A. Proofs for convergence of variational inference

We study convergence of Λ_n to Λ^* in terms of the KL divergence from $p(z)$ to $q(z|\Lambda_n)$. Before proving the convergence rates for (stochastic) variational inference, we first derive a useful bound for the KL divergence, which will be used frequently in the proofs to follow.

Lemma 3. *The KL divergence between two normal distributions $p(z)$ and $q(z|\Lambda_n)$ is upper bounded by their difference in the Frobenius norm:*

$$\begin{aligned}&\mathrm{KL}\left(p(z)\|q(z|\Lambda_n)\right)\\&\le \frac{1}{2}\| \left(\Lambda^*\right)^{-1} \|_2 \cdot \| \left(\Lambda_n\right)^{-1} \|_2 \cdot \|\Lambda^* - \Lambda_n\|_F^2 .\end{aligned}$$

Proof of Lemma 3

$$\begin{aligned}&\mathrm{KL}\left(p(z)\|q(z|\Lambda_n)\right)\\&= \frac{1}{2}\left(-\log\frac{|\Lambda_n|}{|\Lambda^*|} + \mathrm{tr}((\Lambda^*)^{-1}\Lambda_n) - d\right)\\&= \frac{1}{2}\left(-\log\left|(\Lambda^*)^{-1}\Lambda_n\right| + \mathrm{tr}\left((\Lambda^*)^{-1}\Lambda_n - I\right)\right)\end{aligned}$$

Since

$$\log\left|(\Lambda^*)^{-1}\Lambda_n\right| \ge \mathrm{tr}\left(I - \Lambda_n^{-1}\Lambda^*\right),$$

$$\begin{aligned}&\mathrm{KL}\left(p(z)\|q(z|\Lambda_n)\right)\\&\le \frac{1}{2}\mathrm{tr}\left((\Lambda^*)^{-1}\Lambda_n + \Lambda_n^{-1}\Lambda^* - 2I\right)\\&= \frac{1}{2}\mathrm{tr}\left(\left((\Lambda^*)^{-1}(\Lambda_n - \Lambda^*)\right) + \left(\Lambda_n^{-1}(\Lambda^* - \Lambda_n)\right)\right)\\&= \frac{1}{2}\mathrm{tr}\left(\left(\Lambda_n^{-1} - (\Lambda^*)^{-1}\right)(\Lambda^* - \Lambda_n)\right)\\&\le \frac{1}{2}\left\|\Lambda_n^{-1} - (\Lambda^*)^{-1}\right\|_F \cdot \|\Lambda^* - \Lambda_n\|_F\\&\le \frac{1}{2}\| \left(\Lambda^*\right)^{-1} \|_2 \cdot \| \left(\Lambda_n\right)^{-1} \|_2 \cdot \|\Lambda^* - \Lambda_n\|_F^2 .\end{aligned}$$

□

A.1. Proof for “Open-Box” VI Convergence

Proof of Lemma 1 The update rule in equation 13 can be explicitly expressed as:

$$\begin{aligned}\Lambda_n &= \Lambda_{n-1} - h_{n-1}\, g(\Lambda_{n-1})\\&= \Lambda_{n-1} - \frac{h_{n-1}}{2}\left(\Lambda_{n-1} - \Lambda^*\right).\end{aligned}$$

When we take a constant step size $h_k = h = \frac{1}{2}$, we can obtain that

$$\Lambda_n = \left(\frac{1}{2}\right)^n \Lambda_0 + \left(1 - \left(\frac{1}{2}\right)^n\right)\Lambda^*.$$

Therefore,

$$\|\Lambda_n - \Lambda^*\|_F = \left(\frac{1}{2}\right)^n \|\Lambda_0 - \Lambda^*\|_F$$

Using the result of Lemma 3, we can obtian that

$$\begin{aligned}&\mathrm{KL}\left(p(z)\|q(z|\Lambda_n)\right)\\&\leq \frac{1}{2}\|\left(\Lambda^*\right)^{-1}\|_2\cdot\|\left(\Lambda_n\right)^{-1}\|_2\cdot\|\Lambda^*-\Lambda_n\|_F^2.\end{aligned}$$

By Weyl's theorem, we know that the distance from any eigenvalue of Λ_n to the closest eigenvalue of Λ^* is upper bounded by $\|\Lambda_n-\Lambda^*\|_2\leq\|\Lambda_n-\Lambda^*\|_F$. Therefore, $\sigma_{\min}(\Lambda_n)\geq\sigma_{\min}(\Lambda^*)-\left(\frac{1}{2}\right)^n\|\Lambda_0-\Lambda^*\|_F$, resulting in the upper bound for the spectral norm of Λ_n that

$$\|\left(\Lambda_n\right)^{-1}\|_2\leq\frac{1}{\sigma_{\min}(\Lambda^*)-\left(\frac{1}{2}\right)^n\|\Lambda_0-\Lambda^*\|_F}.$$

Therefore, for any

$$n\geq\log_2\frac{2}{\sigma_{\min}(\Lambda^*)}\frac{\|\Lambda_0-\Lambda^*\|_F}{\epsilon},$$

$\mathrm{KL}\left(p(z)\|q(z|\Lambda_n)\right)\leq\epsilon$, for any $\epsilon\leq 1$. □

A.2. Proofs for "Black-Box" VI Convergence

Proof of Theorem 1. We obtain the convergence bound in $\|\Lambda^*-\Lambda_n\|_F^2$ and then incur Lemma 3 to finish the proof.

Lemma 4. *For the stochastic preconditioned gradient descent algorithm described in equation 15, if we take a step size of* $h=\frac{\tilde{\epsilon}}{4\sigma_{\max}^2(\Lambda^*)d\delta^2}$*, we can obtain that when*

$$n\geq\frac{4\sigma_{\max}^2(\Lambda^*)d\delta^2}{\nu\tilde{\epsilon}}\log\frac{2\|\Lambda_0-\Lambda^*\|_F^2}{\nu\tilde{\epsilon}},$$

$\|\Lambda_n-\Lambda^*\|_F^2\leq\tilde{\epsilon}$*, with probability* $1-\nu$*.*

Then by Weyl's theorem, we know that the distance from any eigenvalue of Λ_n to the closest eigenvalue of Λ^* is upper bounded by $\|\Lambda_n-\Lambda^*\|_2\leq\|\Lambda_n-\Lambda^*\|_F$. Therefore, $\sigma_{\min}(\Lambda_n)\geq\sigma_{\min}(\Lambda^*)-\sqrt{\tilde{\epsilon}}$, resulting in the upper bound for the spectral norm of Λ_n that

$$\|\left(\Lambda_n\right)^{-1}\|_2\leq\frac{1}{\sigma_{\min}(\Lambda^*)-\sqrt{\tilde{\epsilon}}}. \tag{20}$$

Applying equation 20 to Lemma 3, we upper bound the KL divergence by $\tilde{\epsilon}$:

$$\mathrm{KL}\left(p(z)\|q(z|\Lambda_n)\right)\leq\frac{1}{2\sigma_{\min}(\Lambda^*)}\frac{\tilde{\epsilon}}{\sigma_{\min}(\Lambda^*)-\sqrt{\tilde{\epsilon}}}.$$

Choosing $\tilde{\epsilon}=\frac{\sigma_{\min}^2(\Lambda^*)}{2}\epsilon$ completes the proof that after

$$n\geq 8\frac{\sigma_{\max}^2(\Lambda^*)\delta^2}{\nu\sigma_{\min}^2(\Lambda^*)}\frac{d}{\epsilon}\cdot\log\frac{4\|\Lambda_0-\Lambda^*\|_F^2}{\nu\sigma_{\min}^2(\Lambda^*)\epsilon}=\tilde{\mathcal{O}}\left(\frac{\sigma_{\max}^2(\Lambda^*)\delta^2}{\sigma_{\min}^2(\Lambda^*)}\frac{d}{\epsilon}\right).$$

number of iterations, $\mathrm{KL}\left(p(z)\|q(z|\Lambda_n)\right)\leq\epsilon$. □

Proof of Lemma 4. We first prove the convergence in $\mathbb{E}\|\Lambda_n-\Lambda^*\|_F^2$ then invoke the Chebychev inequality for the high probability statement.

Since we assumed in equation 14 that $\mathbb{E}\left[\Delta(\Lambda; \mathcal{D}_n)\right] = 0$, and that $\mathbb{E}\left\|\Delta(\Lambda; \mathcal{D}_n)\right\|_F^2 \leq \sigma_{\max}^2(\Lambda^*) d\delta^2$, for any n,

$$
\begin{aligned}
&\mathbb{E}\|\Lambda_n - \Lambda^*\|_F^2 \\
&= \mathbb{E}\left\|\Lambda_{n-1} - \Lambda^* - h_{n-1}\hat{g}(\Lambda)\right\|_F^2 \\
&= \mathbb{E}\left\|\left(1 - \frac{h_{n-1}}{2}\right)(\Lambda_{n-1} - \Lambda^*) + h_{n-1}\Delta(\Lambda_{n-1}; \mathcal{D}_{n-1})\right\|_F^2 \\
&= \left(1 - \frac{h_{n-1}}{2}\right)^2 \mathbb{E}\left\|\Lambda_{n-1} - \Lambda^*\right\|_F^2 + 2h_{n-1}\left(1 - \frac{h_{n-1}}{2}\right)\mathbb{E}\left\langle \Lambda_{n-1} - \Lambda^*, \Delta(\Lambda_{n-1}; \mathcal{D}_{n-1})\right\rangle_F + h_{n-1}^2\mathbb{E}\|\Delta(\Lambda_{n-1}; \mathcal{D}_{n-1})\|_F^2 \\
&= \left(1 - \frac{h_{n-1}}{2}\right)^2 \mathbb{E}\left\|\Lambda_{n-1} - \Lambda^*\right\|_F^2 + h_{n-1}^2\mathbb{E}\|\Delta(\Lambda_{n-1}; \mathcal{D}_{n-1})\|_F^2 \\
&\leq \left(1 - \frac{h_{n-1}}{2}\right) \mathbb{E}\left\|\Lambda_{n-1} - \Lambda^*\right\|_F^2 + h_{n-1}^2\mathbb{E}\|\Delta(\Lambda_{n-1}; \mathcal{D}_{n-1})\|_F^2 \\
&\leq \prod_{i=0}^{n-1}\left(1 - \frac{h_i}{2}\right) \mathbb{E}\left\|\Lambda_0 - \Lambda^*\right\|_F^2 + \sum_{j=0}^{n-1} h_j^2 \left(\prod_{i=j+1}^{n-1}\left(1 - \frac{h_i}{2}\right)\right)\mathbb{E}\|\Delta(\Lambda_{j-1}; \mathcal{D}_{j-1})\|_F^2 \\
&\leq \prod_{i=0}^{n-1}\left(1 - \frac{h_i}{2}\right) \mathbb{E}\left\|\Lambda_0 - \Lambda^*\right\|_F^2 + \sum_{j=0}^{n-1} h_j^2 \prod_{i=j+1}^{n-1}\left(1 - \frac{h_i}{2}\right) \sigma_{\max}^2(\Lambda^*) d\delta^2.
\end{aligned}
$$

When we take a constant step size, $h_k = h$, the above expression simplifies to:

$$
\begin{aligned}
\mathbb{E}\|\Lambda_n - \Lambda^*\|_F^2 &\leq \left(1 - \frac{h_i}{2}\right)^n \mathbb{E}\left\|\Lambda_0 - \Lambda^*\right\|_F^2 + 2h\left(\left(1 - \frac{h_i}{2}\right) - \left(1 - \frac{h_i}{2}\right)^n\right)\sigma_{\max}^2(\Lambda^*) d\delta^2 \\
&\leq \left(1 - \frac{h_i}{2}\right)^n \mathbb{E}\left\|\Lambda_0 - \Lambda^*\right\|_F^2 + 2h\sigma_{\max}^2(\Lambda^*) d\delta^2. \qquad (21)
\end{aligned}
$$

We then invoke the following Chebyshev inequality to obtain the high probability statement:

$$
\mathbb{P}\left(\|\Lambda_n - \Lambda^*\|_F^2 \geq \tilde{\epsilon}\right) \leq \frac{1}{\tilde{\epsilon}}\mathbb{E}\|\Lambda_n - \Lambda^*\|_F^2.
$$

For $\|\Lambda_n - \Lambda^*\|_F^2 \leq \tilde{\epsilon}$ to hold with $1 - \nu$ probability, we need $\|\Lambda_n - \Lambda^*\|_F^2 \leq \nu\tilde{\epsilon}$.

Choosing $h = \frac{\nu\tilde{\epsilon}}{4\sigma_{\max}^2(\Lambda^*) d\delta^2}$, we arrive at our conclusion that $\|\Lambda_n - \Lambda^*\|_F^2 \leq \tilde{\epsilon}$ with probability $1 - \nu$, when

$$
n \geq \frac{4\sigma_{\max}^2(\Lambda^*) d\delta^2}{\nu\tilde{\epsilon}} \log \frac{2\left\|\Lambda_0 - \Lambda^*\right\|_F^2}{\nu\tilde{\epsilon}},
$$

where the log factor can be shaved off by employing a decreasing step size. □

Tightness of the bounds We now demonstrate that the convergence upper bound in Theorem 1 is tight up to a logarithmic factor. We first prove that the Frobenius norm bound in Lemma 4, instead of a spectral norm bound, is indeed necessary to guarantee the convergence in KL divergence.

To this end, we examine an example of the posterior with the precision matrix $\Lambda^* = \frac{1}{4}I$. If the initial distribution has the precision matrix $\Lambda_0 = I$, then $\|\Lambda_0 - \Lambda^*\|_2 = \frac{3}{4}$. However,

$$
\begin{aligned}
&\mathrm{KL}\left(p(z)\|q(z|\Lambda_0)\right) \\
&= \frac{1}{2}\left(-\log\frac{|\Lambda_0|}{|\Lambda^*|} + \mathrm{tr}((\Lambda^*)^{-1}\Lambda_0) - d\right) \geq d,
\end{aligned}
$$

which can be arbitrarily large as dimension d increases.

We then use the same posterior of $\Lambda^* = \frac{1}{4}I$ and take an initial value Λ_0 so that $\|\Lambda_0 - \Lambda^*\|_F^2$ scales inclusively between $\Omega(1)$ and $\mathcal{O}(d)$. Under this mild condition, we demonstrate that the number of iterations, n, required for $\|\Lambda_0 - \Lambda^*\|_F^2$ to decrease to $\|\Lambda_n - \Lambda^*\|_F^2 \leq \frac{1}{2}\|\Lambda_0 - \Lambda^*\|_F^2$ is $n = \Omega(d)$.

We first demonstrate that $\mathbb{E}\|\Delta(\Lambda;\mathcal{D}_n)\|_F^2=\Omega(d)$ for minibatch size $|\mathcal{D}_n|=\mathcal{O}(d)$. From Section 4.3, we know that

$$\begin{aligned}
&\mathbb{E}\|\Delta(\Lambda;\mathcal{D}_n)\|_F^2\\
&=\frac{1}{|\mathcal{D}_n|}\mathbb{E}_{z\sim q}\left[\|v(\Lambda;z)-\mathbb{E}_{\hat{z}\sim q}\left[v(\Lambda;\hat{z})\right]\|_F^2\right]\\
&=\frac{1}{2|\mathcal{D}_n|}\left(\operatorname{tr}(\Lambda-\Lambda^*)\right)^2-\frac{1}{4|\mathcal{D}_n|}\|\Lambda-\Lambda^*\|_F^2\\
&+\frac{1}{|\mathcal{D}_n|}\left(\frac{1}{8}\left\|I-\Lambda^*\Lambda^{-1}\right\|_F^2+\frac{1}{16}\left(\mathrm{KL}\left(q(z|\Lambda)\|p(z)\right)\right)^2\right)\cdot\left(\left(\operatorname{tr}(\Lambda)\right)^2+\operatorname{tr}\left(\Lambda^2\right)\right).
\end{aligned}$$

Since

$$\|\Lambda-\Lambda^*\|_F\le\|\Lambda\|_2\cdot\|I-\Lambda^{-1}\Lambda^*\|_F,$$

we employ Weyl's theorem and obtain that

$$\begin{aligned}
\|I-\Lambda^{-1}\Lambda^*\|_F&\ge\frac{\|\Lambda-\Lambda^*\|_F}{\sigma_{\max}(\Lambda)}\\
&\ge\frac{\|\Lambda-\Lambda^*\|_F}{\sigma_{\max}(\Lambda^*)+\|\Lambda-\Lambda^*\|_F}=\Omega(1).
\end{aligned}$$

Therefore, $\mathbb{E}\|\Delta(\Lambda;\mathcal{D}_n)\|_F^2=\Omega(1)$ for $|\mathcal{D}_n|=\mathcal{O}(d)$ and for $\|\Lambda-\Lambda^*\|_F^2=\Omega(1/d)$ and $\|\Lambda-\Lambda^*\|_F^2=\mathcal{O}(d)$.

We then analyze number of steps n required for $\|\Lambda_n-\Lambda^*\|_F^2\le\frac{1}{2}\|\Lambda_0-\Lambda^*\|_F^2$. In the update rule of equation 15,

$$\begin{aligned}
\Lambda_n&=\Lambda_{n-1}-h_{n-1}\hat{g}(\Lambda_{n-1})\\
&=\Lambda_{n-1}-\frac{h_{n-1}}{2}(\Lambda_{n-1}-\Lambda^*)+h_{n-1}\Delta(\Lambda_{n-1};\mathcal{D}_{n-1}).
\end{aligned}$$

Hence

$$\begin{aligned}
\Lambda_n-\Lambda^*&=\left(1-\frac{h_{n-1}}{2}\right)(\Lambda_{n-1}-\Lambda^*)+h_{n-1}\Delta(\Lambda_{n-1};\mathcal{D}_{n-1})\\
&=\prod_{i=0}^{n-1}\left(1-\frac{h_i}{2}\right)(\Lambda_0-\Lambda^*)+\sum_{j=0}^{n-1}h_j\prod_{i=j+1}^{n-1}\left(1-\frac{h_i}{2}\right)\Delta(\Lambda_j;\mathcal{D}_j).
\end{aligned}$$

Since $\mathcal{D}_n$ are sampled in an i.i.d. fashion and that $\mathbb{E}\left[\Delta(\Lambda;\mathcal{D}_n)\right]=0$ from assumption 14,

$$\mathbb{E}\|\Lambda_n-\Lambda^*\|_F^2=\prod_{i=0}^{n-1}\left(1-\frac{h_i}{2}\right)^2\mathbb{E}\|\Lambda_0-\Lambda^*\|_F^2+\sum_{j=0}^{n-1}h_j^2\prod_{i=j+1}^{n-1}\left(1-\frac{h_i}{2}\right)^2\mathbb{E}\|\Delta(\Lambda_j;\mathcal{D}_j)\|_F^2.$$

Since $\mathbb{E}\|\Delta(\Lambda;\mathcal{D}_n)\|_F^2=\Omega(d)$, to have that $\|\Lambda_n-\Lambda^*\|_F^2\le\frac{1}{2}\|\Lambda_0-\Lambda^*\|_F^2$ with a constant probability, we must require

$$\sum_{j=0}^{n-1}h_j^2\prod_{i=j+1}^{n-1}\left(1-\frac{h_i}{2}\right)^2=\mathcal{O}\left(\frac{1}{d}\right),$$

which implies that $h_j=\mathcal{O}\left(\frac{1}{d}\right),\forall j=0,\cdots,n-1$. On the other hand, to achieve $\|\Lambda_n-\Lambda^*\|_F\le\frac{1}{2}\|\Lambda_0-\Lambda^*\|_F$, we also need

$$\prod_{i=0}^{n-1}\left(1-\frac{h_i}{2}\right)\|\Lambda_0-\Lambda^*\|_F\le\frac{1}{2}\|\Lambda_0-\Lambda^*\|_F,$$

which implies that

$$\begin{aligned}
\sum_{i=0}^{n-1}h_i&\ge\sum_{i=0}^{n-1}\left(\left(1-\frac{h_i}{2}\right)^{-1}-1\right)\\
&\ge\sum_{i=0}^{n-1}\log\left(\left(1-\frac{h_i}{2}\right)^{-1}\right)\\
&\ge\log\frac{\|\Lambda_0-\Lambda^*\|_F}{\|\Lambda_n-\Lambda^*\|_F}=\log(2).
\end{aligned}$$

Since $h_j = \mathcal{O}\left(\frac{1}{d}\right), \forall j$, we need $n = \Omega(d)$ for convergence. $\square$

B. Proofs for convergence of Langevin algorithm

Proof of Lemma 1 Before proving Lemma 1, we first make the assumptions explicit. We are interested in generating samples from $p(\theta) \propto \exp\left(-U(\theta)\right)$, where $U(\theta)$ is L-Lipschitz smooth and m-strongly convex. We further assume, without loss of generality, that U has a fixed point at the origin 0: $\nabla U(0) = 0$.

To prove Lemma 1, we first analyze equation 3 as a discretization scheme of the Langevin diffusion of equation 4. Within each iteration, the ULA update 3 is effectively integrating the following dynamics:

$$\begin{aligned} d\theta_t &= \nabla \log p(\theta_n) dt + \sqrt{2} dW_t \\ &= \nabla \log p(\theta_t) dt + \sqrt{2} dW_t + \left(\nabla \log p(\theta_n) - \nabla \log p(\theta_t)\right) dt, \end{aligned} \tag{22}$$

for $t \in [n\eta, (n+1)\eta]$.

We then analyze the time derivative of the KL divergence $\mathrm{KL}\left(q_t \| p\right)$ within each step:

$$\begin{aligned} \frac{d}{dt}\mathrm{KL}\left(q_t \| p\right) &= -\mathbb{E}\left\langle \nabla \log \frac{q_t(\theta_t)}{p(\theta_t)}, \nabla \log \frac{q_t(\theta_t)}{p(\theta_t)} + \left(\nabla \log p(\theta_n) - \nabla \log p(\theta_t)\right) \right\rangle \\ &= -\mathbb{E}\left\| \nabla \log \frac{q_t(\theta_t)}{p(\theta_t)} \right\|^2 + \mathbb{E}\left\langle \nabla \log \frac{q_t(\theta_t)}{p(\theta_t)}, \nabla \log p(\theta_t) - \nabla \log p(\theta_n) \right\rangle, \end{aligned} \tag{23}$$

where the expectation is taken with respect to the joint distribution of θ_t and θ_n. For the second term in equation 23, we invoke Young's inequality to bound:

$$\begin{aligned} \mathbb{E}\left\langle \nabla \log \frac{q_t(\theta_t)}{p(\theta_t)}, \nabla \log p(\theta_t) - \nabla \log p(\theta_n) \right\rangle &\leq \frac{1}{2}\mathbb{E}\left\| \nabla \log \frac{q_t(\theta_t)}{p(\theta_t)} \right\|^2 + \frac{1}{2}\mathbb{E}\left\| \nabla \log p(\theta_t) - \nabla \log p(\theta_n) \right\|^2 \\ &= \frac{1}{2}\mathbb{E}\left\| \nabla \log \frac{q_t(\theta_t)}{p(\theta_t)} \right\|^2 + \frac{1}{2}\mathbb{E}\left\| \nabla U(\theta_t) - \nabla U(\theta_n) \right\|^2. \end{aligned}$$

Since potential U is L-Lipschitz smooth, $\|\nabla U(\theta_t) - \nabla U(\theta_n)\|^2 \leq L^2 \|\theta_t - \theta_n\|^2$. Also note that we have set $\nabla U(0) = 0$. Hence

$$\begin{aligned} \frac{1}{2}\mathbb{E}\left\| \nabla U(\theta_t) - \nabla U(\theta_n) \right\|^2 &\leq \frac{L^2}{2}\mathbb{E}\|\theta_t - \theta_n\|^2 \\ &= \frac{L^2}{2}\mathbb{E}\left\| -(t - \eta n)\nabla U(\theta_n) + \sqrt{2}\left(W_t - W_{\eta n}\right) \right\|^2 \\ &= \frac{L^2 (t - \eta n)^2}{2}\mathbb{E}_{\theta \sim q_n}\left[\|\nabla U(\theta)\|^2 \right] + L^2 d(t - \eta n) \\ &\leq \frac{L^4 \eta^2}{2}\mathbb{E}_{\theta \sim q_n}\left[\|\theta\|^2 \right] + L^2 d\eta \end{aligned}$$

Applying this result to equation 23, we obtain an upper bound for $\frac{d}{dt}\mathrm{KL}\left(q_t \| p\right)$ within each iteration:

$$\frac{d}{dt}\mathrm{KL}\left(q_t \| p\right) \leq -\frac{1}{2}\mathbb{E}\left\| \nabla \log \frac{q_t(\theta_t)}{p(\theta_t)} \right\|^2 + \frac{L^4 \eta^2}{2}\mathbb{E}_{\theta \sim q_n}\left[\|\theta\|^2 \right] + L^2 d\eta. \tag{24}$$

Since function U is m-strongly convex, we obtain the following log-Sobolev inequality from the Bakry–Emery criterion (see e.g., Bakry & Emery, 1985)

$$\mathbb{E}_{\theta \sim q_t}\left[\left\| \nabla \log \frac{q_t(\theta)}{p(\theta)} \right\|^2 \right] \geq 2m\mathrm{KL}\left(q_t \| p\right).$$

Therefore,

$$\frac{d}{dt}\mathrm{KL}\left(q_t \| p\right) \leq -m\mathrm{KL}\left(q_t \| p\right) + \frac{L^4 \eta^2}{2}\mathbb{E}_{\theta \sim q_n}\left[\|\theta\|^2 \right] + L^2 d\eta. \tag{25}$$

We prove in the Lemma 5 below that $\mathbb{E}_{\theta\sim q_n}\left[\|\theta\|^2\right] \le \frac{4d}{m}$.

Lemma 5. *For step size $\eta \le \frac{1}{L}$, and for q_n following the update of equation 3 $\forall n > 0$, $\mathbb{E}_{\theta\sim q_n}\left[\|\theta\|^2\right] \le \frac{4d}{m}$.*

Plugging this bound into equation 25, we obtain:

$$\frac{d}{dt}\mathrm{KL}\left(q_t\|p\right) \le -m\left(\mathrm{KL}\left(q_t\|p\right) - \left(2\frac{L^4}{m^2}\eta^2 + \frac{L^2}{m}\eta\right)d\right). \tag{26}$$

Invoking Grönwall's inequality, we obtain:

$$\begin{aligned}\mathrm{KL}\left(q_n\|p\right) &\le e^{-m\eta}\mathrm{KL}\left(q_{n-1}\|p\right) + \left(2\frac{L^4}{m^2}\eta^2 + \frac{L^2}{m}\eta\right)d \\ &\le e^{-m\eta n}\mathrm{KL}\left(q_0\|p\right) + \left(2\frac{L^4}{m^2}\eta^2 + \frac{L^2}{m}\eta\right)d.\end{aligned} \tag{27}$$

This means that $\mathrm{KL}\left(q_n\|p\right)$ is converging exponentially to the level of discretization error.

To obtain an accuracy guarantee of ϵ, we choose a step size of $\eta = \frac{m}{4L^2}\frac{\epsilon}{d}$ and have (for $\epsilon \le d$):

$$\mathrm{KL}\left(q_n\|p\right) \le e^{-m\eta n}\mathrm{KL}\left(q_0\|p\right) + \frac{\epsilon}{2}. \tag{28}$$

When $n \ge \frac{1}{m\eta}\log\frac{2\mathrm{KL}(q_0\|p)}{\epsilon}$, $e^{-m\eta n}\mathrm{KL}\left(q_0\|p\right) \le \frac{\epsilon}{2}$, and therefore $\mathrm{KL}\left(q_n\|p\right) \le \epsilon$.

Plugging the setting of η gives us the upper bound for number of iterations:

$$n = 4\frac{L^2}{m^2}\frac{d}{\epsilon}\log\frac{2\mathrm{KL}\left(q_0\|p\right)}{\epsilon} = \widetilde{\mathcal{O}}\left(\frac{L^2}{m^2}\frac{d}{\epsilon}\right).$$

□

Proof of Lemma 5 We prove Lemma 3 by induction. We first see that for the current choice of initialization, $\mathbb{E}_{\theta\sim q_0}\left[\|\theta\|^2\right] \le \frac{d}{m}$. We then assume that $\mathbb{E}_{\theta\sim q_n}\left[\|\theta\|^2\right] \le \frac{d}{m}$ and prove that $\mathbb{E}_{\theta\sim q_{n+1}}\left[\|\theta\|^2\right] \le \frac{d}{m}$.

We know that

$$\theta_{n+1} = \theta_n - \eta\nabla U(\theta_n) + \sqrt{2}\left(W_{\eta(n+1)} - W_{\eta n}\right).$$

To provide a bound on $\|\theta_{n+1}\|$, we first analyze the term: $\theta_n - \eta\nabla U(\theta_n)$. To this end, we construct a function: $V(\theta) = \frac{1}{2}\|\theta\|^2 - \eta U(\theta)$ and prove that it is $(1-m\eta)$-Lipschitz smooth. Since function U is assumed to be m-strongly convex,

$$\begin{aligned}\langle\nabla V(\theta) - \nabla V(\vartheta), \theta - \vartheta\rangle &= \langle(\theta-\vartheta) - \eta\left(\nabla U(\theta) - U(\vartheta)\right), \theta - \vartheta\rangle \\ &= \|\theta - \vartheta\|^2 - \eta\left\langle\left(\nabla U(\theta) - U(\vartheta)\right), \theta - \vartheta\right\rangle \\ &\le (1-m\eta)\|\theta - \vartheta\|^2.\end{aligned}$$

Therefore, function $V(\theta) = \frac{1}{2}\|\theta\|^2 - \eta U(\theta)$ is $(1-m\eta)$-Lipschitz smooth and satisfy $\nabla V(0) = 0$, which means:

$$\|\theta_n - \eta\nabla U(\theta_n)\| = \|\nabla V(\theta_n)\| \le (1-m\eta)\|\theta_n\|.$$

We are now in a position to bound $\mathbb{E}\|\theta_{n+1}\|^2$:

$$\begin{aligned}\mathbb{E}\left[\|\theta_{n+1}\|^2\right] &= \mathbb{E}\left[\left\|\theta_n - \eta\nabla U(\theta_n) + \sqrt{2}\left(W_{\eta(n+1)} - W_{\eta n}\right)\right\|^2\right] \\ &= \mathbb{E}\left[\|\theta_n - \eta\nabla U(\theta_n)\|^2\right] + 2\eta d \\ &\le (1-m\eta)\mathbb{E}\left[\|\theta_n\|^2\right] + 2\eta d \qquad (29) \\ &= \mathbb{E}\left[\|\theta_n\|^2\right] + \eta\left(2d - m\mathbb{E}\left[\|\theta_n\|^2\right]\right). \qquad (30)\end{aligned}$$

By the inductive hypothesis, $\mathbb{E}\left[\|\theta_n\|^2\right] \leq \frac{4d}{m}$. If $\mathbb{E}\left[\|\theta_n\|^2\right] \geq \frac{2d}{m}$, then $\left(2d - m\mathbb{E}\left[\|\theta_n\|^2\right]\right) \leq 0$, $\mathbb{E}\left[\|\theta_{n+1}\|^2\right] \leq \mathbb{E}\left[\|\theta_n\|^2\right] \leq \frac{4d}{m}$. If $\mathbb{E}\left[\|\theta_n\|^2\right] \leq \frac{2d}{m}$ instead, then we use line 29 and that $\eta \leq \frac{1}{L}$ to obtain: $\mathbb{E}\|\theta_{n+1}\|^2 \leq \left(1 - \frac{m}{L}\right)\frac{2d}{m} + \frac{2d}{L} \leq \frac{4d}{m}$.

Therefore, we have proven that for any $n > 0$, $\mathbb{E}_{\theta \sim q_n}\left[\|\theta\|^2\right] \leq \frac{4d}{m}$ by induction. $\square$