A Constructions of the Markov Chain and the "Hard" Function F

A.1 Construction of Markov chains

In this section, we construct a Markov chain that is used for the lower bound proofs. The idea is to construct a chain such that 1) there exist two states between which at least τ steps must be take to transit; 2) the hitting time of the constructed chain is upper bounded by τ. Without loss of generality we assume τ is even.

In particular, we consider a directed cyclic-like chain with self-loops. Denote $s = i$ be the i-th state of the Markov chain for $i \in \{0, 1, \ldots, 2\tau' + 1\}$ with $\tau' = \tau/2$. Then for any $q \in (0, 1/2)$ the transition of the chain is defined as follows:

- $P(s = 2|s = 0) = P(s = 0|s = 0) = P(s = 2|s = 1) = P(s = 1|s = 1) = 1/2$;
- $P(s = \tau' + 3|s = \tau' + 1) = P(s = \tau' + 1|s = \tau' + 1) = P(s = \tau' + 3|s = \tau' + 2) = P(s = \tau' + 2|s = \tau' + 2) = 1/2$;
- $P(s = \tau' + 1|s = \tau') = q, P(s = \tau' + 2|s = \tau') = 1/2 - q, P(s = \tau'|s = \tau') = 1/2$;
- $P(s = 0|s = 2\tau' + 1) = q, P(s = 1|s = 2\tau' + 1) = 1/2 - q, P(s = 2\tau' + 1|s = 2\tau' + 1) = 1/2$;
- $P(s = i + 1|s = i) = P(s = i|s = i) = 1/2, \forall i \notin \{0, 1, \tau', \tau' + 1, \tau' + 2, 2\tau' + 1\}$.

Then letting $v_1^* = 0, v_2^* = 1, w_1^* = \tau' + 1, w_2^* = \tau' + 2$, it is straightforward that the above constructed Markov chain guarantees that transitioning between v_1^* and w_1^* takes at least $\tau' = \tau/2$ steps. Moreover, the hitting time of the chain is $\mathcal{O}(\tau)$ by noting the hitting time of directed cyclic chain with self-loops and n states is $\mathcal{O}(n)$. We denote this chain by P^*, and hence $P^* \in \mathcal{M}(\tau)$.

A.2 Construction of function F

Now we construct a "hard" function that is difficult for any first-order algorithm to search for the critical point. Specifically we consider the following two functions

$$h_1(x) = -\psi(1)\phi([x]_1) + \sum_{i=1}^{\lfloor d/2 \rfloor - 1} (\psi(-[x]_{2i})\phi(-[x]_{2i+1}) - \psi([x]_{2i})\phi([x]_{2i+1})) \tag{10}$$

$$h_2(x) = \sum_{i=1}^{\lfloor d/2 \rfloor} (\psi(-[x]_{2i-1})\phi(-[x]_{2i}) - \psi([x]_{2i-1})\phi([x]_{2i})) \tag{11}$$

where

$$\psi(u) = \begin{cases} 0 & , \quad u \le \frac{1}{2} \\ \exp\left(1 - \frac{1}{(2u-1)^2}\right) & , \quad u > \frac{1}{2} \end{cases}$$

and

$$\phi(u) = \sqrt{e} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt$$

with $u \in \mathbb{R}$.

We denote π_s as the corresponding probability of state s of the stationary distribution π. Then, given the Markov chain P^* constructed in Appendix A.1, we know that at least $\frac{1}{2}\tau$ steps are required to take transiting from v_1^* to w_1^* and vice versa. Then, we construct function F such that $F(x) = \pi_{v^*} h_1(x) + \pi_{w^*} h_2(x)$, where we denote $v^* = \{v_1^*, v_2^*\}$, $w^* = \{w_1^*, w_2^*\}$ and $\pi_{v^*} = \pi_{v_1^*} + \pi_{v_2^*}$, $\pi_{w^*} = \pi_{w_1^*} + \pi_{w_2^*}$. For any x and $i \ge 0$ define $x_{\le i} := ([x]_1, \ldots, [x]_i, 0, \ldots, 0)$ as the truncated version by only keeping the first i coordinates. We also set $x_{\le 0} = x$. Then we have the following properties of F.

Lemma A.1. *Let $F(x) = \pi_{v^*} h_1(x) + \pi_{w^*} h_2(x)$ for h_1, h_2 defined by* (10),(11)*. Then we have the following:*

(1). $F(0) - \inf_x F(x) \le \Delta_0 d$ *for some constant* $\Delta_0 > 0$.

(2). $\|\nabla h_i(x)\|_\infty \le 23$ *and* $\|\nabla h_i(x)\| \le 23\sqrt{d}, i = 1, 2$.

(3). $F(x)$ is l_1-smooth for some constant $l_1 > 0$.

(4). If $\text{prog}_1(x) < d$, $\|\nabla F(x)\| \geq 1$.

(5). $[\nabla h_i(x)]_{\leq prog_{\frac{1}{2}}(x)} = [\nabla h_i(x_{\leq prog_{\frac{1}{2}}(x)})]_{\leq prog_{\frac{1}{2}}(x)}$, $i = 1, 2$.

(6). If $\text{prog}_0(x)$ *is odd,* $\text{prog}_0(\nabla h_1(x)) \leq \text{prog}_{\frac{1}{2}}(x)$, $\text{prog}_0(\nabla h_2(x)) \leq \text{prog}_{\frac{1}{2}}(x) + 1$. *If* $\text{prog}_0(x)$ *is even,* $\text{prog}_0(\nabla h_1(x)) \leq \text{prog}_{\frac{1}{2}}(x) + 1$, $\text{prog}_0(\nabla h_2(x)) \leq \text{prog}_{\frac{1}{2}}(x)$.

(7). If $prog_{\frac{1}{2}}(x)$ *is odd,* $\nabla h_1(x) = \nabla h_1(x_{\leq prog_{\frac{1}{2}}(x)})$, $\nabla h_2(x) = \nabla h_2(x_{\leq 1+prog_{\frac{1}{2}}(x)})$. *If* $prog_{\frac{1}{2}}(x)$ *is even,* $\nabla h_1(x) = \nabla h_1(x_{\leq 1+prog_{\frac{1}{2}}(x)})$, $\nabla h_2(x) = \nabla h_2(x_{\leq prog_{\frac{1}{2}}(x)})$.

Proof. For Part (1), observing that $F(0) < 0$ and noting that $0 \leq \psi(u) \leq e$, $0 \leq \phi(u) \leq \sqrt{2\pi e}$,

$$F(x) \geq -\psi(1)\phi([x]_1) - \sum_{i=2}^{d} \psi([x]_{i-1})\phi([x]_i) \geq -de\sqrt{2\pi e} = -d\Delta_0$$

with $\Delta_0 = e\sqrt{2\pi e}$, which completes its proof.

For Part (2), noting that $0 \leq \psi'(u) \leq \sqrt{54e^{-1}}$ and $0 \leq \phi'(u) \leq \sqrt{e}$, combining with the fact that for each $i = 1, 2$

$$\frac{\partial h_i}{\partial x_j}(x) \geq \psi(-[x]_{j-1})\phi'(-[x]_j) - \psi([x]_{j-1})\phi'([x]_j) - \psi'(-[x]_j)\phi(-[x]_{j+1}) - \psi'([x]_j)\phi([x]_{j+1})$$

yields

$$\left|\frac{\partial h_i}{\partial x_j}(x)\right| \leq e\sqrt{e} + \sqrt{54e^{-1}}\sqrt{2\pi e} \leq 23$$

implying $\|\nabla h_i(x)\| \leq 23$ and $\|\nabla h_i(x)\| \leq 23\sqrt{d}, \forall i = 1, 2$.

Parts (3) and (4) follow directly from [9]. Parts (5)-(7) follow from the observation that

$$\nabla h_1(x) = \nabla h_1([x]_1, \ldots, [x]_{2i+1}, 0, \ldots, 0), \text{ if } |x_{2j}| \leq \frac{1}{2}, \ \forall j \geq i+1$$

$$\nabla h_2(x) = \nabla h_2([x]_1, \ldots, [x]_{2i}, 0, \ldots, 0), \text{ if } |x_{2j-1}| \leq \frac{1}{2}, \ \forall j \geq i+1.$$

□

B Lower Bound for the Smooth Setting

In this section, we show the lower bound of the smooth setting in Theorem 3.1. Based on the contructive F in the last section, we consider the following gradient oracle g: for each i-th coordinate of g

$$\text{if } s \in \{v_1^*, v_2^*\}, \ \ [g(x;s)]_i = [\nabla h_1(x)]_i \cdot \left(1 + \mathbb{1}\{i > \text{prog}_{\frac{1}{2}}(x)\}\left(\frac{\mathbb{1}_{s=v_1^*}}{q} - 1\right)\right),$$

$$\text{if } s \in \{w_1^*, w_2^*\}, \ \ [g(x;s)]_i = [\nabla h_2(x)]_i \cdot \left(1 + \mathbb{1}\{i > \text{prog}_{\frac{1}{2}}(x)\}\left(\frac{\mathbb{1}_{s=w_1^*}}{q} - 1\right)\right),$$

$$\text{otherwise, } g(x;s) = 0. \tag{12}$$

Recalling the definition of P^*, we note $\mathbb{P}(s = v_1^* \mid s \in \{v_1^*, v_2^*\}) = \mathbb{P}(s = w_1^* \mid s \in \{w_1^*, w_2^*\}) = q \in (0, 1/2)$. Then, we have the following lemma.

Lemma B.1. *Considering stochastic gradient* $g(x;s)$ *constructed as* (12)*, the following statements hold:*

(1). For $s \in \{v_1^*, v_2^*, w_1^*, w_2^*\}$*, with probability at least* $1 - q$*,* $prog_0(g(x;s)) \leq prog_{\frac{1}{2}}(x)$ *and* $g(x;s) = g(x_{\leq prog_{\frac{1}{2}}(x)}; s)$ *for all* x.

(2). For $s \notin \{v_1^*, v_2^*, w_1^*, w_2^*\}$, with probability 1, $\text{prog}_0(g(x;s_t)) \le \text{prog}_{\frac{1}{2}}(x)$ and $g(x;s_t) = g(x_{\le \text{prog}_{\frac{1}{2}}(x)}; s_t)$ for all x.

(3). For any s, with probability 1, $\text{prog}_0(g(x;s)) \le 1 + \text{prog}_{\frac{1}{2}}(x)$ and $g(x;s) = g(x_{\le 1+\text{prog}_{\frac{1}{2}}(x)}; s)$ for all x.

(4). $\mathbb{E}_{s\sim\pi}[g(x;s)] = \nabla F(x)$.

Proof. We firstly show Part (3). Note that by (12) and Part (7) of Lemma A.1, for any x, s, $[g(x;s_t)]_i = 0, \forall i > 1+\text{prog}_{\frac{1}{2}}(x)$ in the sense that $[\nabla h_1(x)]_i = [\nabla h_2(x)]_i = 0, \forall i > 1+\text{prog}_{\frac{1}{2}}(x)$, which implies $\text{prog}_0(g(x;s)) \le 1 + \text{prog}_{\frac{1}{2}}(x)$. Moreover, by Part (7) of Lemma A.1, defining $x' := x_{\le 1+\text{prog}_{\frac{1}{2}}(x)}$ gives $\nabla h_1(x) = \nabla h_1(x')$ and $\nabla h_2(x) = \nabla h_2(x')$. Thus, we obtain $g(x;s) = g(x';s)$ for any x, s, implying Part (3).

For Part (1), we note that when $i \ge 1 + \text{prog}_{\frac{1}{2}}(x)$ and $s \in \{v_2^*, w_2^*\}$, $g(x;s) = [\nabla h_j(x)]_{\le \text{prog}_{\frac{1}{2}}(x)}$ for $j = 1, 2$, which implies $\text{prog}_0(g(x;s)) \le \text{prog}_{\frac{1}{2}}(x), \forall s \in \{v_2^*, w_2^*\}$. Further, according to (5) of Lemma A.1, we have $g(x;s) = g(x_{\le \text{prog}_{\frac{1}{2}}(x)}; s)$ for $s \in \{v_2^*, w_2^*\}$ and all x. Since $P(z=0) = 1-q$, hence Part (1) is proved.

Part (2) holds trivially in the sense that $g(x;s_t) = 0$ when $s \notin \{v_1^*, v_2^*, w_1^*, w_2^*\}$. Finally, Part (4) holds since $\mathbb{E}[\mathbb{1}_s/q \mid s \in \{v_1^*, v_2^*\}] = \mathbb{E}[\mathbb{1}_s/q \mid s \in \{w_1^*, w_2^*\}] = 1$. □

Also, we show in the following lemma that g has bounded variance.

Lemma B.2. *For $F(x) = \pi_{v^*} h_1(x) + \pi_{w^*} h_2(x)$ and g defined as* (12)*, then for any Markov chain with stationary distribution π, given any $x \in \mathbb{R}^d$,*

$$\mathbb{E}_{s\sim\pi}\|g(x;s) - \nabla F(x)\|^2 \le a_1 d + a_2 \frac{1-q}{q}$$

for some constant $a_1, a_2 > 0$.

Proof. By Part (4) of Lemma B.1, we know $\mathbb{E}_{s\sim\pi}[g(x;s)] = \nabla F(x)$.

Denote $i^* = 1 + \text{prog}_{\frac{1}{2}}(x)$. For any $s \in \{v_1^*, v_2^*, w_1^*, w_2^*\}$, we have

$$g(x;s) - \nabla F(x) = (0, \dots, 0, [\nabla h_1(x)]_{i^*}(\mathbb{1}_{s=v_1^*}/q - 1), 0, \dots, 0) + (1 - \tilde{\pi}_{v^*})\nabla h_1(x) - \tilde{\pi}_{w^*}\nabla h_2(x), \text{ if } s \in \{v_1^*, v_2^*\}$$
$$g(x;s) - \nabla F(x) = (0, \dots, 0, [\nabla h_2(x)]_{i^*}(\mathbb{1}_{s=w_1^*}/q - 1), 0, \dots, 0) + (1 - \tilde{\pi}_{w^*})\nabla h_2(x) - \tilde{\pi}_{v^*}\nabla h_1(x), \text{ if } s \in \{w_1^*, w_2^*\}.$$

When $i^* - 1$ is odd, from Part (6) of Lemma A.1 we know that $[\nabla h_1(x)]_{i^*} = 0$. Therefore,

$$\|g(x;s) - \nabla F(x)\|^2 \le 2\|\nabla h_1(x)\|^2 + 2\|\nabla h_2(x)\|^2 \le 4 \cdot 23^2 d, \;\; s \in \{v_1^*, v_2^*\}$$

$$\begin{aligned}\|g(x;s) - \nabla F(x)\|^2 &\le 3|[\nabla h_2(x)]_{i^*}|^2(\mathbb{1}_{s=w_1^*}/q - 1)^2 + 3\|\nabla h_1(x)\|^2 + 3\|\nabla h_2(x)\|^2 \\ &\le 3 \cdot 23^2(\mathbb{1}_{s=w_1^*}/q - 1)^2 + 6 \cdot 23^2 d, \;\; s \in \{w_1^*, w_2^*\}\end{aligned}$$

and

$$\|g(x;s_t) - \nabla F(x)\|^2 = \|\nabla F(x)\|^2 \le 4 \cdot 23^2 d, \;\; \text{when } s \notin \{v_1^*, v_2^*, w_1^*, w_2^*\}$$

where we use (2) of Lemma A.1. Combining the above three inequalities, it yields that when $i^* - 1$ is odd, for any Markov chain, any x, $t \ge 0$ and any initial distribution of the chain,

$$\mathbb{E}\|g(x;s_t) - \nabla F(x)\|^2 \le a_1 d + a_2 \frac{1-q}{q}$$

where $a_1 = 6 \cdot 23^2, a_2 = 3 \cdot 23^2$ and we use that $\mathbb{E}[(\mathbb{1}_s/q - 1)^2 \mid s \in \{w_1^*, w_2^*\}] = (1-q)/q$. The case when $i^* - 1$ is even can be derived similarly. □

Then, we are ready to show Lemmas 5.1 and 5.2. We first focus on the case $B = 1$ and then generalize it to $B \geq 1$.

Proof of Lemmas 5.1 and 5.2: Given any $\epsilon > 0$, we consider the following F^*

$$F^*(x) := \frac{L\lambda^2}{l_1} F\left(\frac{x}{\lambda}\right), \text{ where } \lambda = \frac{2l_1}{L}\epsilon. \tag{13}$$

And we consider the following gradient g^*

$$g^*(x; s) := \frac{L\lambda}{l_1} g\left(\frac{x}{\lambda}; s\right)$$

with $g(x; s)$ defined as (12). Since $\nabla F(x) = \mathbb{E}_{s\sim\pi}[g(x; s)]$, $\nabla F^*(x) = \mathbb{E}_{s\sim\pi}[g^*(x; s)]$. We note that

$$\nabla^2 F^*(x) = \frac{L}{l_1}\nabla^2 F\left(\frac{x}{\lambda}\right)$$

which implies that F^* is L-smooth by Part (3) of Lemma A.1. Moreover, by Part (1) of Lemma A.1 we obtain that

$$F^*(0) - \inf_x F^*(x) = \frac{4l_1\epsilon^2}{L}(F(0) - \inf_x F(x)) \leq \frac{4l_1\Delta_0\epsilon^2}{L}d.$$

All the above concludes Lemma 5.1.

To see Lemma 5.2, note by Lemma B.2, we have

$$\mathbb{E}_{s\sim\pi}\|g^*(x; s) - \nabla F^*(x)\|^2 \leq 4a_1 d\epsilon^2 + \frac{4a_2(1-q)}{q}\epsilon^2.$$

Then, define

$$B_t := \mathbb{1}\left\{\exists\, x \,:\, \text{prog}_0(g^*(x; s_t)) = 1 + \text{prog}_{\frac{1}{2}}(x)\right\}.$$

Note that under the construction of the Markov chain P^* and F^* and g^*, for any zero-respecting algorithm $\mathcal{A}$

$$B_{t+k} = 0, \ \forall k = 1, \ldots, \frac{1}{2}\tau, \text{ conditioning on } B_t = 1.$$

That is to say within every $\frac{1}{2}\tau$ iterations B_t can be 1 at most once. And Part (1) of Lemma B.1 indicates that the probability of B_t being 1 is no greater than q. Let $k(t) := \max_{m\in[M]} \max_{l\leq t} \text{prog}_0(x_{l,m}^{\mathcal{A}[O]_{F^*}})$. Then, the above implies that

$$k(t) \leq \sum_{l\leq t} B_l.$$

Also recalling the definition of P^* guarantees for any t in the ideal case the number of possible non-zero B_l can be at most $2t/\tau$ with each being 1 with probability at most q, it implies

$$\sum_{l\leq t} B_l \leq \sum_{i=1}^{\lceil 2t/\tau\rceil} z_i$$

where z_i denotes the Bernoulli random variable with succeeding probability being at most q. Note that z_is are independent in the sense that conditioning on the chain hits $v^* = \{v_1^*, v_2^*\}$ and will hit $w^* = \{w_1^*, w_2^*\}$ at exactly $\tau/2$ steps later, whether w_1^* or w_2^* will be visited is independent of which of v_1^* or v_2^* has been visited. Thus

$$\begin{aligned}
\mathbb{P}(k(t) \geq d) &\leq \mathbb{P}\left(\sum_{l\leq t} B_l \geq d\right) \\
&= \mathbb{P}\left(\exp\left(\sum_{l\leq t} B_l\right) \geq e^d\right) \\
&\leq e^{-d}\mathbb{E}[e^{\sum_{l\leq t} B_l}] \\
&\leq e^{-d}\mathbb{E}[e^{\sum_{i=1}^{\lceil 2t/\tau\rceil} z_i}] \\
&= e^{-d}(1 - q + eq)^{\lceil 2t/\tau\rceil} \\
&\leq e^{\lceil 4t/\tau\rceil q - d}
\end{aligned}$$

Therefore, we conclude that for any $\delta \in (0,1)$ and $q \in (0,1/2)$ with probability at least $1-\delta$,

$$k(t) < d, \ \ \forall t \leq \frac{\tau(d - \log(1/\delta))}{4q}$$

which completes the proof of Lemma 5.2.

Proof of Lemma 5.4: To see the part of the smooth setting of Lemma 5.4, note that for any algorithm $\mathcal{A}$ with $B \geq 1$, we simply observe that by the construction of F^* and g^* and the Markov chain P^*, for any $t \geq 0$ and $m \in [M]$, there exists an algorithm $\tilde{\mathcal{A}}$ with $B = 1$ for which $\text{prog}_0(x_{t,m}^{\mathcal{A}[O_{F^*}]}) \leq \text{prog}_0(x_{t,m}^{\tilde{\mathcal{A}}[O_{F^*}]})$, since multiple samples do not contribute to additional progress of x, which then proves the part of smooth setting (Similar claims can be achieved for the mean-squared setting of Lemma 5.4).

Proof of the smooth setting of Theorem 3.1: Now to show the lower bound for the smooth setting of Theorem 3.1, setting

$$d = \min\left\{ \left\lfloor \frac{L\Delta}{4l_1\Delta_0\epsilon^2} \right\rfloor, \left\lfloor \frac{\sigma^2}{8a_1\epsilon^2} \right\rfloor \right\} \tag{14}$$

and

$$\frac{1}{q} = 1 + \frac{\sigma^2}{8a_2\epsilon^2} \tag{15}$$

yields that $F^* \in \mathcal{F}(\Delta, L)$ and Assumption 2.3 is satisfied. By Part (4) of Lemma A.1 and Lemma 5.2, choosing $\delta = 1/2$ renders that for any $m \in [M]$ with probability at least $1/2$,

$$\|\nabla F^*(x_{t,m}^{\mathcal{A}[O_{F^*}]})\| \geq 2\epsilon, \ \ \forall t \leq \frac{\tau(d-1)}{4q}$$

which implies that

$$\mathbb{E}\|\nabla F^*(x_{t,m}^{\mathcal{A}[O_{F^*}]})\| \geq \epsilon, \ \ \forall t \leq \frac{\tau(d-1)}{4q}.$$

Therefore, we conclude that

$$N_s^\epsilon(M, \Delta, L, \sigma^2, \tau) \geq \frac{\tau(d-1)}{4q} \succeq \frac{\tau L\Delta}{\epsilon^2} + \frac{\tau\sigma^2}{\epsilon^4}\min\left\{c_1\sigma^2, c_2 L\Delta\right\}$$

by the selections of d, q as (14),(15) for some constants $c_1, c_2 > 0$.

C Lower Bound for the Mean-squared Setting

In this section, we show the lower bound of the mean-squared smooth setting in Theorem 3.1. The idea is similar to the proof for the smooth setting, except that we replace the indicator function in (12) by its smoothed surrogate:

$$\Theta_i(x) := \Gamma\left(1 - \left(\sum_{k=i}^{d} \Gamma^2(|x_k|)\right)^{1/2}\right) \tag{16}$$

where Γ is defined by

$$\Gamma(t) = \frac{\int_{1/4}^{t} \Delta(\tau)d\tau}{\int_{1/4}^{1/2} \Delta(\tau)d\tau}$$

with

$$\Delta(t) = \begin{cases} 0, & t \leq 1/4 \text{ or } t \geq 1/2 \\ \exp\left(1/(100(t-1/4)(t-1/2))\right), & 1/4 < t < 1/2. \end{cases}$$

Then, we consider the following stochastic gradient $\bar{g}$:

$$\begin{aligned} &\text{if } s \in \{v_1^*, v_2^*\}, \ \ [\bar{g}(x;s)]_i = [\nabla h_1(x)]_i \cdot \left(1 + \Theta_i(x)\left(\frac{\mathbb{1}_{s=v_1^*}}{q} - 1\right)\right), \\ &\text{if } s \in \{w_1^*, w_2^*\}, \ \ [\bar{g}(x;s)]_i = [\nabla h_2(x)]_i \cdot \left(1 + \Theta_i(x)\left(\frac{\mathbb{1}_{s=w_1^*}}{q} - 1\right)\right), \\ &\text{otherwise, } \bar{g}(x;s) = 0. \end{aligned} \tag{17}$$

It is straightforward to see $\mathbb{E}_{s\sim\pi}[\bar{g}(x;s)] = \nabla F(x)$. According to Observation 1 of [5], we know $\Theta_i(x) = 0, \forall i \le \text{prog}_{\frac{1}{2}}(x)$ and hence $[\bar{g}(x;s)]_i = [\nabla h_j(x)]_i, j = 1, 2, \forall i \le \text{prog}_{\frac{1}{2}}(x)$ when $s \in \{v_1^*, v_2^*, w_1^*, w_2^*\}$. Moreover, according to Part (6) of Lemma A.1, we have for any $s \in \mathcal{S}$

$$[\bar{g}(x;s)]_i = 0, \ \ \forall i > 1 + \text{prog}_{\frac{1}{2}}(x).$$

Defining $\delta(x;s) := \bar{g}(x;s) - \nabla F(x)$, it yields that there is eactly one non-zero coordinate of $\delta(x;s)$, which is the $1 + \text{prog}_{\frac{1}{2}}(x)$-th coordinate. Moreover, we have the following results for $\bar{g}$.

Lemma C.1. *Consider $\bar{g}$ defined as* (17)*. Then,*

$$\mathbb{E}_{s\sim\pi}\|\bar{g}(x;s) - \nabla F(x)\|^2 \le a_3 d + a_4 \frac{1-q}{q} \tag{18}$$

for some constants $a_3, a_4 > 0$. And for some constant $\bar{l}_1 > 0$

$$\mathbb{E}_{s\sim\pi}\|\bar{g}(x;s) - \bar{g}(y;s)\|^2 \le \frac{\bar{l}_1^2}{q}\|x-y\|^2, \ \forall x, y \in \mathbb{R}^d. \tag{19}$$

Proof. Similar to the proof of Lemma B.2, we can easily obtain (18). To show (19), note that

$$\begin{aligned}\mathbb{E}\|\bar{g}(x;s) - \bar{g}(y;s)\|^2 &= \mathbb{E}\|\delta(x;s) - \delta(y;s)\|^2 + \|\nabla F(x) - \nabla F(y)\|^2 \\ &= \sum_{i\in\{i_x^*, i_y^*\}} \mathbb{E}([\delta(x;s)]_i - [\delta(y;s)]_i)^2 + \|\nabla F(x) - \nabla F(y)\|^2\end{aligned}$$

where $i_x^* = 1 + \text{prog}_{\frac{1}{2}}(x), i_y^* = 1 + \text{prog}_{\frac{1}{2}}(y)$. Since

$$\begin{aligned}\mathbb{E}([\delta(x;s)]_i - [\delta(y;s)]_i)^2 &= ([\nabla F(x)]_i\Theta_i(x) - [\nabla F(y)]_i\Theta_i(y))^2\frac{1-q}{q} \\ &= ([\nabla F(x)]_i(\Theta_i(x) - \Theta_i(y)) + [\nabla F(x) - \nabla F(y)]_i\Theta_i(y))^2\frac{1-q}{q} \\ &\le 2([\nabla F(x)]_i^2(\Theta_i(x) - \Theta_i(y))^2 + [\nabla F(x) - \nabla F(y)]_i^2\Theta_i(y)^2)\frac{1}{q}\end{aligned}$$

and by Observation 1.3 of [5] $|\Theta_i(x) - \Theta_i(y)| \le 36\|x-y\|$ and noting $|\Theta_i(x)| \le 1$, $\|\nabla F(x)\|_\infty \le 23$, we obtain

$$\mathbb{E}([\delta(x;s)]_i - [\delta(y;s)]_i)^2 \le \frac{2}{q}(23^2 \cdot 36^2\|x-y\|^2 + \|\nabla F(x) - \nabla F(y)\|^2).$$

Finally leveraging Part (3) of Lemma A.1 gives

$$\mathbb{E}\|\bar{g}(x;s) - \bar{g}(y;s)\|^2 \le \frac{\bar{l}_1^2}{q}\|x-y\|^2$$

with $\bar{l}_1^2 = 4 \cdot 23^2 \cdot 36^2 + 5l_1^2$. □

Then we show the lower bound corresponding to the mean-squared smooth setting of Theorem 3.1.

Proof of the mean-squared smooth setting of Theorem 3.1 and Lemma 5.3: Noting that $\bar{L}$-mean-squared smoothness implies $\bar{L}$ smoothness, we thus consider the case $L \le \bar{L}$ with L to be determined later. Consider the same F^* as (13) and let $\bar{g}^*(x;s) = (L\lambda/l_1)\bar{g}(x/\lambda;s)$. Similarly, we have F^* is L-smooth. Also

$$\mathbb{E}_{s\sim\pi}\|\bar{g}^*(x;s) - \nabla F^*(x)\|^2 \le \left(\frac{L\lambda}{l_1}\right)^2 (a_3 d + a_4(1-q)/q)$$

and

$$\begin{aligned}\mathbb{E}_{s\sim\pi}\|\bar{g}^*(x;s) - \bar{g}^*(y;s)\| &= \left(\frac{L\lambda}{l_1}\right)^2 \mathbb{E}_{s\sim\pi}\|\bar{g}(x/\lambda;s) - \bar{g}(y/\lambda;s)\|^2 \\ &\le \left(\frac{L\bar{l}_1}{l_1\sqrt{q}}\right)^2 \|x-y\|^2.\end{aligned}$$

Then taking

$$d = \min\left\{\left\lfloor \frac{L\Delta}{4l_1\Delta_0\epsilon^2}\right\rfloor, \left\lfloor \frac{\sigma^2}{8a_3\epsilon^2}\right\rfloor\right\}$$

$$\frac{1}{q} = \max\left\{1 + \frac{\sigma^2}{8a_4\epsilon^2}, \frac{\bar{l}_1^2}{l_1^2}\right\}$$

$$L = \frac{\bar{L}l_1\sqrt{q}}{\bar{l}_1} \leq \bar{L}$$

we guarantee that $F^* \in \mathcal{F}(L, \Delta)$ and Assumptions 2.3, 2.4 are satisfied. Similar to the proof of the smooth setting, we can easily obtain Lemma 5.3 and then conclude that

$$N^\epsilon(M, \Delta, \bar{L}^2, \sigma^2, \tau) \succeq \frac{\tau\bar{L}\Delta}{\epsilon} + \frac{\tau\bar{L}\Delta\sigma^2}{\epsilon^3}$$

which completes the proof.

D Convergence Analysis of MaC-SPIDER

In this section, we provide the proof for Section 4. We first present the following technical lemma.

Lemma D.1. *We have the following claims:*

- $d_{TV}(\mu P^{t+1}, \pi) \leq d_{TV}(\mu P^t, \pi)$.
- *For* $k \geq 2$, $t_{mix}(2^{-k}) \leq (k-1)\tau_{mix}$.
- *Moreover,*

$$\sum_{k=0}^{T} d_{TV}(\mu P^k, \pi) \leq 3\tau_{mix}, \ \ \forall T \geq 0.$$

Proof. The first two claims are directly from [24].

To see the third claim, we note that

$$\begin{aligned}
\sum_{k=0}^{T} d_{TV}(\mu P^k, \pi) &\leq \sum_{k=0}^{\infty} d_{TV}(\mu P^k, \pi) \\
&\leq \sum_{l=0}^{\tau_{mix}} d_{TV}(\mu P^l, \pi) + \sum_{k=0}^{\infty} \sum_{l=t_{mix}(2^{-k})+1}^{t_{mix}(2^{-(k+1)})} d_{TV}(\mu P^l, \pi) \\
&\leq d_{TV}(\mu, \pi)\tau_{mix} + \sum_{k=2}^{\infty} (t_{mix}(2^{-(k+1)}) - t_{mix}(2^{-k}))2^{-k} \\
&\leq d_{TV}(\mu, \pi)\tau_{mix} + \sum_{k=2}^{\infty} k2^{-k}\tau_{mix} \\
&\leq d_{TV}(\mu, \pi)\tau_{mix} + 2\tau_{mix}
\end{aligned}$$

which completes the proof with $d_{TV}(\mu, \pi) + 2 \leq 3$. □

D.1 Proof of Lemma 4.1 and Proposition 4.3

Proof of Lemma 4.1: Let $\tilde{h}_t^i = h(s_{t+i}) - h_\pi$ and $\tilde{h}(s) = h(s) - h_\pi$. We have

$$\mathbb{E}_t \left\| \frac{1}{M}\sum_{i=1}^{M} h(s_{t+i}) - h_\pi \right\|^2 = \frac{1}{M^2}\sum_{i=1}^{M} \mathbb{E}_t\|\tilde{h}_t^i\|^2 + \frac{2}{M^2}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} \mathbb{E}_t\langle \tilde{h}_t^i, \tilde{h}_t^j\rangle.$$

First, we show the following useful bound: for any $s \in \mathcal{S}$, given $t \geq 0$ and $1 \leq i < j$,

$$\begin{aligned}
&\left(\sum_{s' \in \mathcal{S}} |P(s_{t+j} = s' \mid s_{t+i} = s) - \pi(s')| \|\tilde{h}(s')\|\right)^2 \\
&= \left(\sum_{s' \in \mathcal{S}} \pi(s')^{-1/2} |P(s_{t+j} = s' \mid s_{t+i} = s) - \pi(s')| \sqrt{\pi(s')} \|\tilde{h}(s')\|\right)^2 \\
&\leq \sum_{s' \in \mathcal{S}} \pi_{min}^{-1} |P(s_{t+j} = s' \mid s_{t+i} = s) - \pi(s')|^2 \sum_{s' \in \mathcal{S}} \pi(s') \|\tilde{h}(s')\|^2 \\
&\leq \left(\sum_{s' \in \mathcal{S}} \pi_{min}^{-1/2} |P(s_{t+j} = s' \mid s_{t+i} = s) - \pi(s')|\right)^2 \sigma^2 \\
&\leq 4\sigma^2 \pi_{min}^{-1} (\max_s d_{TV}(P^{j-i}(\cdot \mid s_t = s) - \pi))^2 \qquad (20)
\end{aligned}$$

Then, noting that for any $1 \leq i < j \leq M$,

$$\begin{aligned}
\mathbb{E}_t \langle \tilde{h}_t^i, \tilde{h}_t^j \rangle &= \sum_{s \in \mathcal{S}} P(s_{t+i} = s \mid s_t) \sum_{s' \in \mathcal{S}} P(s_{t+j} = s' \mid s_{t+i} = s) \langle \tilde{h}(s), \tilde{h}(s') \rangle \\
&= \sum_{s \in \mathcal{S}} (P(s_{t+i} = s \mid s_t) - \pi(s)) \sum_{s' \in \mathcal{S}} (P(s_{t+j} = s' \mid s_{t+i} = s) - \pi(s')) \langle \tilde{h}(s), \tilde{h}(s') \rangle \\
&\quad + \sum_{s \in \mathcal{S}} \pi(s) \sum_{s' \in \mathcal{S}} (P(s_{t+j} = s' \mid s_{t+i} = s) - \pi(s')) \langle \tilde{h}(s), \tilde{h}(s') \rangle \\
&\quad + \sum_{s \in \mathcal{S}} P(s_{t+i} = s \mid s_t) \sum_{s' \in \mathcal{S}} \pi(s') \langle \tilde{h}(s), \tilde{h}(s') \rangle \\
&\leq \sum_{s \in \mathcal{S}} |P(s_{t+i} = s \mid s_t) - \pi(s)| \sum_{s' \in \mathcal{S}} |P(s_{t+j} = s' \mid s_{t+i} = s) - \pi(s')| \|\tilde{h}(s)\| \|\tilde{h}(s')\| \\
&\quad + \sum_{s \in \mathcal{S}} \pi(s) \sum_{s' \in \mathcal{S}} |P(s_{t+j} = s' \mid s_{t+i} = s) - \pi(s')| \|\tilde{h}(s)\| \|\tilde{h}(s')\| \\
&\leq 4\sigma^2 \pi_{min}^{-1} \max_s d_{TV}(P^i(\cdot \mid s), \pi) \cdot \max_s d_{TV}(P^{j-i}(\cdot \mid s), \pi) \\
&\quad + 2\sigma^2 \pi_{min}^{-1/2} \max_s d_{TV}(P^{j-i}(\cdot \mid s), \pi)
\end{aligned}$$

where we use (20) and note $\sum_{s'} \pi(s') \langle \tilde{h}(s), \tilde{h}(s') \rangle = \langle \tilde{h}(s), \sum_{s'} \pi(s') \tilde{h}(s') \rangle = \langle \tilde{h}(s), \mathbb{E}_{s' \sim \pi}[\tilde{h}(s')] \rangle = 0$. Further using the fact $\sum_{t=1}^{T} \max_s d_{TV}(P^t(\cdot \mid s), \pi) \leq 3\tau_{mix}$ by Lemma D.1, we obtain

$$\frac{2}{M^2} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \mathbb{E}_t \langle \tilde{h}_t^i, \tilde{h}_t^j \rangle \leq \frac{72\tau_{mix}^2 \sigma^2}{\pi_{min} M^2} + \frac{6\tau_{mix}\sigma^2}{\sqrt{\pi_{min}} M}.$$

Similarly we note that

$$\begin{aligned}
\mathbb{E}_t \|\tilde{h}_t^i\|^2 &= \sum_{s \in \mathcal{S}} P(s_{t+i} = s \mid s_t) \|\tilde{h}(s)\|^2 \\
&= \sum_{s \in \mathcal{S}} (P(s_{t+i} = s \mid s_t) - \pi(s)) \|\tilde{h}(s)\|^2 + \sum_{s \in \mathcal{S}} \pi(s) \|\tilde{h}(s)\|^2 \\
&\leq \sum_{s \in \mathcal{S}} |P(s_{t+i} = s \mid s_t) - \pi(s)| \|\tilde{h}(s)\|^2 + \sigma^2 \\
&\leq \pi_{min}^{-1} \sum_{s \in \mathcal{S}} |P(s_{t+i} = s \mid s_t) - \pi(s)| \sum_{s \in \mathcal{S}} \pi(s) \|\tilde{h}(s)\|^2 + \sigma^2 \\
&\leq 2\sigma^2 \pi_{min}^{-1} \max_s d_{TV}(P^i(\cdot \mid s_t), \pi) + \sigma^2
\end{aligned}$$

which then implies by Lemma D.1

$$\frac{1}{M^2}\sum_{i=1}^{M}\mathbb{E}_t\|\tilde{h}_t^i\|^2 \le \frac{\sigma^2}{M} + \frac{6\tau_{mix}\sigma^2}{\pi_{min}M^2}.$$

Combining all above gives (7).

To show (9), we note that

$$\begin{aligned}
&\mathbb{E}_t\left\|\frac{1}{M}\sum_{i=1}^{M}h(s_{t+i}) - h_\pi\right\|^2 \\
&= \mathbb{E}_t\left\|\frac{1}{M}\sum_{i=1}^{M}h(s_{t+i}) - \mathbb{E}_t\left(\frac{1}{M}\sum_{i=1}^{M}h(s_{t+i})\right) + \mathbb{E}_t\left(\frac{1}{M}\sum_{i=1}^{M}(h(s_{t+i}) - h_\pi)\right)\right\|^2 \\
&\le 2\mathbb{E}_t\left\|\frac{1}{M}\sum_{i=1}^{M}h(s_{t+i})\right\|^2 + 2\left\|\mathbb{E}_t\left(\frac{1}{M}\sum_{i=1}^{M}(h(s_{t+i}) - h_\pi)\right)\right\|^2
\end{aligned}$$

where we use $(a+b)^2 \le 2a^2 + 2b^2$ and $\mathbb{E}[X - \mathbb{E}(X)]^2 \le \mathbb{E}[X]^2$.

We note that by replacing $\tilde{h}(s)$ in (20) by $h(s)$,

$$\begin{aligned}
\left\|\mathbb{E}_t\left(\frac{1}{M}\sum_{i=1}^{M}(h(s_{t+i}) - h_\pi)\right)\right\| &= \left\|\frac{1}{M}\sum_{i=1}^{M}\sum_{s\in\mathcal{S}}(P(s_{t+i} = s \mid s_t) - \pi(s))h(s)\right\| \\
&\le \frac{1}{M}\sum_{i=1}^{M}\sum_{s\in\mathcal{S}}|P(s_{t+i} = s \mid s_t) - \pi(s)|\|h(s)\| \\
&\le \frac{2B}{\sqrt{\pi_{min}}M}\sum_{i=1}^{M}\max_s d_{TV}(P^i(\cdot \mid s), \pi) \\
&\le \frac{6\tau_{mix}B}{\sqrt{\pi_{min}}M}
\end{aligned}$$

where we use $\mathbb{E}_{s\sim\pi}\|h(s)\| \le (\mathbb{E}_{s\sim\pi}\|h(s)\|^2)^{1/2} \le B$ and this concludes (8). Moreover, similar to the analysis of $\mathbb{E}_t\langle\tilde{h}_t^i, \tilde{h}_t^j\rangle$, we have

$$\begin{aligned}
\mathbb{E}_t\left\|\frac{1}{M}\sum_{i=1}^{M}h(s_{t+i})\right\|^2 &= \frac{B^2}{M} + \frac{6\tau_{mix}B^2}{\pi_{min}M^2} + \frac{2}{M}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\mathbb{E}_t\langle h(s_{t+i}), h(s_{t+j})\rangle \\
&\le \frac{B^2}{M} + \frac{6\tau_{mix}B^2}{\pi_{min}M^2} + \frac{6\tau_{mix}B^2}{\sqrt{\pi_{min}}M} + \frac{72\tau_{mix}^2B^2}{\pi_{min}M^2} \\
&\le \frac{7\tau_{mix}B^2}{\sqrt{\pi_{min}}M} + \frac{78\tau_{mix}^2B^2}{\pi_{min}M^2}
\end{aligned}$$

and hence

$$\mathbb{E}_t\left\|\frac{1}{M}\sum_{i=1}^{M}h(s_{t+i}) - h_\pi\right\|^2 \le \frac{14\tau_{mix}B^2}{\sqrt{\pi_{min}}M} + \frac{228\tau_{mix}^2B^2}{\pi_{min}M^2}$$

which concludes (9).

Proof of Proposition 4.3: We denote $\mathbb{E}_t(\cdot)$ as the expectation conditioning on filtration $\mathcal{F}_t$.

Note that for $t \bmod r = 0$,

$$\begin{aligned}\mathbb{E}_t\|v_t - \nabla F(x_t)\|^2 &= \mathbb{E}_t\left\|\frac{1}{M_1}\sum_{i=1}^{M_1} g(x_t; s_{N_t+i}) - \nabla F(x_t)\right\|^2 \\ &\le \frac{7\tau_{mix}\sigma^2}{\sqrt{\pi_{min}}M_1} + \frac{78\tau_{mix}^2\sigma^2}{\pi_{min}M_1^2} \\ &\le \frac{\epsilon^2}{8}\end{aligned}$$

by (7) of Lemma 4.1 and noting $M_1 = 112\tau_{mix}\pi_{min}^{-1/2}\epsilon^{-2}\max\{\sigma, \sigma^2\}$.

For $t \bmod r \neq 0$, conditioning on $\mathcal{F}_{t+1}$ and letting $\tilde{g}_{t+1}^{M_2} = \frac{1}{M_2}\sum_{i=1}^{M_2} g(x_{t+1}; s_{N_{t+1}+i}) - g(x_t; s_{N_{t+1}+i})$ yields

$$\begin{aligned}\mathbb{E}_{t+1}\|v_{t+1} - \nabla F(x_{t+1})\|^2 &= \mathbb{E}_{t+1}\left\|v_t - \nabla F(x_t) + \tilde{g}_{t+1}^{M_2} - \nabla F(x_{t+1}) + \nabla F(x_t)\right\|^2 \\ &= \|v_t - \nabla F(x_t)\|^2 + \mathbb{E}_{t+1}\|\tilde{g}_{t+1}^{M_2} - \nabla F(x_{t+1}) + \nabla F(x_t)\|^2 \\ &\quad + 2\langle v_t - \nabla F(x_t), \mathbb{E}_{t+1}[\tilde{g}_{t+1}^{M_2} - \nabla F(x_{t+1}) + \nabla F(x_t)]\rangle \\ &\le \|v_t - \nabla F(x_t)\|^2 + \mathbb{E}_{t+1}\|\tilde{g}_{t+1}^{M_2} - \nabla F(x_{t+1}) + \nabla F(x_t)\|^2 \\ &\quad + 2\|v_t - \nabla F(x_t)\|\|\mathbb{E}_{t+1}[\tilde{g}_{t+1}^{M_2} - \nabla F(x_{t+1}) + \nabla F(x_t)]\|\end{aligned}$$

Noting that $\mathbb{E}_{t+1,s\sim\pi}\|g(x_{t+1}; s) - g(x_t; s)\| \le L\|x_{t+1} - x_t\|, \forall s \in \mathcal{S}$ and $\mathbb{E}_{t+1,s\sim\pi}[g(x_{t+1}; s) - g(x_t; s)] = \nabla F(x_{t+1}) - \nabla F(x_t)$, combining with (8) and (9) of Lemma 4.1 gives

$$\mathbb{E}_{t+1}\|v_{t+1} - \nabla F(x_{t+1})\|^2 \le \|v_t - \nabla F(x_t)\|^2 + \frac{12\tau_{mix}B}{\sqrt{\pi_{min}}M_2}\|v_t - \nabla F(x_t)\| + \frac{14\tau_{mix}B^2}{\sqrt{\pi_{min}}M_2} + \frac{228\tau_{mix}^2B^2}{\pi_{min}M_2^2}.$$

where $B := L\max_t\|x_{t+1} - x_t\|$. Further noting

$$M_2 = \frac{16\tau_{mix}}{\sqrt{\pi_{min}}\epsilon}, \;\; B \le L\max_t\{\eta_t\|v_t\|\} \le \frac{\epsilon}{4}$$

we have

$$\begin{aligned}\mathbb{E}_{t+1}\|v_{t+1} - \nabla F(x_{t+1})\|^2 &\le \|v_t - \nabla F(x_t)\|^2 + \frac{\epsilon^2}{4}\|v_t - \nabla F(x_t)\| + \frac{\epsilon^3}{16} + \frac{\epsilon^4}{16} \\ &\le \|v_t - \nabla F(x_t)\|^2 + \frac{\epsilon}{2}\|v_t - \nabla F(x_t)\|^2 + \frac{1}{8}(\epsilon^3 + \epsilon^4/2)\end{aligned}$$

where we use $\frac{\epsilon^2}{4}\|v_t - \nabla F(x_t)\| \le \frac{\epsilon}{2}\|v_t - \nabla F(x_t)\|^2 + \frac{1}{2\epsilon}(\frac{\epsilon^2}{4})^2$ in the second inequality. Then noting that for $rt_0 \le t < r(t_0+1)$ given any $t_0 \ge 0$, we have for $r = 1/\epsilon$

$$\begin{aligned}\mathbb{E}\|v_t - \nabla F(x_t)\|^2 &\le (1+\epsilon/2)^r\mathbb{E}\|v_{rt_0} - \nabla F(x_{rt_0})\|^2 + (1+\epsilon/2)^r(\epsilon^3 + \epsilon^4/2)\cdot\frac{1}{4\epsilon} \\ &\le 2\cdot\frac{\epsilon^2}{8} + \frac{\epsilon^2}{2} + \frac{\epsilon^3}{4} \\ &\le \frac{3\epsilon^2}{4} + \frac{\epsilon^3}{4}\end{aligned}$$

where we use the fact $(1+\epsilon/2)^{1/\epsilon} \le \sqrt{e} \le 2$ and $\mathbb{E}\|v_{rt_0} - \nabla F(x_{rt_0})\|^2 \le \epsilon^2/8$.

D.2 Proof of Theorem 4.4

Noting $\bar{L}$-mean-squared smoothness implies $\bar{L}$-smoothness of F, we have

$$\begin{aligned}F(x_{t+1}) - F(x_t) &\le \langle\nabla F(x_t), x_{t+1} - x_t\rangle + \frac{\bar{L}}{2}\|x_{t+1} - x_t\|^2 \\ &= -\eta_t\langle\nabla F(x_t), v_t\rangle + \frac{\bar{L}\eta_t^2}{2}\|v_t\|^2 \\ &= -\eta_t\langle\nabla F(x_t) - v_t, v_t\rangle + \frac{\bar{L}\eta_t^2}{2}\|v_t\|^2 - \eta_t\|v_t\|^2 \\ &\le -\frac{\eta_t}{2}\left(1 - \bar{L}\eta_t\right)\|v_t\|^2 + \frac{\eta_t}{2}\|v_t - \nabla F(x_t)\|^2.\end{aligned}$$

Noting that $\eta_t \leq \frac{1}{2\bar{L}}$, then using the fact that $\min\{|x|, x^2/2\} \geq |x| - 2, \forall x \in \mathbb{R}$,

$$\begin{aligned}
\frac{\eta_t}{2}(1 - \bar{L}\eta_t)\|v_t\|^2 &\geq \frac{\eta_t}{4}\|v_t\|^2 \\
&= \frac{\epsilon^2}{16L}\min\left\{\frac{\|v_t\|^2}{2\epsilon^2}, \frac{\|v_t\|}{\epsilon}\right\} \\
&\geq \frac{\epsilon^2}{16L}\left(\frac{\|v_t\|}{\epsilon} - 2\right) \\
&= \frac{\epsilon}{16L}\|v_t\| - \frac{\epsilon^2}{8L}
\end{aligned}$$

which hence induces

$$F(x_{t+1}) - F(x_t) \leq -\frac{\epsilon}{16\bar{L}}\|v_t\| + \frac{\epsilon^2}{8\bar{L}} + \frac{1}{4\bar{L}}\|v_t - \nabla F(x_t)\|^2.$$

Taking expectation on both sides and using Lemma 4.3 gives

$$\frac{\epsilon}{16\bar{L}}\mathbb{E}\|v_t\| \leq \mathbb{E}[F(x_t) - F(x_{t+1})] + \frac{3\epsilon^2}{8\bar{L}}$$

which indicates

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|v_t\| &\leq \frac{16\bar{L}\mathbb{E}[F(x_0) - F(x_T)]}{T\epsilon} + 6\epsilon \\
&\leq \frac{16\bar{L}\Delta_0}{T\epsilon} + 6\epsilon.
\end{aligned}$$

By $T = 16\bar{L}\Delta_0\epsilon^{-2}$ we conclude

$$\begin{aligned}
\mathbb{E}\|\nabla F(\tilde{x}_T)\| &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\| \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}\|v_t\| + \mathbb{E}\|v_t - \nabla F(x_t)\|) \\
&\leq 7\epsilon.
\end{aligned}$$

It is straightforward to see that the total number of samples is upper bounded by

$$\left\lceil \frac{T}{r} \right\rceil (M_1 + M_2 r) = \mathcal{O}(\tau_{mix}\pi_{min}^{-1/2}\epsilon^{-3}).$$