A. Proof of Main Theory

In this section, we present our theoretical analysis of the proposed SVR-HMC algorithm. Before we present the proof of our main theorem, we introduce some notations for the ease of our presentation. We use notation $\mathcal{S}_\eta$ to denote the one-step SVR-HMC update in (3.1) with step size η, i.e., $\mathbf{x}_{k+1} = \mathcal{S}_\eta \mathbf{x}_k$ and $\mathbf{v}_{k+1} = \mathcal{S}_\eta \mathbf{v}_k$. Similarly, We define an operator $\mathcal{G}_\eta$ which also performs one step update with step size η, but replaces the semi-stochastic gradient in (3.1) with the full gradient. Specifically, we have

$$\begin{aligned} \mathcal{G}_\eta \mathbf{v}_k &= \mathbf{v}_k - \gamma\eta\mathbf{v}_k - \eta u \nabla f(\mathbf{x}_k) + \boldsymbol{\epsilon}_k^v, \\ \mathcal{G}_\eta \mathbf{x}_k &= \mathbf{x}_k + \eta \mathbf{v}_k + \boldsymbol{\epsilon}_k^x, \end{aligned} \tag{A.1}$$

for any $\mathbf{x}_k, \mathbf{v}_k \in \mathbb{R}^d$, where $\boldsymbol{\epsilon}_k^v$ and $\boldsymbol{\epsilon}_k^x$ are the same as defined in Algorithm 1. Next, we define an operator $\mathcal{L}_\eta$ which represents the integration over a time interval of length η on the continuous dynamics (1.3). Specifically, for any starting point $\boldsymbol{V}_0$ and $\boldsymbol{X}_0$, integrating (1.3) over time interval $(0, \eta)$ yields the following equations:

$$\boldsymbol{V}_t = \mathcal{L}_t \boldsymbol{V}_0 = \boldsymbol{V}_0 e^{-\gamma t} - u\bigg(\int_0^t e^{-\gamma(t-s)} \nabla f(\boldsymbol{X}_t) \mathrm{d}s\bigg) + \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)} \mathrm{d}\boldsymbol{B}_s, \tag{A.2}$$

$$\boldsymbol{X}_t = \mathcal{L}_t \boldsymbol{X}_0 = \boldsymbol{X}_0 + \int_0^t \boldsymbol{V}_s \mathrm{d}s. \tag{A.3}$$

(A.2) and (A.3) give out an implicit solution of dynamics (1.3), which can be easily verified by taking derivatives of these two equations (Cheng et al., 2017). The following lemma characterizes the mean value and covariance of the Brownian motion terms.

Lemma A.1. (Cheng et al., 2017) The additive Brownian motion in (A.2), denoted by $\boldsymbol{\epsilon}^v = \sqrt{2\gamma u}\int_0^t e^{-\gamma(t-s)} d\boldsymbol{B}_s$, has mean $\mathbf{0}$ and covariance matrix

$$\mathbb{E}[\boldsymbol{\epsilon}^v (\boldsymbol{\epsilon}^v)^\top] = 2\gamma u \mathbb{E}\bigg[\int_0^t e^{-\gamma(t-s)} \mathrm{d}\boldsymbol{B}_s \int_0^t e^{-\gamma(t-s)} \mathrm{d}\boldsymbol{B}_s^\top\bigg] = u(1 - e^{-2\gamma t}) \cdot \mathbf{I}_{d\times d}.$$

Note that there also exists a hidden Brownian motion term in (A.3), which comes from the velocity $\boldsymbol{V}_s$, denoted by $\boldsymbol{\epsilon}^x = \sqrt{2\gamma u}\int_0^t\int_0^s e^{-\gamma(s-r)} \mathrm{d}\boldsymbol{B}_r \mathrm{d}t$, having mean $\mathbf{0}$ and covariance matrix

$$\mathbb{E}[\boldsymbol{\epsilon}^x (\boldsymbol{\epsilon}^x)^\top] = 2\gamma u \mathbb{E}\bigg[\int_0^t\int_0^s e^{-\gamma(s-r)} \mathrm{d}B_r \mathrm{d}s \int_0^t\int_0^s e^{-\gamma(s-r)} \mathrm{d}\boldsymbol{B}_r^\top \mathrm{d}s\bigg] = \frac{u}{\gamma^2}(2\gamma t + 4e^{-\gamma t} - e^{-2\gamma t} - 3) \cdot \mathbf{I}_{d\times d}.$$

In addition, $\boldsymbol{\epsilon}^v$ and $\boldsymbol{\epsilon}^x$ have the following cross-covariance

$$\mathbb{E}[\boldsymbol{\epsilon}^v (\boldsymbol{\epsilon}^x)^\top] = 2\gamma u \mathbb{E}\bigg[\int_0^t e^{-\gamma(t-s)} \mathrm{d}\boldsymbol{B}_s \int_0^t\int_0^s e^{-\gamma(s-r)} \mathrm{d}\boldsymbol{B}_r^\top \mathrm{d}s\bigg] = \frac{u}{\gamma}(1 - 2e^{-\gamma t} + e^{-2\gamma t}) \cdot \mathbf{I}_{d\times d}.$$

Recall the independent Gaussian random vectors $\boldsymbol{\epsilon}_k^v$ and $\boldsymbol{\epsilon}_k^x$ used in each iteration of Algorithm 1. They all have zero mean and the covariance matrices defined in (3.3) have exactly the same form with the covariance matrices in Lemma A.1 when $t = \eta$. Due to this property, we will use a synchronous coupling technique that ensures the Gaussian random vectors in each one-step update of the discrete algorithm, i.e., $\mathcal{S}_\eta \mathbf{x}$ and $\mathcal{S}_\eta \mathbf{v}$, are exactly the same additive Brownian motion terms in the one-step integration of the continuous dynamics $\mathcal{L}_\eta \mathbf{x}$ and $\mathcal{L}_\eta \mathbf{v}$. The shared Brownian motions between $\mathcal{S}_\eta \mathbf{v}$ and $\mathcal{L}_\eta \mathbf{v}$ ($\mathcal{S}_\eta \mathbf{x}$ and $\mathcal{L}_\eta \mathbf{x}$) are pivotal to our analysis. Similar coupling techniques are also used in Eberle et al. (2017); Cheng et al. (2017).

A.1. Proof of Theorem 4.3

We first lay down some technical lemmas that are useful in our proof. The first lemma characterizes the discretization error between the full gradient-based HMC update in (A.1) and the solutions of continuous Hamiltonian dynamics (1.3).

Lemma A.2. Under Assumptions 4.1 and 4.2, consider one-step discrete update (A.1) and Langevin diffusion (A.2)-(A.3) starting from point $(\mathbf{x}_k, \mathbf{v}_k)$. Then the discretization error for velocity and position are bounded by

$$\mathbb{E}[\|\mathcal{G}_\eta \mathbf{x}_k - \mathcal{L}_\eta \mathbf{x}_k\|_2^2] \leq \eta^4 \bigg[\bigg(\frac{2\gamma^2 + 2uL}{3}\bigg) U_v + \frac{4u^2 L}{3} U_f + \frac{8u^2 L\gamma d\eta}{3}\bigg] \triangleq D_1 \eta^4,$$

$$\mathbb{E}[\|\mathcal{G}_\eta \mathbf{v}_k - \mathcal{L}_\eta \mathbf{v}_k\|_2^2] \leq \eta^4 \bigg[\bigg(\frac{3\gamma^4}{4} + u^2 L^2\bigg) U_v + \bigg(\frac{3u^2\gamma^2 L}{2} + 4u^3 L^2\bigg) U_f + 4u^3 L^2 \eta\gamma d\bigg] \triangleq D_2 \eta^4,$$

where parameters U_v and U_f are in the order of $O(d/\mu)$ and $O(d\kappa)$ respectively.

The difference between our SVR-HMC update and the full gradient-based HMC update in (A.1) can be characterized by the following lemma.

Lemma A.3. Under Assumptions 4.1 and 4.2, for any $\mathbf{x}_k, \mathbf{v}_k \in \mathbb{R}^d$, we have

$$\mathbb{E}[\|\mathcal{S}_\eta \mathbf{x}_k - \mathcal{G}_\eta \mathbf{x}_k\|_2^2] = 0, \tag{A.4}$$

$$\mathbb{E}[\|\mathcal{S}_\eta \mathbf{v}_k - \mathcal{G}_\eta \mathbf{v}_k\|_2^2] \le 2\eta^4 m^2 u^2 L^2 (U_v + \gamma u d) \triangleq D_3 u^2 L^2 m^2 \eta^4. \tag{A.5}$$

The following lemma shows the contraction property for the diffusion operator in terms of the coupled ℓ_2 norm.

Lemma A.4. (Cheng et al., 2017) Under Assumptions 4.1 and 4.2, let $\mathbf{z} = (\mathbf{x}^\top, (\mathbf{x}+\mathbf{v})^\top)^\top \in \mathbb{R}^{2d}$ and $\mathcal{L}_t \mathbf{z} = ((\mathcal{L}_t\mathbf{x})^\top, (\mathcal{L}_t\mathbf{x} + \mathcal{L}_t\mathbf{v})^\top)^\top$. Set $\gamma = 2$ and $u = 1/L$ in (A.2)-(A.3). Starting from two different points $\mathbf{z}_1$ and $\mathbf{z}_2$, the continuous-time dynamics after time t satisfy

$$\mathbb{E}[\|\mathcal{L}_t\mathbf{z}_1 - \mathcal{L}_t\mathbf{z}_2\|_2^2] \le e^{-t/\kappa}\mathbb{E}[\|\mathbf{z}_1 - \mathbf{z}_2\|_2^2],$$

where the diffusion operators on $\mathbf{z}_1$ and $\mathbf{z}_2$ share the same Brownian motion, and $\kappa = L/\mu$ denotes the condition number.

For the operators $\mathcal{L}_\eta$, we denote $\mathcal{L}_\eta^k \mathbf{x} = \mathcal{L}_\eta \circ \mathcal{L}_\eta \circ \cdots \circ \mathcal{L}_\eta \mathbf{x}$ as the result after $\mathcal{L}_\eta$ operates k times starting at $\mathbf{x}$. We have the following lemma which is useful to characterize the distance $\mathbb{E}[\|\mathbf{z}_k - \mathcal{L}_\eta^k \mathbf{z}^\pi\|_2^2]$ based on some recursive arguments, where $\mathbf{z}^\pi = \big((\mathbf{x}^\pi)^\top, (\mathbf{x}^\pi + \mathbf{v}^\pi)^\top\big)^\top$.

Lemma A.5. (Dalalyan & Karagulyan, 2017) Let A, B and C be given non-negative numbers such that $A \in (0,1)$. Assume that the sequence of non-negative numbers $\{x_k\}_{k=0,1,2,\ldots}$ satisfies the recursive inequality

$$x_{k+1}^2 \le [(1-A)x_k + C]^2 + B^2,$$

for every integer $k \ge 0$. Then, for all integers $k \ge 0$,

$$x_k \le (1-A)^k x_0 + \frac{C}{A} + \frac{B}{\sqrt{A}}.$$

Lemma A.6. For any two random vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^d$, the following holds

$$\mathbb{E}[\|\boldsymbol{X} + \boldsymbol{Y}\|_2^2] \le \left(\sqrt{\mathbb{E}[\|\boldsymbol{X}\|_2^2]} + \sqrt{\mathbb{E}[\|\boldsymbol{Y}\|_2^2]}\right)^2.$$

Based on all the above lemmas, we are now ready to prove Theorem 4.3.

Proof of Theorem 4.3. Let $\mathbf{z}^\pi$ denote the random variable satisfying distribution $\pi_{\mathbf{z}}$, then we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{z}_{k+1} - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\|_2^2] &= \mathbb{E}[\|\mathbf{z}_{k+1} - \mathcal{G}_\eta\mathbf{z}_k + \mathcal{G}_\eta\mathbf{z}_k - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\|_2^2] \\ &= \mathbb{E}[\|\mathbf{z}_{k+1} - \mathcal{G}_\eta\mathbf{z}_k\|_2^2 + 2\langle \mathbf{z}_{k+1} - \mathcal{G}_\eta\mathbf{z}_k, \mathcal{G}_\eta\mathbf{z}_k - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\rangle + \|\mathcal{G}_\eta\mathbf{z}_k - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\|_2^2] \\ &= \mathbb{E}[\|\mathbf{z}_{k+1} - \mathcal{G}_\eta\mathbf{z}_k\|_2^2 + \|\mathcal{G}_\eta\mathbf{z}_k - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\|_2^2], \qquad (A.6)\end{aligned}$$

where the last equality follows from the fact that $\mathbb{E}\big[\langle \mathbf{z}_{k+1} - \mathcal{G}_\eta\mathbf{z}_k, \mathcal{G}_\eta\mathbf{z}_k - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\rangle\big] = \mathbb{E}\big[\mathbb{E}_{i_k}[\langle \mathbf{z}_{k+1} - \mathcal{G}_\eta\mathbf{z}_k, \mathcal{G}_\eta\mathbf{z}_k - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\rangle]\big] = \mathbf{0}$ and $\mathbb{E}_{i_k}[\mathbf{z}_{k+1}] = \mathcal{G}_\eta\mathbf{z}_k$. Note that $\mathbf{z}_{k+1} = (\mathbf{x}_{k+1}^\top, (\mathbf{x}_{k+1} + \mathbf{v}_{k+1})^\top)^\top$, thus

$$\begin{aligned}\mathbb{E}[\|\mathbf{z}_{k+1} - \mathcal{G}_\eta\mathbf{z}_k\|_2^2 &= \mathbb{E}[\|\mathbf{x}_{k+1} - \mathcal{G}_\eta\mathbf{x}_k\|_2^2 + \mathbb{E}[\|\mathbf{x}_{k+1} + \mathbf{v}_{k+1} - \mathcal{G}_\eta(\mathbf{x}_k + \mathbf{v}_k)\|_2^2 \\ &= \mathbb{E}[\|\mathbf{v}_{k+1} - \mathcal{G}_\eta\mathbf{v}_k\|_2^2 \qquad (A.7) \\ &\le D_3 m^2\eta^4, \qquad (A.8)\end{aligned}$$

where the second equality follows from $\mathbf{x}_{k+1} = \mathcal{G}_\eta \mathbf{x}_k$, the inequality follows from Lemma A.3 and the fact that $uL = 1$. The second term on the R.H.S of (A.6) can be further bounded as follows,

$$\begin{aligned}\mathbb{E}[\|\mathcal{G}_\eta \mathbf{z}_k - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\|_2^2] &= \mathbb{E}[\|\mathcal{G}_\eta \mathbf{z}_k - \mathcal{L}_\eta \mathbf{z}_k + \mathcal{L}_\eta \mathbf{z}_k - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\|_2^2]\\ &\le \left(\sqrt{\mathbb{E}[\|\mathcal{G}_\eta \mathbf{z}_k - \mathcal{L}_\eta \mathbf{z}_k\|_2^2]} + \sqrt{\mathbb{E}[\|\mathcal{L}_\eta \mathbf{z}_k - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\|_2^2]}\right)^2\\ &\le \left(\sqrt{\mathbb{E}[\|\mathcal{G}_\eta \mathbf{z}_k - \mathcal{L}_\eta \mathbf{z}_k\|_2^2]} + e^{-\eta/(2\kappa)}\sqrt{\mathbb{E}[\|\mathbf{z}_k - \mathcal{L}_\eta^{k}\mathbf{z}^\pi\|_2^2]}\right)^2, \qquad \text{(A.9)}\end{aligned}$$

where the first inequality holds due to Lemma A.6 and the second inequality follows from Lemma A.4. We further have

$$\begin{aligned}\sqrt{\mathbb{E}[\|\mathcal{G}_\eta \mathbf{z}_k - \mathcal{L}_\eta \mathbf{z}_k\|_2^2]} &= \sqrt{\mathbb{E}[\|\mathcal{G}_\eta (\mathbf{v}_k + \mathbf{x}_k) - \mathcal{L}_\eta (\mathbf{v}_k + \mathbf{x}_k)\|_2^2] + \mathbb{E}[\|\mathcal{G}_\eta \mathbf{x}_k - \mathcal{L}_\eta \mathbf{x}_k\|_2^2]}\\ &\le \sqrt{\mathbb{E}[\|\mathcal{G}_\eta (\mathbf{v}_k + \mathbf{x}_k) - \mathcal{L}_\eta (\mathbf{v}_k + \mathbf{x}_k)\|_2^2} + \sqrt{\mathbb{E}[\|\mathcal{G}_\eta \mathbf{x}_k - \mathcal{L}_\eta \mathbf{x}_k\|_2^2]}\\ &\le 2\sqrt{\mathbb{E}[\|\mathcal{G}_\eta \mathbf{x}_k - \mathcal{L}_\eta \mathbf{x}_k\|_2^2]} + \sqrt{\mathbb{E}[\|\mathcal{G}_\eta \mathbf{v}_k - \mathcal{L}_\eta \mathbf{v}_k\|_2^2]} \qquad \text{(A.10)}\\ &\le 2\sqrt{D_1}\eta^2 + \sqrt{D_2}\eta^2, \qquad \text{(A.11)}\end{aligned}$$

where the first inequality is due to $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, the second inequality is due to (B.8) and the last inequality comes from Lemma A.2. Here D_1, D_2 are constants which are both in the order of $O(d/\mu)$. Denote $w_{k+1}^2 = \mathbb{E}[\|\mathbf{z}_{k+1} - \mathcal{L}_\eta^{k+1}\mathbf{z}^\pi\|_2^2]$. Submitting (A.8), (A.9) and (A.11) into (A.6) yields

$$w_{k+1}^2 \le \left(e^{-\eta/(2\kappa)} w_k + 2\sqrt{D_1}\eta^2 + \sqrt{D_2}\eta^2\right)^2 + D_3 m^2 \eta^4. \qquad \text{(A.12)}$$

Then, by Lemma A.5, w_k can be bounded by

$$w_k \le e^{-k\eta/(2\kappa)} w_0 + \frac{2\sqrt{D_1}\eta^2 + \sqrt{D_2}\eta^2}{1 - e^{-\eta/(2\kappa)}} + \frac{\sqrt{D_3} m \eta^2}{\sqrt{1 - e^{-\eta/(2\kappa)}}}.$$

Note that the above results rely on the shared Brownian motion in the discrete update and continuous Langevin diffusion, i.e., we assume identical Brownian motion sequences are used in the updates $\mathbf{z}_k = \mathcal{S}_\eta^k \mathbf{z}_0$ and $\mathcal{L}_\eta^k \mathbf{z}^\pi$. Since $\mathbf{z}^\pi$ satisfies the stationary distribution $\pi_{\mathbf{z}}$, $\mathcal{L}_\eta^k \mathbf{z}^\pi$ satisfies $\pi_{\mathbf{z}}$ as well. According to the definition of 2-Wasserstein distance, we have

$$\begin{aligned}\mathcal{W}_2\big(P(\mathbf{z}_k), \pi_{\mathbf{z}}\big) &= \left(\inf_{\zeta \in \Gamma(\mathbf{z}_k, \mathcal{L}_\eta^k \mathbf{z}^\pi)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{z}_k - \mathcal{L}_\eta^k \mathbf{z}^\pi\|_2^2 \mathrm{d}\zeta(\mathbf{z}_k, \mathcal{L}_\eta^k \mathbf{z}^\pi)\right)^{1/2}\\ &\le \sqrt{\mathbb{E}[\|\mathbf{z}_k - \mathcal{L}_\eta^k \mathbf{z}^\pi\|_2^2]}\\ &= w_k,\end{aligned}$$

which further implies that

$$\mathcal{W}_2\big(P(\mathbf{z}_K), \pi_{\mathbf{z}}\big) \le w_K \le e^{-K\eta/(2\kappa)} w_0 + \frac{2\sqrt{D_1}\eta^2 + \sqrt{D_2}\eta^2}{1 - e^{-\eta/(2\kappa)}} + \frac{\sqrt{D_3} m \eta^2}{\sqrt{1 - e^{-\eta/(2\kappa)}}}. \qquad \text{(A.13)}$$

Let $K\eta = T$, and note that $1 - e^{-\eta/(2\kappa)} \ge \eta/(4\kappa)$ when assuming $0 < \eta/\kappa \le 1$. Therefore, we have

$$\mathcal{W}_2\big(P(\mathbf{z}_K), \pi_{\mathbf{z}}\big) \le w_K \le e^{-T/(2\kappa)} w_0 + 4\eta\kappa(2\sqrt{D_1} + \sqrt{D_2}) + 2\sqrt{\kappa D_3} m \eta^{3/2}. \qquad \text{(A.14)}$$

Moreover, note that

$$\begin{aligned}\mathcal{W}_2\big(P(\mathbf{x}_K), \pi\big) &= \mathbb{E}_{\Gamma(\mathbf{x}_K, \mathbf{x}^\pi)}[\|\mathbf{x}_K - \mathbf{x}^\pi\|_2^2]\\ &\le \mathbb{E}_{\Gamma(\mathbf{x}_K, \mathbf{v}_K, \mathbf{x}^\pi, \mathbf{v}^\pi)}[\|\mathbf{x}_K - \mathbf{x}^\pi\|_2^2 + \|\mathbf{x}_K + \mathbf{v}_K - \mathbf{x}^\pi - \mathbf{v}^\pi\|_2^2]\\ &= \mathcal{W}_2\big(P(\mathbf{z}_K), \pi_{\mathbf{z}}\big).\end{aligned}$$

Substituting the above into (A.14) directly yields the argument in Theorem 4.3. □

A.2. Proof of Corollary 4.5

Now we present the calculation of gradient complexity of our algorithm.

We first present the following Lemma that characterizes the expectation $\mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}^*\|]$, where $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$ is the global minimizer of function f.

Lemma A.7 (Proposition 1 in Durmus & Moulines (2016b))**.** Let $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$ denote the global minimizer of function f, and $\mathbf{x}^\pi$ be the random vector satisfying distribution $\pi \propto e^{-f(\mathbf{x})}$, the following holds,

$$\mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}^*\|_2^2] \leq \frac{d}{\mu}.$$

Then we are going to prove Corollary 4.5.

Proof of Corollary 4.5. We first let $w_0 e^{-T/2\kappa} = \epsilon/3$, which implies that $T = 2\kappa\log(3w_0/\epsilon)$. Note that $\mathbf{x}^*$ is the minimizer of f and by assumption we have $\|\mathbf{x}_0 - \mathbf{x}^*\|_2 \leq R$. Recall the definition of w_k, we have

$$w_0 = \mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}^\pi\|_2^2] = \mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}^* + \mathbf{x}^* - \mathbf{x}^\pi\|_2^2] \leq 2\mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}^*\|_2^2] + 2\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \leq \frac{2d}{\mu} + 2R,$$

where the last inequality comes from Lemma A.7. Then we obtain $T = \widetilde{O}(\kappa)$, where $\widetilde{O}(\cdot)$ notation hides the logarithmic term of ϵ, d, μ and R. We then rewrite (A.14) as follow,

$$\mathcal{W}_2\big(P(\mathbf{z}_K), \pi_{\mathbf{z}}\big) \leq e^{-T/(2\kappa)} w_0 + \widetilde{C}_2\eta + \widetilde{C}_3 m\eta^{3/2}, \tag{A.15}$$

where $\widetilde{C}_2 = O(\kappa(d/\mu)^{1/2})$ and $\widetilde{C}_3 = O\big((\kappa d/\mu)^{1/2}\big)$. We then let

$$\widetilde{C}_2\eta = \frac{\epsilon}{3}, \quad \text{and} \quad \widetilde{C}_3 m\eta^{3/2} = \frac{\epsilon}{3},$$

and solve for η, which leads to

$$\eta = \min\left\{\frac{\epsilon}{3\widetilde{C}_2}, \frac{\epsilon^{2/3}}{(3\widetilde{C}_3 m)^{2/3}}\right\}.$$

Thus, the total iteration number satisfies

$$K = \frac{T}{\eta} \leq \frac{3T\widetilde{C}_2}{\epsilon} + \frac{T(3\widetilde{C}_3 m)^{2/3}}{\epsilon^{2/3}}.$$

In terms of gradient complexity, we have

$$T_g = K + n\left(1 \vee \frac{K}{m}\right) \leq K + \frac{Kn}{m} + n.$$

Substituting $\widetilde{C}_2$, $\widetilde{C}_3$, T into the above equation, and let $m = n$, we obtain

$$T_g \leq 2K + n = \widetilde{O}\left(\frac{\kappa^2(d/\mu)^{1/2}}{\epsilon} + \frac{\kappa^{4/3}(d/\mu)^{1/3}n^{2/3}}{\epsilon^{2/3}} + n\right). \tag{A.16}$$

When μ and L appear individually, they can be treated as constants. Thus we arrive at the result in Corollary 4.5. □

A.3. Proof of Theorem 4.8

In this section, we prove the convergence result of SVR-HMC for sampling from a general log-concave distribution. Note that for a μ-strongly log-concave distribution $\pi \propto e^{-f}$, it must satisfy a logarithmic Sobolev inequality with constant $C_{LS} = 1/\mu$ (Raginsky et al., 2017). We first present the following two useful lemmas.

Lemma A.8. (Dalalyan, 2014) Let f and $\bar{f}$ be two functions such that $f(\mathbf{x}) \leq \bar{f}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$, suppose e^{-f} and $e^{-\bar{f}}$ are both integratable. Then the Kullback-Leibler (KL) divergence between distribution $\pi \propto e^{-f}$ and $\bar{\pi} \propto e^{-\bar{f}}$ satisfies

$$\mathrm{KL}(\pi || \bar{\pi}) \leq \frac{1}{2} \int_{\mathbb{R}^d} \big(\bar{f}(\mathbf{x}) - f(\mathbf{x})\big)^2 \mathrm{d}\pi(\mathbf{x}).$$

Lemma A.9. (Bakry et al., 2013) If a stationary distribution π_1 satisfies a logarithmic Sobolev inequality with constant C_{LS}, for any probability measure π_2, it follows that

$$\mathcal{W}_2(\pi_1, \pi_2) \leq \sqrt{2C_{LS}\mathrm{KL}(\pi_2||\pi_1)}.$$

In what follows, we are going to leverage the above two lemmas to analyze the convergence rate of SVR-HMC for sampling from general log-concave distributions. Based on Assumption 4.7, we have

$$\int_{\mathbb{R}^d} \big(\bar{f}(\mathbf{x}) - f(\mathbf{x})\big)^2 \mathrm{d}\pi(\mathbf{x}) = \frac{\lambda^2}{4} \int_{\mathbb{R}^d} \|\mathbf{x}\|_2^4 \mathrm{d}\pi(\mathbf{x}) \leq \frac{\lambda^2 \bar{U} d^2}{4},$$

where $\bar{U} > 0$ is an absolute constant. Then, by Lemma A.8, we immediately have

$$\mathrm{KL}(\pi||\bar{\pi}) \leq \frac{\lambda^2 \bar{U} d^2}{8}.$$

From Lemma A.9, the 2-Wasserstein distance $W_2(\bar{\pi}, \pi)$ is upper bounded by

$$\mathcal{W}_2(\bar{\pi}, \pi) \leq \sqrt{2C_{LS}\mathrm{KL}(\pi||\bar{\pi})} \leq \frac{\sqrt{\lambda \bar{U} d^2}}{2}, \tag{A.17}$$

where we use the fact that the probability measure $\bar{\pi}$ satisfies a logarithmic Sobolev with constant $C_{LS} = 1/\lambda$ due to the strong convexity of $\bar{f}$. By triangle inequality in 2-Wasserstein distance, for any distribution p, we have $\mathcal{W}_2(p, \pi) \leq \mathcal{W}_2(p, \bar{\pi}) + \mathcal{W}_2(\bar{\pi}, \pi)$. Thus, we can perform our algorithms over distribution $\bar{\pi} \propto e^{-\bar{f}}$, and obtain an approximate sampling $\boldsymbol{X}$ which achieves the ϵ-precision requirement in $\mathcal{W}_2(P(\boldsymbol{X}), \pi)$, as long as ensuring $\mathcal{W}_2(P(\boldsymbol{X}), \bar{\pi}) \leq \epsilon/2$ and $\mathcal{W}_2(\bar{\pi}, \pi) \leq \epsilon/2$. According to (A.17), the requirement $\mathcal{W}_2(\bar{\pi}, \pi) \leq \epsilon/2$ suggests that the parameter λ should be selected such that $\lambda \leq \epsilon^2/(\bar{U}d^2) = O(\epsilon^2/d^2)$. Based on the above discussion, we are ready to prove Theorem 4.8 as follows.

Proof of Theorem 4.8. Note that we perform Algorithm 1 on the approximate density $\bar{\pi} \propto e^{-\bar{f}}$, where $\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \lambda\|\mathbf{x} - \mathbf{x}^*\|_2^2/2$, and $\lambda = O(\epsilon^2/d^2)$. It can be readily seen that function $\bar{f}(\mathbf{x})$ is an $(L + \lambda)$-smooth and λ-strongly convex function. Thus, we can directly replace the parameter μ in (A.16) with λ, and obtain

$$T_g = \widetilde{O}\bigg(\frac{d^{1/2}}{\epsilon\lambda^{5/2}} + \frac{d^{1/3}n^{2/3}}{\epsilon^{2/3}\lambda^{5/3}} + n\bigg),$$

where we treat the smoothness parameter $L + \lambda$ as constant of order $O(1)$ when it appears individually. Plugging the fact $\lambda = O(\epsilon^2/d^2)$ into the above equation, we have

$$T_g = \widetilde{O}\bigg(n + \frac{d^{11/2}}{\epsilon^6} + \frac{d^{11/3}n^{2/3}}{\epsilon^4}\bigg),$$

which completes the proof. □

B. Proof of Technical Lemmas

In this section, we prove the technical lemmas used in the proof of our main theorems. We first present some useful lemmas that will be used in our analysis.

Lemma B.1. Under Assumptions 4.1 and 4.2, the solution of Hamiltonian Langevin dynamics in (A.2)-(A.3) satisfies

$$\mathbb{E}[\|\boldsymbol{V}_t\|_2^2] \leq 2u\big[f(\boldsymbol{X}_0) - f(\mathbf{x}^*) + \gamma dt\big] + \|\boldsymbol{V}_0\|_2^2,$$
$$\mathbb{E}[f(\boldsymbol{X}_t)] \leq f(\boldsymbol{X}_0) + \frac{\|\boldsymbol{V}_0\|_2^2}{2u} + \gamma dt,$$
$$\mathbb{E}[\|\nabla f(\boldsymbol{X}_t)\|_2^2] \leq 2L\bigg(f(\boldsymbol{X}_0) - f(\mathbf{x}^*) + \frac{\|\boldsymbol{V}_0\|_2^2}{2u} + \gamma dt\bigg),$$

where $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$ denotes the global minimizer of function $f(\mathbf{x})$.

Lemma B.2. Consider iterates $\mathbf{x}_k$ and $\mathbf{z}_k$ in Algorithm 1. Let $u = 1/L$ and $\gamma = 2$, choose $\eta = \widetilde{O}(1/\kappa \wedge 1/(\kappa^{1/3}m^{2/3}))$, and assume that $\eta^2 \leq \log(2)/(36\kappa K)$, we have the following union bounds on $\mathbb{E}[\|\mathbf{x} - \mathbf{x}^*\|_2^2]$, $\mathbb{E}[f(\mathbf{x}_k)] - f(\mathbf{x}^*)$ and $\mathbb{E}[\|\mathbf{v}_k\|_2^2]$,

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|_2^2] &\leq \frac{42d}{\mu} + 24\|\mathbf{x}^*\|_2^2 + \frac{8d}{L} \triangleq U_x,\\
\mathbb{E}[f(\mathbf{x}_k)] - f(\mathbf{x}^*) &\leq 21d\kappa + 12L\|\mathbf{x}^*\|_2^2 + 4d \triangleq U_f,\\
\mathbb{E}[\|\mathbf{v}_k\|_2^2] &\leq \frac{80d}{\mu} + 48\|\mathbf{x}^*\|_2^2 + \frac{18d}{L} \triangleq U_v.
\end{aligned}
$$

Moreover, it can be seen that U_x and U_v are both in the order of $O(d/\mu)$, and U_f is in the order of $O(d\kappa)$.

B.1. Proof of Lemma A.2

Proof of Lemma A.2. In discrete update (3.1), the added Gaussian noises $\boldsymbol{\epsilon}_k^x$ and $\boldsymbol{\epsilon}_k^v$ have mean $\mathbf{0}$ and satisfy

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{\epsilon}_k^v(\boldsymbol{\epsilon}_k^v)^\top] &= u(1 - e^{-2\gamma\eta}) \cdot \mathbf{I}_{d\times d},\\
\mathbb{E}[\boldsymbol{\epsilon}_k^x(\boldsymbol{\epsilon}_k^x)^\top] &= \frac{u}{\gamma^2}(2\gamma\eta + 4e^{-\gamma\eta} - e^{-2\gamma\eta} - 3) \cdot \mathbf{I}_{d\times d},\\
\mathbb{E}[\boldsymbol{\epsilon}_k^v(\boldsymbol{\epsilon}_k^x)^\top] &= \frac{u}{\gamma}(1 - 2e^{-\gamma\eta} + e^{-2\gamma\eta}) \cdot \mathbf{I}_{d\times d},
\end{aligned}
\tag{B.1}
$$

which are identical to those of the Brownian motions in Langevin diffusion (A.2) and (A.3) with time $t = \eta$ by Lemma A.1. Note that when $0 < x < 1$, we have $1 - x \leq \exp(-x) \leq 1 - x + x^2/2$. Thus assuming $2\gamma\eta \leq 1$, we obtain

$$
\mathbb{E}[\|\boldsymbol{\epsilon}_k^v\|_2^2] \leq 2\gamma u\eta d, \quad \mathbb{E}[\|\boldsymbol{\epsilon}_k^x\|_2^2] \leq 2u\eta^2 d, \text{ and } \quad \mathbb{E}[\langle \epsilon_k^v, \epsilon_k^x\rangle] \leq 2\gamma u\eta^2 d. \tag{B.2}
$$

Therefore, we are able to apply synchronous coupling argument, i.e., considering shared Brownian terms in both discrete update (3.1) and Langevin diffusion.

Firstly, we are going to bound the discretization error in the velocity variable $\mathbf{v}$. Let $\boldsymbol{X}_0 = \mathbf{x}_k$, $\boldsymbol{V}_0 = \mathbf{v}_k$, $\boldsymbol{X}_s = \mathcal{L}_s\mathbf{x}_k$ and $\boldsymbol{V}_s = \mathcal{L}_s\mathbf{v}_k$. Based on (3.1) and (A.2), we have

$$
\begin{aligned}
\mathbb{E}[\|\mathcal{G}_\eta\mathbf{v}_k - \mathcal{L}_\eta\mathbf{v}_k\|_2^2] &= \mathbb{E}\bigg[\bigg\|\mathbf{v}_k(1 - \gamma\eta - e^{-\gamma\eta}) + \frac{u}{\gamma}(1 - \gamma\eta - e^{-\gamma\eta})\nabla f(\mathbf{x}_k)\\
&\qquad + u\int_0^\eta e^{-\gamma(\eta - s)}\big[\nabla f(\boldsymbol{X}_s) - \nabla f(\boldsymbol{X}_0)\big]\mathrm{d}s\bigg\|_2^2\bigg]\\
&\leq 3\mathbb{E}\bigg[\frac{\gamma^4\eta^4}{4}\|\mathbf{v}_k\|_2^2 + \frac{u^2\gamma^2\eta^4}{4}\|\nabla f(\mathbf{x}_k)\|_2^2 + u^2\bigg\|\int_0^\eta e^{-\gamma(\eta-s)}\big[\nabla f(\boldsymbol{X}_s) - \nabla f(\boldsymbol{X}_0)\big]\mathrm{d}s\bigg\|_2^2\bigg],
\end{aligned}
\tag{B.3}
$$

where the inequality follows from facts that $|1-x-e^{-x}| \leq x^2/2$ when $0 \leq x \leq 1$ and $\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|_2^2 \leq 3(\|\mathbf{x}\|_2^2+\|\mathbf{y}\|_2^2+\|\mathbf{z}\|_2^2)$. In terms of the third term on the R.H.S of (B.3), we have

$$
\begin{aligned}
\mathbb{E}\bigg[\bigg\|\int_0^\eta e^{-\gamma(\eta-s)}\big[\nabla f(\boldsymbol{X}_s) - \nabla f(\boldsymbol{X}_0)\big]\mathrm{d}s\bigg\|_2^2\bigg] &\leq \eta\mathbb{E}\bigg[\int_0^\eta \big\|e^{-\gamma(\eta-s)}\big[\nabla f(\boldsymbol{X}_s) - \nabla f(\boldsymbol{X}_0)\big]\big\|_2^2\mathrm{d}s\bigg]\\
&\leq \eta\mathbb{E}\bigg[\int_0^\eta \big\|\nabla f(\boldsymbol{X}_s) - \nabla f(\boldsymbol{X}_0)\big\|_2^2\mathrm{d}s\bigg]\\
&\leq \eta L^2\bigg[\int_0^\eta \mathbb{E}\big\|\boldsymbol{X}_s - \boldsymbol{X}_0\big\|_2^2\mathrm{d}s\bigg],
\end{aligned}
$$

where the first inequality follows from inequality $\|\int_0^t \mathbf{x}(s)ds\|_2^2 \leq t\int_0^t \|\mathbf{x}(s)\|_2^2 ds$, the second inequality is due to $\exp(-x) \leq 1$, and the last inequality follows from Assumption 4.1. Note that $\mathrm{d}\boldsymbol{X}_s = \boldsymbol{V}_s\mathrm{d}s$, we further have

$$
\begin{aligned}
\eta L^2\bigg[\int_0^\eta \mathbb{E}\big\|\boldsymbol{X}_s - \boldsymbol{X}_0\big\|_2^2\mathrm{d}s\bigg] &= \eta L^2\bigg[\int_0^\eta \mathbb{E}\bigg\|\int_0^s \boldsymbol{V}_r dr\bigg\|_2^2\mathrm{d}s\bigg]\\
&\leq \eta L^2\bigg[\int_0^\eta s\int_0^s \mathbb{E}\|\boldsymbol{V}_r\|_2^2\mathrm{d}r\mathrm{d}s\bigg],
\end{aligned}
$$

where the last inequality is due to the fact that $\|\int_0^t \mathbf{x}(s)\mathrm{d}s\|_2^2 \le t\int_0^t \|\mathbf{x}(s)\|_2^2\mathrm{d}s$. By Lemma B.1, we know that

$$\mathbb{E}[\|\boldsymbol{V}_r\|_2^2] \le 2u\mathbb{E}[f(\boldsymbol{X}_0) - f(\mathbf{x}^*) + 2\eta\gamma d] + \mathbb{E}[\|\boldsymbol{V}_0\|_2^2]$$

for $r \le \eta$. Thus, it follows that

$$\eta L^2\bigg[\int_0^\eta s\int_0^s \mathbb{E}\|\boldsymbol{V}_r\|_2^2\mathrm{d}r\mathrm{d}s\bigg] \le \frac{\eta^4L^2\big[2u\mathbb{E}[f(\boldsymbol{X}_0)-f(\mathbf{x}^*)+2\eta\gamma d]\big]+\mathbb{E}[\|\boldsymbol{V}_0\|_2^2]}{3}.$$

Substituting the above into (B.3), we obtain

$$\begin{aligned}
&\mathbb{E}[\|\mathcal{G}_\eta\mathbf{v}_k - \mathcal{L}_\eta\mathbf{v}_k\|_2^2]\\
&\le \eta^4\mathbb{E}\bigg[\frac{3\gamma^4}{4}\|\boldsymbol{V}_0\|_2^2 + \frac{3u^2\gamma^2}{4}\|\nabla f(\boldsymbol{X}_0)\|_2^2 + u^2L^2\big[2u[f(\boldsymbol{X}_0)-f(\mathbf{x}^*)+2\eta\gamma d] + \|\boldsymbol{V}_0\|_2^2\big]\bigg]\\
&\le \eta^4\bigg[\Big(\frac{3\gamma^4}{4}+u^2L^2\Big)\mathbb{E}[\|\mathbf{v}_k\|_2^2] + \Big(\frac{3u^2\gamma^2L}{2}+2u^3L^2\Big)\mathbb{E}\big[f(\mathbf{x}_k)-f(\mathbf{x}^*)\big] + 4u^3L^2\eta\gamma d\bigg],
\end{aligned} \tag{B.4}$$

where the second inequality is by facts that $\boldsymbol{V}_0 = \mathbf{v}_k$, $\boldsymbol{X}_0 = \mathbf{x}_k$ and $\|\nabla f(\mathbf{x})\|_2^2 \le 2L\big(f(\mathbf{x}) - f(\mathbf{x}^*)\big)$. Next, we are going to bound the discretization error in the position variable $\mathbf{x}$. Note that the randomness of $\boldsymbol{X}_\eta$ comes from the Brownian term in the velocity variation, and can be also regarded as an additive Gaussian noise, i.e., $\sqrt{2\gamma u}\int_0^\eta \mathrm{d}t\int_0^t e^{-\gamma(t-s)}\mathrm{d}\boldsymbol{B}_s$. Note that we utilize the identical random variable in the discrete update (3.1), which implies that the coupling technique can still be used in the discretization error computation in $\mathbf{x}_k$. Let $\widetilde{\boldsymbol{V}}_t = \boldsymbol{V}_0e^{-\gamma t} - u\int_0^t e^{-\gamma(t-s)}\nabla f(\boldsymbol{X}_t)\mathrm{d}s$, we have

$$\begin{aligned}
\mathbb{E}[\|\mathcal{G}_\eta\mathbf{x}_k - \mathcal{L}_\eta\mathbf{x}_k\|_2^2] &= \mathbb{E}\bigg[\bigg\|\int_0^\eta\big(\boldsymbol{V}_0 - \widetilde{\boldsymbol{V}}_t\big)\mathrm{d}t\bigg\|_2^2\bigg]\\
&\le \eta\int_0^\eta \mathbb{E}[\|\boldsymbol{V}_0 - \widetilde{\boldsymbol{V}}_t\|_2^2]\mathrm{d}t\\
&= \eta\int_0^\eta \mathbb{E}\bigg[\bigg\|\boldsymbol{V}_0(1-e^{-\gamma t}) + u\int_0^t e^{-\gamma(t-s)}\nabla f(\boldsymbol{X}_s)\mathrm{d}s\bigg\|_2^2\bigg]\mathrm{d}t\\
&\le \eta\int_0^\eta\bigg\{2\gamma^2t^2\mathbb{E}[\|\boldsymbol{V}_0\|_2^2] + 2u^2\mathbb{E}\bigg[\bigg\|\int_0^t e^{-\gamma(t-s)}\nabla f(\boldsymbol{X}_s)\mathrm{d}s\bigg\|_2^2\bigg]\bigg\}\mathrm{d}t\\
&\le \frac{2\gamma^2\eta^4}{3}\mathbb{E}[\|\boldsymbol{V}_0\|_2^2] + 2u^2\eta\int_0^\eta t\int_0^t \mathbb{E}[\|\nabla f(\boldsymbol{X}_s)\|_2^2]\mathrm{d}s\mathrm{d}t.
\end{aligned}$$

From Lemma B.1, it can be seen that

$$\mathbb{E}\|\nabla f(\boldsymbol{X}_s)\|_2^2 \le 2L\bigg(\mathbb{E}[f(\boldsymbol{X}_0) - f(\mathbf{x}^*)] + \frac{\mathbb{E}[\|\mathbf{v}_k\|_2^2]}{2u} + 2\gamma d\eta\bigg)$$

for any $s \le \eta$, thus we have

$$2u^2\eta\int_0^\eta t\int_0^t \mathbb{E}[\|\nabla f(\boldsymbol{X}_s)\|_2^2]\mathrm{d}s\mathrm{d}t \le \frac{4u^2L\eta^4}{3}\bigg(\mathbb{E}[f(\boldsymbol{X}_0) - f(\mathbf{x}^*)] + \frac{\mathbb{E}[\|\mathbf{v}_k\|_2^2]}{2u} + 2\gamma d\eta\bigg).$$

Then, replacing $\boldsymbol{V}_0$ and $\boldsymbol{X}_0$ by $\mathbf{v}_k$ and $\mathbf{x}_k$ respectively, the discretization error in
xb_k is bounded by

$$\mathbb{E}[\|\mathcal{G}_\eta\mathbf{x}_k - \mathcal{L}_\eta\mathbf{x}_k\|_2^2] \le \eta^4\bigg[\bigg(\frac{2\gamma^2+2uL}{3}\bigg)\mathbb{E}[\|\mathbf{v}_k\|_2^2] + \frac{4u^2L}{3}\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] + \frac{8u^2L\gamma d\eta}{3}\bigg]. \tag{B.5}$$

Finally, by Lemma B.2, we have uniform bounds U_v and U_f on $\mathbb{E}[\|\mathbf{v}_k\|_2^2]$ and $\mathbb{E}[f(\mathbf{x}_k)] - f(\mathbf{x}^*)$, substituting these bounds into (B.4) and (B.5), we are able to complete the proof. □

B.2. Proof of Lemma A.3

Proof of Lemma A.3. Note that the update for $\mathbf{x}_k$ does not contain the gradient term, which implies $\mathcal{S}_\eta \mathbf{x}_k = \mathcal{G}_\eta \mathbf{x}_k$ and $\mathbb{E}[\|\mathcal{S}_\eta \mathbf{x}_k - \mathcal{G}_\eta \mathbf{x}_k\|_2^2] = 0$. In the sequel, we mainly consider the velocity variable. Applying coupling argument, it can be directly observed that

$$\begin{aligned}\mathbb{E}[\|\mathcal{S}_\eta \mathbf{v}_k - \mathcal{G}_\eta \mathbf{v}_k\|_2^2] &= \eta^2 u^2 \mathbb{E}\big[\big\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}}_j) - \big(\nabla f(\mathbf{x}_k) - \nabla f(\widetilde{\mathbf{x}}_j)\big)\big\|_2^2\big] \\ &\le \eta^2 u^2 \mathbb{E}\big[\big\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}}_j)\big\|_2^2\big] \\ &\le \eta^2 u^2 L^2 \mathbb{E}\big[\big\|\mathbf{x}_k - \widetilde{\mathbf{x}}_j\big\|_2^2\big], \end{aligned} \tag{B.6}$$

where the first inequality is by the fact that $\mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|_2^2] \le \mathbb{E}[\|\mathbf{x}\|_2^2]$, and the second inequality follows from Assumption 4.1. Note that by (3.1), we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_k - \widetilde{\mathbf{x}}_j\|_2^2] &= \mathbb{E}\Bigg[\Bigg\|\sum_{r=jm}^{jm+l-1} \eta \mathbf{v}_r + \boldsymbol{\epsilon}_r^x\Bigg\|_2^2\Bigg] \\ &\le 2\mathbb{E}\Bigg[\Bigg\|\sum_{r=jm}^{jm+l-1} \eta \mathbf{v}_r\Bigg\|_2^2\Bigg] + 2\mathbb{E}\Bigg[\Bigg\|\sum_{r=jm}^{jm+l-1} \boldsymbol{\epsilon}_r^x\Bigg\|_2^2\Bigg] \\ &= 2\eta^2 \mathbb{E}\Bigg[\Bigg\|\sum_{r=jm}^{jm+l-1} \mathbf{v}_r\Bigg\|_2^2\Bigg] + 2\sum_{r=jm}^{jm+l-1} \mathbb{E}[\|\boldsymbol{\epsilon}_r^x\|_2^2] \\ &\le 2l\eta^2 \sum_{r=jm}^{jm+l-1} \mathbb{E}[\|\mathbf{v}_r\|_2^2] + 4lu\eta^2 d, \end{aligned} \tag{B.7}$$

where the first inequality is due to $(a+b)^2 \le 2(a^2+b^2)$, the second equation is due to the independence among Gaussian random variables ϵ_r^x and the last inequality is due to $\big(\sum_{i=1}^n a_i\big)^2 \le n\sum_{i=1}^n a_i^2$ and (B.2). Let U_v denote the union upper bound of $\mathbb{E}[\|\mathbf{v}_k\|_2^2]$ for all $0 \le k \le K$, (B.7) can be further relaxed as follows

$$\mathbb{E}\|\mathbf{x}_k - \widetilde{\mathbf{x}}_j\|_2^2 \le 2\eta^2(l^2 U_v + 2lud) \le 2\eta^2(m^2 U_v + 2mud).$$

Since $m \le m^2$, we are able to complete the proof by submitting the above inequality into (B.6) and setting $D_3 = 2(U_v + 2ud)$, i.e.,

$$\mathbb{E}[\|\mathcal{S}_\eta \mathbf{v}_k - \mathcal{G}_\eta \mathbf{v}_k\|_2^2] \le 2\eta^4 u^2 L^2 m^2 (U_v + 2ud) \triangleq D_3 u^2 L^2 m^2 \eta^4.$$

□

B.3. Proof of Lemma A.6

Proof of Lemma A.6. Note that for random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, we have

$$\big(\mathbb{E}[\langle \boldsymbol{X}, \boldsymbol{Y}\rangle]\big)^2 = \bigg(\sum_{i=1}^d \mathbb{E}\boldsymbol{X}_i \boldsymbol{Y}_i\bigg)^2 \le \bigg(\sum_{i=1}^d (\mathbb{E}\boldsymbol{X}_i^2)^{1/2}\mathbb{E}(\boldsymbol{Y}_i^2)^{1/2}\bigg)^2 \le \bigg(\sum_{i=1}^d \mathbb{E}\boldsymbol{X}_i^2\bigg)\bigg(\sum_{i=1}^d \mathbb{E}\boldsymbol{Y}_i^2\bigg) = \mathbb{E}[\|\boldsymbol{X}\|_2^2]\mathbb{E}[\|\boldsymbol{Y}\|_2^2],$$

where the first and second inequalities are by Hölder's inequality and Cauchy-Schwarz inequality respectively. Thus, it follows that

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{X} + \boldsymbol{Y}\|_2^2] &= \mathbb{E}[\|\boldsymbol{X}\|_2^2 + \|\boldsymbol{Y}\|_2^2 + 2\langle \boldsymbol{X}, \boldsymbol{Y}\rangle] \\ &\le \mathbb{E}[\|\boldsymbol{X}\|_2^2] + \mathbb{E}[\|\boldsymbol{Y}\|_2^2] + 2\sqrt{\mathbb{E}[\|\boldsymbol{X}\|_2^2]\mathbb{E}[\|\boldsymbol{Y}\|_2^2]} = \Big(\sqrt{\mathbb{E}[\|\boldsymbol{X}\|_2^2]} + \sqrt{\mathbb{E}[\|\boldsymbol{Y}\|_2^2]}\Big)^2, \end{aligned} \tag{B.8}$$

which completes the proof. □

C. Proof of Auxiliary Lemmas

In this section, we prove extra lemmas used in our proof.

C.1. Proof of Lemma B.1

Proof. We consider the Lyapunov function $\mathcal{E}_t = \mathbb{E}[f(\boldsymbol{X}_t) + \|\boldsymbol{V}_t\|_2^2/(2u)]$, which corresponds to the expected total energy of such dynamic system. By Itô's lemma, we have

$$\begin{aligned}\frac{\mathrm{d}\mathcal{E}_t}{\mathrm{d}t} &= \frac{1}{\mathrm{d}t}\big[\mathbb{E}\langle\nabla_{\boldsymbol{V}_t}\mathcal{E}_t, d\boldsymbol{V}_t\rangle + \mathbb{E}\langle\nabla_{\boldsymbol{X}_t}\mathcal{E}_t, d\boldsymbol{X}_t\rangle\big] + \gamma u\mathbb{E}\langle\nabla^2_{\boldsymbol{V}_t}\mathcal{E}_t, \mathbf{I}\rangle\\ &= \frac{1}{u}\mathbb{E}[-\gamma\|\boldsymbol{V}_t\|_2^2 - u\langle\boldsymbol{V}_t, \nabla f(\boldsymbol{X}_t)\rangle] + \mathbb{E}[\langle\nabla f(\boldsymbol{X}_t), \boldsymbol{V}_t\rangle] + \gamma d\\ &= \gamma d - \frac{\gamma}{u}\mathbb{E}[\|\boldsymbol{V}_t\|_2^2]\\ &\le \gamma d,\end{aligned}$$

where in the third equation we use the martingale property of $\mathrm{d}\boldsymbol{B}_t$. Thus, we have $\mathcal{E}_t \le t\gamma d + \mathcal{E}_0$. Adding the term $-f(\mathbf{x}^*)$ on the both sides, we have

$$\mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)] + \mathbb{E}[\|\boldsymbol{V}_t\|_2^2]/(2u) \le t\gamma d + f(\boldsymbol{X}_0) - f(\mathbf{x}^*) + \|\boldsymbol{V}_0\|_2^2/(2u). \tag{C.1}$$

Note that both terms $\mathbb{E}[\|\boldsymbol{V}_t\|_2^2]$ and $\mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)]$ are positive, which immediately implies that

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{V}_t\|_2^2] &\le 2u\big[f(\boldsymbol{X}_0) - f(\mathbf{x}^*) + \gamma dt\big] + \|\boldsymbol{V}_0\|_2^2,\\ \mathbb{E}[f(\boldsymbol{X}_t)] &\le f(\boldsymbol{X}_0) + \frac{\|\boldsymbol{V}_0\|_2^2}{2u} + \gamma dt.\end{aligned}$$

Moreover, note that $\mathbf{x}^* = \mathrm{argmin}\, f(\mathbf{x})$ and thus

$$\begin{aligned}f(\mathbf{x}^*) - f(\mathbf{x}) &\le f(\mathbf{x} - \nabla f(\mathbf{x})/L) - f(\mathbf{x})\\ &\le \langle\nabla f(\mathbf{x}), -\nabla f(\mathbf{x})/L\rangle + \|\nabla f(\mathbf{x})\|_2^2/(2L)\\ &= -\|\nabla f(\mathbf{x})\|_2^2/(2L),\end{aligned}$$

where the second inequality is due to Assumption 4.1, which further implies that

$$\mathbb{E}[\|\nabla f(\boldsymbol{X}_t)\|_2^2] \le 2L\mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)].$$

By (C.1) we have

$$\mathbb{E}[f(\boldsymbol{X}_t) - f(\mathbf{x}^*)] \le t\gamma d + f(\boldsymbol{X}_0) - f(\mathbf{x}^*) + \|\boldsymbol{V}_0\|_2^2/(2u), \tag{C.2}$$

which further indicates

$$\mathbb{E}[\|\nabla f(\boldsymbol{X}_t)\|_2^2] \le 2L\bigg(f(\boldsymbol{X}_0) - f(\mathbf{x}^*) + \frac{\|\boldsymbol{V}_0\|_2^2}{2u} + \gamma dt\bigg).$$

Thus, we complete the proof. □

C.2. Proof of Lemma B.2

To prove Lemma B.2, we need the following lemma.

Lemma C.1. Under Assumptions 4.1 and 4.2, when $\eta \le 1/(2\gamma)$, expectations $\mathbb{E}[f(\mathbf{x}_k)]$ and $\mathbb{E}[\|\mathbf{v}_k\|_2^2]$ are upper bounded as follows,

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}_k)] &\le \frac{1}{1-\gamma\eta}\bigg[e^{G_1T\eta}\mathcal{E}_0 + T(\gamma d + G_0\eta) + \frac{1}{2}T^2\eta G_1 e^{G_1T\eta}(\gamma d + \eta G_0)\bigg],\\ \mathbb{E}[\|\mathbf{v}_k\|_2^2] &\le 2u\bigg[e^{G_1T\eta}\mathcal{E}_0 + T(\gamma d + G_0\eta) + \frac{1}{2}T^2\eta G_1 e^{G_1T\eta}(\gamma d + \eta G_0) - f(\mathbf{x}^*)\bigg],\end{aligned}$$

where $T = K\eta$ denotes the length of time, $G_0 = 6u\mathbb{E}[\|\nabla f_i(\mathbf{x}^*)\|] + 2L\gamma ud - 18uL\kappa f(\mathbf{x}^*)$, and $G_1 = 36uL\kappa$.

Proof of Lemma B.2. We first prove the upper bound for $\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^\pi\|_2^2]$. Applying triangle inequality yields

$$\begin{aligned} \mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|_2^2] &\le 2\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^\pi\|_2^2] + 2\mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}^*\|_2^2] \\ &\le 2w_k^2 + \frac{2d}{\mu}, \end{aligned} \tag{C.3}$$

where the second inequality comes from Lemma A.7 and $w_k = \big(\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^\pi\|_2^2] + \mathbb{E}[\|\mathbf{x}_k + \mathbf{v}_k - \mathbf{x}^\pi - \mathbf{v}^\pi\|_2^2]\big)^{1/2}$. According to (A.6), (A.7),(A.9) and (A.10), we have

$$w_{k+1}^2 \le \Big(e^{-\eta/(2\kappa)} w_k + 2\sqrt{\mathbb{E}[\|\mathcal{G}_\eta \mathbf{x}_k - \mathcal{L}_\eta \mathbf{x}_k\|_2^2]} + \sqrt{\mathbb{E}[\|\mathcal{G}_\eta \mathbf{v}_k - \mathcal{L}_\eta \mathbf{v}_k\|_2^2]}\Big)^2 + \mathbb{E}[\|\mathcal{S}_\eta \mathbf{v}_k - \mathcal{G}_\eta \mathbf{v}_k\|_2^2].$$

By (B.4), (B.5), (B.6) and (B.7), we have

$$\begin{aligned} \mathbb{E}[\|\mathcal{G}_\eta \mathbf{x}_k - \mathcal{L}_\eta \mathbf{x}_k\|_2^2] &\le \eta^4 \bigg[\Big(\frac{2\gamma^2 + 2uL}{3}\Big)\widetilde{U}_v + \frac{4u^2 L}{3}\widetilde{U}_f + \frac{8u^2 L\gamma d\eta}{3}\bigg] = \widetilde{D}_1 \eta^4, \\ \mathbb{E}[\|\mathcal{G}_\eta \mathbf{v}_k - \mathcal{L}_\eta \mathbf{v}_k\|_2^2] &\le \eta^4 \bigg[\Big(\frac{3\gamma^4}{4} + u^2 L^2\Big)\widetilde{U}_v + \Big(\frac{3u^2\gamma^2 L}{2} + 4u^3 L^2\Big)\widetilde{U}_f + 4u^3 L^2 \eta\gamma d\bigg] = \widetilde{D}_2 \eta^4, \\ \mathbb{E}[\|\mathcal{S}_\eta \mathbf{v}_k - \mathcal{G}_\eta \mathbf{v}_k\|_2^2] &\le 2(\widetilde{U}_v + 2ud) m^2 u^2 L^2 \eta^2 = \widetilde{D}_3 m^2 u^2 L^2 \eta^2, \end{aligned} \tag{C.4}$$

where $\widetilde{U}_v$ and $\widetilde{U}_f$ denote any uniform upper bounds for $\mathbb{E}[\|\mathbf{v}_k\|_2^2]$ and $\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)]$ respectively. Applying Lemma A.5 yields

$$\begin{aligned} w_k &\le e^{-k\eta/(2\kappa)} w_0 + \frac{2\sqrt{\widetilde{D}_1}\eta^2 + \sqrt{\widetilde{D}_2}\eta^2}{1 - e^{-\eta/(2\kappa)}} + \frac{\sqrt{\widetilde{D}_3} m\eta^2}{\sqrt{1 - e^{-\eta/(2\kappa)}}} \\ &\le w_0 + 4\eta\kappa\Big(2\sqrt{\widetilde{D}_1} + \sqrt{\widetilde{D}_2}\Big) + 2\sqrt{\kappa \widetilde{D}_3} m\eta^{3/2}, \end{aligned} \tag{C.5}$$

where we use the fact that $e^{-k\eta/(2\kappa)} < 1$ and $1 - e^{-\eta/(2\kappa)} \ge \eta/(4\kappa)$ when $0 < \eta/\kappa \le 1$. It is then left to show the order of $\widetilde{D}_1$, $\widetilde{D}_2$ and $\widetilde{D}_3$. To this end, we need to find uniform upper bounds for $\mathbb{E}[\|\mathbf{v}_k\|_2^2]$ and $\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)]$ by (C.4), namely, we need to find the order of $\widetilde{U}_v$ and $\widetilde{U}_f$. In the following, we will show this by applying Lemma C.1. Denote T as $T = k\eta$ and consider sufficiently small η such that $G_1 T\eta \le \log(2)$, $G_0\eta \le \gamma d$ and $\gamma\eta \le 1/2$, by Lemma C.1 we obtain the following upper bounds for $\mathbb{E}[\|\mathbf{v}_k\|_2^2]$ and $\mathbb{E}[f(\mathbf{x}_k)] - f(\mathbf{x}^*)$

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_k)] - f(\mathbf{x}^*) &\le 2\big(2\mathcal{E}_0 + 2T\gamma d + 2\log(2)T\gamma d\big) + |f(\mathbf{x}^*)| \le 4(\mathcal{E}_0 + 2T\gamma d) + |f(\mathbf{x}^*)| = \widetilde{U}_f, \\ \mathbb{E}[\|\mathbf{v}_k\|_2^2] &\le 2u\big(2\mathcal{E}_0 + 4T\gamma d + |f(\mathbf{x}^*)|\big) \le u\widetilde{U}_f = \widetilde{U}_v. \end{aligned}$$

In addition, since $u = 1/L$ and $\gamma = 2$, we can write $\widetilde{U}_f = O(Td) = O(\kappa d\log(1/\epsilon))$ and $\widetilde{U}_v = O(d\log(1/\epsilon)/\mu)$. Recall the definition of $\widetilde{D}_1$, $\widetilde{D}_2$ and $\widetilde{D}_3$ in (C.4), for sufficiently small $\eta < 1/2\gamma = 1/4$, we have

$$\begin{aligned} \widetilde{D}_1 &\le 10\widetilde{U}_v/3 + 4u\widetilde{U}_f/3 + 4ud/3 \le 5\widetilde{U}_v + 2ud, \\ \widetilde{D}_2 &\le 13\widetilde{U}_v + 10u\widetilde{U}_f + 2ud = 23\widetilde{U}_v + 2ud, \\ \widetilde{D}_3 &\le 2\widetilde{U}_v + 4ud. \end{aligned} \tag{C.6}$$

We choose step size η in (C.5) such that $4\eta\kappa\Big(2\sqrt{\widetilde{D}_1} + \sqrt{\widetilde{D}_2}\Big) \le \sqrt{d/\mu}$ and $2\sqrt{\kappa\widetilde{D}_3} m\eta^{3/2} \le \sqrt{d/\mu}$. To this end, we let

$$\eta \le \min\bigg\{\frac{1}{4\kappa(2\sqrt{\widetilde{D}_1\mu/d} + \sqrt{\widetilde{D}_2\mu/d})}, \bigg(\frac{1}{2n\sqrt{\kappa\widetilde{D}_3\mu/d}}\bigg)^{3/2}\bigg\} = \widetilde{O}(1/\kappa \wedge 1/(\kappa^{1/3}m^{2/3})),$$

where the equation is calculated based on (C.6). Then by (C.5) we have

$$w_k^2 \le \Big(w_0 + 2\sqrt{d/\mu}\Big)^2 \le 2w_0^2 + \frac{8d}{\mu}.$$

Now we deal with w_0. Note that $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{v}_0 = \mathbf{0}$. By the definition of w_k, we have

$$\begin{aligned} w_0^2 &= \mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}_0\|_2^2 + \|\mathbf{x}^\pi + \mathbf{v}^\pi - \mathbf{x}_0 - \mathbf{v}_0\|_2^2] \\ &\le 3\mathbb{E}[\|\mathbf{x}^\pi\|_2^2] + 2\mathbb{E}[\|\mathbf{v}^\pi\|_2^2] \\ &\le \frac{6d}{\mu} + 6\|\mathbf{x}^*\|_2^2 + \frac{2d}{L}, \end{aligned}$$

where the first inequality comes form triangle inequality and in the second inequality we use facts that $\mathbb{E}[\|\mathbf{x}^\pi\|_2^2] = 2\mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}^*\|_2^2] + 2\|\mathbf{x}^*\|_2^2$, $\mathbb{E}[\|\mathbf{v}^\pi\|_2^2] = 1/\sqrt{(2\pi)^d}\int_{\mathbb{R}^d}\|\mathbf{v}\|_2^2\exp(-\|\mathbf{v}\|_2^2/2u)d\mathbf{v} = ud = d/L$ and $\mathbb{E}[\|\mathbf{x}^\pi - \mathbf{x}^*\|_2^2] \le d/\mu$ by Lemma A.7. Applying (C.3) we further have

$$\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|_2^2] \le 2w_k^2 + \frac{2d}{\mu} \le 2\Big(2w_0^2 + \frac{8d}{\mu}\Big) + \frac{2d}{\mu} = \frac{42d}{\mu} + 24\|\mathbf{x}^*\|_2^2 + \frac{8d}{L},$$

which completes the proof for the upper bound of $\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|_2^2]$. Moreover, according to Assumption 4.1, we have

$$\mathbb{E}[f(\mathbf{x}_k)] - f(\mathbf{x}^*) \le \frac{L\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|_2^2]}{2} \le 21d\kappa + 12L\|\mathbf{x}^*\|_2^2 + 4d.$$

In the following, we are going to prove the union upper bound on $\mathbb{E}[\|\mathbf{v}_k\|_2^2]$. Similar to the proof of U_x, we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{v}_k\|_2^2] &= \mathbb{E}[\|\mathbf{v}_k - \mathbf{v}^\pi + \mathbf{v}^\pi\|] \\ &\le 2\mathbb{E}[\|\mathbf{v}^\pi\|_2^2] + 2\mathbb{E}[\|\mathbf{v}^\pi - \mathbf{v}_k\|_2^2] \\ &\le 2\mathbb{E}[\|\mathbf{v}^\pi\|_2^2] + 4\mathbb{E}[\|\mathbf{v}^\pi - \mathbf{v}_k + \mathbf{x}^* - \mathbf{x}_k\|_2^2] + 4\mathbb{E}[\|\mathbf{x}^* - \mathbf{x}_k\|_2^2] \\ &= 2\mathbb{E}[\|\mathbf{v}^\pi\|_2^2] + 4w_k^2. \end{aligned}$$

Note that $w_k^2 \le 2w_0^2 + 8d/\mu \le 20d/\mu + 12\|\mathbf{x}^*\|_2^2 + 4d/L$ and $\mathbb{E}[\|\mathbf{v}^\pi\|_2^2] = d/L$, we have

$$\mathbb{E}[\|\mathbf{v}_k\|_2^2] \le \frac{80d}{\mu} + \frac{18d}{L} + 48\|\mathbf{x}^*\|_2^2 \triangleq U_v,$$

which completes our proof. □

C.3. Proof of Lemma C.1

Proof. Recall the discrete update form (3.1) and the proposed SVR-HMC algorithm. Let $k = jm + l$, we first rewrite the l-th update in the j-th epoch as follows,

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k + \eta\mathbf{v}_k + \boldsymbol{\epsilon}_k^x, \\ \mathbf{v}_{k+1} &= \mathbf{v}_k - \gamma\eta\mathbf{v}_k - \eta u\mathbf{g}_k + \boldsymbol{\epsilon}_k^v, \end{aligned} \tag{C.7}$$

where $\mathbf{g}_k = \nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}}_j) + \nabla f(\widetilde{\mathbf{x}}_j)$.

In order to show the upper bounds of $\mathbb{E}[f(\mathbf{x}_k)]$ and $\mathbb{E}[\|\mathbf{v}_k\|_2^2]$, we consider the Lyapunov function $\mathcal{E}_k = \mathbb{E}[(1-\gamma\eta)f(\mathbf{x}_k) + \|\mathbf{v}_k\|_2^2/(2u)]$. In what follows, we aim to establish the relationship between $\mathcal{E}_{k+1}$ and $\mathcal{E}_k$. To begin with, we deal with $\mathbb{E}[f(\mathbf{x}_{k+1})]$, which can be upper bounded by

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_{k+1})] &\le \mathbb{E}\bigg[f(\mathbf{x}_k) + \eta\langle\mathbf{v}_k, \nabla f(\mathbf{x}_k)\rangle + \frac{L\|\eta\mathbf{v}_k + \boldsymbol{\epsilon}_k^x\|_2^2}{2}\bigg] \\ &= \mathbb{E}\bigg[f(\mathbf{x}_k) + \eta\langle\mathbf{v}_k, \nabla f(\mathbf{x}_k)\rangle + \frac{L\eta^2\|\mathbf{v}_k\|_2^2}{2}\bigg] + \frac{L}{2}\mathbb{E}[\|\boldsymbol{\epsilon}_k^x\|_2^2]. \end{aligned} \tag{C.8}$$

In terms of $\mathbb{E}\|\mathbf{v}_{k+1}\|_2^2$, we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{v}_{k+1}\|_2^2] &= \mathbb{E}\|\mathbf{v}_k - \gamma\eta\mathbf{v}_k - \eta u\mathbf{g}_k + \boldsymbol{\epsilon}_k^v\|_2^2 \\ &= \mathbb{E}[\|\mathbf{v}_k - \gamma\eta\mathbf{v}_k - \eta u\mathbf{g}_k\|_2^2] + \mathbb{E}[\|\boldsymbol{\epsilon}_k^v\|_2^2]. \end{aligned} \tag{C.9}$$

As for the first term on the R.H.S of the above equation, we have

$$\mathbb{E}[\|\mathbf{v}_k - \gamma\eta\mathbf{v}_k - \eta u\mathbf{g}_k\|_2^2] = \mathbb{E}[\|(1-\gamma\eta)\mathbf{v}_k\|_2^2] - 2(1-\gamma\eta)\eta u\mathbb{E}[\langle\mathbf{v}_k, \mathbf{g}_k\rangle] + \eta^2u^2\mathbb{E}[\|\mathbf{g}_k\|_2^2].$$

Note that

$$\mathbb{E}[\langle\mathbf{v}_k, \mathbf{g}_k\rangle] = \mathbb{E}[\langle\mathbf{v}_k, \mathbb{E}_{i_k}\mathbf{g}_k\rangle] = \mathbb{E}[\langle\mathbf{v}_k, \nabla f(\mathbf{x}_k)\rangle],$$

which immediately implies

$$\begin{aligned}\mathbb{E}[\|\mathbf{v}_k - \gamma\eta\mathbf{v}_k - \eta u\mathbf{g}_k\|_2^2] &= (1-\gamma\eta)^2\mathbb{E}[\|\mathbf{v}_k\|_2^2] - 2(1-\gamma\eta)\eta u\mathbb{E}\big[\langle\mathbf{v}_k, \nabla f(\mathbf{x}_k)\rangle\big] \\ &\quad + \eta^2u^2\mathbb{E}[\|\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\widetilde{\mathbf{x}}_j) + \nabla f(\widetilde{\mathbf{x}}_j)\|_2^2] \\ &\le (1-\gamma\eta)\mathbb{E}[\|\mathbf{v}_k\|_2^2] - 2(1-\gamma\eta)\eta u\mathbb{E}[\langle\mathbf{v}_k, \nabla f(\mathbf{x}_k)\rangle] \\ &\quad + 3\eta^2u^2\mathbb{E}[\|\nabla f_{i_k}(\mathbf{x}_k)\|_2^2 + \|\nabla f_{i_k}(\widetilde{\mathbf{x}}_j)\|_2^2 + \|\nabla f(\widetilde{\mathbf{x}}_j)\|_2^2], \end{aligned} \tag{C.10}$$

where the first inequality follows from the fact that $(a+b+c)^3 \le 3(a^2+b^2+c^2)$ and that $1-\eta\gamma < 1$. Combining (C.8), (C.9) and (C.10), we obtain

$$\begin{aligned}\mathcal{E}_{k+1} &= \mathbb{E}\bigg[(1-\gamma\eta)f(\mathbf{x}_{k+1}) + \frac{\|\mathbf{v}_{k+1}\|_2^2}{2u}\bigg] \\ &\le (1-\gamma\eta)\mathbb{E}f(\mathbf{x}_k) + \frac{1-\gamma\eta + Lu\eta^2(1-\gamma\eta)}{2u}\mathbb{E}\|\mathbf{v}_k\|_2^2 + \frac{3\eta^2u}{2}\mathbb{E}[\|\nabla f_{i_k}(\mathbf{x}_k)\|_2^2 + \|\nabla f_{i_k}(\widetilde{\mathbf{x}}_j)\|_2^2 + \|\nabla f(\widetilde{\mathbf{x}}_j)\|_2^2] \\ &\quad + \frac{(1-\gamma\eta)L}{2}\mathbb{E}[\|\boldsymbol{\epsilon}_k^x\|_2^2] + \frac{\mathbb{E}[\|\boldsymbol{\epsilon}_k^v\|_2^2]}{2u}. \end{aligned} \tag{C.11}$$

From (B.2), we know that

$$\mathbb{E}[\|\boldsymbol{\epsilon}_k^v\|_2^2] \le 2\gamma ud\eta, \quad \text{and} \quad \mathbb{E}[\|\boldsymbol{\epsilon}_k^x\|_2^2] \le 2ud\eta^2.$$

We bound the gradient norm term as follows.

$$\begin{aligned}\mathbb{E}[\|\nabla f_i(\mathbf{x}_k)\|_2^2] &\le 2\mathbb{E}[\|\nabla f_i(\mathbf{x}_k) - f_i(\mathbf{x}^*)\|_2^2] + 2\mathbb{E}[\|\nabla f_i(\mathbf{x}^*)\|_2^2] \\ &\le 2L^2\mathbb{E}[\|\mathbf{x} - \mathbf{x}^*\|_2^2] + 2\mathbb{E}[\|\nabla f_i(\mathbf{x}^*)\|_2^2] \\ &\le \frac{4L^2}{\mu}\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] + 2\mathbb{E}[\|\nabla f_i(\mathbf{x}^*)\|_2^2]. \end{aligned}$$

Upper bounds of $\|\nabla f_{i_k}(\widetilde{\mathbf{x}}_j)\|_2^2$ and $\|\nabla f(\widetilde{\mathbf{x}}_j)\|_2^2$ can be established in the same way. Then (C.11) can be further bounded by

$$\begin{aligned}\mathcal{E}_{k+1} &\le (1-\gamma\eta)\mathbb{E}[f(\mathbf{x}_k)] + \frac{1-\gamma\eta+Lu\eta^2}{2u}\mathbb{E}[\|\mathbf{v}_k\|_2^2] \\ &\quad + 6\eta^2uL\kappa\mathbb{E}[f(\mathbf{x}_k) + 2f(\widetilde{\mathbf{x}}_j) - 3f(\mathbf{x}^*)] + 6\eta^2u\mathbb{E}[\|\nabla f_i(\mathbf{x}^*)\|] + d\eta(\gamma + Lu\eta) \\ &\le \big(1-\gamma\eta+6\eta^2uL\kappa\big)\mathbb{E}[f(\mathbf{x}_k)] + \frac{1-\gamma\eta+Lu\eta^2}{2u}\mathbb{E}[\|\mathbf{v}_k\|_2^2] + 12\eta^2uL\kappa\mathbb{E}[f(\widetilde{\mathbf{x}}_j)] \\ &\quad + \eta\gamma d + \eta^2\big[6u\mathbb{E}[\|\nabla f_i(\mathbf{x}^*)\|_2^2] + Lud - 18uL\kappa f(\mathbf{x}^*)\big]. \end{aligned} \tag{C.12}$$

Note that we have assumed $\gamma\eta \le 1/2$, which further implies that

$$\begin{aligned}(1-\gamma\eta+6\eta^2uL\kappa)\mathbb{E}[f(\mathbf{x}_k)] + \frac{1-\gamma\eta+Lu\eta^2}{2u}\mathbb{E}[\|\mathbf{v}_k\|_2^2] &\le \max\bigg\{\frac{1-\gamma\eta+6\eta^2Lu\kappa}{1-\gamma\eta}, 1-\gamma\eta+Lu\eta^2\bigg\}\mathcal{E}_k \\ &\le (1+12\eta^2uL\kappa)\mathcal{E}_k, \end{aligned}$$

where in the second inequality we use the fact that $(1-\gamma\eta+a)/(1-\gamma\eta) \le 1+2a$ for any $a>0$ and $0<\gamma\eta\le 1/2$. Moreover, since $0<\gamma\eta\le 1/2$, we have $\mathbb{E}[f(\widetilde{\mathbf{x}}_j)] \le 2(1-\gamma\eta)\mathbb{E}[f(\widetilde{\mathbf{x}}_j)] + \mathbb{E}[\|\widetilde{\mathbf{v}}_j\|_2^2]/(u) = 2\mathcal{E}_{jm}$, where we used the fact that $\widetilde{\mathbf{x}}_j = \mathbf{x}_{jm}$. Therefore (C.12) turns to

$$\mathcal{E}_{k+1} \le (1+12\eta^2uL\kappa)\mathcal{E}_k + 24\eta^2uL\kappa\mathcal{E}_{jm} + \eta\gamma d + \eta^2G_0, \tag{C.13}$$

where $G_0 = 6u\mathbb{E}[\|\nabla f_i(\mathbf{x}^*)\|] + 2Lud - 18uL\kappa f(\mathbf{x}^*)$. Note that the inequality (C.13) can be relaxed by

$$\mathcal{E}_{k+1} \le (1 + 36\eta^2 uL\kappa)\max\{\mathcal{E}_k, \mathcal{E}_{jm}\} + \eta\gamma d + \eta^2 G_0. \tag{C.14}$$

We then consider two cases: $\mathcal{E}_k \ge \mathcal{E}_{jm}$ and $\mathcal{E}_{jm} > \mathcal{E}_k$ and analyze the upper bound of $\mathcal{E}_{k+1}$ respectively.

Case I: $\mathcal{E}_k \ge \mathcal{E}_{jm}$. The inequality (C.14) reduces to

$$\mathcal{E}_{k+1} \le (1 + 36\eta^2 uL\kappa)\mathcal{E}_k + \eta\gamma d + \eta^2 G_0,$$

which immediately implies that

$$\begin{aligned}\mathcal{E}_k &\le (1 + 36\eta^2 uL\kappa)^k \mathcal{E}_0 + (\eta\gamma d + \eta^2 G_0)\sum_{i=0}^{k-1}(1 + 36\eta^2 uL\kappa)^i \\ &= (1 + 36\eta^2 uL\kappa)^k \mathcal{E}_0 + (\eta\gamma d + \eta^2 G_0)\frac{(1 + 36\eta^2 uL\kappa)^k - 1}{36\eta^2 uL\kappa}.\end{aligned}$$

Let $G_1 = 36uL\kappa$, and it is easy to verify the following fact for any $0 < G_1\eta^2$.

$$(1 + G_1\eta^2)^k = \exp\big(k\log(1 + G_1\eta^2)\big) \le \exp\big(kG_1\eta^2\big).$$

Then, $\mathcal{E}_{k+1}$ can be further bounded as

$$\begin{aligned}\mathcal{E}_k &\le (1 + G_1\eta^2)^k \mathcal{E}_0 + (\eta\gamma d + \eta^2 G_0)\frac{(1 + G_1\eta^2)^k - 1}{G_1\eta^2} \\ &\le e^{G_1 k\eta^2}\mathcal{E}_0 + (\eta\gamma d + \eta^2 G_0)\frac{e^{G_1 k\eta^2} - 1}{G_1\eta^2} \\ &\le e^{G_1 k\eta^2}\mathcal{E}_0 + (\eta\gamma d + \eta^2 G_0)\frac{G_1 k\eta^2 + e^{G_1 k\eta^2}G_1^2 k^2\eta^4/2}{G_1\eta^2} \\ &= e^{G_1 k\eta^2}\mathcal{E}_0 + k\eta\gamma d + k\eta^2 G_0 + \frac{1}{2}k^2\eta^3 G_1 e^{G_1 k\eta^2}(\gamma d + \eta G_0),\end{aligned} \tag{C.15}$$

where the third inequality holds because $h(y) \le h(0) + h'(0)y + \max_{s\in[0,y]} h''(s)y^2/2$ holds for any $\mathcal{C}^2$ function h.

Case II: $\mathcal{E}_{jm} > \mathcal{E}_k$. In order to obtain the upper bound of $\mathcal{E}_k$, we still need to recursively call (C.14) many times. However, note that $jm \le k$, which implies that we only need to perform recursions less than k times. Thus, (C.15) remains true.

Finally, using facts that $\mathbb{E}[f(\mathbf{x}_k)] - f(\mathbf{x}^*) \ge 0$, $\mathbb{E}[\|\mathbf{v}_k\|_2^2] \ge 0$ and the definition of $\mathcal{E}_k$, replacing k in (C.15) by K, we arrive at the arguments proposed in this lemma. □