A Proof of the Main Results

Let $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$ be the discrete-time time-inhomogeneous Markov chain generated by Algorithm 1, and let $\boldsymbol{X}(k\eta)$ be the Langevin dynamics (1.1) at time $k\eta$, which satisfies $\boldsymbol{X}(0) = \boldsymbol{X}_0$. Consider the target distribution $\pi = \exp(-\gamma F(\mathbf{x}))/\int \exp(-\gamma F(\mathbf{x}))d\mathbf{x}$, we decompose the 2-Wasserstein distance $\mathcal{W}_2\big(P(\boldsymbol{X}_k), \pi\big)$ into the following two terms based on triangle inequality.

$$\mathcal{W}_2\big(P(\boldsymbol{X}_k), \pi\big) \leq \mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) + \mathcal{W}_2\big(P(\boldsymbol{X}(k\eta), \pi)\big). \tag{A.1}$$

The first term in (A.1) stands for the discretization error between the continuous-time Langevin dynamics at time $k\eta$ and the k-th iteration of SVRG-LD in 2-Wasserstein distance. The second term describes the convergence of the probability density of Markov process $\{\boldsymbol{X}(k\eta)\}_{t\geq 0}$ to its stationary distribution, and is referred to as the ergordicity of a Markov process. In what follows, we aim at establishing upper bounds for these two terms, respectively.

A.1 Proof of Theorem 4.3

We first study the discretization error between the distribution of continuous Markov process at time $k\eta$ and that of the discrete iterate at the k-th update in Algorithm 1.

Lemma A.1. Under Assumptions 4.1 and 4.2, consider $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$ generated by Algorithm 1 with initial point $\boldsymbol{X}_0 = \mathbf{0}$. The 2-Wasserstein distance between distributions of the iterate $\boldsymbol{X}_k$ in Algorithm 1 and the point $\boldsymbol{X}(k\eta)$ in the Langevin dynamic sequence (1.1) is upper bounded by

$$\mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) \leq D_A\bigg[\bigg(\frac{6\gamma m^2 M^2(n-B)}{B(n-1)} + 3\gamma M^2\bigg)(M^2 D_B^2 + G^2)k\eta^3 + \bigg(\frac{2dmM^2(n-B)}{B(n-1)} + M^2 d\bigg)k\eta^2\bigg]^{1/4},$$

where $D_A = 4\sqrt{3/2 + (2b + d/\gamma)k\eta}$ and $D_B = \sqrt{2(1 + 1/b)(a + G^2 + d/\gamma)}$.

In what follows, we show that the continuous-time process $\{\boldsymbol{X}(t)\}_{t\geq 0}$ converges to its stationary distribution with linear rate.

Lemma A.2. Under Assumptions 4.1 and 4.2, the continuous-time Markov chain $\boldsymbol{X}(t)$ generated by Langevin dynamics (1.1) converges exponentially to the stationary distribution π, i.e.,

$$\mathcal{W}_2\big(P(\boldsymbol{X}(t), \pi)\big) \leq D_4 e^{-\frac{t}{\gamma D_5}},$$

where both D_4 and D_5 are in the order of $\exp\big(\widetilde{O}(\gamma + d)\big)$.

It can be seen that the 2-Wasserstein distance diminishes exponentially fast, and the crucial factor that determines the rate is the parameter in the exponential term, i.e., D_5. It is worth noting that D_5 has an exponential dependence on γ and d.

Proof of Theorem 4.3. In previous parts, we have shown the upper bounds on terms $\mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big)$ and $\mathcal{W}_2\big(P(\boldsymbol{X}(k\eta), \pi)\big)$ in (A.1), thus we are ready to prove the main theorem 4.3. It can be seen that by combining Lemmas A.1 and A.2, together with the triangle inequality and the fact that $(n-B)/(n-1) \leq 1$, we have

$$\begin{aligned}
\mathcal{W}_2\big(P(\boldsymbol{X}_k), \pi\big) &\leq \mathcal{W}_2\big(P(\mathbf{x}_k), P(\boldsymbol{X}(k\eta))\big) + \mathcal{W}_2\big(P(\boldsymbol{X}(k\eta), \pi)\big) \\
&\leq D_1\bigg[D_2\bigg(\frac{2m^2}{B} + 1\bigg)k\eta^3 + D_3\bigg(\frac{2m}{B} + 1\bigg)k\eta^2\bigg]^{1/4} + D_4 e^{-\frac{k\eta}{\gamma D_5}},
\end{aligned}$$

where

$$\begin{aligned}
&D_1 = D_A = 4\sqrt{3/2 + (2b + d/\gamma)k\eta}, \\
&D_2 = 3\gamma M^2(M^2 D_B^2 + G^2) = 3\gamma M^2\big(2M^2(1 + 1/b)(a + G^2 + d/\gamma) + G^2\big), \\
&D_3 = M^2 d.
\end{aligned}$$

This completes the proof. □

A.2 Proof of Theorem 4.6

Similar to the proof of SVRG-LD, we first present the following lemma that characterizes the discretization error between the continuous Markov process at time $k\eta$ and the discrete iterate at the k-th update in Algorithm 2.

Lemma A.3. Under Assumptions 4.1 and 4.2, consider $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$ generated by Algorithm 2 with initial point $\boldsymbol{X}_0 = \mathbf{0}$. The 2-Wasserstein distance between distributions of the iterate $\boldsymbol{X}_k$ in Algorithm 1 and the point $\boldsymbol{X}(k\eta)$ in the Langevin dynamic sequence (1.1) is upper bounded by

$$\mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) \leq D_A\bigg[\bigg(\frac{144n^2(n-B)M^2}{B^3(n-1)} + 3M^2\bigg)(M^2D_B^2 + G^2)\gamma k\eta^3 + \bigg(\frac{4n(n-B)M^2d}{B^2(n-1)} + M^2d\bigg)k\eta^2\bigg]^{1/4},$$

where $D_A = 4\sqrt{3/2 + (2b + d/\gamma)k\eta}$ and $D_B = \sqrt{2(1 + 1/b)(a + G^2 + d/\gamma)}$.

In terms of the sequence of continuous-time Langevin dynamics $\{\boldsymbol{X}(t)\}_{t\geq 0}$, Lemma A.2 is also applicable. Thus we are able to complete the proof by combining Lemmas A.2 and A.3.

Proof of Theorem 4.6. Straightforwardly, combining Lemmas A.3 and A.2 together with triangle inequality, and use the fact that $(n-B)/(n-1) \leq 1$, we obtain

$$\begin{aligned}\mathcal{W}_2\big(P(\boldsymbol{X}_k), \pi\big) &\leq \mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) + \mathcal{W}_2\big(P(\boldsymbol{X}(k\eta), \pi)\big)\\ &\leq D_1\bigg[D_2\bigg(\frac{48n^2}{B^3} + 1\bigg)k\eta^3 + D_3\bigg(\frac{4n}{B^2} + 1\bigg)k\eta^2\bigg]^{1/4} + D_4e^{-\frac{k\eta}{\gamma D_5}},\end{aligned}$$

where D_1, D_2, D_3, D_4 and D_5 are identical to those in Theorem 4.3. This completes the proof. □

B Proof of Corollaries

In this section, we provide the proofs of our corollaries in Section 4.

Proof of Corollary 4.4. In order to ensure the ϵ-accuracy in 2-Wasserstein distance, we set

$$\begin{aligned}&D_1\bigg[D_2\Big(\frac{2m^2}{B} + 1\Big)k\eta^3 + D_3\Big(\frac{2m}{B} + 1\Big)k\eta^2\bigg]^{1/4} = \frac{\epsilon}{2},\\ &D_4e^{-\frac{k\eta}{\gamma D_5}} = \frac{\epsilon}{2}.\end{aligned} \tag{B.1}$$

Based on the second equation in (B.1), it can be derived that

$$T \triangleq k\eta = \gamma D_5 \log\Big(\frac{2D_4}{\epsilon}\Big).$$

Then, note that if we have $a + b = c$ for positive constants a, b and c, it either follows that $c \leq 2a$ or $c \leq 2b$. Then we have the following according to the first equation in (B.1),

$$\eta \geq \min\bigg\{\sqrt{\frac{\epsilon^4}{32D_1^4D_2(2m^2/B + 1)T}}, \frac{\epsilon^4}{32D_1^4D_3(2m/B + 1)T}\bigg\}.$$

Combine the above two results, we have

$$k = \frac{T}{\eta} \leq \gamma D_5 \log\Big(\frac{2D_4}{\epsilon}\Big)\bigg(\sqrt{\frac{32D_1^4D_2(2m^2/B + 1)T}{\epsilon^4}} + \frac{32D_1^4D_3(2m/B + 1)T}{\epsilon^4}\bigg).$$

From Lemma A.2, we know that $D_5 = \exp\big(\widetilde{O}(\gamma + d)\big)$, thus the required iteration number k exponentially depends on dimension d and inverse temperature γ. Then, we focus on figuring out dependence on ϵ. Ignoring constants that have no dependence in ϵ and only polynomially depends on γ and d, we have

$$k = \widetilde{O}\Big(\frac{m/B^{1/2} + 1}{\epsilon^2} + \frac{m/B + 1}{\epsilon^4}\Big) \cdot \exp\big(\widetilde{O}(\gamma + d)\big).$$

Note that we have to compute full gradient for k/m times, thus the total gradient complexity is

$$T_g \leq kB + n(k/m \vee 1) \leq kB + \frac{kn}{m} + n.$$

Obviously, minimizing T_g requires $mB = n$. Then, the gradient complexity becomes

$$T_g \leq 2kB + n = \widetilde{O}\Big(\frac{nB^{-1/2}}{\epsilon^2} + \frac{n/B + B}{\epsilon^4} + n\Big) \cdot \exp\big(\widetilde{O}(\gamma + d)\big).$$

Let $B = O(n^{1/2})$, we straightforwardly obtain

$$T_g = \widetilde{O}\Big(n + \frac{n^{3/4}}{\epsilon^2} + \frac{n^{1/2}}{\epsilon^4}\Big) \cdot \exp\big(\widetilde{O}(\gamma + d)\big),$$

which completes the proof. □

Proof of Corollary 4.7. Analogous to the proof of Corollary 4.4, we set

$$D_1\bigg[D_2\bigg(\frac{48n^2}{B^3} + 1\bigg)k\eta^3 + D_3\bigg(\frac{4n}{B^2} + 1\bigg)k\eta^2\bigg]^{1/4} = \frac{\epsilon}{2}$$
$$D_4 e^{-\frac{k\eta}{\gamma D_5}} = \frac{\epsilon}{2}.$$

From the second equation, we obtain

$$k\eta = \gamma D_5 \log\bigg(\frac{2D_4}{\epsilon}\bigg).$$

Let $T = k\eta$, the first equation yields that

$$\eta \geq \min\bigg\{\sqrt{\frac{\epsilon^4}{32D_1^4 D_2(48n^2/B^3 + 1)T}}, \frac{\epsilon^4}{32D_1^4 D_3(4n/B^2 + 1)T}\bigg\}.$$

Then the required number of iterations satisfies

$$k = \frac{T}{\eta} \leq \gamma D_5 \log\bigg(\frac{2D_4}{\epsilon}\bigg)\bigg(\sqrt{\frac{32D_1^4 D_2(48n^2/B^3 + 1)T}{\epsilon^4}} + \frac{32D_1^4 D_3(4n/B^2 + 1)T}{\epsilon^4}\bigg).$$

From Lemma A.2, we know that $D_5 = \exp\big(\widetilde{O}(\gamma + d)\big)$, the complexity k must exponentially depends on dimension d and inverse temperature γ. Then, we focus on figuring out the dependence on ϵ. Ignoring constants that have no dependence in ϵ, we have

$$k = \widetilde{O}\bigg(\frac{n/B^{3/2} + 1}{\epsilon^2} + \frac{n/B^2 + 1}{\epsilon^4}\bigg).$$

Then the corresponding gradient complexity is

$$T_g = n + kB = \widetilde{O}\bigg(n + \frac{n/B^{1/2}}{\epsilon^2} + \frac{n/B + B}{\epsilon^4}\bigg).$$

Plugging the dependence on d and γ, we complete the proof.

□

C Proof of Technical Lemmas

In this section, we prove the technical lemmas in Appendix A.

C.1 Proof of Lemma A.1

We first lay out the following 5 lemmas which is useful for proving Lemma A.1.

Lemma C.1. For all $\mathbf{x} \in \mathbb{R}^d$ and $i = 1, \ldots, n$, we have

$$\|\nabla f_i(\mathbf{x})\|_2 \leq M\|\mathbf{x}\|_2 + G.$$

Moreover, it follows that

$$\|\nabla f_i(\mathbf{x})\|_2^2 \leq 2M\|\mathbf{x}\|_2^2 + 2G.$$

Lemma C.2. Under Assumptions 4.1 and 4.2, for sufficiently small step size η, suppose the initial point is chosen at $\boldsymbol{X}_0 = \mathbf{0}$, the expectation of the ℓ^2 norm of the iterates generated by Algorithm 1 is bounded by

$$\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] \leq 2\bigg(1 + \frac{1}{b}\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg) \triangleq D_B.$$

Lemma C.3. (Bolley and Villani, 2005) For any two probability measures P and Q, if they have finite second moments, the following holds,

$$\mathcal{W}_2(Q, P) \leq \Lambda(\sqrt{D_{KL}(Q||P)} + \sqrt[4]{D_{KL}(Q||P)}),$$

where $\Lambda = 2\inf_{\lambda>0}\sqrt{1/\lambda(3/2 + \log \mathbb{E}_P[e^{\lambda\|\mathbf{x}\|_2^2}])}$, where $\mathbf{x}$ satisfies probability measure P.

Lemma C.4. Under Assumptions 4.1 and 4.2, for sufficiently small step size η and $\beta \geq 2/m$, we have

$$\log \mathbb{E}[\exp(\|\boldsymbol{X}(t)\|_2^2)] \leq \|\mathbf{X}_0\|_2^2 + (2b + d/\gamma)k\eta,$$

where we consider the fact that $\eta \leq 1$, and require that $\gamma > 4$.

Lemma C.5. Under Assumption 4.1, we have the following upper bound on the variance of semi-stochastic gradient $\widetilde{\nabla}_k$ in the SVRG-LD update,

$$\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\boldsymbol{X}_k)\|_2^2] \leq \frac{M^2(n-B)}{B(n-1)}\mathbb{E}\|\boldsymbol{X}_k - \widetilde{\boldsymbol{X}}\|_2^2.$$

In order to analyze the long-time behaviour of the error between the discrete-time algorithm and continuous-time Langevin dynamics, we follow the similar technique used in Dalalyan (2016); Raginsky et al. (2017); Xu et al. (2018), in which a continuous-time Markov process $\{\boldsymbol{D}(t)\}_{t\geq 0}$ is introduced to describe the numerical approximation sequence $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$. Define

$$d\boldsymbol{D}(t) = -b(\boldsymbol{D}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t), \tag{C.1}$$

where $b(\boldsymbol{D}(t)) = \sum_{k=0}^{\infty}\widetilde{\nabla}_k \mathbb{1}\{t \in \big[\eta k, \eta(k+1)\big)\}$. Integrating (C.1) on interval $\big[\eta k, \eta(k+1)\big)$ yields

$$\boldsymbol{D}(\eta(k+1)) = \boldsymbol{D}(\eta k) - \eta\nabla F(\boldsymbol{D}(\eta k)) + \sqrt{2\eta\gamma^{-1}}\cdot\boldsymbol{\epsilon}_k,$$

where $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \mathbf{I}_{d\times d})$ and $\widetilde{\nabla}_k$ is the semi-stochastic gradient at k-th iteration of VR-SGLD. This implies that the distribution of random vector $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k, \ldots)$ is equivalent to that of $(\boldsymbol{D}(\eta), \ldots, \boldsymbol{D}(\eta k), \ldots)$. Note that (C.1) is not a time-homogeneous Markov chain since the semi-stochastic gradient $b(\boldsymbol{D}(t))$ also depends on some historical iterates. However, Gyöngy (1986) showed that one can construct an alternative Markov chain which enjoys the same one-time marginal distribution as that of $\boldsymbol{D}(t)$, which is formulated as follows,

$$d\widetilde{\boldsymbol{D}}(t) = -\widetilde{b}(\widetilde{\boldsymbol{D}}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t),$$

where $\widetilde{b}(\widetilde{\boldsymbol{D}}(t)) = \mathbb{E}[b(\boldsymbol{D}(t))|\widetilde{D}(t) = D(t)]$. Then we let $\mathbb{P}_t$ denote the distribution of $\widetilde{\boldsymbol{D}}(t)$, which is identical to that of $\boldsymbol{D}(t)$. Recall the SDE of Langevin dynamics, i.e.,

$$d\boldsymbol{X}(t) = -q(\boldsymbol{X}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t),$$

where $q(\boldsymbol{X}(t)) = \nabla F(\boldsymbol{X}(t))$ and define by $\mathbb{Q}_t$ the distribution of $\boldsymbol{X}(t)$. Now, we have constructed two continuous continuous process. Thus, the Radon-Nykodim derivative of $\mathbb{P}_t$ with respective to $\mathbb{Q}_t$ can be obtained by the Girsanov formula

$$\frac{d\mathbb{P}_t}{d\mathbb{Q}_t}(\widetilde{\boldsymbol{D}}(s)) = \exp\bigg\{\int_0^t (q(\widetilde{\boldsymbol{D}}(s)) - b(\widetilde{\boldsymbol{D}}(s)))^\top d\boldsymbol{B}(s) - \frac{\gamma}{4}\int_0^t \|q(\widetilde{\boldsymbol{D}}(s)) - b(\widetilde{\boldsymbol{D}}(s))\|_2^2 ds\bigg\}.$$

This suggests that the KL divergence between $\mathbb{P}_t$ and $\mathbb{Q}_t$ has the following form

$$D_{KL}(\mathbb{Q}_t||\mathbb{P}_t) = -\mathbb{E}\bigg[\log\bigg(\frac{d\mathbb{P}_t}{d\mathbb{Q}_t}(\widetilde{\boldsymbol{D}}(s))\bigg)\bigg] = \frac{\gamma}{4}\int_0^t \mathbb{E}\big[\|q(\widetilde{\boldsymbol{D}}(s)) - b(\widetilde{\boldsymbol{D}}(s))\|_2^2\big] ds. \tag{C.2}$$

This result gives us an opportunity to estimate the 2-Wasserstein distance $\mathcal{W}_2\big(P(\boldsymbol{X}_k, P(\boldsymbol{X}(k\eta))\big)$, since we are able to apply KL divergence $D_{KL}(\mathbb{Q}_{k\eta}||\mathbb{P}_{k\eta})$ to generate an upper bound based on Lemma C.3. Now, we are going to complete the proof for Lemma A.1 in the following.

Proof of Lemma A.1. Denote P_k, Q_k as the probability density functions of $\boldsymbol{X}_k$ and $\boldsymbol{X}(k\eta)$ respectively. By Lemma C.3, we know that the 2-Wasserstein distance is upper bounded as follows,

$$\mathcal{W}_2(Q_k, P_k) \le \Lambda(\sqrt{D_{KL}(Q_k||P_k)} + \sqrt[4]{D_{KL}(Q_k||P_k)}).$$

Moreover, by data-processing theorem in terms of KL divergence, we have

$$\begin{aligned} D_{KL}(Q_k||P_k) \le D_{KL}(\mathbb{Q}_{k\eta}||\mathbb{P}_{k\eta}) &= \frac{\gamma}{4}\int_0^{k\eta} \mathbb{E}\big[\|q(\widetilde{\boldsymbol{D}}(s)) - b(\widetilde{\boldsymbol{D}}(s))\|_2^2\big] ds \\ &= \frac{\gamma}{4}\int_0^{k\eta} \mathbb{E}\big[\|q(\boldsymbol{D}(s)) - b(\boldsymbol{D}(s))\|_2^2\big] ds, \end{aligned}$$

where the second equality holds due to the fact that $\widetilde{\boldsymbol{D}}(s)$ and $\boldsymbol{D}(s)$ have same one-time distribution. Note that $\boldsymbol{D}(k\eta)$ is generated based on $\boldsymbol{X}_k$. By definition, we know that $b(\boldsymbol{D}(s))$ is a step function and remains constant when $s \in [\eta k, \eta(k+1))$ for any k, and $q(\boldsymbol{D}(s))$ is a continuous function for any s. Based on this observation, it follows that

$$\begin{aligned} &\int_0^{\eta k} \mathbb{E}\big[\|q(\boldsymbol{D}(s)) - b(\boldsymbol{D}(s))\|_2^2\big] ds \\ &\quad= \sum_{v=0}^{k-1}\int_{v\eta}^{\eta(v+1)} \mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{D}(s))\|_2^2\big] ds \\ &\quad\le 2\eta\sum_{v=0}^{k-1}\mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2\big] + 2\sum_{v=0}^{k-1}\int_{v\eta}^{\eta(v+1)} \mathbb{E}\big[\|\nabla F(\boldsymbol{D}(v\eta)) - \nabla F(\boldsymbol{D}(s))\|_2^2\big] ds, \end{aligned}$$

where the second inequality is due to Jensen's inequality and the convexity of function $\|\cdot\|_2^2$, and $\nabla F(\boldsymbol{X}_v) = \nabla F(\boldsymbol{D}(v\eta))$ denotes the gradient of $F(\cdot)$ at $\boldsymbol{X}_v$. Combine the above results we obtain

$$\begin{aligned} D_{KL}(Q_k||P_k) \le{}& \frac{\gamma\eta}{2}\sum_{v=0}^{k-1}\mathbb{E}[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2] \\ &+ \frac{\gamma}{2}\sum_{v=0}^{k-1}\int_{v\eta}^{\eta(v+1)} \mathbb{E}\big[\|\nabla F(\boldsymbol{D}(v\eta)) - \nabla F(\boldsymbol{D}(s))\|_2^2\big] ds, \end{aligned} \tag{C.3}$$

where the first term on the R.H.S. can be further bounded by

$$\frac{\gamma\eta}{2}\sum_{v=0}^{k-1}\mathbb{E}[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2] \le \frac{\gamma\eta}{2}\sum_{i=0}^{s}\sum_{j=0}^{m-1}\mathbb{E}[\|\widetilde{\nabla}_{im+j} - \nabla F(\boldsymbol{X}_{im+j})\|_2^2],$$

where we use the fact that $k = sm + \ell \leq (s+1)m$ for some $\ell = 0, 1, \ldots, m-1$. Applying Lemma C.5, the inner summation satisfies

$$\sum_{j=0}^{m-1} \mathbb{E}[\|\widetilde{\nabla}_{im+j} - \nabla F(\boldsymbol{X}_{im+j})\|_2^2] \leq \sum_{j=0}^{m-1} \frac{M^2(n-B)}{B(n-1)} \mathbb{E}\big\|\boldsymbol{X}_{im+j} - \widetilde{\boldsymbol{X}}^{(i)}\big\|_2^2. \tag{C.4}$$

Note that we have

$$\begin{aligned}
&\mathbb{E}\big\|\boldsymbol{X}_{im+j} - \widetilde{\boldsymbol{X}}^{(i)}\big\|_2^2 \\
&\quad= \mathbb{E}\bigg\|\sum_{u=0}^{j-1} \eta\big(\nabla f_{i_{im+u}}(\boldsymbol{X}_{im+u}) - \nabla f_{i_{im+u}}(\widetilde{\boldsymbol{X}}^{(i)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(i)})\big) - \sum_{u=0}^{j-1} \sqrt{\frac{2\eta}{\gamma}}\epsilon_j\bigg\|_2^2 \\
&\quad\leq j\sum_{u=0}^{j-1} \mathbb{E}\big[2\eta^2\big\|\nabla f_{i_{im+u}}(\boldsymbol{X}_{im+u}) - \nabla f_{i_{im+u}}(\widetilde{\boldsymbol{X}}^{(i)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(i)})\big\|_2^2\big] + \sum_{u=0}^{j-1} \frac{4\eta d}{\gamma} \\
&\quad\leq j\sum_{u=0}^{j-1} \mathbb{E}\big[6\eta^2\big(\|\nabla f_{i_{im+u}}(\boldsymbol{X}_{im+u})\|_2^2 + \|\nabla f_{i_{im+u}}(\widetilde{\boldsymbol{X}}^{(i)})\|_2^2 + \|\nabla F(\widetilde{\boldsymbol{X}}^{(i)})\|_2^2\big)\big] + \sum_{u=0}^{j-1} \frac{4\eta d}{\gamma} \\
&\quad\leq 36j^2\eta^2(M^2D_B^2 + G^2) + \frac{4j\eta d}{\gamma},
\end{aligned} \tag{C.5}$$

where the first and second inequalities follow from Young's inequality and the last one follows from Lemma C.1 and Lemma C.2, and $D_B = \sqrt{2(1+1/b)(a + G^2 + d/\gamma)}$. Submit (C.5) back into (C.4) we have

$$\begin{aligned}
\sum_{j=0}^{m-1} \mathbb{E}[\|\widetilde{\nabla}_{im+j} - \nabla F(\boldsymbol{X}_{im+j})\|_2^2] &\leq \sum_{j=0}^{m-1} \frac{4M^2(n-B)}{B(n-1)}\bigg(9j^2\eta^2(M^2D_B^2 + G^2) + \frac{j\eta d}{\gamma}\bigg) \\
&\leq \frac{4M^2(n-B)}{B(n-1)}\bigg(3m^3\eta^2(M^2D_B^2 + G^2) + \frac{m^2\eta d}{\gamma}\bigg).
\end{aligned} \tag{C.6}$$

Submitting (C.6) into (C.3) yields

$$\sum_{v=0}^{k-1} \mathbb{E}[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2] \leq \frac{4kM^2(n-B)}{B(n-1)}\bigg(3m^2\eta^2(M^2D_B^2 + G^2) + \frac{m\eta d}{\gamma}\bigg). \tag{C.7}$$

Next, we are going to upper bound the second term on the R.H.S of (C.3). According to the smoothness assumption on $F(\mathbf{x})$, we have

$$\mathbb{E}[\|\nabla F(\boldsymbol{D}(v\eta)) - \nabla F(\boldsymbol{D}(s))\|_2^2] \leq M^2\mathbb{E}[\|\boldsymbol{D}(s) - \boldsymbol{D}(v\eta)\|_2^2],$$

which yields that

$$\begin{aligned}
&\sum_{v=0}^{k-1} \int_{v\eta}^{\eta(v+1)} \mathbb{E}\big[\|\nabla F(\boldsymbol{X}_v) - \nabla F(\boldsymbol{N}(s))\|_2^2\big]ds \\
&\qquad\leq \sum_{v=0}^{k-1} \int_{v\eta}^{\eta(v+1)} M^2\mathbb{E}\big[\|\boldsymbol{D}(s) - \boldsymbol{D}(v\eta)\|_2^2\big]ds \\
&\qquad= \sum_{v=0}^{k-1} \int_{v\eta}^{\eta(v+1)} M^2\bigg((s-v\eta)^2\mathbb{E}[\|\widetilde{\nabla}_v\|_2^2] + \frac{2(s-v\eta)d}{\gamma}\bigg)ds \\
&\qquad\leq \frac{M^2\eta^3}{3}\sum_{v=0}^{k-1} \mathbb{E}[\|\widetilde{\nabla}_v\|_2^2] + \frac{2kM^2\eta^2 d}{\gamma}.
\end{aligned} \tag{C.8}$$

By Lemma C.1, we know that

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla}_v\|_2^2] &= \mathbb{E}\bigg[\bigg\|1/B\sum_{i_k\in I_k}\Big(\nabla f_{i_k}(\boldsymbol{X}_v) - \nabla f_{i_k}(\widetilde{\boldsymbol{X}}^{(s)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(s)})\Big)\bigg\|_2^2\bigg] \\
&\le 3\mathbb{E}\big[(M\|\boldsymbol{X}_v\|_2+G)^2 + 2(M\|\widetilde{\boldsymbol{X}}^{(s)}\|_2+G)^2\big] \\
&\le 6M^2\mathbb{E}[\|\boldsymbol{X}_v\|_2^2] + 12N^2\mathbb{E}[\|\widetilde{\boldsymbol{X}}^{(s)}\|_2^2] + 18G^2 \\
&\le 18M^2D_B^2 + 18G^2,
\end{aligned}$$

where $D_B = \sqrt{2(1+1/b)(a+G^2+d/\gamma)}$ is defined in Lemma C.2, and the last second inequality follows from the fact that $(M\|\boldsymbol{X}_v\|_2+G)^2 \le 2M^2\|\boldsymbol{X}_v\|_2^2 + 2G^2$. Thus, combining (C.3), (C.8), (C.7), we arrive at

$$\begin{aligned}
D_{KL}(Q_k||P_k) &\le \frac{2k\eta\gamma M^2(n-B)}{B(n-1)}\bigg(3m^2\eta^2(M^2D_B^2+G^2) + \frac{m\eta d}{\gamma}\bigg) \\
&\quad + \frac{\gamma}{2}\bigg(6M^2k\eta^3(M^2D_B^2+G^2) + \frac{2kM^2\eta^2 d}{\gamma}\bigg) \\
&= \bigg(\frac{6m^2M^2(n-B)}{B(n-1)} + 3\gamma M^2\bigg)(M^2D_B^2+G^2)k\eta^3 \\
&\quad + \bigg(\frac{2dmM^2(n-B)}{B(n-1)} + M^2d\bigg)k\eta^2 \qquad \text{(C.9)}
\end{aligned}$$

Combining (C.9) and Lemma C.3, assume that $D_{KL}(Q_k||P_k) \le 1$, and choose $\lambda = 1$ in Lemma C.3 we obtain

$$\begin{aligned}
&\mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) \\
&\le D_A\bigg[\bigg(\frac{6\gamma m^2M^2(n-B)}{B(n-1)} + 3\gamma M^2\bigg)(M^2D_B^2+G^2)k\eta^3 + \bigg(\frac{2dmM^2(n-B)}{B(n-1)} + M^2d\bigg)k\eta^2\bigg]^{1/4},
\end{aligned}$$

where $D_A = 2\Lambda = 4\sqrt{3/2 + (2b+d/\gamma)k\eta}$ since $\|\boldsymbol{X}_0\|_2 = 0$. □

C.2 Proof of Lemma A.2

In the following, we adopt the method in Bakry et al. (2013) to show the exponential ergodicity of Langevin diffusion (1.1). In detail, following Bakry et al. (2013), we show the exponential decay in terms of KullbackLeibler divergence (KL divergence) between the probability measure $P_L^t(\cdot)$ and the stationary distribution π, characterize the convergence rate of Langevin dynamics, and link the 2-Wasserstein distance and KL divergence using Otto-Villani theorem (Bakry et al., 2013). We first present the following lemma, which is necessary for the estimation of constant D_3 in Lemma A.2.

Lemma C.6. (Raginsky et al. (2017)) Consider Langevin diffusion (1.1), under Assumptions 4.1 and 4.2, its stationary distribution π satisfies the logarithmic Sobolev inequality with constant C, i.e., for any function h such that $\int_{\mathbb{R}^d} h|\log h|d\pi < \infty$ and $\int_{\mathbb{R}^d} h^2 d\pi = 1$, we have

$$\int_{\mathbb{R}^d} 2h^2\log h d\pi \le 2\Gamma\cdot\int_{\mathbb{R}}^d \|\nabla h\|_2^2 d\pi, \qquad \text{(C.10)}$$

where $\Gamma = \exp\big(\widetilde{O}(\gamma+d)\big)$.

Proof of Lemma A.2. By Lemma C.6, the stationary distribution π satisfies logarithmic Sobolev inequality with constant Γ. According to Bakry et al. (2013) (Theorem 5.2.1), we know that for Langevin diffusion (1.1), the KL divergence between probability measure of $\boldsymbol{X}(t)$ and the stationary distribution π satisfies the following inequality for any $t \ge 0$,

$$D\big(P_L^t(\cdot)\|\pi\big) \le D\big(P_L^0(\cdot)\|\pi\big)e^{-\frac{2t}{\gamma\Gamma}}. \qquad \text{(C.11)}$$

where Γ is the constant in logarithmic Sobolev inequality. It can be seen that the above result gives the form of exponential decay, and the corresponding rate relies on the constant Γ, which is specified in Lemma C.6.

Moreover, according to Bakry et al. (2013) (Theorem 9.6.1), it can be seen that if (C.10) holds for stationary distribution π with constant Γ, we have the following hold for probability measure $P_L^t(\cdot)$,

$$\mathcal{W}_2(P_L^t(\cdot),\pi) \leq \sqrt{2\Gamma \cdot D\big(P_L^t(\cdot)\|\pi\big)}, \tag{C.12}$$

where $\mathcal{W}_2(u,v)$ is the 2-Wasserstein distance between probability measures u and v. Submit (C.12) into (C.11), we have the following

$$\mathcal{W}_2(P_L^t(\cdot),\pi) \leq \sqrt{2\Gamma \cdot D\big(P_L^0(\cdot)\|\pi\big)}e^{-\frac{t}{\gamma\Gamma}}.$$

Let $D_4 = \sqrt{2\Gamma \cdot D\big(P(\boldsymbol{X}(0))\|\pi\big)}$ and $D_5 = \Gamma$, we have

$$\mathcal{W}_2(P_L^t(\cdot),\pi) \leq D_4 e^{-\frac{t}{\gamma D_5}},$$

which completes the proof.

□

C.3 Proof of Lemma A.3

We first lay out the following Lemmas which will be used to prove Lemma A.3

Lemma C.7. Under Assumptions 4.1 and 4.2, for sufficiently small step size η, suppose the initial point is $\boldsymbol{X}_0 = \boldsymbol{0}$, the expectation of the squared ℓ^2 norm of the iterates generated by Algorithm 2 is bounded by

$$\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] \leq 2\bigg(1+\frac{1}{b}\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg) = D_B.$$

Lemma C.8. Under Assumption 4.1, we have the following upper bound on the variance of semi-stochastic gradient $\widetilde{\nabla}_k$ in the SAGA-LD update,

$$\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\boldsymbol{X}_k)\|_2^2] \leq \frac{n-B}{B(n-1)}\mathbb{E}\|\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\mathbf{G}}_{i_k}\|_2^2.$$

Similar to the proof of Lemma A.1, we have two continuous Markov chains, one of them is generated by the Langevin dynamics, i.e.,

$$d\boldsymbol{X}(t) = -q(\boldsymbol{X}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t),$$

where $q(\boldsymbol{X}(t)) = \nabla f(\boldsymbol{X}(t))$, and the other one, denoted as $\{\boldsymbol{H}(t)\}_{t\geq 0}$, follows from the iterate sequence $\{\boldsymbol{X}_k\}_{k=0,1,\ldots,K}$ generated by Algorithm 2, and takes the following form

$$d\boldsymbol{H}(t) = -h(\boldsymbol{H}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t),$$

where the drift term $h(\boldsymbol{H}(t)) = \widetilde{\nabla}_k$ is defined in Algorithm 2. Similar to SVRG-LD, $\{\boldsymbol{H}(t)\}_{t\geq 0}$ does not form a Markov Chain since the drift term $h(\boldsymbol{H}(t))$ depends on some history iterates $\{\boldsymbol{H}(\tau), \tau \leq t\}$. However, we can again construct a Markov chain $\{\widetilde{\boldsymbol{H}}(t)\}_{t\geq 0}$ which possesses the identical one-time distribution of $\{\boldsymbol{H}(t)\}_{t\geq 0}$. $\{\widetilde{\boldsymbol{H}}(t)\}_{t\geq 0}$ is defined by the following SDE

$$d\widetilde{\boldsymbol{H}}(t) = -\widetilde{h}(\widetilde{\boldsymbol{H}}(t))dt + \sqrt{2\gamma^{-1}}d\boldsymbol{B}(t),$$

where $\widetilde{h}(\widetilde{\boldsymbol{H}}(t)) = \mathbb{E}[h(\boldsymbol{H}(t))|\widetilde{\boldsymbol{H}}(t) = \boldsymbol{H}(t)]$. Let $\mathbb{P}_t$ and $\mathbb{Q}_t$ denote the distributions of $\widetilde{\boldsymbol{H}}(t)$ and $\boldsymbol{X}(t)$ respectively. Using the Radon-Nykodim derivative of $\mathbb{P}_t$ with respect to $\mathbb{Q}_t$, we obtain the following formula in terms of the KL divergence between $\mathbb{P}_t$ and $\mathbb{Q}_t$,

$$D_{KL}(\mathbb{Q}_t||\mathbb{P}_t) = -\mathbb{E}\bigg[\log\bigg(\frac{d\mathbb{P}_t}{d\mathbb{Q}_t}(\widetilde{\boldsymbol{H}}(t))\bigg)\bigg] = \frac{\gamma}{4}\int_0^t \mathbb{E}\big[\|q(\widetilde{\boldsymbol{H}}(s)) - h(\widetilde{\boldsymbol{H}}(s)\|\big].$$

Then we are going to complete the proof.

Proof of Lemma A.3. Note that $h(\boldsymbol{H}(s))$ is a step function and remains constant when $s \in [v\eta, (v+1)\eta]$ for any v, then we have

$$\begin{aligned}
\int_0^{k\eta} \mathbb{E}\big[\|q(\widetilde{\boldsymbol{H}}(s)) - h(\widetilde{\boldsymbol{H}}(s))\|_2^2\big] &= \int_0^{k\eta} \mathbb{E}\big[\|q(\boldsymbol{H}(s)) - h(\boldsymbol{H}(s))\|_2^2\big] \\
&= \sum_{v=0}^{k-1} \int_{v\eta}^{(v+1)\eta} \mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{H}(s))\|_2^2\big] \\
&\le 2\sum_{v=0}^{k-1} \int_{v\eta}^{(v+1)\eta} \mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2\big] \\
&\quad + 2\sum_{v=0}^{k-1} \int_{v\eta}^{(v+1)\eta} \mathbb{E}\big[\|\nabla F(\boldsymbol{H}(v\eta)) - \nabla F(\boldsymbol{H}(s))\|_2^2\big], \qquad \text{(C.13)}
\end{aligned}$$

where the first equality holds since $\widetilde{\boldsymbol{H}}(s)$ and $\boldsymbol{H}(s)$ has identical distribution, and the inequality is by Young's inequality and the fact that $\boldsymbol{X}_v = \boldsymbol{H}(v\eta)$. In terms of the first term on the R.H.S of the above inequality, the following holds according to Lemma C.8,

$$\mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2\big] \le \frac{n-B}{B(n-1)} \mathbb{E}\big[\|\nabla f_{i_v}(\boldsymbol{X}_v) - \widetilde{\boldsymbol{G}}_{i_v}\|_2^2\big].$$

Note that $\widetilde{\boldsymbol{G}}_{i_v} = \nabla f_{i_v}(\boldsymbol{X}_u)$ for some u satisfying $0 \le u < v$. Then we have

$$\begin{aligned}
\mathbb{E}\big[\|\widetilde{\nabla}_v - \nabla F(\boldsymbol{X}_v)\|_2^2\big] &= \frac{(n-B)}{B(n-1)} \mathbb{E}\big[\|\nabla f_{i_v}(\boldsymbol{X}_v) - \nabla f_{i_v}(\boldsymbol{X}_u)\|_2^2\big] \\
&\le \frac{(n-B)M^2}{B(n-1)} \mathbb{E}[\|\boldsymbol{X}_v - \boldsymbol{X}_u\|_2^2]. \qquad \text{(C.14)}
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{X}_u - \boldsymbol{X}_v\|_2^2 | u] &= \mathbb{E}\bigg[\bigg\|\sum_{j=u}^{v-1} \eta\widetilde{\nabla}_j + \sum_{j=u}^{v-1} \sqrt{\frac{2\eta}{\gamma}}\boldsymbol{\epsilon}_j\bigg\|_2^2\bigg] \\
&\le 2(u-v)\sum_{j=u}^{v-1} \mathbb{E}[\|\widetilde{\nabla}_j\|_2^2] + \frac{4(u-v)\eta d}{\gamma} \\
&\le 36(u-v)^2\eta^2(M^2D_B^2 + G^2) + \frac{4(u-v)\eta d}{\gamma},
\end{aligned}$$

where the first inequality follows from Jensen's inequality, and the second inequality is by Young's inequality and Lemma C.7, where $D_B = \sqrt{2(1+1/b)(a + G^2 + d/\gamma)}$. Then we have

$$\mathbb{E}[\|\boldsymbol{X}_v - \boldsymbol{X}_u\|_2^2] = \mathbb{E}\mathbb{E}[\|\boldsymbol{X}_u - \boldsymbol{X}_v\|_2^2 | u, v] \le \mathbb{E}\big[36(u-v)^2\eta^2(M^2D_B^2 + G^2) + \frac{4(u-v)\eta d}{\gamma}\big].$$

Let $q = 1 - (1-1/n)^B$ be the probability of choosing a particular index, then

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{X}_v - \boldsymbol{X}_u\|_2^2] &\le 36\eta^2(M^2D_B^2 + G^2)\mathbb{E}[(u-v)^2] + \frac{4\eta d}{\gamma}\mathbb{E}[(u-v)] \\
&= 36\eta^2(M^2D_B^2 + G^2)\sum_{t=0}^{v-1}(v-t)^2(1-q)^{v-t-1}q + \frac{4\eta d}{\gamma}\sum_{t=0}^{v-1}(v-t)(1-q)^{v-t-1}q \\
&\le 36\eta^2(M^2D_B^2 + G^2)\sum_{t=0}^{\infty} t^2(1-q)^{t-1}q + \frac{4\eta d}{\gamma}\sum_{t=0}^{\infty} t(1-q)^{t-1}q \\
&\le \frac{72\eta^2(M^2D_B^2 + G^2)}{q^2} + \frac{4\eta d}{q\gamma}.
\end{aligned}$$

From Dubey et al. (2016) we know that $q = 1-(1-1/n)^B \geq B/(2n)$, thus

$$\mathbb{E}[\|\boldsymbol{X}_v - \boldsymbol{X}_u\|_2^2] \leq \frac{288n^2\eta^2(M^2D_B^2+G^2)}{B^2} + \frac{8n\eta d}{B\gamma}. \tag{C.15}$$

In the following, we are going to bound the second term on the R.H.S of (C.13). Based on the definition of $\boldsymbol{H}(s)$, the following holds,

$$\begin{aligned}
\int_{v\eta}^{(v+1)\eta} \mathbb{E}\big[\|\nabla F(N(v\eta)) - \nabla F(\boldsymbol{H}(s))\|_2^2\big] ds &\leq \int_{v\eta}^{(v+1)\eta} M^2\mathbb{E}[\|\boldsymbol{H}(s) - \boldsymbol{H}(v\eta)\|_2^2] ds \\
&= \int_{v\eta}^{(v+1)\eta} M^2\bigg((s-v\eta)^2\mathbb{E}[\|\widetilde{\nabla}_v\|_2^2] + \frac{2(s-v\eta)d}{\gamma}\bigg) ds \\
&\leq \frac{M^2\eta^3}{3}\mathbb{E}[\|\widetilde{\nabla}_v\|_2^2] + \frac{2M^2\eta^2 d}{\gamma} \\
&\leq 6M^2\eta^3(M^2D_B^2+G^2) + \frac{2M^2\eta^2 d}{\gamma}.
\end{aligned} \tag{C.16}$$

where the last inequality follows from Lemma C.7. Then, plugging (C.16), (C.15), (C.14) into (C.13), we arrive at

$$\begin{aligned}
D_{KL}(Q_k||P_k) &\leq \frac{\gamma}{4}\int_0^{k\eta} \mathbb{E}\big[\|q(\boldsymbol{H}(s) - h(\boldsymbol{H}(s))\|_2^2\big] \\
&\leq \frac{\gamma}{2}\bigg[\frac{k\eta^2 n(n-B)M^2}{B^2(n-1)}\bigg(\frac{288n\eta(M^2D_B^2+G^2)}{B} + \frac{8d}{\gamma}\bigg) + k\eta^2M^2\bigg(6\eta(M^2D_B^2+G^2) + \frac{2d}{\gamma}\bigg)\bigg] \\
&= \bigg(\frac{144n^2(n-B)M^2}{B^3(n-1)} + 3M^2\bigg)(M^2D_B^2+G^2)\gamma k\eta^3 + \bigg(\frac{4n(n-B)M^2d}{B^2(n-1)} + M^2d\bigg)k\eta^2.
\end{aligned}$$

Apply Lemma C.3, and choose $\lambda = 1$ in Lemma C.3 we obtain

$$\begin{aligned}
&\mathcal{W}_2\big(P(\boldsymbol{X}_k), P(\boldsymbol{X}(k\eta))\big) \\
&\leq D_A\bigg[\bigg(\frac{144n^2(n-B)M^2}{B^3(n-1)} + 3M^2\bigg)(M^2D_B^2+G^2)\gamma k\eta^3 + \bigg(\frac{4n(n-B)M^2d}{B^2(n-1)} + M^2d\bigg)k\eta^2\bigg]^{1/4},
\end{aligned}$$

where $D_A = 4\sqrt{3/2 + (2b+d/\gamma)k\eta}$.

□

D Proof of Auxiliary Lemmas in Appendix C

In this section, we prove the technical lemmas in Appendix C.

D.1 Proof of Lemma C.1

Proof. Let $G = \max_{i=1,\ldots,n}\|f_i(\mathbf{0})\|$, then we have

$$\|\nabla f_i(\mathbf{x})\|_2 \leq \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{0})\|_2 + \|\nabla f_i(\mathbf{0})\|_2 \leq M\|\mathbf{x}\|_2 + G,$$

where the first inequality follows from triangle inequality and the second inequality follows from Assumption 4.1. This completes the proof. □

D.2 Proof of Lemma C.2

Proof. We prove the bound for $\mathbb{E}[\|\boldsymbol{X}_k\|_2^2]$ by mathematical induction. Since $\widetilde{\nabla}_k = 1/B\sum_{i_k\in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k) - \nabla f_{i_k}(\widetilde{\boldsymbol{X}}^{(s)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(s)})\big)$, we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] &= \mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] + \sqrt{\frac{8\eta}{\gamma}}\mathbb{E}[\langle\boldsymbol{X}_k - \eta\widetilde{\nabla}_k, \boldsymbol{\epsilon}_k\rangle] + \frac{2\eta}{\gamma}\mathbb{E}[\|\boldsymbol{\epsilon}_k\|_2^2] \\
&= \mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] + \frac{2\eta d}{\gamma},
\end{aligned} \tag{D.1}$$

where the second equality follows from the fact that $\boldsymbol{\epsilon}_k$ is independent of $\boldsymbol{X}_k$ and standard Gaussian.

We prove it by induction. First, consider the case when $k = 1$. Since we choose the initial point at $\boldsymbol{X}_0 = \mathbf{0}$, we immediately have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_1\|_2^2] &= \mathbb{E}[\|\boldsymbol{X}_0 - \eta\widetilde{\nabla}_0\|_2^2] + \sqrt{\frac{8\eta}{\gamma}}\mathbb{E}[\langle \boldsymbol{X}_0 - \eta\widetilde{\nabla}_0, \boldsymbol{\epsilon}_0\rangle] + \frac{2\eta}{\gamma}\mathbb{E}[\|\boldsymbol{\epsilon}_0\|_2^2] \\ &= \eta^2\mathbb{E}[\|\nabla F(\boldsymbol{X}_0)\|_2^2] + \frac{2\eta d}{\gamma} \\ &\leq \eta^2 G^2 + \frac{2\eta d}{\gamma},\end{aligned}$$

where the second equality holds due to the fact that $\widetilde{\nabla}_0 = \nabla F(\boldsymbol{X}_0)$ and the inequality follows from Lemma C.1. For sufficiently small η we can easily make the conclusion holds for $\mathbb{E}[\|\boldsymbol{X}_1\|_2^2]$.

Now assume that the conclusion holds for all iteration from 1 to k, then for the $(k+1)$-th iteration, by (D.1) we have,

$$\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] = \mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] + \frac{2\eta d}{\gamma}, \tag{D.2}$$

For the first term on the R.H.S of (D.2) we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] &= \mathbb{E}[\|\boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k)\|_2^2] + 2\eta\mathbb{E}\langle \boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k), \nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\rangle \\ &\quad + \eta^2\mathbb{E}[\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2] \\ &= \underbrace{\mathbb{E}[\|\boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k)\|_2^2]}_{T_1} + \underbrace{\eta^2\mathbb{E}[\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2]}_{T_2},\end{aligned} \tag{D.3}$$

where the second equality holds due to the fact that $\mathbb{E}[\widetilde{\nabla}_k] = \nabla F(\boldsymbol{X}_k)$. For term T_1, we can further bound it by

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k)\|_2^2] &= \mathbb{E}[\|\boldsymbol{X}_k\|_2^2] - 2\eta\mathbb{E}[\langle \boldsymbol{X}_k, \nabla F(\boldsymbol{X}_k)\rangle] + \eta^2\mathbb{E}[\|\nabla F(\boldsymbol{X}_k)\|_2^2] \\ &\leq \mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + 2\eta(a - b\mathbb{E}[\|\boldsymbol{X}_k\|_2^2]) + 2\eta^2(M^2\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + G^2) \\ &= (1 - 2\eta b + 2\eta^2 M^2)\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + 2\eta a + 2\eta^2 G^2,\end{aligned} \tag{D.4}$$

where the inequality follows from Lemma C.1 and triangle inequality. For term T_2, by Lemma C.5 we have

$$\mathbb{E}\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2 \leq \frac{M^2(n-B)}{B(n-1)}\mathbb{E}\big\|\boldsymbol{X}_k - \widetilde{\boldsymbol{X}}^{(s)}\big\|_2^2 \leq \frac{2M^2(n-B)}{B(n-1)}\Big(\mathbb{E}\big\|\boldsymbol{X}_k\big\|_2^2 + \mathbb{E}\big\|\widetilde{\boldsymbol{X}}^{(s)}\big\|_2^2\Big).$$

Submit the above bound back into (D.1) we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] \leq &\bigg(1 - 2\eta b + 2\eta^2 M^2\Big(1 + \frac{n-B}{B(n-1)}\Big)\bigg)\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] \\ &+ \frac{2\eta^2 M^2(n-B)}{B(n-1)}\mathbb{E}\big\|\widetilde{\boldsymbol{X}}^{(s)}\big\|_2^2 + 2\eta a + 2\eta^2 G^2 + \frac{2\eta d}{\gamma}.\end{aligned} \tag{D.5}$$

Note that by assumption we have $\mathbb{E}\big\|\boldsymbol{X}_j\big\|_2^2 \leq C_\psi$ for all $j = 1, \ldots, k$ where $C_\psi = 2\big(1 + 1/b\big)\big(a + G^2 + d/\gamma\big)$, thus (D.5) can be further bounded as:

$$\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] \leq \underbrace{\bigg(1 - 2\eta b + 2\eta^2 M^2\Big(1 + \frac{2(n-B)}{B(n-1)}\Big)\bigg)}_{C_\lambda} C_\psi + 2\eta a + 2\eta^2 G^2 + \frac{2\eta d}{\gamma}. \tag{D.6}$$

For sufficient small η that satisfies

$$\eta \leq \min\left(1, \frac{b}{2M^2\big(1 + 2(n-B)/(B(n-1))\big)}\right),$$

there are only two cases we need to take into account:
If $C_\lambda \leq 0$, then from (D.6) we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] &\leq 2\eta a + 2\eta^2 G^2 + \frac{2\eta d}{\gamma} \\ &\leq 2\bigg(a + G^2 + \frac{d}{\gamma}\bigg).\end{aligned} \tag{D.7}$$

If $0 < C_\lambda \leq 1$, then iterate (D.6) and we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] &\leq C_\lambda^{k+1}\|\boldsymbol{X}_0\|_2^2 + \frac{\eta a + \eta^2 G^2 + \frac{\eta d}{\gamma}}{\eta b - \eta^2 M^2\Big(1 + \frac{2(n-B)}{B(n-1)}\Big)} \\ &\leq \frac{2}{b}\bigg(a + G^2 + \frac{d}{\gamma}\bigg).\end{aligned} \tag{D.8}$$

Combining (D.7) and (D.8), we have

$$\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] \leq 2\bigg(1 + \frac{1}{b}\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg).$$

Thus we show that when $\mathbb{E}[\|\boldsymbol{X}_j\|_2^2], j = 1, \ldots, k$ are bounded, $\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2]$ is also bounded. By mathematical induction we complete the proof. □

D.3 Proof of Lemma C.5

Proof. Since by Algorithm 1 we have $\widetilde{\nabla}_k = (1/B)\sum_{i_k \in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k) - \nabla f_{i_k}(\widetilde{\boldsymbol{X}}^{(s)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(s)})\big)$, therefore,

$$\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\boldsymbol{X}_k)\|_2^2] = \mathbb{E}\bigg\|\frac{1}{B}\sum_{i_k \in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k) - \nabla f_{i_k}(\widetilde{\boldsymbol{X}}^{(s)}) + \nabla F(\widetilde{\boldsymbol{X}}^{(s)}) - \nabla F(\boldsymbol{X}_k)\big)\bigg\|_2^2.$$

Let $\mathbf{v}_i = \nabla F(\mathbf{x}_k) - \nabla F(\widetilde{\mathbf{x}}^{(s)}) - \big(\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}}^{(s)})\big)$.

$$\begin{aligned}\mathbb{E}\bigg\|\frac{1}{B}\sum_{i \in I_k}\mathbf{v}_i(\mathbf{x})\bigg\|_2^2 &= \frac{1}{B^2}\mathbb{E}\bigg[\sum_{i \neq i', \{i,i'\} \in I_k}\mathbf{v}_i(\mathbf{x})^\top \mathbf{v}_{i'}(\mathbf{x})\bigg] + \frac{1}{B}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2 \\ &= \frac{B-1}{Bn(n-1)}\mathbb{E}\bigg[\sum_{i \neq i'}\mathbf{v}_i(\mathbf{x})^\top \mathbf{v}_{i'}(\mathbf{x})\bigg] + \frac{1}{B}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2 \\ &= \frac{B-1}{Bn(n-1)}\mathbb{E}\bigg[\sum_{i,i'}\mathbf{v}_i(\mathbf{x})^\top \mathbf{v}_{i'}(\mathbf{x})\bigg] - \frac{B-1}{B(n-1)}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2 + \frac{1}{B}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2 \\ &= \frac{n-B}{B(n-1)}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2,\end{aligned} \tag{D.9}$$

where the last equality is due to the fact that $\frac{1}{n}\sum_{i=1}^n \mathbf{v}_i(\mathbf{x}) = 0$.

Therefore, we have

$$\begin{aligned}\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\mathbf{x}_k)\|_2^2] &\leq \frac{n-B}{B(n-1)}\mathbb{E}\|\mathbf{v}_i\|_2^2 \\ &= \frac{n-B}{B(n-1)}\mathbb{E}\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}}) - \mathbb{E}[\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}})]\|_2^2 \\ &\leq \frac{n-B}{B(n-1)}\mathbb{E}\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}})\|_2^2 \\ &\leq \frac{M^2(n-B)}{B(n-1)}\mathbb{E}\|\mathbf{x}_k - \widetilde{\mathbf{x}}\|_2^2,\end{aligned} \tag{D.10}$$

where the second inequality holds due to the fact that $\mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|_2^2] \leq \mathbb{E}[\|\mathbf{x}\|_2^2]$ and the last inequality follows from Assumption 4.1. This completes the proof. □

D.4 Proof of Lemma C.6

Although the similar proof has been shown in Raginsky et al. (2017), we provide a refined version to make this paper self-contained.

In order to prove Lemma C.6, we need the following three lemmas.

Lemma D.1. (Raginsky et al. (2017)) In terms of the Langevin dynamics (1.1), under Assumption 4.2, we have the following upper bound on the expectation $\mathbb{E}[\|\boldsymbol{X}(t)\|_2^2]$

$$\mathbb{E}\big[\|\boldsymbol{X}(t)\|_2^2\big] \leq e^{-2bt}\|\boldsymbol{X}(0)\|_2^2 + \frac{a+d/\gamma}{b}(1-e^{-2bt}).$$

Lemma D.2. (Bakry et al. (2008)). Suppose that there exists constants $k_0, \lambda_0 > 0$, $R \geq 0$ and a C^2 function $V : \mathbb{R}^d \to [1, \infty)$ such that

$$\mathcal{L}V(\mathbf{w}) \leq -\lambda_0 V(\mathbf{w}) + k_0 \mathbf{1}\{\|\mathbf{w}\|_2 \leq R\},$$

where the operator $\mathcal{L}$ is Itô differential operator. Then the stationary distribution, i.e., π, satisfies a Poincaré inequality with constant

$$c_p \leq \frac{1}{\lambda_0}\bigg(1 + C_p k_0 R^2 e^{Osc_R(g)}\bigg),$$

where $C_p > 0$ is a universal constant and $Osc_R(f) := \max_{\|\mathbf{w}\|_2 \leq R} f(\mathbf{w}) - \min_{\|\mathbf{w}\|_2 \leq R} f(\mathbf{w})$.

Lemma D.3. (Cattiaux et al. (2010)) Suppose the following conditions hold:

1. There exist constants $k, \lambda > 0$ and a C^2 function $V : \mathbb{R}^d \to [1, \infty)$ such that
$$\frac{\mathcal{L}V(\mathbf{w})}{V(\mathbf{w})} \leq k - \lambda\|\mathbf{w}\|_2^2$$
for all $\mathbf{w} \in \mathbb{R}^d$.
2. π satisfies a Poincaré inequality with constant c_p.
3. There exists some constant $K \geq 0$, such that $\nabla^2 f \geq -K\mathbf{I}$.

Let $\widetilde{C}_1$ and $\widetilde{C}_2$ be defined, for some $\epsilon > 0$, by

$$\widetilde{C}_1 = \frac{2}{\lambda}\bigg(\frac{1}{\epsilon} + \frac{K}{2}\bigg) + \epsilon \quad \text{and} \quad \widetilde{C}_2 = \frac{2}{\lambda}\bigg(\frac{1}{\epsilon} + \frac{K}{2}\bigg)\bigg(K + \lambda\int_{\mathbb{R}^d}\|\mathbf{w}\|_2^2 d\pi\bigg).$$

Then π satisfies a logarithmic Sobolev inequality with constant $\Gamma = \widetilde{C}_1 + (\widetilde{C}_2 + 2)c_p$.

Based on the above two lemmas, we are able to complete the proof.

Proof of Lemma C.6. We first give the upper bound of the constant c_p in Poincaré inequality. Following from Lemma D.2, we can establish a Lyapunov function $V(\mathbf{w})$ and then derive a upper bound of c_p. In this proof, we apply the same Lyapunov as Raginsky et al. (2017). Let $V(\mathbf{w}) = e^{-b\gamma\|\mathbf{w}\|_2^2/4}$, and we have

$$\begin{aligned}
\mathcal{L}V(\mathbf{w}) &= -\gamma\langle\nabla V, \nabla F\rangle + \nabla^2 V : \mathbf{I} \\
&= \bigg(-\frac{b\gamma^2}{2}\langle\mathbf{w}, \nabla F\rangle + \frac{b\gamma d}{2} + \frac{(b\gamma)^2}{4}\|\mathbf{w}\|_2^2\bigg)V \\
&\leq \bigg(\frac{b\gamma(d+a\gamma)}{2} - \frac{(b\gamma)^2}{4}\|\mathbf{w}\|_2^2\bigg)V, \qquad \text{(D.11)}
\end{aligned}$$

where the last inequality follows from Assumption 4.2. Thus, let $R^2 = 4(d+a\gamma)/(b\gamma)$, we have

$$\mathcal{L}V(\mathbf{w}) \leq -\frac{b\gamma(d+a\gamma)}{2}V(\mathbf{w}) + \max_{\|\mathbf{w}\|_2 \leq R}\bigg(\frac{b\gamma(d+a\gamma)}{2} - \frac{(b\gamma)^2}{4}\|\mathbf{w}\|_2^2\bigg)V(\mathbf{w})\mathbf{1}\{\|\mathbf{w}\|_2 \leq R\}.$$

Let

$$\lambda_0 = \frac{b\gamma(d+a\gamma)}{2}, \quad \text{and} \quad k_0 = \frac{b\gamma(d+a\gamma)e^{b\gamma R^2/4}}{2},$$

we immediately have

$$\mathcal{L}V(\mathbf{w}) \leq -\lambda_0 V(\mathbf{w}) + k_0 \mathbf{1}\{\|\mathbf{w}\|_2 \leq R\}.$$

Under Assumption 4.1, it follows that

$$F(\mathbf{x}) - F(\mathbf{y}) \leq \langle \nabla F(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{M}{2}\|\mathbf{x}-\mathbf{y}\|_2^2,$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. By taking $\mathbf{y} = \mathbf{0}$, we obtain that there exists a constant $K_0 > 0$, such that

$$F(\mathbf{x}) \leq F(0) + \langle \nabla F(0), \mathbf{x}\rangle + \frac{M}{2}\|\mathbf{x}\|_2^2 \leq K_0(1 + \|\mathbf{x}\|_2^2), \tag{D.12}$$

where

$$K_0 = \max\left\{F(0) + \frac{1}{2}\|\nabla F(0)\|_2^2, \frac{M+1}{2}\right\}.$$

By (D.12), we have

$$Osc_R(\gamma F) \leq 2\gamma K_0(1+R^2).$$

Thus, based on Lemma D.2, the stationary distribution π satisfies a Poincaré inequality with constant

$$c_p \leq \frac{1}{b\gamma(d+a\gamma)} + \frac{4C_p(d+a\gamma)}{b\gamma}\exp\left(2\gamma K_0 + \frac{(8K_0+b)(d+a\gamma)}{b}\right).$$

Next, we are going to prove the upper bound of constant Γ in logarithmic Sobolev inequality. According to (D.11), we know that

$$\frac{\mathcal{L}V(\mathbf{w})}{V(\mathbf{w})} \leq k - \lambda\|\mathbf{w}\|_2^2$$

holds with

$$k = \frac{b\gamma(d+a\gamma)}{2}, \quad \text{and} \quad \lambda = \frac{(b\gamma)^2}{4}.$$

In addition, for function $f(\mathbf{x}) = \gamma F(\mathbf{x})$, we have $\nabla^2 f \geq -M\gamma\mathbf{I}$ according to Assumption 4.1. Then substitute the above parameters into Lemma D.3, choose $\epsilon = \frac{2}{M}$, we obtain

$$\widetilde{C}_1 = \frac{2b^2 + 8M^2}{M\gamma b^2}.$$

Moreover, from Lemma D.1, constant $\widetilde{C}_2$ is bounded by

$$\widetilde{C}_2 \leq \frac{6M(d+a\gamma)}{b},$$

Submitting $\widetilde{C}_1$ and $\widetilde{C}_2$ back to Lemma D.3, we have

$$C \leq \frac{2b^2+8M^2}{b^2M\gamma} + c_p\left(\frac{6M(d+a\gamma)}{b} + 2\right),$$

note that $c_p = e^{\widetilde{O}(\gamma+d)}$, we also have $\Gamma = e^{\widetilde{O}(\gamma+d)}$, which completes the proof. □

D.5 Proof of Lemma C.7

Proof of Lemma C.7. The proof for Lemma C.7 is quite similar to that for Lemma C.2. Based on the update form of $\boldsymbol{X}_k$ in Algorithm 2, we have

$$\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] = \mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k + \sqrt{2\eta/\gamma}\boldsymbol{\epsilon}_k\|_2^2] = \mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] + \frac{2\eta d}{\gamma}.$$

Similar to (D.3), we further have

$$\mathbb{E}[\|\boldsymbol{X}_k - \eta\widetilde{\nabla}_k\|_2^2] = \mathbb{E}[\|\boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k)\|_2^2] + \eta^2\mathbb{E}[\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2]. \tag{D.13}$$

Compared with the argument in (D.3), the first term on the R.H.S of the above inequality can be upper bounded in the same way as we did in (D.4), which is stated as follows,

$$\mathbb{E}[\|\boldsymbol{X}_k - \eta\nabla F(\boldsymbol{X}_k)\|_2^2] \leq (1 - 2\eta b + 2\eta^2 M^2)\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + 2\eta a + 2\eta^2 G^2.$$

Regarding the second term on the R.H.S of (D.13), we have the following based on Lemma C.5

$$\mathbb{E}[\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2] \leq \frac{n-B}{B(n-1)}\mathbb{E}[\|\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\boldsymbol{G}}_{i_k}\|_2^2] = \frac{n-B}{B(n-1)}\mathbb{E}[\|\nabla f_{i_k}(\boldsymbol{X}_k) - \nabla f_{i_k}(\boldsymbol{X}_u)\|_2^2],$$

where u is an index satisfying $u < k$. Applying smoothness assumption we have

$$\begin{aligned}\mathbb{E}[\|\nabla F(\boldsymbol{X}_k) - \widetilde{\nabla}_k\|_2^2] &\leq \frac{(n-B)M^2}{B(n-1)}\mathbb{E}[\|\boldsymbol{X}_k - \boldsymbol{X}_u\|_2^2] \\ &\leq 2\big(\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + \mathbb{E}[\|\boldsymbol{X}_u\|_2^2]\big),\end{aligned}$$

where the second inequality follows from Young's inequality and the fact that $B \geq 1$. Now, we are able to upper bound $\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2]$ as follows

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] &\leq (1 - 2\eta b + 2\eta^2 M^2)\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + 2\eta a + 2\eta^2 G^2 + 2\eta^2(\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] + \mathbb{E}[\|\boldsymbol{X}_u\|_2^2]) + \frac{2\eta d}{\gamma} \\ &\leq \big(1 - 2\eta b + 2\eta^2(M^2+4)\big)\max\{\mathbb{E}[\|\boldsymbol{X}_k\|_2^2], \mathbb{E}[\|\boldsymbol{X}_u\|_2^2]\} + 2\eta(a + d/\gamma) + 2\eta^2 G^2\end{aligned}$$

Then we apply induction to prove that $\mathbb{E}[\|\boldsymbol{X}_k\|_2^2] \leq 2(1+1/b)(a+G^2+d/\gamma)$. It is easy to verify that $\mathbb{E}[\|\boldsymbol{X}_0\|_2^2] = 0$ satisfies the argument. Then we assume that the argument holds for all iterates from 0 to k. Note that $u < k$, which implies that

$$\max\{\mathbb{E}[\|\boldsymbol{X}_k\|_2^2], \mathbb{E}[\|\boldsymbol{X}_u\|_2^2]\} \leq 2\bigg(1 + \frac{1}{b}\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg).$$

Then, for sufficiently small η such that

$$\eta \leq \min\bigg\{1, \frac{b}{2(M^2+4)}\bigg\},$$

it follows that

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2] &\leq 2\bigg[(1 - \eta b)\bigg(1 + \frac{1}{b}\bigg) + \eta\bigg]\bigg(a + G^2 + \frac{d}{\gamma}\bigg) \\ &\leq 2\bigg(1 + \frac{1}{b} - \eta b\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg) \\ &\leq 2\bigg(1 + \frac{1}{b}\bigg)\bigg(a + G^2 + \frac{d}{\gamma}\bigg),\end{aligned}$$

which indicates that $\mathbb{E}[\|\boldsymbol{X}_{k+1}\|_2^2]$ also satisfies the argument. Thus we are able to complete the proof. □

D.6 Proof of Lemma C.8

Proof. Since by Algorithm 2 we have $\widetilde{\nabla}_k = \frac{1}{B}\sum_{i_k \in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\boldsymbol{G}}_{i_k} + \widetilde{\boldsymbol{g}}_k\big)$, therefore,

$$\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\boldsymbol{X}_k)\|_2^2] = \mathbb{E}\bigg\|\frac{1}{B}\sum_{i_k \in I_k}\big(\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\mathbf{G}}_{i_k} + \widetilde{\mathbf{g}}_k - \nabla F(\boldsymbol{X}_k)\big)\bigg\|_2^2.$$

Let $\mathbf{v}_i = \nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\mathbf{G}}_{i_k} + \widetilde{\mathbf{g}}_k - \nabla F(\boldsymbol{X}_k)$, following the same procedure in (D.9) we have

$$\mathbb{E}\bigg\|\frac{1}{B}\sum_{i \in I_k}\mathbf{v}_i(\mathbf{x})\bigg\|_2^2 = \frac{n-B}{B(n-1)}\mathbb{E}\|\mathbf{v}_i(\mathbf{x})\|_2^2.$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla}_k - \nabla F(\mathbf{x}_k)\|_2^2] &= \frac{n-B}{B(n-1)}\mathbb{E}\|\mathbf{v}_i\|_2^2 \\
&= \frac{n-B}{B(n-1)}\mathbb{E}\|\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\mathbf{G}}_{i_k} - (\nabla F(\boldsymbol{X}_k) - \widetilde{\mathbf{g}}_k)\|_2^2 \\
&\leq \frac{n-B}{B(n-1)}\mathbb{E}\|\nabla f_{i_k}(\boldsymbol{X}_k) - \widetilde{\mathbf{G}}_{i_k}\|_2^2,
\end{aligned}$$

where the inequality holds due to the fact that $\mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|_2^2] \leq \mathbb{E}[\|\mathbf{x}\|_2^2]$, which completes the proof. □