}&\tilde{P}^{EM}(\boldsymbol{\theta}_{k+1},\boldsymbol{r}_{k+1}|\boldsymbol{\theta}_k,\boldsymbol{r}_k)\\=&\delta(\boldsymbol{\theta}_{k+1}-(\boldsymbol{\theta}_k+\boldsymbol{r}_k h))\sum_{i=1}^n p_i\Phi\left(\frac{\boldsymbol{r}_{k+1}-\boldsymbol{r}_k+h\gamma\boldsymbol{r}_k+hn\nabla V_i(\boldsymbol{\theta}_k)}{\sigma\sqrt{h}}\right)\frac{1}{\sigma\sqrt{h}}\\=&\delta(\boldsymbol{\theta}_{k+1}-(\boldsymbol{\theta}_k+\boldsymbol{r}_k h))\sum_{i=1}^n\frac{\tilde{p}_i}{\hat{Z}}\Phi\left(\frac{\boldsymbol{r}_{k+1}-\boldsymbol{r}_k+h\gamma\boldsymbol{r}_k+hn\nabla V_i(\boldsymbol{\theta}_k)}{\sigma\sqrt{h}}\right)\frac{1}{\sigma\sqrt{h}}\\=&\frac{1}{\hat{Z}}\delta(\boldsymbol{\theta}_{k+1}-(\boldsymbol{\theta}_k+\boldsymbol{r}_k h))\sum_{i=1}^n\exp\left\{-\frac{\|\boldsymbol{x}+\sum_{j=1}^n\boldsymbol{a}_j\|^2}{2}+\frac{\|\boldsymbol{x}+n\boldsymbol{a}_i\|^2}{2}\right\}\frac{1}{\sqrt{(2\pi)^d}}\exp\left\{-\frac{\|\boldsymbol{x}+n\boldsymbol{a}_i\|^2}{2}\right\}\frac{1}{\sigma\sqrt{h}}\\=&\frac{n}{\hat{Z}\sqrt{(2\pi)^d}}\delta(\boldsymbol{\theta}_{k+1}-(\boldsymbol{\theta}_k+\boldsymbol{r}_k h))\exp\left\{-\frac{\|\boldsymbol{x}+\sum_{j=1}^n\boldsymbol{a}_j\|^2}{2}\right\}\frac{1}{\sigma\sqrt{h}}\end{aligned}$$

Recall the transition kernel of EM integrator with full gradient is

$$\begin{aligned}P^{EM}(\boldsymbol{\theta}_{k+1},\boldsymbol{r}_{k+1}|\boldsymbol{\theta}_k,\boldsymbol{r}_k)=&\delta(\boldsymbol{\theta}_{k+1}-(\boldsymbol{\theta}_k+\boldsymbol{r}_k h))\Phi\left(\frac{\boldsymbol{r}_{k+1}-\boldsymbol{r}_k+h\gamma\boldsymbol{r}_k+h\nabla V(\boldsymbol{\theta}_k)}{\sigma\sqrt{h}}\right)\frac{1}{\sigma\sqrt{h}}\\=&\delta(\boldsymbol{\theta}_{k+1}-(\boldsymbol{\theta}_k+\boldsymbol{r}_k h))\frac{1}{\sqrt{(2\pi)^d}}\exp\left\{-\frac{\|\boldsymbol{x}+\sum_{j=1}^n\boldsymbol{a}_j\|^2}{2}\right\}\frac{1}{\sigma\sqrt{h}}\end{aligned}$$

As both transition kernels are proportional to

$$\delta(\boldsymbol{\theta}_{k+1}-(\boldsymbol{\theta}_k+\boldsymbol{r}_k h))\exp\left\{-\frac{\|\boldsymbol{x}+\sum_{j=1}^n\boldsymbol{a}_j\|^2}{2}\right\}$$

We therefore conclude that

$$P^{EM}(\boldsymbol{\theta}_{k+1},\boldsymbol{r}_{k+1}|\boldsymbol{\theta}_k,\boldsymbol{r}_k)=\tilde{P}^{EM}(\boldsymbol{\theta}_{k+1},\boldsymbol{r}_{k+1}|\boldsymbol{\theta}_k,\boldsymbol{r}_k)$$

■

C PROOF OF THEOREM 3

Proof: Let $\boldsymbol{b}_i = n\nabla V_i$ and assume $\|\boldsymbol{b}_i\|_2 \leq R$ for some constant R. Denote $B = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_n]$. For any probability distribution $\boldsymbol{p}$ over $\{1, \cdots, n\}$, we have

$$\begin{aligned}
&\mathrm{cov}_{I\sim\boldsymbol{p}}[\boldsymbol{b}_I|\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n] \\
=&\sum_{i=1}^n p_i \boldsymbol{b}_i \boldsymbol{b}_i^T - \left(\sum_{i=1}^n p_i \boldsymbol{b}_i\right)\left(\sum_{i=1}^n p_i \boldsymbol{b}_i\right)^T \\
=&\sum_{i=1}^n p_i \boldsymbol{b}_i \boldsymbol{b}_i^T \sum_{i=1}^n p_i - \left(\sum_{i=1}^n p_i \boldsymbol{b}_i\right)\left(\sum_{i=1}^n p_i \boldsymbol{b}_i\right)^T \\
=&\sum_{i<j} (\boldsymbol{b}_i - \boldsymbol{b}_j)(\boldsymbol{b}_i - \boldsymbol{b}_j)^T p_i p_j
\end{aligned}$$

Therefore we let

$$\begin{aligned}
f(B) :=& \mathrm{Tr}\left[\sum_{i<j} (\boldsymbol{b}_i - \boldsymbol{b}_j)(\boldsymbol{b}_i - \boldsymbol{b}_j)^T p_i p_j - \sum_{i<j} (\boldsymbol{b}_i - \boldsymbol{b}_j)(\boldsymbol{b}_i - \boldsymbol{b}_j)^T \frac{1}{n^2}\right] \\
=& \sum_{i<j} \|\boldsymbol{b}_i - \boldsymbol{b}_j\|^2 p_i p_j - \sum_{i<j} \|\boldsymbol{b}_i - \boldsymbol{b}_j\|^2 \frac{1}{n^2} \qquad (\mathrm{Tr}[AB] = \mathrm{Tr}[BA])
\end{aligned}$$

and use it to compare the trace of covariance matrix of uniform- and nonuniform- subsamplings.

First of all,

$$\begin{aligned}
&\mathbb{E}[f(B)] \\
=&\mathbb{E}[\|\boldsymbol{b}_i - \boldsymbol{b}_j\|^2] \sum_{i<j} \left(p_i p_j - \frac{1}{n^2}\right) \\
=&\mathbb{E}[\|\boldsymbol{b}_i - \boldsymbol{b}_j\|^2] \left(\sum_{i<j} p_i p_j - \frac{n-1}{2n}\right) \\
=&\mathbb{E}[\|\boldsymbol{b}_i - \boldsymbol{b}_j\|^2] \left(\frac{1 - \sum_{i=1}^n p_i^2}{2} - \frac{n-1}{2n}\right) \\
\leq&\mathbb{E}[\|\boldsymbol{b}_i - \boldsymbol{b}_j\|^2] \left(\frac{1 - \frac{1}{n}}{2} - \frac{n-1}{2n}\right) \\
=&0
\end{aligned}$$

where the inequality is due to Cauchy-Schwarz and it is a strict inequality unless all p_i's are equal, which means uniform subsampling on average has larger variablity than a non-uniform scheme measured by the trace of covariance matrix.

Moreover, concentration inequality can help show $f(B)$ is negative with high probability if h is small. To this end, plug $\boldsymbol{x} = \mathcal{O}(\sqrt{h})$ in and rewrite

$$p_i = \frac{1}{Z} \exp\left\{Fh\left[\frac{\|\boldsymbol{y} + \frac{1}{n}\sum_{i=1}^n \boldsymbol{b}_i\|^2}{2} - \frac{\|\boldsymbol{y} + \boldsymbol{b}_i\|^2}{2}\right]\right\}$$

where $\boldsymbol{y} = \frac{\sigma}{\sqrt{h}}\boldsymbol{x} = \mathcal{O}(1)$, $F = -\frac{1}{\sigma^2}$ and Z is the normalization constant. Denote the unnormalized probability by

$$\widetilde{p}_i = \exp\left\{Fh\left[\frac{\|\boldsymbol{y} + \frac{1}{n}\sum_{i=1}^n \boldsymbol{b}_i\|^2}{2} - \frac{\|\boldsymbol{y} + \boldsymbol{b}_i\|^2}{2}\right]\right\}$$

and we have

$$f(B) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{b}_i - \boldsymbol{b}_j\|^2\left(p_i p_j - \frac{1}{n^2}\right)$$
$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{b}_i - \boldsymbol{b}_j\|^2\frac{\widetilde{p}_i\widetilde{p}_j}{[\sum_{k=1}^{n}\widetilde{p}_k]^2} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{b}_i - \boldsymbol{b}_j\|^2\frac{1}{n^2}$$

To prove concentration results, it is useful to estimate

$$C_i = \sup_{\substack{\boldsymbol{b}_1,\cdots,\boldsymbol{b}_n\in B(\mathbf{0},R)\\ \widehat{\boldsymbol{b}}_i\in B(\mathbf{0},R)}} |f(\boldsymbol{b}_1,\cdots,\boldsymbol{b}_i,\cdots,\boldsymbol{b}_n)$$
$$-f(\boldsymbol{b}_1,\cdots,\widehat{\boldsymbol{b}}_i,\cdots,\boldsymbol{b}_n)|$$

where $B(\mathbf{0},R)$ is a ball centered at origin with radius R in $\mathbb{R}^d$.

Due to the mean value theorem, we have $C_i \le 2R\sup|\frac{\partial f}{\partial \boldsymbol{b}_i}|$. By symmetry, it suffices to compute $\sup|\frac{\partial f}{\partial \boldsymbol{b}_1}|$ to upper bound C_1. Note that

$$\frac{\partial \widetilde{p}_j}{\partial \boldsymbol{b}_1} = 2\widetilde{p}_j Fh[\frac{1}{n}(\boldsymbol{y} + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{b}_i) - (\boldsymbol{y}+\boldsymbol{b}_j)\delta_{1j}] = \mathcal{O}(h)\widetilde{p}_j$$

where δ_{1j} is the Kronecker delta function. Thus

$$\frac{\partial f}{\partial \boldsymbol{b}_1} = \sum_{j=1}(\boldsymbol{b}_1-\boldsymbol{b}_j)\frac{\widetilde{p}_1\widetilde{p}_j}{[\sum_{k=1}^{n}\widetilde{p}_k]^2} - \sum_{j=1}^{n}(\boldsymbol{b}_1-\boldsymbol{b}_j)\frac{1}{n^2} + \sum_{i,j=1}^{n}\|\boldsymbol{b}_1-\boldsymbol{b}_j\|^2\frac{\mathcal{O}(h)\widetilde{p}_i\widetilde{p}_j}{[\sum_{k=1}^{n}\widetilde{p}_k]^2}$$
$$-2\sum_{i,j=1}^{n}\|\boldsymbol{b}_1-\boldsymbol{b}_j\|^2\frac{\widetilde{p}_i\widetilde{p}_j}{[\sum_{k=1}^{n}\widetilde{p}_k]^3}\sum_{k=1}^{n}\widetilde{p}_k\mathcal{O}(h)$$
$$= \widetilde{p}_1\sum_{j=1}^{n}(\boldsymbol{b}_1-\boldsymbol{b}_j)\frac{\widetilde{p}_j}{[\sum_{k=1}^{n}\widetilde{p}_k]^2} - \sum_{j=1}^{n}(\boldsymbol{b}_1-\boldsymbol{b}_j)\frac{1}{n^2} + \frac{\mathcal{O}(n^2)\mathcal{O}(h)}{\mathcal{O}(n^2)} + \frac{\mathcal{O}(n^2)}{\mathcal{O}(n^3)}\mathcal{O}(n)\mathcal{O}(h)$$
$$= \mathcal{O}(\frac{h}{n}) + \mathcal{O}(h) + \mathcal{O}(h)$$
$$= \mathcal{O}(h)$$

where $\mathcal{O}(\frac{h}{n})$ in the 2nd last equation comes from the difference of the first two terms in the 3rd last equation. This estimation shows that $C_i \le 2R\mathcal{O}(h) = \mathcal{O}(h)$.

Therefore, by McDiarmid's inequality, we conclude for any $\epsilon > 0$,

$$\mathbb{P}(|f - \mathbb{E}[f]| > \epsilon) \le 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n}C_i^2}\right) = 2\exp\left(\frac{-2\epsilon^2}{n\mathcal{O}(h^2)}\right).$$

Any choice of $h(n) = o(n^{-1/2})$ will render this probability asymptotically vanishing as n grows, which means that f will be negative with high probability, which is equivalent to reduced variance per step. ■

D PROOF OF THEOREM 4

Proof: We rewrite the generator of underdamped Langevin with full gradient as

$$\mathcal{L}f(\boldsymbol{X}) = \boldsymbol{F}(\boldsymbol{X})^T\begin{bmatrix}\nabla_{\boldsymbol{\theta}}f(\boldsymbol{X})\\ \nabla_{\boldsymbol{r}}f(\boldsymbol{X})\end{bmatrix} + \frac{1}{2}A : \nabla\nabla f(\boldsymbol{X})$$

where

$$\boldsymbol{F}(\boldsymbol{X}) = \begin{bmatrix}\boldsymbol{r}\\ -\gamma\boldsymbol{r} - \nabla V(\boldsymbol{\theta})\end{bmatrix}, \quad A = GG^T \text{and} \quad G = \begin{bmatrix}O_{d\times d} & O_{d\times d}\\ O_{d\times d} & \sqrt{2\gamma}I_{d\times d}\end{bmatrix}$$

Rewrite the discretized underdamped Langevin with stochastic gradient in variable $\boldsymbol{X}$

$$\boldsymbol{X}_{k+1}^E - \boldsymbol{X}_k^E = h\boldsymbol{F}_k(\boldsymbol{X}_k^E) + \sqrt{h}G_k\boldsymbol{\eta}_{k+1}$$

where

$$\boldsymbol{F}_k(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{r} \\ -\gamma\boldsymbol{r} - n\nabla V_{I_k}(\boldsymbol{\theta}) \end{bmatrix}, \qquad G_k = G = \begin{bmatrix} O_{d\times d} & O_{d\times d} \\ O_{d\times d} & \sqrt{2\gamma}I_{d\times d} \end{bmatrix}$$

and $\boldsymbol{\eta}_{k+1}$ is a $2d$ dimensional standard Gaussian random vector. Note that this representation include both SGHMC and EWSG, for SGHMC I_k follows uniform distribution and for EWSG, I_k follows the MCMC-approximated exponentially weighted distribution.

Denote the generator associated with stochastic gradient underdamped Langevin at the k-th iteration by

$$\mathcal{L}_k f(\boldsymbol{X}) = \boldsymbol{F}_k(\boldsymbol{X})^T \begin{bmatrix} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{X}) \\ \nabla_{\boldsymbol{r}} f(\boldsymbol{X}) \end{bmatrix} + \frac{1}{2}A : \nabla\nabla f(\boldsymbol{X})$$

and the difference of the generators of full gradient and stochastic gradient underdamped Langevin at k-th interation is denoted by

$$\Delta\mathcal{L}_k f(\boldsymbol{X}) = (\mathcal{L}_k - \mathcal{L})f(\boldsymbol{X}) = (\boldsymbol{F}_k(\boldsymbol{X}) - \boldsymbol{F}(\boldsymbol{X}))^T \begin{bmatrix} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{X}) \\ \nabla_{\boldsymbol{r}} f(\boldsymbol{X}) \end{bmatrix} = \langle \nabla V(\boldsymbol{\theta}) - n\nabla V_{I_k}(\boldsymbol{\theta}), \nabla_{\boldsymbol{r}} f(\boldsymbol{X})\rangle$$

For brevity, we write $\phi_k = \phi(\boldsymbol{X}_k^E)$, $\boldsymbol{F}_k^E = \boldsymbol{F}_k(\boldsymbol{X}_k^E)$, $\psi_k = \psi(\boldsymbol{X}_k^E)$ and $D^l\phi_k = (D^l\psi)(\boldsymbol{X}_k^E)$ where $(D^l\psi)(z)$ is the l-th order derivative. We write $(D^l\psi)[\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_l]$ for derivative evaluated in the direction $\boldsymbol{s}_j, j = 1, 2, \cdots, l$. Define

$$\boldsymbol{\delta}_k = \boldsymbol{X}_{k+1}^E - \boldsymbol{X}_k^E = h\boldsymbol{F}_k^E + \sqrt{h}G_k\boldsymbol{\eta}_{k+1}$$

Under the assumptions of Theorem 4, we show that the vector field F_k^E also has bounded momentum up to p-th order.

Lemma 5 *Under the assumption of Theorem 4, there exists a constant M such that up to $\frac{p}{2}$-th order moments of random vector field $\boldsymbol{F}_k^E$ are bounded*

$$\mathbb{E}\|\boldsymbol{F}_k^E\|_2^j \le M,\ \forall j = 0, 1, 2, \cdots, \frac{p}{2},\ \forall k = 0, 1, 2\cdots,$$

Proof: It suffices to bound the highest moment, as all other lower order moments are bounded by the highest one by Holder's inequality.

First notice that

$$\|\boldsymbol{F}_k^E\|_2 = \left\| \begin{bmatrix} \boldsymbol{r}_k^E \\ -\gamma\boldsymbol{r}_k^E - \nabla V_{I_k}(\boldsymbol{\theta}_k^E) \end{bmatrix} \right\|_2 \le \sqrt{1+\gamma^2}\|\boldsymbol{r}_k^E\|_2 + \|\nabla V_{I_k}(\theta_k^E)\|_2$$

Hence

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{F}_k^E\|_2^{\frac{p}{2}} \le& \mathbb{E}\left(\sqrt{1+\gamma^2}\|\boldsymbol{r}_k^E\|_2 + \|\nabla V_{I_k}(\theta_k^E)\|_2\right)^{\frac{p}{2}} \\
=& \mathbb{E}\left\{\sum_{i=0}^{\frac{p}{2}} \binom{\frac{p}{2}}{i} \|\boldsymbol{r}_k^E\|_2^i \|\nabla V_{I_k}(\theta_k^E)\|_2^{\frac{p}{2}-i}\right\} \\
=& \sum_{i=0}^{\frac{p}{2}} \binom{\frac{p}{2}}{i} \mathbb{E}\left[\|\nabla V_{I_k}(\theta_k^E)\|_2^{\frac{p}{2}-i}\|\boldsymbol{r}_k^E\|_2^i\right] \\
\le& \sum_{i=0}^{\frac{p}{2}} \binom{\frac{p}{2}}{i} \sqrt{\mathbb{E}\left[\|\nabla V_{I_k}(\theta_k^E)\|_2^{p-2i}\right]}\sqrt{\mathbb{E}\left[\|\boldsymbol{r}_k^E\|_2^{2i}\right]} \quad \text{(Cauchy-Schwarz inequality)}
\end{aligned}$$

By assumption, we know each $\mathbb{E}\left[\|\nabla V_{I_k}(\theta_k^E)\|_2^l\right], \mathbb{E}\|\boldsymbol{r}_k^E\|_2^l, l = 0, 1, \cdots, p$ is bounded, so we conclude there exists a constant $M > 0$ that bounds the $\frac{p}{2}$-th order moment of $\boldsymbol{F}_k^E, \forall k = 0, 1, \cdots,$ ■

Using Taylor's expansion for ψ, we have

$$\psi_{k+1} = \psi_k + D\psi_k[\boldsymbol{\delta}_k] + \frac{1}{2}D^2\psi_k[\boldsymbol{\delta}_k, \boldsymbol{\delta}_k] + \frac{1}{6}D^3\psi_k[\boldsymbol{\delta}_k, \boldsymbol{\delta}_k, \boldsymbol{\delta}_k] + R_{k+1}$$

where

$$R_{k+1} = \left(\frac{1}{6}\int_0^1 s^3 D^4\psi(s\boldsymbol{X}_k^E + (1-s)\boldsymbol{X}_{k+1}^E)ds\right)[\boldsymbol{\delta}_k, \boldsymbol{\delta}_k, \boldsymbol{\delta}_k, \boldsymbol{\delta}_k]$$

is the remainder term. Therefore, we have

$$\begin{aligned}\psi_{k+1} =& \psi_k + h\mathcal{L}_k\psi_k + h^{\frac{1}{2}}D\psi_k[G_k\boldsymbol{\eta}_{k+1}] + h^{\frac{3}{2}}D^2\psi_k[\boldsymbol{F}_k^E, G_k\boldsymbol{\eta}_{k+1}] \\ &+\frac{1}{2}h^2D^2\psi_k[\boldsymbol{F}_k^E, \boldsymbol{F}_k^E] + \frac{1}{6}D^3\psi_k[\boldsymbol{\delta}_k, \boldsymbol{\delta}_k, \boldsymbol{\delta}_k] + r_{k+1} + R_{k+1}\end{aligned} \tag{7}$$

where

$$r_{k+1} = \frac{h}{2}\left(D^2\psi_k[G_k\boldsymbol{\eta}_{k+1}, G_k\boldsymbol{\eta}_{k+1}] - A : \nabla\nabla\psi_k\right)$$

Summing Equation (7) ove the first K terms, dividing by Kh and use Poisson equation, we have

$$\frac{1}{Kh}(\psi_K - \psi_0) = \frac{1}{K}\sum_{k=0}^{K-1}(\phi_k - \bar{\phi}) + \frac{1}{K}\sum_{k=0}^{K-1}\Delta\mathcal{L}_k\psi_k + \frac{1}{Kh}\sum_{i=1}^{3}(M_{i,K} + S_{i,K}), \tag{8}$$

where

$$M_{1,K} = \sum_{k=0}^{K-1} r_{k+1},\ M_{2,K} = h^{\frac{1}{2}}\sum_{k=0}^{K-1} D\psi_k[G_k\boldsymbol{\eta}_{k+1}],\ M_{3,K} = h^{\frac{3}{2}}\sum_{k=0}^{K-1} D^2\psi_k[\boldsymbol{F}_k^E, G_k\boldsymbol{\eta}_{k+1}],$$
$$S_{1,K} = \frac{h^2}{2}\sum_{k=0}^{K-1} D^2\psi_k[\boldsymbol{F}_k^E, \boldsymbol{F}_k^E],\ S_{2,K} = \sum_{k=0}^{K-1} R_{k+1},\ S_{3,K} = \frac{1}{6}\sum_{k=0}^{K-1} D^3\psi_k[\boldsymbol{\delta}_k, \boldsymbol{\delta}_k, \boldsymbol{\delta}_k]$$

Furthermore, it will be convenient to decompose

$$S_{3,K} = M_{0,K} + S_{0,K}$$

where

$$\begin{aligned}S_{0,K} =& h^2\sum_{k=0}^{K-1}\left(hD^3\psi_k[\boldsymbol{F}_k^E, \boldsymbol{F}_k^E, \boldsymbol{F}_k^E] + 3D^3\psi_k[\boldsymbol{F}_k^E, G_k\boldsymbol{\eta}_{k+1}, G_k\boldsymbol{\eta}_{k+1}]\right) \\ M_{0,K} =& h^{\frac{3}{2}}\sum_{k=0}^{K-1}\left(D^3\psi_k[G_k\boldsymbol{\eta}_{k+1}, G_k\boldsymbol{\eta}_{k+1}, G_k\boldsymbol{\eta}_{k+1}] + 3hD^3\psi_k[\boldsymbol{F}_k^E, \boldsymbol{F}_k^E, G_k\boldsymbol{\eta}_{k+1}]\right)\end{aligned}$$

Rearrange terms in Equation (7), square on both sides, use Cauchy-Schwarz inequality and take expectation, we have

$$\begin{aligned}\mathbb{E}\left(\hat{\phi}_K - \bar{\phi}\right)^2 \leq& C_1\left[\mathbb{E}\frac{(\psi_K - \psi_0)^2}{(Kh)^2} + \frac{1}{K^2}\mathbb{E}\left(\sum_{k=0}^{K-1}(\Delta\mathcal{L}_k\psi_k)\right)^2 + \frac{1}{(Kh)^2}\sum_{i=0}^{2}\mathbb{E}S_{i,K}^2 + \frac{1}{(Kh)^2}\sum_{i=0}^{3}\mathbb{E}M_{i,K}^2\right] \\ =& C_1\left[\mathbb{E}\frac{(\psi_K - \psi_0)^2}{T^2} + \frac{1}{K^2}\mathbb{E}\left(\sum_{k=0}^{K-1}(\Delta\mathcal{L}_k\psi_k)\right)^2 + \frac{1}{T^2}\sum_{i=0}^{2}\mathbb{E}S_{i,K}^2 + \frac{1}{T^2}\sum_{i=0}^{3}\mathbb{E}M_{i,K}^2\right]\end{aligned} \tag{9}$$

where $T = kh$, the corresponding time of the underlying continuous dynamics.

We now show how each term is bounded. By boundedness of ψ, we have

$$\mathbb{E}\frac{(\psi_K - \psi_0)^2}{T^2} \leq \frac{4\|\psi\|_\infty^2}{T^2} = \mathcal{O}(\frac{1}{T^2})$$

The second term $\frac{1}{K^2}\mathbb{E}\big(\sum_{k=0}^{K-1}(\Delta\mathcal{L}_k\psi_k)\big)^2$ is critical in showing the advantage of EWSG, and we will show how to derive its bound in detail later.

The technique we use to bound $\frac{1}{T^2}\mathbb{E}S_{i,K}^2, i=0,1,2$ are all similar, we will first show an upper bound for $|S_{i,K}|$ in terms of powers of $\|\boldsymbol{F}_k^E\|$, then take square and expectation, and finally expand squares and use Lemma 5 extensively to derive bounds. As a concrete example, we will show how to bound $\frac{1}{T^2}\mathbb{E}S_{0,K}^2$. Other bounds follow in a similar fashion and details are omitted.

To bound the term containing $S_{0,K}$, we first note that

$$\begin{aligned}|S_{0,K}| \leq& h^2\sum_{k=0}^{K-1}\big(h|D^3\psi_k[\boldsymbol{F}_k^E,\boldsymbol{F}_k^E,\boldsymbol{F}_k^E]|+3|D^3\psi_k[\boldsymbol{F}_k^E,G_k\boldsymbol{\eta}_{k+1},G_k\boldsymbol{\eta}_{k+1}]|\big)\\ \leq& h^2\|D^3\psi\|_\infty\sum_{k=0}^{K-1}\big(h\|\boldsymbol{F}_k^E\|_2^3+3\|\boldsymbol{F}_k^E\|_2\|G_k\boldsymbol{\eta}_{k+1}\|_2^2\big)\end{aligned}$$

Square both sides of the above inequality and take expectation, we obtain

$$\begin{aligned}&\frac{1}{T^2}\mathbb{E}|S_{0,K}|^2 \qquad (10)\\ \leq&\frac{h^4}{T^2}\|D^3\psi\|_\infty^2\mathbb{E}\big(\sum_{k=0}^{K-1}h\|\boldsymbol{F}_k^E\|_2^3+3\|\boldsymbol{F}_k^E\|_2\|G_k\boldsymbol{\eta}_{k+1}\|_2^2\big)^2\\ \leq&\frac{h^4}{T^2}\|D^3\psi\|_\infty^2K\sum_{k=0}^{K-1}\mathbb{E}(h\|\boldsymbol{F}_k^E\|_2^3+3\|\boldsymbol{F}_k^E\|_2\|G_k\boldsymbol{\eta}_{k+1}\|_2^2)^2 \quad \text{(Cauchy-Schwarz inequality)}\\ =&\frac{h^4}{T^2}\|D^3\psi\|_\infty^2K\sum_{k=0}^{K-1}\mathbb{E}[h^2\|\boldsymbol{F}_k^E\|_2^6+6\|\boldsymbol{F}_k^E\|_2^4\|G_k\boldsymbol{\eta}_{k+1}\|_2^2+9\|\boldsymbol{F}_k^E\|_2^2\|G_k\boldsymbol{\eta}_{k+1}\|_4^2]\\ =&\frac{h^4}{T^2}\|D^3\psi\|_\infty^2K\sum_{k=0}^{K-1}h^2\mathbb{E}\|\boldsymbol{F}_k^E\|_2^6+6\mathbb{E}\|\boldsymbol{F}_k^E\|_2^4\mathbb{E}\|G_k\boldsymbol{\eta}_{k+1}\|_2^2+9\mathbb{E}\|\boldsymbol{F}_k^E\|_2^2\mathbb{E}\|G_k\boldsymbol{\eta}_{k+1}\|_4^2\\ =&\frac{1}{T^2}\mathcal{O}(K^2h^4)\\ =&\mathcal{O}(h^2)\end{aligned}$$

To bound the term containing $S_{1,K}$ and $S_{2,K}$, we have

$$\begin{aligned}|S_{1,K}| \leq&\frac{h^2}{2}\sum_{k=0}^{K-1}\|D^2\psi\|_\infty\|\boldsymbol{F}_k^E\|_2^2\\ |S_{2,K}| \leq&\frac{1}{24}\|D^4\psi\|_\infty\sum_{k=0}^{K-1}\|\boldsymbol{\delta}_k\|_2^4 \leq \frac{1}{24}h^2\|D^4\psi\|_\infty\sum_{k=0}^{K-1}\|\sqrt{h}\boldsymbol{F}_k^E+G_k\boldsymbol{\eta}_{k+1}\|_2^4\end{aligned}$$

Then we can obtain the following bound in a similar fashion as in Equation (10)

$$\begin{aligned}\frac{1}{T^2}\mathbb{E}S_{1,K}^2 =&\mathcal{O}(h^2)\\ \frac{1}{T^2}\mathbb{E}S_{2,K}^2 =&\mathcal{O}(h^2)\end{aligned}$$

Now we will use martingale argument to bound $\frac{1}{T^2}\mathbb{E}M_{i,K}^2, i=0,1,2,3$. There are two injected randomness at k-th iteration, the Gaussian noise $\boldsymbol{\eta}_{k+1}$ and the stochastic gradient term determined by the stochastic index I_k. Denote the sigma algebra at k-th iteration by $\mathcal{F}_k$. For both SGHMC and EWSG we have

$$\boldsymbol{\eta}_{k+1}\perp\mathcal{F}_k \text{ and } I_k\perp\boldsymbol{\eta}_{k+1}$$

hence

$$\begin{aligned}\mathbb{E}[\boldsymbol{\eta}_{k+1}|\mathcal{F}_k] =&\mathbf{0}\\ \mathbb{E}[D^3\psi_k[G_k\boldsymbol{\eta}_{k+1}, G_k\boldsymbol{\eta}_{k+1}, G_k\boldsymbol{\eta}_{k+1}]|\mathcal{F}_k] =&0\\ \mathbb{E}[D^2\psi_k[\boldsymbol{F}_k^E, G_k\boldsymbol{\eta}_{k+1}]|\mathcal{F}_k] =&0\\ \mathbb{E}[D^3\psi_k[\boldsymbol{F}_k^E, \boldsymbol{F}_k^E, G_k\boldsymbol{\eta}_{k+1}]|\mathcal{F}_k] =&0\end{aligned}$$

Therefore, it is clear that $M_{i,K}, i = 0, 1, 2, 3$ are all martingales. Due to martingale properties, we have

$$\frac{1}{T^2}\mathbb{E}M_{0,K}^2 = \frac{h^3}{T^2}\sum_{k=0}^{K-1}\mathbb{E}\big(D^3\psi_k[G_k\boldsymbol{\eta}_{k+1}, G_k\boldsymbol{\eta}_{k+1}, G_k\boldsymbol{\eta}_{k+1}]+3hD^3\psi_k[\boldsymbol{F}_k^E, \boldsymbol{F}_k^E, G_k\boldsymbol{\eta}_{k+1}]\big)^2 = \frac{1}{T^2}\mathcal{O}(h^3K) = \mathcal{O}(\frac{h^2}{T})$$

$$\frac{1}{T^2}\mathbb{E}M_{1,K}^2 = \frac{1}{T^2}\sum_{k=0}^{K-1}\mathbb{E}r_{k+1}^2 = \frac{1}{T^2}\mathcal{O}(h^2K) = \mathcal{O}(\frac{h}{T})$$

$$\frac{1}{T^2}\mathbb{E}M_{2,K}^2 = \frac{h}{T^2}\sum_{k=0}^{K-1}\mathbb{E}(D\psi_k[G_k\boldsymbol{\eta}_{k+1}])^2 = \frac{1}{T^2}\mathcal{O}(hK) = \mathcal{O}(\frac{1}{T})$$

$$\frac{1}{T^2}\mathbb{E}M_{3,K}^2 = \frac{1}{T^2}h^3\sum_{k=0}^{K-1}\mathbb{E}(D^2\psi_k[\boldsymbol{F}_k^E, G_k\boldsymbol{\eta}_{k+1}])^2 = \frac{1}{T^2}\mathcal{O}(h^3K) = \mathcal{O}(\frac{h^2}{T})$$

We now collect all bounds derived so far and obtain

$$\begin{aligned}\mathbb{E}\big(\hat{\phi}_K - \bar{\phi}\big)^2 \leq& C_1\left[\mathcal{O}(\frac{1}{T^2}) + \frac{1}{K^2}\mathbb{E}\left(\sum_{k=0}^{K-1}(\Delta\mathcal{L}_k\psi_k)\right)^2 + \mathcal{O}(h^2) + \mathcal{O}(\frac{h}{T}) + \mathcal{O}(\frac{1}{T}) + \mathcal{O}(\frac{h^2}{T})\right]\\ \leq& C_1\left[\mathcal{O}(\frac{1}{T}) + \frac{1}{K^2}\mathbb{E}\left(\sum_{k=0}^{K-1}(\Delta\mathcal{L}_k\psi_k)\right)^2 + \mathcal{O}(h^2)\right]\\ \leq& C_2\left[\frac{1}{T} + \frac{1}{K^2}\mathbb{E}\left(\sum_{k=0}^{K-1}(\Delta\mathcal{L}_k\psi_k)\right)^2 + h^2\right]\end{aligned} \tag{11}$$

where $C_2 > 0$ is a constant. In the above inequality, we use $\frac{1}{T^2} < \frac{1}{T}$ and $\frac{h}{T} \leq \frac{1}{T}, \frac{h^2}{T} \leq \frac{1}{T}$ as typically we assume $T \gg 1$ and $h \ll 1$ in non-asymptotic analysis.

Now we focus on the remaining term $\frac{1}{K^2}\mathbb{E}\big(\sum_{k=0}^{K-1}\Delta\mathcal{L}_k\psi_k\big)^2$. For SGHMC, we have that $\mathbb{E}[\Delta\mathcal{L}_k\psi_k|\mathcal{F}_k] = 0$, hence $\sum_{k=0}^{K-1}\Delta\mathcal{L}_k\psi_k$ is a martingale. By martingale property, we have

$$\frac{1}{K^2}\mathbb{E}\left(\sum_{k=0}^{K-1}\Delta\mathcal{L}_k\psi_k\right)^2 = \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2$$

For EWSG, $\sum_{k=0}^{K-1}\Delta\mathcal{L}_k\psi_k$ is no longer a martingale, but we still have the following

$$\begin{aligned}\frac{1}{K^2}\mathbb{E}\left(\sum_{k=0}^{K-1}\Delta\mathcal{L}_k\psi_k\right)^2 =&\frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2 + \frac{2}{K^2}\sum_{i<j}\mathbb{E}(\Delta\mathcal{L}_i\psi_i)(\Delta\mathcal{L}_j\psi_j)\\ =&\frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2 + \frac{2}{K^2}\sum_{i<j}\mathbb{E}[(\Delta\mathcal{L}_i\psi_i)\mathbb{E}[\Delta\mathcal{L}_j\psi_j|\mathcal{F}_j]]\end{aligned} \tag{12}$$

For the term $\mathbb{E}[\Delta\mathcal{L}_j\psi_j|\mathcal{F}_j]$, we have

$$\mathbb{E}[\Delta\mathcal{L}_j\psi_j|\mathcal{F}_j] = \mathbb{E}[\langle\nabla V(\boldsymbol{\theta}_j^E) - n\nabla V_{I_j}(\boldsymbol{\theta}_j^E), \nabla_{\boldsymbol{r}}\psi_j\rangle|\mathcal{F}_j] = \langle\mathbb{E}[\nabla V(\boldsymbol{\theta}_j^E) - n\nabla V_{I_j}(\boldsymbol{\theta}_j^E)|\mathcal{F}_j], \nabla_{\boldsymbol{r}}\psi_j\rangle$$

as $\psi_j \in \mathcal{F}_j$. Then by Cauchy-Schwarz inequality, boundedness of ψ and the fact $\|\nabla V(\boldsymbol{\theta}_j^E) - \mathbb{E}[n\nabla V_{I_j}(\boldsymbol{\theta}_j^E)|\mathcal{F}_j]\|_2 = \mathcal{O}(h)$ as shown in the proof of Theorem 3, we conclude $\mathbb{E}[\Delta\mathcal{L}_j\psi_j|\mathcal{F}_j] = \mathcal{O}(h)$.

Now plug the above result in Equation (12), we have

$$\begin{aligned}
\frac{1}{K^2}\mathbb{E}\left(\sum_{k=0}^{K-1}\Delta\mathcal{L}_k\psi_k\right)^2 =& \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2 + \frac{2}{K^2}\sum_{i<j}\mathbb{E}[(\Delta\mathcal{L}_i\psi_i)\mathbb{E}[\Delta\mathcal{L}_j\psi_j|\mathcal{F}_j]] \\
=& \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2 + \frac{2}{K^2}\sum_{i<j}\mathbb{E}[\Delta\mathcal{L}_i\psi_i]\mathcal{O}(h) \\
=& \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2 + \frac{2}{K^2}\sum_{i<j}\mathcal{O}(h^2) \\
=& \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2 + \frac{2}{K^2}\sum_{i<j}\mathcal{O}(h^2) \\
=& \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2 + \mathcal{O}(h^2)
\end{aligned}$$

Combine both cases of SGHMC and EWSG, we obtain

$$\frac{1}{K^2}\mathbb{E}\left(\sum_{k=0}^{K-1}\Delta\mathcal{L}_k\psi_k\right)^2 = \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2 + \mathcal{O}(h^2)$$

Note that $\mathcal{O}(h^2)$ term will later be combined with other error terms with the same order.

The final piece is to bound $\frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2$, and we have

$$\begin{aligned}
\frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}(\Delta\mathcal{L}_k\psi_k)^2 &= \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}\langle\nabla V(\boldsymbol{\theta}_k^E) - n\nabla V_{I_k}(\boldsymbol{\theta}_k^E), \nabla_{\boldsymbol{r}}\psi_k\rangle^2 \\
&\le \frac{1}{K^2}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla V(\boldsymbol{\theta}_k^E) - n\nabla V_{I_k}(\boldsymbol{\theta}_k^E)\|_2^2 \cdot \|\nabla_{\boldsymbol{r}}\psi_k\|_2^2] \quad \text{(Cauchy-Schwarz inequality)} \\
&\le \frac{M_3^2}{K^2}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla V(\boldsymbol{\theta}_k^E) - n\nabla V_{I_k}(\boldsymbol{\theta}_k^E)\|_2^2] \\
&= \frac{M_3^2}{K^2}\sum_{k=0}^{K-1}\mathbb{E}[\mathbb{E}[\|\nabla V(\boldsymbol{\theta}_k^E) - n\nabla V_{I_k}(\boldsymbol{\theta}_k^E)\|_2^2 \,|\, \mathcal{F}_k]] \\
&\le \frac{2M_3^2}{K^2}\sum_{k=0}^{K-1}\mathbb{E}[\underbrace{\mathbb{E}[\,\|\nabla V(\boldsymbol{\theta}_k^E) - \mathbb{E}[n\nabla V_{I_k}(\boldsymbol{\theta}_k^E)\,|\,\mathcal{F}_k]\|^2\,|\,\mathcal{F}_k]}_{Q_1}] \\
&+ \underbrace{\mathbb{E}[\,\|\mathbb{E}[n\nabla V_{I_k}(\boldsymbol{\theta}_k^E)\,|\,\mathcal{F}_k] - n\nabla V_{I_k}(\boldsymbol{\theta}_k^E)\|_2^2\,|\,\mathcal{F}_k]}_{Q_2}
\end{aligned}$$

The term Q_1 captures the bias of stochastic gradient. For SGHMC, uniform gradient subsamping leads to an unbiased gradient estimator, so $Q_1 = 0$ for SGHMC. For EWSG, same as in the proof of Theorem 2, we have that

$$\mathbb{E}\left[\,\|\nabla V(\boldsymbol{\theta}_k^E) - \mathbb{E}[n\nabla V_{I_k}(\boldsymbol{\theta}_k^E)\,|\,\mathcal{F}_k]\|^2\,|\,\mathcal{F}_k\right] = \mathcal{O}(h^2)$$

Combining two cases, we have

$$Q_1 = \mathcal{O}(h^2)$$

For a random vector $\boldsymbol{v}$ with mean $\mathbb{E}[\boldsymbol{v}] = \mathbf{0}$, we have

$$\mathbb{E}[\|\boldsymbol{v}\|^2] = \mathbb{E}\left[\mathrm{Tr}[\boldsymbol{v}\boldsymbol{v}^T]\right] = \mathrm{Tr}\left[\mathbb{E}[\boldsymbol{v}\boldsymbol{v}^T]\right] = \mathrm{Tr}\left[\mathrm{cov}(\boldsymbol{v})\right]$$

where $\mathrm{cov}(\boldsymbol{v})$ is the covariance matrix of random vector $\boldsymbol{v}$. Therefore, we have that

$$Q_2 = \mathrm{Tr}\left[\mathrm{cov}(n\nabla V_{I_k}|\mathcal{F}_k)\right],$$

i.e., Q_2 is the trace of the covariance matrix of stochastic gradient estimate conditioned on current filtration $\mathcal{F}_k$.

Combining Q_1 and Q_2, we have that

$$\begin{aligned}\frac{1}{K^2}\mathbb{E}\left(\sum_{k=0}^{K-1}\Delta\mathcal{L}_k\psi_k\right)^2 \leq& \frac{2M_3^2}{K^2}\sum_{k=0}^{K-1}\left[\mathbb{E}[\mathrm{Tr}[\mathrm{cov}(n\nabla V_{I_k}|\mathcal{F}_k)]] + \mathcal{O}(h^2)\right] \\ =& \frac{2M_3^2 h}{T}\frac{\sum_{k=0}^{K-1}\mathbb{E}[\mathrm{Tr}[\mathrm{cov}(n\nabla V_{I_k}|\mathcal{F}_k)]]}{K} + \mathcal{O}(\frac{h^3}{T})\end{aligned}$$

Now plug this bound into Equation (11) and we obtain

$$\mathbb{E}\big(\hat{\phi}_K - \bar{\phi}\big)^2 \leq C\left[\frac{1}{T} + \frac{h}{T}\frac{\sum_{k=0}^{K-1}\mathbb{E}\left[\mathrm{Tr}[\mathrm{cov}(n\nabla V_{I_k}|\mathcal{F}_k)]\right]}{K} + h^2\right]$$

for some constant $C > 0$. ■

E MINI BATCH VERSION OF EWSG

When mini batch size $b > 1$, for each mini batch $\{i_1, i_2, \cdots, i_b\}$, we use $\frac{n}{b}\sum_{j=1}^{b}\nabla V_{i_j}$ to approximate full gradient ∇V, and assign the mini batch $\{i_1, i_2, \cdots, i_b\}$ probability $p_{i_1 i_2, \cdots, i_b}$. We can easily extend the transition probability of $b = 1$ to general b, simply by replacing $n\nabla V_i$ with $\frac{n}{b}\sum_{j=1}^{b}\nabla V_{i_j}$ and end up with

$$\tilde{P}(\boldsymbol{\theta}_{k+1}, \boldsymbol{r}_{k+1}|\boldsymbol{\theta}_k, \boldsymbol{r}_k) = \delta(\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \boldsymbol{r}_k h)\times$$

$$\sum_{i_1, i_2, \cdots, i_b} p_{i_1 i_2 \cdots i_b}\Phi\left(\boldsymbol{x} + n\boldsymbol{a}_{i_1 i_2 \cdots i_b}\right)\frac{1}{\sigma\sqrt{h}}$$

where

$$\boldsymbol{x} = \frac{\boldsymbol{r}_{k+1} - \boldsymbol{r}_k + h\gamma\boldsymbol{r}_k}{\sigma\sqrt{h}}, \quad \boldsymbol{a}_{i_1 i_2 \cdots i_b} = \frac{\sqrt{h}}{\sigma}\frac{1}{b}\sum_{j=1}^{b}\nabla V_{i_j}(\boldsymbol{\theta}_k)$$

Therefore, to match the transition probability of underdamped Langevin dynamics with stochastic gradient and full gradient, we let $p_{i_1 i_2 \cdots i_b} =$

$$\frac{1}{Z}\exp\left\{\frac{1}{2}\left[\|\boldsymbol{x} + n\boldsymbol{a}_{i_1 i_2 \cdots i_b}\|^2 - \|\boldsymbol{x} + \sum_{i_1 i_2 \cdots i_b}\boldsymbol{a}_{i_1 i_2 \cdots i_b}\|^2\right]\right\}$$

where Z is a normalization constant.

To sample multidimensional random data indices $I_1, \cdots, I_b$ from $p_{i_1 i_2 \cdots i_b}$, we again use a Metropolis chain, whose acceptance probability only depends on $a_{i_1 i_2 \cdots i_b}$ and $a_{j_1 j_2 \cdots j_b}$ but not the full gradient.

F EWSG VERSION FOR OVERDAMPED LANGEVIN

Overdamped Langevin equation is the following SDE

$$d\boldsymbol{\theta}_t = -\nabla V(\boldsymbol{\theta}_t)dt + \sqrt{2}d\boldsymbol{B}_t$$

where $V(\boldsymbol{\theta}) = \sum_{i=1}^{n} V_i(\boldsymbol{\theta})$ and B_t is a d-dimensional Brownian motion. The Euler-Maruyama discretization is

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - h\nabla V(\boldsymbol{\theta}_k) + \sqrt{2h}\boldsymbol{\xi}_{k+1}$$

where $\boldsymbol{\xi}_{k+1}$ is a d-dimensional random Gaussian vector. When stochastic gradient is used, the above numerical schedme turns to

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - h\nabla V_{I_k}(\boldsymbol{\theta}_k) + \sqrt{2h}\boldsymbol{\xi}_{k+1}$$

where I_k is the datum index used in k-th iteration to estimate the full gradient.

Denote $\boldsymbol{x} = \frac{\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k}{\sqrt{2h}}$ and $\boldsymbol{a}_i = \frac{\sqrt{h}\nabla V_i(\boldsymbol{\theta}_k)}{\sqrt{2}}$. If we set

$$p_i = \mathbb{P}(I_k = i) \propto \exp\Big\{ - \frac{\|\boldsymbol{x} + \sum_{j=1}^{n} \boldsymbol{a}_j\|^2}{2} + \frac{\|\boldsymbol{x} + n\boldsymbol{a}_i\|^2}{2} \Big\}$$

and follow the same steps in the proof of Theorem 2, we will see the transition kernel of full gradient and the transition kernel of stochastic gradient are matched up.

G VARIANCE REDUCTION (VR)

We have seen that when step size h is large, EWSG still introduces extra variance. To further mitigate this inaccuracy, we provide in this section a complementary variance reduction technique.

Locally (i.e., conditioned on the state of the system at the current step), we have increased variance

$$\begin{aligned} \operatorname{cov}[\boldsymbol{r}_{k+1}|\boldsymbol{r}_k] &= \mathbb{E}[\operatorname{cov}[\boldsymbol{r}_{k+1}|I]] + \operatorname{cov}[\mathbb{E}[\boldsymbol{r}_{k+1}|I]] \\ &= h(\Sigma_{k+1}^2 + h\operatorname{cov}[n\nabla V_I(\boldsymbol{\theta}_k)]) \end{aligned} \tag{13}$$

where $\Sigma_{k+1}^2 = \frac{1}{h}\mathbb{E}[\operatorname{cov}[\boldsymbol{r}_{k+1}|I]]$. The extra randomness due to the randomness of the index I enters the parameter space through the coupling of $\boldsymbol{\theta}$ and $\boldsymbol{r}$ and eventually deviates the stationary distribution from that of the original dynamics. Adopting the perspective of modified equation (Borkar & Mitter, 1999; Mandt et al., 2017; Li et al., 2017), we model this as an enlarged diffusion coefficient. To correct for this enlargement and still sample from the correct distribution, we can either, in each step, shrink the size of intrinsic noise to $\Sigma_k \in \mathbb{R}^{d\times d}$ such that $\sigma^2 I = \Sigma_k^2 + h\operatorname{cov}[n\nabla V_I(\boldsymbol{\theta}_{k-1})]$, or alternatively increase the dissipation. More precisely, due to the matrix version fluctuation dissipation theorem $\Sigma^2 = 2\Gamma T$, one could instead increase the friction coefficient $\Gamma \in \mathbb{R}^{d\times d}$ rather than shrinking the intrinsic noise. The second approach is computationally more efficient because it no longer requires square-rooting / Cholesky decomposition of (possibly large-scale) matrices. Therefore, in each step, we set

$$\Gamma_k = \frac{1}{2T}(\sigma^2 I + h\operatorname{cov}[n\nabla V_I(\boldsymbol{\theta}_{k-1})]).$$

Accurately computing $\operatorname{cov}[n\nabla V_I(\boldsymbol{\theta}_{k-1})]$ is expensive as it requires running I through $1, \cdots, n$, which defeats the purpose of introducing a stochastic gradient. To downscale the computation cost from $\mathcal{O}(n)$ to $\mathcal{O}(1)$, we use an SVRG type estimation of the this variance instead. More specifically, we periodically compute $\operatorname{cov}[n\nabla V_I(\boldsymbol{\theta}_{k-1})]$ only every L data passes, in an outer loop. In every iteration of an inner loop, which integrates the Langevin, an estimate of $\operatorname{cov}[n\nabla V_I(\boldsymbol{\theta}_{k-1})]$ is updated in an SVRG fashion.

See Algorithm 2 for detailed description. We refer variance reduced variant of EWSG as EWSG-VR.

To demonstrate the performance of EWSG-VR, we reuse the setup of simple Gaussian example in subsection 5.1. As shown in Algorithm 2, the only hyper-parameter of EWSG-VR additional to EWSG is the period of variance calibration, for which we set $L = 1$. All other hyper-parameters (e.g. step size h, friction coefficient γ) are set the same as EWSG. We also run underdamped Langevin dynamics with full gradient (FG) using the same hyper-parameters of EWSG. We plot the KL divergence in Figure 4. We see that EWSG-VR further reduces variance and achieves better statistical accuracy measured in KL divergence. Although EWSG-VR periodically use full data set to calibrate variance estimation, it is still significantly faster than the full gradient version. Note that KL divergence of SGLD, pSGLD and SGHMC are too large so that we can not even see them in Figure 4

Algorithm 2 EWSG-VR

Input: {number of data terms n, gradient functions $\nabla V_i(\cdot)$, step size h, number of data passes K, period of variance calibration L, index chain length M, friction and noise coefficients γ and σ}
initialize $\boldsymbol{\theta}_0, \boldsymbol{r}_0, \gamma_0 = \gamma$
initialize inner loop index $k = 0$
for $l = 1, 2, \cdots, K$ **do**
  if $(l-1) \mod L = 0$ **then**
    compute $\boldsymbol{m}_1 \leftarrow \mathbb{E}_I[n\nabla V_I(\boldsymbol{\theta}_k)]$, $\boldsymbol{m}_2 \leftarrow \mathbb{E}_I[n^2 \nabla V_I(\boldsymbol{\theta}_k)\nabla V_I(\boldsymbol{\theta}_k)^T]$
    $\boldsymbol{\omega} \leftarrow \boldsymbol{\theta}_k$
  else
    for $t = 1, 2, \cdots, \lceil \frac{n}{M+1} \rceil$ **do**
      $i \leftarrow$ uniformly sampled from $1, \cdots, n$, compute and store $n\nabla V_i(\boldsymbol{\theta}_k)$
      for $m = 1, 2, \cdots, M$ **do**
        $j \leftarrow$ uniformly sampled from $1, \cdots, n$, compute and store $n\nabla V_j(\boldsymbol{\theta}_k)$
        $i \leftarrow j$ with probability in Equation 5
      end for
      update $(\boldsymbol{\theta}_{k+1}, \boldsymbol{r}_{k+1}) \leftarrow (\boldsymbol{\theta}_k, \boldsymbol{r}_k)$ according to Equation 3, using $n\nabla V_i(\boldsymbol{\theta}_k)$ as gradient and Γ_k as friction
      $\boldsymbol{m}_1 \leftarrow \boldsymbol{m}_1 + \nabla V_i(\boldsymbol{\theta}_k) - \nabla V_i(\boldsymbol{\omega})$
      $\boldsymbol{m}_2 \leftarrow \boldsymbol{m}_2 + n\nabla V_i(\boldsymbol{\theta}_k)\nabla V_i(\boldsymbol{\theta}_k)^T - n\nabla V_i(\boldsymbol{\omega})\nabla V_i(\boldsymbol{\omega})^T$
      covar $\leftarrow \boldsymbol{m}_2 - \boldsymbol{m}_1\boldsymbol{m}_1^T$
      $\Gamma_{k+1} \leftarrow \frac{1}{2T}(\sigma^2 \mathrm{I} + h\,\mathrm{covar})$
      $k \leftarrow k + 1$
    end for
  end if
end for

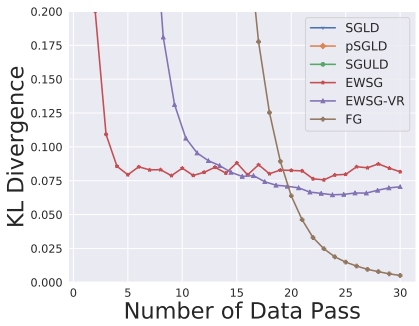


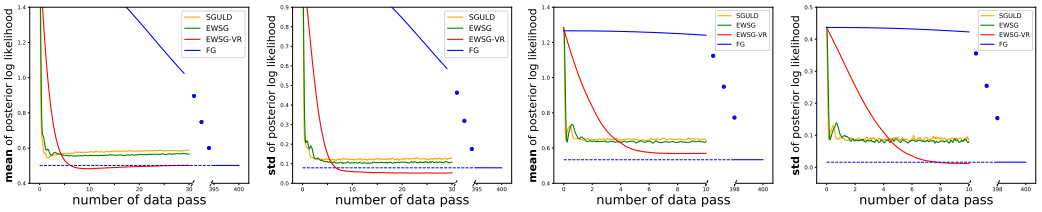
Figure 4: KL divergence

We also consider applying EWSG-VR to Bayesian logistic regression problems. We run experiments on two standard classification data sets `parkinsons` [7], `pima`[8] from UCI repository (Lichman et al., 2013).

From Figure 5, we see stochastic gradient methods (SGHMC, EWSG and EWSG-VR) only take tens of data passes to converge while full gradient version (FG) requires hundreds of data passes to converge. Compared with SGHMC, EWSG produces closer results to FG for which we treat as ground truth, in terms of statistical accuracy. With variance reduction, EWSG-VR is able to achieve even better performance, significantly improving the accuracy of the prediction of mean and standard deviation of log likelihood. It, however, converges slower than EWSG without VR.

One downside of EWSG-VR is that it periodically use whole data set to calibrate variance estimation, so it may not be suitable for very large data sets (e.g. Covertype data set used in subsection 5.2) for which stochastic gradient methods could converge within one data pass.

[7] https://archive.ics.uci.edu/ml/datasets/parkinsons
[8] https://archive.ics.uci.edu/ml/datasets/diabetes



(a) parkinsons (b) pima

Figure 5: Posterior prediction of mean (*left*) and standard deviation (*right*) of log likelihood on test data set generated by SGHMC, EWSG and EWSG-VR on two Bayesian logistic regression tasks. Statistics are computed based on 1000 independent simulations. Minibatch size $b = 1$ for all methods except FG. $M = 1$ for EWSG and EWSG-VR.

H Additional Experiments

H.1 A Misspecified Gaussian Case

In this subsection, we follow the same setup as in (Bardenet et al., 2017) and study a misspecified Gaussian model where one fits a one-dimensional normal distribution $p(\theta) = \mathcal{N}(\theta|\mu_0, \sigma_0^2)$ to 10^5 i.i.d points drawn according to $X_i \sim \log \mathcal{N}(0, 1)$, and flat prior is assigned $p(\mu_0, \log \sigma_0) \propto 1$. It was shown in (Bardenet et al., 2017) that FlyMC algorithm behaves erratically in this case, as "bright" data points with large values are rarely updated and they drive samples away from the target distribution. Consequently the chain mixes very slowly. One important commonality FlyMC shares with EWSG is that in each iteration, both algorithms select a subset of data in a non-uniform fashion. Therefore, it is interesting to investigate the performance of EWSG in this misspecified model.

For FlyMC[9], a tight lower bound based on Taylor's expansion is used to minimize "bright" data points used per iteration. At each iteration, 10% data points are resampled and turned "on/off" accordingly and the step size is adaptively adjusted. FlyMC algorithm is run for 10000 iterations. Figure 6a shows the histogram of number of data points used in each iteration for FlyMC algorithm. On average, FlyMC consumes 10.9% of all data points per iteration. For fair comparison, the minibatch size of EWSG is hence set $10^5 \times 10.9\% = 10900$ and we run EWSG for 1090 data passes. We set step size $h = 1 \times 10^{-4}$ and friction coefficient $\gamma = 300$ for EWSG. An isotropic random walk Metropolis Hasting (MH) is also run for sufficiently long and serves as the ground truth.

Figure 6b shows the autocorrelation of three algorithms. The autocorrelation of FlyMC decays very slowly, samples that are even 500 iterations away still show strong correlation. The autocorrelation of EWSG, on the other hand, decays much faster, suggesting EWSG explores parameter space efficiently than FlyMC does. Figure 6c and 6d show the samples (the first 1000 samples are discarded as burn-in) generated by EWSG and FlyMC respectively. The samples of EWSG center around the mode of the target distribution while the samples of FlyMC are still far away from the true posterior. The experiment shows EWGS works quite well even in misspecified models, and hence is an effective candidate in combining importance sampling with scalable Bayesian inference.

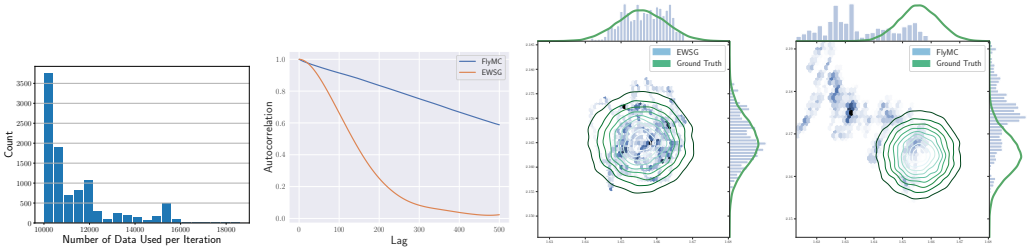
H.2 Additional Results of BNN Experiment

We report the test error of various SG-MCMC methods after 200 epochs in Table 2. For both MLP and CNN architecture, EWSG outperforms its uniform counterpart SGHMC as well as other benchmarks SGLD, pSGLD and CP-SGHMC. The results clearly demonstrate the effectiveness of the proposed EWSG on deep models.

H.3 Additional Experiment on BNN: Tuning M

In each iteration of EWSG, we run an index Markov chain of length M and select a "good" minibatch to estimate gradient, therefore EWSG essentially uses $b \times (M + 1)$ data points per iteration

[9] https://github.com/rbardenet/2017JMLR-MCMCForTallData



(a) Histogram (b) Autocorrelation (c) Samples of EWSG (d) Samples of FlyMC

Figure 6: (a) Histogram of data used in each iteration for FlyMC algorithm. (b) Autocorrelation plot of FlyMC, EWSG and MH. (c) Samples of EWSG. (d) Samples of FlyMC.

Table 2: Test error (mean $\pm$ standard deviation) after 200 epoches.

Method	Test Error(%), MLP	Test Error(%), CNN
SGLD	1.976 ± 0.055	0.848 ± 0.060
pSGLD	1.821 ± 0.061	0.860 ± 0.052
SGHMC	1.833 ± 0.073	0.778 ± 0.040
CP-SGHMC	1.835 ± 0.047	0.772 ± 0.055
EWSG	$\mathbf{1.793 \pm 0.100}$	$\mathbf{0.753 \pm 0.035}$

where b is minibatch size. How does EWSG compare with its uniform gradient subsampling counterpart with a larger minibatch size ($b \times (M+1)$)?

We empirically answer this question in the context of BNN with MLP architecture. We use the same step size for SGHMC and EWSG and experiment a large range of values of minibatch size b and index chain length M. Each algorithm is run for 200 data passes and 10 independent samples are drawn to estimate test error. The results are shown in Table 3. We find that EWSG beats SGHMC with larger minibatch in 8 out of 9 comparison groups, which suggests in general EWSG could be a better way to consuming data compared to increasing minibatch size and may shed light on other areas where stochastic gradient methods are used (e.g. optimization).

b	$M+1=2$	$M+1=5$	$M+1=10$
100	**1.86%** 1.94%	**1.83%** 1.92%	**1.80%** 1.97%
200	1.90% **1.87%**	**1.87%** 1.97%	**1.80%** 2.07%
500	**1.79%** 1.97%	**2.01%** 2.17%	**2.36%** 2.37%

Table 3: Test errors of EWSG (top of each cell) and SGHMC (bottom of each cell) after 200 epoches. b is minibatch size for EWSG, and minibatch size of SGHMC is set as $b \times (M+1)$ to ensure the same number of data used per parameter update for both algorithms. Step size is set $h = \frac{10}{b(M+1)}$ as suggested in (Chen et al., 2014), different from that used to produce Table 2. Results with smaller test error is highlighted in boldface.

I EWSG DOES NOT NECESSARILY CHANGE THE SPEED OF CONVERGENCE SIGNIFICANTLY

Changing the weights of stochastic gradient from uniform to non-uniform, as we saw, can increase the statistical accuracy of the sampling; however, it does not necessarily increase or decrease the speed of convergence to the (altered) limiting distribution. Numerical examples already demonstrated this fact, but on the theoretical side, we note the non-asymptotic bound provided by Theorem 4 may not provide a tight enough quantification of speed of convergence due to its generality. Therefore, here we quantify the convergence speed on a simple quadratic example:

Consider $V_i(\theta) = \frac{1}{n}(\theta - \mu_i)^2/2$ where μ_i's are constant scalars. Assume without loss of generality that $\sum_i \mu_i = 0$, and thus $V(\theta) = \sum_{i=1}^n V_i(\theta) = \theta^2/2 +$ some constant. We will show the convergence speed of $\mathbb{E}\theta$ is comparable for uniform and a class of non-uniform SG-MCMC (including EWSG) applied to second-order Langevin equation (overdamped Langevin will be easier and thus omitted):

Theorem 6 *Consider, for $0 < \gamma < 2$, respectively SGHMC and EWSG,*

$$\begin{cases} \theta'_{k+1} &= \theta'_k + hr'_k \\ r'_{k+1} &= r'_k - h\gamma r'_k - h(\theta'_k - \mu_{I'_k}) + \sqrt{h}\sigma\xi'_{k+1} \end{cases} \quad \textit{and} \quad \begin{cases} \theta_{k+1} &= \theta_k + hr_k \\ r_{k+1} &= r_k - h\gamma r_k - h(\theta_k - \mu_{I_k}) + \sqrt{h}\sigma\xi_{k+1} \end{cases},$$

where I'_k are i.i.d. uniform random variable on $[n]$, I_k are $[\theta, r]$ dependent random variable on $[n]$ satisfying $\mathbb{P}(I_k = i) = 1/n + \mathcal{O}(h^p)$, and ξ_{k+1}, ξ'_{k+1} are standard i.i.d. Gaussian random variables. Denote by $\bar{\theta'}_k = \mathbb{E}\theta'_k$, $\bar{r'}_k = \mathbb{E}r'_k$, $\bar{\theta}_k = \mathbb{E}\theta_k$, $\bar{r}_k = \mathbb{E}r_k$, $x'_k = [\bar{\theta'}_k, \bar{r'}_k]^T$, and $x_k = [\bar{\theta}_k, \bar{r}_k]^T$, then

$$x'_k = (I + Ah)^k x'_0, \quad \textit{where } A = \begin{bmatrix} 0 & 1 \\ -1 & -\gamma \end{bmatrix}, \tag{14}$$

for small enough h, $\|x'_k\|$ converges to 0 exponentially with $k \to \infty$, and x_k converges at a comparable speed in the sense that $\|x_k - x'_k\| = \mathcal{O}(h^p)$ if $x_0 = x'_0$.

Proof: Taking the expectation of the $[\theta', r']$ iteration and using the fact that $\sum_i \mu_i = 0$ and hence $\mathbb{E}\mu_{I'_k} = 0$, one easily obtains (14). The geometric convergence of x'_k thus follows from the fact that eigenvalues of $I + Ah$ have less than 1 modulus for small enough h.

Let $e_k = [0, \mathbb{E}\mu_{I_k}]^T$ and then

$$e_k = [0, \sum_{i=1}^n \mathbb{P}(I_k = i)\mu_i]^T = [0, \mathcal{O}(h^p)]^T$$

Now we take the expectation of both sides of the $[\theta, r]$ iteration and obtain $x_{k+1} = (I+Ah)x_k + he_k$. Therefore

$$x_k = (I+Ah)^k x_0 + (I+Ah)^{k-1} he_0 + \cdots + (I+Ah)he_{k-2} + he_{k-1} = x'_k + h\big((I+Ah)^{k-1}e_0 + \cdots + (I+Ah)e_{k-2} + e_{k-1}\big)$$

To bound the difference, note $I + Ah$ is diagonalizable with complex eigenvalues $\lambda_{1,2}$ satisfying

$$|\lambda_1| = |\lambda_2| = \sqrt{1 - h\gamma + h^2} = 1 - \gamma h/2 + \mathcal{O}(h^2).$$

Projecting e_j to the corresponding eigenspaces via $e_j = v_{1,j} + v_{2,j}$, we can get

$$\begin{aligned} h\|(I + Ah)^{k-1}e_0 + \cdots + e_{k-1}\| &\le h\left(\|(I + Ah)^{k-1}e_0\| + \cdots + \|e_{k-1}\|\right) \\ &= h\left(|\lambda_1|^{k-1}\|v_{1,0}\| + |\lambda_2|^{k-1}\|v_{2,0}\| + \cdots + \|v_{1,k-1}\| + \|v_{2,k-1}\|\right) \\ &\le hCh^p(|\lambda_1|^{k-1} + \cdots + 1) = hCh^p\frac{1 - |\lambda_1|^k}{1 - |\lambda_1|} \le hCh^p\frac{1}{1 - |\lambda_1|} \\ &\le \hat{C}h^p \end{aligned}$$

for some constant C and $\hat{C}$. ■

Important to note is, although this is already a nonlinear example for EWSG (as nonlinearity enters through the μ_{I_k} term), it is a linear example for SGHMC. We do not have a tight quantification for the fully nonlinear cases, for which whether EWSG converges faster or comparably like suggested by the experiments remains to be an open theoretical challenge.