A Proof of the Main Theory

In this section, we present the proofs of our main theorems. Let us first recall the notations used in Algorithm 1. $\mathbf{v}_t^s$ and $\mathbf{U}_t^s$ are the semi-stochastic gradient and Hessian defined in (3.1) and (3.2) respectively. $\mathbf{x}_t^s$'s are the iterates and $\widehat{\mathbf{x}}^s$'s are the reference points used in Algorithm 1. b_g and b_h are the batch sizes of semi-stochastic gradient and Hessian. S and T are the number of epochs and epoch length of Algorithm 1. We set $M_{s,t} := M = C_M\rho$ as suggested by Theorems 4.7 and 5.3, where $C_M > 0$ is a constant. $\mathbf{h}_t^s$ is the exact minimizer of $m_t^s(\mathbf{h})$, where $m_t^s(\mathbf{h})$ is defined in (3.3). $\widetilde{\mathbf{h}}_t^s$ is the inexact minimizer defined in (5.1).

In order to prove Theorems 4.7 and 5.3, we first lay down the following useful technical lemmas.

Lemma A.1. (Nesterov & Polyak, 2006) We have the following basic results:

$$\mathbf{v}_t^s + \mathbf{U}_t^s\mathbf{h}_t^s + \frac{M}{2}\|\mathbf{h}_t^s\|_2\mathbf{h}_t^s = 0, \tag{A.1}$$

$$\mathbf{U}_t^s + \frac{M}{2}\|\mathbf{h}_t^s\|_2\mathbf{I} \succeq 0, \tag{A.2}$$

$$\langle \mathbf{v}_t^s, \mathbf{h}_t^s\rangle + \frac{1}{2}\langle \mathbf{U}_t^s\mathbf{h}_t^s, \mathbf{h}_t^s\rangle + \frac{M}{6}\|\mathbf{h}_t^s\|_2^3 \le -\frac{M}{12}\|\mathbf{h}_t^s\|_2^3. \tag{A.3}$$

Next we have two lemmas which we use to control the variance of $\mathbf{v}_t^s$ and $\mathbf{U}_t^s$. They play important roles in our proof:

Lemma A.2. For the semi-stochastic gradient $\mathbf{v}_t^s$ defined in (3.1), we have

$$\mathbb{E}_{i_t}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} \le \frac{\rho^{3/2}}{b_g^{3/4}}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3,$$

where $\mathbb{E}_{i_t}$ is the expectation over all $i_t \in I_g$.

Lemma A.3. Let $\mathbf{U}_t^s$ be the semi-stochastic Hessian defined in (3.2). If the batch size satisfy $b_h \ge 400\log d$, then we have

$$\mathbb{E}_{j_t}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3 \le 1200\rho^3\left(\frac{\log d}{b_h}\right)^{3/2}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3,$$

where $\mathbb{E}_{j_t}$ is the expectation over all $j_t \in I_h$.

Lemma A.4. For any $\mathbf{h} \in \mathbb{R}^d$, we have

$$\langle \nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s, \mathbf{h}\rangle \le \frac{M}{27}\|\mathbf{h}\|_2^3 + \frac{2\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2}}{M^{1/2}},$$
$$\big\langle\big(\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big)\mathbf{h}, \mathbf{h}\big\rangle \le \frac{2M}{27}\|\mathbf{h}\|_2^3 + \frac{27}{M^2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3.$$

Lemma A.5. For any $\mathbf{h} \in \mathbb{R}^d$, if $C_M \ge 100$, then we have

$$\mu(\mathbf{x}_t^s + \mathbf{h}) \le 9C_M^{3/2}\Big[M^{3/2}\|\mathbf{h}\|_2^3 + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} + M^{-3/2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3$$
$$+ \|\nabla m_t^s(\mathbf{h})\|_2^{3/2} + M^{3/2}\big|\|\mathbf{h}\|_2 - \|\mathbf{h}_t^s\|_2\big|^3\Big].$$

Lemma A.6. If $T \ge 1$, then for any $\mathbf{h} \in \mathbb{R}^d$, we have

$$\|\mathbf{x}_t^s + \mathbf{h} - \widehat{\mathbf{x}}^s\|_2^3 \le 2T^2\|\mathbf{h}\|_2^3 + (1 + 3/T)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3. \tag{A.4}$$

Lemma A.7. We define constant series $c_t, 0 \le t \le T$ as the following: $c_T = 0,\ c_t = c_{t+1}(1+3/T) + M(500T^3)^{-1},\ 0 \le t \le T-1$. Then we have for any $1 \le t \le T$,

$$M/24 - 2c_tT^2 \ge 0. \tag{A.5}$$

A.1 Proof of Theorem 4.7

Proof of Theorem 4.7. We first upper bound $F(\mathbf{x}_{t+1}^s)$ as follows

$$\begin{aligned}
F(\mathbf{x}_{t+1}^s) &\le F(\mathbf{x}_t^s) + \langle \nabla F(\mathbf{x}_t^s), \mathbf{h}_t^s\rangle + \frac{1}{2}\big\langle \nabla^2 F(\mathbf{x}_t^s)\mathbf{h}_t^s, \mathbf{h}_t^s\big\rangle + \frac{\rho}{6}\|\mathbf{h}_t^s\|_2^3 \\
&= F(\mathbf{x}_t^s) + \langle \mathbf{v}_t^s, \mathbf{h}_t^s\rangle + \frac{1}{2}\langle \mathbf{U}_t^s\mathbf{h}_t^s, \mathbf{h}_t^s\rangle + \frac{M}{6}\|\mathbf{h}_t^s\|_2^3 + \langle \nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s, \mathbf{h}_t^s\rangle \\
&\quad + \frac{1}{2}\big\langle \big(\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big)\mathbf{h}_t^s, \mathbf{h}_t^s\big\rangle - \frac{M-\rho}{6}\|\mathbf{h}_t^s\|_2^3 \\
&\le F(\mathbf{x}_t^s) - \frac{M}{12}\|\mathbf{h}_t^s\|_2^3 + \bigg(\frac{M}{27}\|\mathbf{h}_t^s\|_2^3 + \frac{2\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2}}{M^{1/2}}\bigg) \\
&\quad + \frac{1}{2}\bigg(\frac{2M}{27}\|\mathbf{h}_t^s\|_2^3 + \frac{27}{M^2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3\bigg) - \frac{M-\rho}{6}\|\mathbf{h}_t^s\|_2^3 \\
&\le F(\mathbf{x}_t^s) - \frac{M}{12}\|\mathbf{h}_t^s\|_2^3 + \frac{2}{M^{1/2}}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} + \frac{27}{M^2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3,
\end{aligned} \tag{A.6}$$

where the first inequality follows from Lemma 4.2 and the second inequality holds due to Lemmas A.1 and A.4. We define

$$R_t^s = \mathbb{E}\big[F(\mathbf{x}_t^s) + c_t\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3\big], \tag{A.7}$$

where c_t is defined in the Lemma A.7. Then by Lemma A.6, we have

$$c_{t+1}\|\mathbf{x}_{t+1}^s - \widehat{\mathbf{x}}^s\|_2^3 \le 2c_{t+1}T^2\|\mathbf{h}_t^s\|_2^3 + c_{t+1}(1+3/T)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3. \tag{A.8}$$

Applying Lemma A.5 with $\mathbf{h} = \mathbf{h}_t^s$, we have

$$\begin{aligned}
\big(240C_M^2\rho^{1/2}\big)^{-1}\mu(\mathbf{x}_{t+1}^s) &\le \frac{M}{24}\|\mathbf{h}_t^s\|_2^3 + \frac{\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2}}{24M^{1/2}} + \frac{\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3}{24M^2} \\
&\quad + \frac{\|\nabla m_t^s(\mathbf{h}_t^s)\|_2^{3/2}}{24M^{1/2}} + \frac{M}{24}\big|\|\mathbf{h}_t^s\|_2 - \|\mathbf{h}_t^s\|_2\big|^3 \\
&= \frac{M}{24}\|\mathbf{h}_t^s\|_2^3 + \frac{\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2}}{24M^{1/2}} + \frac{\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3}{24M^2},
\end{aligned} \tag{A.9}$$

where the equality is due to Lemma A.1.

Adding (A.6) with (A.8) and (A.9) and taking total expectation, we have

$$\begin{aligned}
&R_{t+1}^s + \big(240C_M^2\rho^{1/2}\big)^{-1}\mathbb{E}\mu(\mathbf{x}_{t+1}^s) \\
&= \mathbb{E}\Big[F(\mathbf{x}_{t+1}^s) + c_{t+1}\|\mathbf{x}_{t+1}^s - \widehat{\mathbf{x}}^s\|_2^3 + \big(240C_M^2\rho^{1/2}\big)^{-1}\mu(\mathbf{x}_{t+1}^s)\Big] \\
&\le \mathbb{E}\Big[F(\mathbf{x}_t^s) + c_{t+1}(1+3/T)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3 - \|\mathbf{h}_t^s\|_2^3\big(M/24 - 2c_{t+1}T^2\big)\Big] \\
&\quad + \mathbb{E}\Big[3M^{-1/2}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} + 28M^{-2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3\Big] \\
&\le \mathbb{E}\Big[F(\mathbf{x}_t^s) + c_{t+1}(1+3/T)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3\Big] + \mathbb{E}\Big[3M^{-1/2}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} + 28M^{-2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3\Big],
\end{aligned} \tag{A.10}$$

where the third inequality holds due to Lemma A.7. To further bound (A.10), we have

$$\frac{3}{M^{1/2}}\mathbb{E}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} \le \frac{3\rho^{3/2}}{M^{1/2}b_g^{3/4}}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3 \le \frac{M}{1000T^3}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3, \tag{A.11}$$

where the first inequality holds due to Lemma A.2, the second inequality holds due to $M \ge 100\rho$ and $b_g \ge 5T^4$ from the condition of Theorem 4.7. We also have

$$\frac{28}{M^2}\mathbb{E}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3 \le \frac{28 \times 15000\rho^3}{M^2(b_h/\log d)^{3/2}}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3 \le \frac{M}{1000T^3}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3, \tag{A.12}$$

where the first inequality holds due to Lemma A.3, where we have $b_h \geq 100T^2 \log d \geq 400 \log d$, and the second inequality holds due to $M \geq 100\rho$ and $b_h \geq 100T^2 \log d$ from the assumption of Theorem 4.7. Thus, submitting (A.11) and (A.12) into (A.10), we have

$$\begin{aligned} R^s_{t+1} + \big(240C_M^2\rho^{1/2}\big)^{-1}\mathbb{E}\mu(\mathbf{x}^s_{t+1}) &\leq \mathbb{E}\bigg[F(\mathbf{x}^s_t) + \|\mathbf{x}^s_t - \widehat{\mathbf{x}}^s\|_2^3\bigg(c_{t+1}(1+3/T) + \frac{M}{500T^3}\bigg)\bigg] \\ &= \mathbb{E}\big[F(\mathbf{x}^s_t) + c_t\|\mathbf{x}^s_t - \widehat{\mathbf{x}}^s\|_2^3\big] \\ &= R^s_t, \end{aligned} \tag{A.13}$$

where the first equality holds due to the definition of c_t in Lemma A.7. Telescoping (A.13) from $t = 0$ to $T-1$, we have

$$R^s_0 - R^s_T \geq \sum_{t=1}^{T} \big(240C_M^2\rho^{1/2}\big)^{-1}\mathbb{E}\mu(\mathbf{x}^s_t).$$

By the definition of c_T in Lemma A.7, we have $c_T = 0$, then $R^s_T = \mathbb{E}\big[F(\mathbf{x}^s_T) + c_T\|\mathbf{x}^s_T - \widehat{\mathbf{x}}^s\|_2^3\big] = \mathbb{E}F(\widehat{\mathbf{x}}^{s+1})$; meanwhile by the definition of $\mathbf{x}^s_0$, we have $\mathbf{x}^s_0 = \widehat{\mathbf{x}}^s$. Thus we have $R^s_0 = \mathbb{E}\big[F(\mathbf{x}^s_0) + c_0\|\mathbf{x}^s_0 - \widehat{\mathbf{x}}^s\|_2^3\big] = \mathbb{E}F(\widehat{\mathbf{x}}^s)$, which implies

$$\mathbb{E}F(\widehat{\mathbf{x}}^s) - \mathbb{E}F(\widehat{\mathbf{x}}^{s+1}) = R^s_0 - R^s_T \geq \big(240C_M^2\rho^{1/2}\big)^{-1}\sum_{t=1}^{T}\mathbb{E}\mu(\mathbf{x}^s_t). \tag{A.14}$$

Finally, telescoping (A.14) from $s = 1$ to S yields

$$\Delta_F \geq \sum_{s=1}^{S}\mathbb{E}F(\widehat{\mathbf{x}}^s) - \mathbb{E}F(\widehat{\mathbf{x}}^{s+1}) \geq \big(240C_M^2\rho^{1/2}\big)^{-1}\sum_{s=1}^{S}\sum_{t=1}^{T}\mathbb{E}\mu(\mathbf{x}^s_t).$$

By the definition about choice of $\mathbf{x}_{\text{out}}$, we complete the proof. □

A.2 Proof of Corollary 4.10

Proof. We can verify that the parameter setting in Corollary 4.10 satisfies the requirement of Theorem 4.7. Thus, submitting the choice of parameters into Theorem 4.7, the output of Algorithm 1 $\mathbf{x}_{\text{out}}$ satisfies that

$$\mathbb{E}[\mu(\mathbf{x}_{\text{out}})] \leq \frac{240C_M^2\rho^{1/2}\Delta_F}{ST} \leq \epsilon^{3/2}, \tag{A.15}$$

which indeed implies that $\mathbf{x}_{\text{out}}$ is an $(\epsilon, \sqrt{\rho\epsilon})$-approximate local minimum. Next we calculate how many SO calls and CSO calls are needed. Algorithm 1 needs to calculate full gradient $\mathbf{g}_s$ and full Hessian $\mathbf{H}_s$ at the beginning of each epoch, with n SO calls. In each epoch, Algorithm 1 needs to calculate $\mathbf{v}^s_t$ and $\mathbf{U}^s_t$ with $b_g + b_h$ SO calls at each iteration. Thus, the total amount of SO calls is

$$\begin{aligned} Sn + (ST)(b_g + b_h) &\leq n + C_1\Delta_F\rho^{1/2}n^{4/5}\epsilon^{-3/2} + C_1\Delta_F\rho^{1/2}\epsilon^{-3/2}(5n^{4/5} + 1000n^{2/5}\log d) \\ &= \widetilde{O}\bigg(n + \frac{\Delta_F\sqrt{\rho}n^{4/5}}{\epsilon^{3/2}}\bigg), \end{aligned}$$

where $C_1 = 240C_M^2$. For the CSO calls, Algorithm 1 needs to solve cubic subproblem at each single iteration. Thus, the total amount of CSO calls is

$$ST \leq C_1\Delta_F\rho^{1/2}\epsilon^{-3/2} = O\bigg(\frac{\Delta_F\sqrt{\rho}}{\epsilon^{3/2}}\bigg).$$

□

A.3 Proof of Theorem 5.3

Proof of Theorem 5.3. Similar to (A.6) in the proof of Theorem 4.7, we have

$$
\begin{aligned}
F(\mathbf{x}_{t+1}^s) &\le F(\mathbf{x}_t^s) + \big\langle \nabla F(\mathbf{x}_t^s), \widetilde{\mathbf{h}}_t^s \big\rangle + \frac{1}{2}\big\langle \nabla^2 F(\mathbf{x}_t^s)\widetilde{\mathbf{h}}_t^s, \widetilde{\mathbf{h}}_t^s \big\rangle + \frac{\rho}{6}\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 \\
&= F(\mathbf{x}_t^s) + \big\langle \mathbf{v}_t^s, \widetilde{\mathbf{h}}_t^s \big\rangle + \frac{1}{2}\big\langle \mathbf{U}_t^s\widetilde{\mathbf{h}}_t^s, \widetilde{\mathbf{h}}_t^s \big\rangle + \frac{M}{6}\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 + \big\langle \nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s, \widetilde{\mathbf{h}}_t^s \big\rangle \\
&\quad + \frac{1}{2}\big\langle (\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s)\widetilde{\mathbf{h}}_t^s, \widetilde{\mathbf{h}}_t^s \big\rangle - \frac{M-\rho}{6}\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 \\
&\le F(\mathbf{x}_t^s) - \frac{M}{12}\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 + \delta + \bigg(\frac{M}{27}\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 + \frac{2\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2}}{M^{1/2}} \bigg) \\
&\quad + \frac{1}{2}\bigg(\frac{2M}{27}\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 + \frac{27}{M^2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3 \bigg) - \frac{M-\rho}{6}\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 \\
&\le F(\mathbf{x}_t^s) - \frac{M}{12}\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 + \frac{2}{M^{1/2}}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} + \frac{27}{M^2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3 + \delta, \qquad \text{(A.16)}
\end{aligned}
$$

where the second inequality holds because $\widetilde{\mathbf{h}}_t^s$ is an inexact solver satisfying Condition 5.1. By Lemma A.6 with $\mathbf{h} = \widetilde{\mathbf{h}}_t^s$, we have

$$
c_{t+1}\|\mathbf{x}_{t+1}^s - \widehat{\mathbf{x}}^s\|_2^3 = c_{t+1}\big\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s + \widetilde{\mathbf{h}}_t^s\big\|_2^3 \le 2c_{t+1}T^2\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 + c_{t+1}(1 + 3/T)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3. \qquad \text{(A.17)}
$$

By Lemma A.5, we also have

$$
\begin{aligned}
&\big(240C_M^2\rho^{1/2}\big)^{-1}\mu(\mathbf{x}_{t+1}^s) \\
&= \big(240C_M^2\rho^{1/2}\big)^{-1}\mu\big(\mathbf{x}_t^s + \widetilde{\mathbf{h}}_t^s\big) \\
&\le \frac{M}{24}\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 + \frac{\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2}}{24M^{1/2}} + \frac{\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3}{24M^2} + \frac{\big\|\nabla m_t^s\big(\widetilde{\mathbf{h}}_t^s\big)\big\|_2^{3/2}}{24M^{1/2}} + \frac{M\big|\big\|\widetilde{\mathbf{h}}_t^s\big\|_2 - \|\mathbf{h}_t^s\|_2\big|^3}{24}, \qquad \text{(A.18)}
\end{aligned}
$$

Since $\widetilde{\mathbf{h}}_t^s$ is an inexact solver satisfying Condition 5.1, we have

$$
\frac{\big\|\nabla m_t^s\big(\widetilde{\mathbf{h}}_t^s\big)\big\|_2^{3/2}}{24M^{1/2}} + \frac{M\big|\big\|\widetilde{\mathbf{h}}_t^s\big\|_2 - \|\mathbf{h}_t^s\|_2\big|^3}{24} \le \frac{\delta}{24} + \frac{\delta}{24} < \delta. \qquad \text{(A.19)}
$$

Submitting (A.19) into (A.18), we have

$$
\big(240C_M^2\rho^{1/2}\big)^{-1}\mu(\mathbf{x}_{t+1}^s) \le \frac{M}{24}\big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3 + \frac{\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2}}{24M^{1/2}} + \frac{\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3}{24M^2} + \delta. \qquad \text{(A.20)}
$$

Then adding (A.16), (A.17) and (A.20) up, we have

$$
\begin{aligned}
&R_{t+1}^s + \big(240C_M^2\rho^{1/2}\big)^{-1}\mathbb{E}\mu(\mathbf{x}_{t+1}^s) \\
&= \mathbb{E}\Big[F(\mathbf{x}_{t+1}^s) + c_{t+1}\|\mathbf{x}_{t+1}^s - \widehat{\mathbf{x}}^s\|_2^3 + \big(240C_M^2\rho^{1/2}\big)^{-1}\mu(\mathbf{x}_{t+1}^s)\Big] \\
&\le \mathbb{E}\Big[F(\mathbf{x}_t^s) + c_{t+1}(1 + 3/T)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3 - \big\|\widetilde{\mathbf{h}}_t^s\big\|_2^3\big(M/24 - 2c_{t+1}T^2\big)\Big] \\
&\quad + \mathbb{E}\bigg[\frac{3}{M^{1/2}}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} + \frac{28}{M^2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3\bigg] + 2\delta \\
&\le \mathbb{E}\Big[F(\mathbf{x}_t^s) + c_{t+1}(1 + 3/T)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3\Big] + \mathbb{E}\bigg[\frac{3}{M^{1/2}}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} + \frac{28}{M^2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3\bigg] + 2\delta. \qquad \text{(A.21)}
\end{aligned}
$$

Since the parameter setting is the same as Theorem 4.7, by (A.11) and (A.12), we have

$$
\frac{3}{M^{1/2}}\mathbb{E}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} \le \frac{M}{1000T^3}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3, \qquad \text{(A.22)}
$$

and

$$\frac{28}{M^2}\mathbb{E}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3 \le \frac{M}{1000T^3}\mathbb{E}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3. \tag{A.23}$$

Submitting (A.22) and (A.23) into (A.21) yields

$$\begin{aligned} R_{t+1}^s + \big(240C_M^2\rho^{1/2}\big)^{-1}\mathbb{E}\mu(\mathbf{x}_{t+1}^s) &\le \mathbb{E}\bigg[F(\mathbf{x}_t^s) + \|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3\bigg(c_{t+1}(1+3/T) + \frac{M}{500T^3}\bigg)\bigg] + 2\delta \\ &= \mathbb{E}\big[F(\mathbf{x}_t^s) + c_t\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3\big] + 2\delta \\ &= R_t^s + 2\delta, \end{aligned} \tag{A.24}$$

where the first equality holds due to the definition of c_t in Lemma A.7. Telescoping (A.13) from $t = 0$ to $T-1$, we have

$$R_0^s - R_T^s \ge \sum_{t=1}^{T}\Big[\big(240C_M^2\rho^{1/2}\big)^{-1}\mathbb{E}\mu(\mathbf{x}_t^s) - 2\delta\Big].$$

By the definition of c_T in Lemma A.7, we have $c_T = 0$, then $R_T^s = \mathbb{E}\big[F(\mathbf{x}_T^s) + c_T\|\mathbf{x}_T^s - \widehat{\mathbf{x}}^s\|_2^3\big] = \mathbb{E}F(\widehat{\mathbf{x}}^{s+1})$; meanwhile by the definition of $\mathbf{x}_0^s$, we have $\mathbf{x}_0^s = \widehat{\mathbf{x}}^s$. Thus we have $R_0^s = \mathbb{E}\big[F(\mathbf{x}_0^s) + c_0\|\mathbf{x}_0^s - \widehat{\mathbf{x}}^s\|_2^3\big] = \mathbb{E}F(\widehat{\mathbf{x}}^s)$, which further implies

$$\mathbb{E}F(\widehat{\mathbf{x}}^s) - \mathbb{E}F(\widehat{\mathbf{x}}^{s+1}) = R_0^s - R_T^s \ge \sum_{t=1}^{T}\Big[\big(240C_M^2\rho^{1/2}\big)^{-1}\mathbb{E}\mu(\mathbf{x}_t^s) - 2\delta\Big]. \tag{A.25}$$

Finally, telescoping (A.25) from $s = 1$ to S, we obtain

$$\Delta_F \ge \sum_{s=1}^{S}\mathbb{E}F(\widehat{\mathbf{x}}^s) - \mathbb{E}F(\widehat{\mathbf{x}}^{s+1})\sum_{s=1}^{S}\sum_{t=1}^{T}\Big[\big(240C_M^2\rho^{1/2}\big)^{-1}\mathbb{E}\mu(\mathbf{x}_t^s) - 2\delta\Big].$$

By the definition about choice of $\mathbf{x}_{\text{out}}$, we finish the proof.

□

A.4 Proof of Corollary 5.5

Proof of Corollary 5.5. We can verify that under the parameter choice in Corollary 5.5,

$$\mathbb{E}[\mu(\mathbf{x}_{\text{out}})] \le \frac{240C_M^2\rho^{1/2}\Delta_F}{ST} + 480C_M^2\rho^{1/2}\delta \le \epsilon^{3/2}/2 + \epsilon^{3/2}/2 = \epsilon^{3/2}. \tag{A.26}$$

Thus, $\mathbf{x}_{\text{out}}$ is an $(\epsilon, \sqrt{\rho\epsilon})$-approximate local minimum. By the proof of Corollary 4.10, the total amount of SO calls is

$$\begin{aligned} Sn + (ST)(b_g + b_h) &\le n + C_1\Delta_F\rho^{1/2}n^{4/5}\epsilon^{-3/2} + C_1\Delta_F\rho^{1/2}\epsilon^{-3/2}(5n^{4/5} + 1000n^{2/5}\log d) \\ &= \widetilde{O}\bigg(n + \frac{\Delta_F\sqrt{\rho}n^{4/5}}{\epsilon^{3/2}}\bigg), \end{aligned}$$

where $C_1 = 480C_M^2$. For the CSO calls, Algorithm 1 needs to solve cubic subproblem at each single iteration. Thus, the total amount of CSO calls is

$$ST \le C_1\Delta_F\rho^{1/2}\epsilon^{-3/2} = O\bigg(\frac{\Delta_F\sqrt{\rho}}{\epsilon^{3/2}}\bigg).$$

□

B Proof of Technical Lemmas

Now we prove the technical lemmas used in Section A.

B.1 Proof of Lemma A.1

The result of Lemma A.1 is typical in the literature of cubic regularization (Nesterov & Polyak, 2006; Cartis et al., 2011a;b), but no exactly the same result has been shown in any formal way. Thus we present the proof here for self-containedness.

Proof of Lemma A.1. For simplicity, we let $\mathbf{g} = \mathbf{v}_t^s, \mathbf{H} = \mathbf{U}_t^s, \theta = M_t$ and $\mathbf{h}_{\text{opt}} = \mathbf{h}_t^s$. Then we need to prove

$$\begin{align}
\mathbf{g} + \mathbf{H}\mathbf{h}_{\text{opt}} + \frac{\theta}{2}\|\mathbf{h}_{\text{opt}}\|_2\mathbf{h}_{\text{opt}} = \mathbf{0}, \tag{B.1}\\
\mathbf{H} + \frac{\theta}{2}\|\mathbf{h}_{\text{opt}}\|_2\mathbf{I} \succeq \mathbf{0}, \tag{B.2}\\
\langle \mathbf{g}, \mathbf{h}_{\text{opt}}\rangle + \frac{1}{2}\langle \mathbf{H}\mathbf{h}_{\text{opt}}, \mathbf{h}_{\text{opt}}\rangle + \frac{\theta}{6}\|\mathbf{h}_{\text{opt}}\|_2^3 \le -\frac{\theta}{12}\|\mathbf{h}_{\text{opt}}\|_2^3. \tag{B.3}
\end{align}$$

Let $\lambda = \theta\|\mathbf{h}_{\text{opt}}\|_2/2$. Note that $\mathbf{h}_{\text{opt}} = \operatorname{argmin} m(\mathbf{h})$, then the necessary condition $\nabla m(\mathbf{h}_{\text{opt}}) = \mathbf{0}$ and $\nabla^2 m(\mathbf{h}_{\text{opt}}) \succeq \mathbf{0}$ can be written as

$$\begin{align}
\nabla m(\mathbf{h}_{\text{opt}}) = \mathbf{g} + \mathbf{H}\mathbf{h}_{\text{opt}} + \lambda\mathbf{h}_{\text{opt}} = \mathbf{0}, \tag{B.4}\\
\mathbf{w}^\top\nabla^2 m(\mathbf{h}_{\text{opt}})\mathbf{w} = \mathbf{w}^\top\left(\mathbf{H} + \lambda\mathbf{I} + \lambda\left(\frac{\mathbf{h}_{\text{opt}}}{\|\mathbf{h}_{\text{opt}}\|_2}\right)\left(\frac{\mathbf{h}_{\text{opt}}}{\|\mathbf{h}_{\text{opt}}\|_2}\right)^\top\right)\mathbf{w} \ge 0, \forall \mathbf{w} \in \mathbb{R}^d. \tag{B.5}
\end{align}$$

Apparently, (B.4) directly implies (B.1). To prove (B.2), we adapt the proof of Lemma 5.1 in Agarwal et al. (2017). Note that if $\langle \mathbf{w}, \mathbf{h}_{\text{opt}}\rangle = 0$, then (B.5) directly implies (B.2). So we only need to focus on the case that $\langle \mathbf{w}, \mathbf{h}_{\text{opt}}\rangle \neq 0$.

Since $\langle \mathbf{w}, \mathbf{h}_{\text{opt}}\rangle \neq 0$, there exists $\eta \neq 0$ such that $\|\mathbf{h}_{\text{opt}} + \eta\mathbf{w}\|_2 = \|\mathbf{h}_{\text{opt}}\|_2$. (In fact, we can find $\eta = -2\langle \mathbf{w}, \mathbf{h}_{\text{opt}}\rangle/\|\mathbf{w}\|_2^2$ satisfies the requirement). Next we will take a close look at the difference $m(\mathbf{h}_{\text{opt}} + \eta\mathbf{w}) - m(\mathbf{h}_{\text{opt}})$. On one hand, we have

$$\begin{align}
m(\mathbf{h}_{\text{opt}} + \eta\mathbf{w}) - m(\mathbf{h}_{\text{opt}}) &= \mathbf{g}^\top[(\mathbf{h}_{\text{opt}} + \eta\mathbf{w}) - \mathbf{h}_{\text{opt}}] + \frac{(\mathbf{h}_{\text{opt}} + \eta\mathbf{w})^\top\mathbf{H}(\mathbf{h}_{\text{opt}} + \eta\mathbf{w})}{2} - \frac{\mathbf{h}_{\text{opt}}^\top\mathbf{H}\mathbf{h}_{\text{opt}}}{2} \notag\\
&= -[(\mathbf{h}_{\text{opt}} + \eta\mathbf{w}) - \mathbf{h}_{\text{opt}}]^\top(\mathbf{H} + \lambda\mathbf{I})\mathbf{h}_{\text{opt}} + \frac{(\mathbf{h}_{\text{opt}} + \eta\mathbf{w})^\top\mathbf{H}(\mathbf{h}_{\text{opt}} + \eta\mathbf{w})}{2} - \frac{\mathbf{h}_{\text{opt}}^\top\mathbf{H}\mathbf{h}_{\text{opt}}}{2} \tag{B.6}\\
&= \frac{\lambda\eta^2}{2}\|\mathbf{w}\|_2^2 + [\mathbf{h}_{\text{opt}} - (\mathbf{h}_{\text{opt}} + \eta\mathbf{w})]^\top\mathbf{H}\mathbf{h}_{\text{opt}} + \frac{(\mathbf{h}_{\text{opt}} + \eta\mathbf{w})^\top\mathbf{H}(\mathbf{h}_{\text{opt}} + \eta\mathbf{w})}{2} - \frac{\mathbf{h}_{\text{opt}}^\top\mathbf{H}\mathbf{h}_{\text{opt}}}{2} \tag{B.7}\\
&= \frac{\lambda\eta^2}{2}\|\mathbf{w}\|_2^2 + \frac{\mathbf{h}_{\text{opt}}^\top\mathbf{H}\mathbf{h}_{\text{opt}}}{2} - (\mathbf{h}_{\text{opt}} + \eta\mathbf{w})^\top\mathbf{H}\mathbf{h}_{\text{opt}} + \frac{(\mathbf{h}_{\text{opt}} + \eta\mathbf{w})^\top\mathbf{H}(\mathbf{h}_{\text{opt}} + \eta\mathbf{w})}{2} \notag\\
&= \frac{\lambda\eta^2}{2}\|\mathbf{w}\|_2^2 + \frac{\eta^2}{2}\mathbf{w}^\top\mathbf{H}\mathbf{w} = \frac{\eta^2}{2}\mathbf{w}^\top(\mathbf{H} + \lambda\mathbf{I})\mathbf{w}, \notag
\end{align}$$

where (B.6) holds due to (B.4) and (B.7) holds due to the definition of η. On the other hand, by the definition of $\mathbf{h}_{\text{opt}}$, $m(\mathbf{h}_{\text{opt}} + \eta\mathbf{w}) - m(\mathbf{h}_{\text{opt}}) \ge 0$. Thus, we have proved (B.2). Finally, we prove (B.3) by showing that

$$\begin{align}
\langle \mathbf{g}, \mathbf{h}_{\text{opt}}\rangle + \frac{1}{2}\langle \mathbf{H}\mathbf{h}_{\text{opt}}, \mathbf{h}_{\text{opt}}\rangle + \frac{\theta}{6}\|\mathbf{h}_{\text{opt}}\|_2^3 &= \left\langle \mathbf{g} + \mathbf{H}\mathbf{h}_{\text{opt}} + \frac{\theta}{2}\|\mathbf{h}_{\text{opt}}\|_2\mathbf{h}_{\text{opt}}, \mathbf{h}_{\text{opt}}\right\rangle - \frac{1}{2}\mathbf{h}_{\text{opt}}^\top(\mathbf{H} + \lambda\mathbf{I})\mathbf{h}_{\text{opt}} - \frac{\theta}{12}\|\mathbf{h}_{\text{opt}}\|_2^3 \notag\\
&= -\frac{1}{2}\mathbf{h}_{\text{opt}}^\top(\mathbf{H} + \lambda\mathbf{I})\mathbf{h}_{\text{opt}} - \frac{\theta}{12}\|\mathbf{h}_{\text{opt}}\|_2^3 \tag{B.8}\\
&\le -\frac{\theta}{12}\|\mathbf{h}_{\text{opt}}\|_2^3, \tag{B.9}
\end{align}$$

where (B.8) holds due to (B.1) and (B.9) holds due to (B.2). □

B.2 Proof of Lemma A.2

In order to prove Lemma A.2, we need the following useful lemma.

Lemma B.1. Suppose $\mathbf{a}_1, \ldots, \mathbf{a}_N$ are i.i.d. and $\mathbb{E}\mathbf{a}_i = 0$, then

$$\mathbb{E}\bigg\|\frac{1}{N}\sum_{i=1}^N \mathbf{a}_i\bigg\|_2^{3/2} \le \frac{1}{N^{3/4}}\big(\mathbb{E}\|\mathbf{a}_i\|_2^2\big)^{3/4}.$$

Proof of Lemma A.2. For simplification, we use $\mathbb{E}$ to replace $\mathbb{E}_{\mathbf{v}_{i_t}}$. We have

$$\begin{aligned}
&\mathbb{E}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2}\\
&= \mathbb{E}\bigg\|\frac{1}{b_g}\sum\big[\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s)\big] + \mathbf{g}^s - \bigg[\frac{1}{b_g}\sum \nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s) - \mathbf{H}^s\bigg](\mathbf{x}_t^s - \widehat{\mathbf{x}}^s) - \nabla F(\mathbf{x}_t^s)\bigg\|_2^{3/2}\\
&= \mathbb{E}\bigg\|\frac{1}{b_g}\sum\big[\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s) - \nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s) - \nabla F(\mathbf{x}_t^s) + \nabla F(\widehat{\mathbf{x}}^s) + \nabla^2 F(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s)\big]\bigg\|_2^{3/2}.
\end{aligned}$$

Now we set the parameters in Lemma B.1 as

$$N = b_g, \mathbf{a}_{i_t} = \nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s) - \nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s) - \nabla F(\mathbf{x}_t^{s+1}) + \nabla F(\widehat{\mathbf{x}}^s) + \nabla^2 F(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s).$$

We can check that $\mathbf{a}_{i_t}$ satisfy the assumption of Lemma B.1. Thus, by Lemma B.1, we have

$$\begin{aligned}
\mathbb{E}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} \le \frac{1}{b_g^{3/4}}\Big(&\mathbb{E}\big\|\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s) - \nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s)\\
&- \nabla F(\mathbf{x}_t^s) + \nabla F(\widehat{\mathbf{x}}^s) + \nabla^2 F(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s)\big\|_2^2\Big)^{3/4}. \qquad \text{(B.10)}
\end{aligned}$$

By Assumption 4.1, we have

$$\begin{aligned}
&\big\|\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s) - \nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s) - \nabla F(\mathbf{x}_t^s) + \nabla F(\widehat{\mathbf{x}}^s) + \nabla^2 F(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s)\big\|_2\\
&\le \big\|\nabla f_{i_t}(\mathbf{x}_t^s) - \nabla f_{i_t}(\widehat{\mathbf{x}}^s) - \nabla^2 f_{i_t}(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s)\big\|_2 + \big\|\nabla F(\mathbf{x}_t^s) - \nabla F(\widehat{\mathbf{x}}^s) - \nabla^2 F(\widehat{\mathbf{x}}^s)(\mathbf{x}_t^s - \widehat{\mathbf{x}}^s)\big\|_2\\
&\le \frac{\rho}{2}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^2 + \frac{\rho}{2}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^2\\
&= \rho\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^2. \qquad \text{(B.11)}
\end{aligned}$$

Plugging (B.11) into (B.10) yields

$$\mathbb{E}\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} \le \frac{1}{b_g^{3/4}}\big(\rho^2\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^4\big)^{3/4} = \frac{\rho^{3/2}}{b_g^{3/4}}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3.$$

□

B.3 Proof of Lemma A.3

In order to prove Lemma A.3, we need the following supporting lemma.

Lemma B.2. Suppose that $q \ge 2, p \ge 2$, and fix $r \ge \max\{q, 2\log p\}$. Consider $\mathbf{Y}_1, ..., \mathbf{Y}_N$ of i.i.d. random self-adjoint matrices with dimension $p \times p$, $\mathbb{E}\mathbf{Y}_i = \mathbf{0}$, then

$$\bigg[\mathbb{E}\bigg\|\sum_{i=1}^N \mathbf{Y}_i\bigg\|_2^q\bigg]^{1/q} \le 2\sqrt{er}\bigg\|\bigg(\sum_{i=1}^N \mathbb{E}\mathbf{Y}_i^2\bigg)^{1/2}\bigg\|_2 + 4er\big(\mathbb{E}\max_i \|\mathbf{Y}_i\|_2^q\big)^{1/q}.$$

Proof of Lemma A.3. In the following proof, we replace $\mathbb{E}_{j_t}$ with $\mathbb{E}$ for simplification. We have

$$\begin{aligned}
\mathbb{E}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_2^3 &= \mathbb{E}\bigg\|\nabla^2 F(\mathbf{x}_t^s) - \frac{1}{b_h}\bigg(\sum\big(\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s) + \mathbf{H}^s\big)\bigg)\bigg\|_2^3\\
&= \mathbb{E}\bigg\|\frac{1}{b_h}\bigg[\sum\big[\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s) + \mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s)\big]\bigg]\bigg\|_2^3. \qquad \text{(B.12)}
\end{aligned}$$

We apply Lemma B.2 with parameters

$$q = 3, p = d, r = 2\log p, \mathbf{Y}_{j_t} = \nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s) + \mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s), N = b_h.$$

It can be easily checked that these parameters satisfy the assumption of Lemma B.2. Meanwhile, by Assumption 4.1, we have the following upper bound for $\mathbf{Y}_{j_t}$:

$$\begin{aligned}\big\|\mathbf{Y}_{j_t}\big\|_2 &= \Big\|\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s) + \mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s)\Big\|_2 \\ &\le \Big\|\nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s)\Big\|_2 + \Big\|\mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s)\Big\|_2 \\ &\le \rho\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2 + \rho\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2 \\ &= 2\rho\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2. \end{aligned} \tag{B.13}$$

By Lemma B.2, we have

$$\bigg[\mathbb{E}\Big\|\sum \mathbf{Y}_{j_t}\Big\|_2^3\bigg]^{1/3} \le 2\sqrt{er}\Big\|\Big(\sum \mathbb{E}\mathbf{Y}_{j_t}^2\Big)^{1/2}\Big\|_2 + 4er\big(\mathbb{E}\max_i \|\mathbf{Y}_i\|_2^3\big)^{1/3}. \tag{B.14}$$

The first term in RHS of (B.14) can be bounded as

$$\begin{aligned} 2\sqrt{er}\Big\|\Big(\sum \mathbb{E}\mathbf{Y}_{j_t}^2\Big)^{1/2}\Big\|_2 &= 2\sqrt{er}\Big\|\sum \mathbb{E}\mathbf{Y}_{j_t}^2\Big\|_2^{1/2} \\ &= 2\sqrt{Ner}\Big\|\mathbb{E}\mathbf{Y}_{j_t}^2\Big\|_2^{1/2} \\ &\le 2\sqrt{Ner}\Big(\mathbb{E}\Big\|\mathbf{Y}_{j_t}^2\Big\|_2\Big)^{1/2} \\ &= 2\sqrt{Ner}\Big(\mathbb{E}\|\mathbf{Y}_{j_t}\|_2^2\Big)^{1/2} \\ &\le 4\rho\sqrt{Ner}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2, \end{aligned} \tag{B.15}$$

where the first inequality holds due to Jensen's inequality, the third equality holds because $\big\|\mathbf{Y}_{j_t}^2\big\|_2 = \|\mathbf{Y}_{j_t}\|_2^2$ and the last inequality holds due to (B.13). The second term in RHS of (B.14) can be bounded as

$$4er\big(\mathbb{E}\max_i \|\mathbf{Y}_i\|_2^3\big)^{1/3} \le 4er[(2\rho\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2)^3]^{1/3} = 8\rho er\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2. \tag{B.16}$$

Submitting (B.15), (B.16) into (B.14), we have

$$\bigg[\mathbb{E}\Big\|\sum \mathbf{Y}_{j_t}\Big\|_2^3\bigg]^{1/3} \le 4\rho\sqrt{Ner}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2 + 8\rho er\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2,$$

which immediately implies

$$\mathbb{E}\Big\|\frac{1}{N}\sum \mathbf{Y}_{j_t}\Big\|_2^3 \le 64\rho^3\bigg(\sqrt{\frac{er}{N}} + \frac{2er}{N}\bigg)^3\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3. \tag{B.17}$$

Submitting (B.17) into (B.12) with $\mathbf{Y}_{j_t} = \nabla^2 f_{j_t}(\mathbf{x}_t^s) - \nabla^2 f_{j_t}(\widehat{\mathbf{x}}^s) + \mathbf{H}^s - \nabla^2 F(\mathbf{x}_t^s),\ r = 2\log d,\ N = b_h$, we have

$$\begin{aligned}\mathbb{E}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3 &\le 64\rho^3\bigg(\sqrt{\frac{2e\log d}{b_h}} + \frac{4e\log d}{b_h}\bigg)^3\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3 \\ &\le 1200\rho^3\bigg(\frac{\log d}{b_h}\bigg)^{3/2}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3,\end{aligned}$$

where the last inequality holds due to $b_h \ge 400\log d$.

□

B.4 Proof of Lemma A.4

Proof. we have

$$
\begin{aligned}
\langle \nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s, \mathbf{h} \rangle &\le \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2 \cdot \|\mathbf{h}\|_2 \\
&= \bigg(\frac{M^{1/3}}{9^{1/3}} \|\mathbf{h}\|_2 \bigg) \cdot \bigg(\frac{9^{1/3}}{M^{1/3}} \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2 \bigg) \\
&\le \frac{1}{3} \bigg(\frac{M^{1/3}}{9^{1/3}} \|\mathbf{h}\|_2 \bigg)^3 + \frac{2}{3} \bigg(\frac{9^{1/3}}{M^{1/3}} \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2 \bigg)^{3/2} \\
&= \frac{M}{27} \|\mathbf{h}\|_2^3 + \frac{2\|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2}}{M^{1/2}},
\end{aligned}
$$

where the second inequality holds due to Young's inequality. Meanwhile, we have

$$
\begin{aligned}
\big\langle \big(\nabla^2 F(\mathbf{x}_t^s) - \mathbf{H}^s\big)\mathbf{h}, \mathbf{h} \big\rangle &\le \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{H}^s\big\|_2 \cdot \|\mathbf{h}\|_2^2 \\
&= \bigg(\frac{M^{2/3}}{9^{2/3}} \|\mathbf{h}\|_2^2 \bigg) \cdot \bigg(\frac{9^{2/3}}{M^{2/3}} \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{H}^s\big\|_2 \bigg) \\
&\le \frac{2}{3} \bigg(\frac{M^{2/3}}{9^{2/3}} \|\mathbf{h}\|_2^2 \bigg)^{3/2} + \frac{1}{3} \bigg(\frac{9^{2/3}}{M^{2/3}} \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{H}^s\big\|_2 \bigg)^3 \\
&= \frac{2M}{27} \|\mathbf{h}\|_2^3 + \frac{27}{M^2} \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3,
\end{aligned}
$$

where the second inequality holds due to Young's inequality. □

B.5 Proof of Lemma A.5

In order to prove Lemma A.5, we need to the following two useful lemmas.

Lemma B.3. Under Assumption 4.1, if $M \ge 2\rho$, then we have

$$
\big\|\nabla F(\mathbf{x}_t^s + \mathbf{h})\big\|_2 \le M\|\mathbf{h}\|_2^2 + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2 + \frac{1}{M}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^2 + \big\|\nabla m_t^s(\mathbf{h})\big\|_2.
$$

Lemma B.4. Under Assumption 4.1, if $M \ge 2\rho$, then we have

$$
-\lambda_{\min}\big(\nabla^2 F(\mathbf{x}_t^s + \mathbf{h})\big) \le M\|\mathbf{h}\|_2 + \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2 + M\big|\|\mathbf{h}\|_2 - \|\mathbf{h}_t^s\|_2\big|.
$$

Proof of Lemma A.5. By the definition of μ, we can bound $\|\nabla F(\mathbf{x}_t^s + \mathbf{h})\|_2^{3/2}$ and $0 \vee -\rho^{-3/2}\big[\lambda_{\min}\big(\nabla^2 F(\mathbf{x}_t^s + \mathbf{h})\big)\big]^3$ separately. To bound $\|\nabla F(\mathbf{x}_t^s + \mathbf{h})\|_2^{3/2}$, applying Lemma B.3 we have

$$
\begin{aligned}
\big\|\nabla F(\mathbf{x}_t^s + \mathbf{h})\big\|_2^{3/2} &\le \Big[M\|\mathbf{h}\|_2^2 + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2 + \frac{1}{M}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^2 + \big\|\nabla m_t^s(\mathbf{h})\big\|_2\Big]^{3/2} \\
&\le 2\Big[M^{3/2}\|\mathbf{h}\|_2^3 + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} + M^{-3/2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3 + \big\|\nabla m_t^s(\mathbf{h})\big\|_2^{3/2}\Big],
\end{aligned}
$$

where the second inequality holds due to the following basic inequality $(a+b+c+d)^{3/2} \le 2(a^{3/2} + b^{3/2} + c^{3/2} + d^{3/2})$. To bound $-\lambda_{\min}\big(\nabla^2 F(\mathbf{x}_t^s + \mathbf{h})\big)$, applying Lemma A.3, we have

$$
\begin{aligned}
-\rho^{-3/2}\big[\lambda_{\min}\big(\nabla^2 F(\mathbf{x}_t^s + \mathbf{h})\big)\big]^3 &= -C_M^{3/2} M^{-3/2}\big[\lambda_{\min}\big(\nabla^2 F(\mathbf{x}_t^s + \mathbf{h})\big)\big]^3 \\
&\le C_M^{3/2} M^{-3/2}\Big[M\|\mathbf{h}\|_2 + \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2 + M\big|\|\mathbf{h}\|_2 - \|\mathbf{h}_t^s\|_2\big|\Big]^3 \\
&\le 9C_M^{3/2}\bigg[M^{3/2}\|\mathbf{h}\|_2^3 + M^{-3/2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3 + M^{3/2}\big|\|\mathbf{h}\|_2 - \|\mathbf{h}_t^s\|_2\big|^3\bigg],
\end{aligned}
$$

where the second inequality holds due to $(a+b+c)^3 \leq 9(a^3+b^3+c^3)$. Since $9C_M^{3/2} > 2$, we have

$$\begin{aligned}
&\mu(\mathbf{x}_t^s + \mathbf{h}) \\
&= \max\Big\{\|\nabla F(\mathbf{x}_t^s + \mathbf{h})\|_2^{3/2}, -\rho^{-3/2}\big[\lambda_{\min}\big(\nabla^2 F(\mathbf{x}_t^s + \mathbf{h})\big)\big]^3\Big\} \\
&\leq 9C_M^{3/2}\Big[M^{3/2}\|\mathbf{h}\|_2^3 + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2^{3/2} + M^{-3/2}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^3 + \big\|\nabla m_t^s(\mathbf{h})\big\|_2^{3/2} + M^{3/2}\big|\|\mathbf{h}\|_2 - \|\mathbf{h}_t^s\|_2\big|^3\Big],
\end{aligned}$$

which completes the proof. □

B.6 Proof of Lemma A.6

Proof. We have

$$\begin{aligned}
&\|\mathbf{x}_t^s + \mathbf{h} - \widehat{\mathbf{x}}^s\|_2^3 \\
&\leq \big(\|\mathbf{h}\|_2 + \|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2\big)^3 \\
&= \|\mathbf{h}\|_2^3 + 3\|\mathbf{h}\|_2^2 \cdot \|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2 + 3\|\mathbf{h}\|_2 \cdot \|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^2 + \|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3 \\
&= \|\mathbf{h}\|_2^3 + 3\big(T^{1/3}\|\mathbf{h}\|_2^2\big) \cdot \frac{\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2}{T^{1/3}} + 3\big(T^{2/3}\|\mathbf{h}\|_2\big) \cdot \frac{\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^2}{T^{2/3}} + \|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3 \\
&\leq \|\mathbf{h}\|_2^3 + 3\bigg(\frac{2}{3}\big(T^{1/3}\|\mathbf{h}\|_2^2\big)^{3/2} + \frac{1}{3}\bigg(\frac{\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2}{T^{1/3}}\bigg)^3\bigg) + 3\bigg(\frac{1}{3}\big(T^{2/3}\|\mathbf{h}\|_2\big)^3 + \frac{2}{3}\bigg(\frac{\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^2}{T^{2/3}}\bigg)^{3/2}\bigg) + \|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3 \\
&= \|\mathbf{h}\|_2^3 + \bigg(2T^{1/2}\|\mathbf{h}\|_2^3 + \frac{1}{T}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3\bigg) + \bigg(T^2\|\mathbf{h}\|_2^3 + \frac{2}{T}\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3\bigg) + \|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|_2^3 \\
&\leq 2T^2\|\mathbf{h}\|_2^3 + \bigg(1 + \frac{3}{T}\bigg)\|\mathbf{x}_t^s - \widehat{\mathbf{x}}^s\|^3, \qquad \text{(B.18)}
\end{aligned}$$

where the second inequality holds due to Young's inequality, the last inequality holds because $T \geq 1$. □

B.7 Proof of Lemma A.7

Proof. By induction, we have for any $0 \leq t \leq T$,

$$c_t = M\frac{(1+3/T)^{T-t} - 1}{1500T^2}.$$

Then for any $0 \leq t \leq T$,

$$2c_tT^2 \leq M\frac{2(1+3/T)^T}{1500} \leq M\frac{2 \cdot 27}{1500} < \frac{M}{24}.$$

□

C Proof of Auxiliary Lemmas

In this section, we prove auxillary lemmas used in Appendix B.

C.1 Proof of Lemma B.1

Proof. We have

$$\mathbb{E}\bigg\|\frac{1}{N}\sum_{i=1}^N \mathbf{a}_i\bigg\|_2^{3/2} = \frac{\mathbb{E}\|\sum_{i=1}^N \mathbf{a}_i\|_2^{3/2}}{N^{3/2}} \leq \frac{(\mathbb{E}\|\sum_{i=1}^N \mathbf{a}_i\|_2^2)^{3/4}}{N^{3/2}} = \frac{(\sum_{i=1}^N \mathbb{E}\|\mathbf{a}_i\|_2^2)^{3/4}}{N^{3/2}} = \frac{\big(\mathbb{E}\|\mathbf{a}_i\|_2^2\big)^{3/4}}{N^{3/4}}. \qquad \text{(C.1)}$$

The first inequality holds due to Lemma D.1, where we set s, t in Lemma D.1 as $s = 3/2, t = 2$; the second equality holds due to $\mathbb{E}\mathbf{a}_i = 0$ and that $\mathbf{a}_i$ are identically independently distributed. □

C.2 Proof of Lemma B.2

Proof. This proof is mainly adapted from Chen et al. (2012); Tropp (2016). First, Let $\{\mathbf{Y}_i{}' : i = 1, \ldots, N\}$ be an independent copy of the sequence $\{\mathbf{Y}_i : i = 1, \ldots, N\}$. We denote $\mathbb{E}_{\mathbf{Y}'}$ to be the expectation over the independent copy $\mathbf{Y}'$. Then $\mathbb{E}_{\mathbf{Y}'}\mathbf{Y}_i{}' = 0$, then

$$\mathbb{E}\bigg\|\sum_{i=1}^{N}\mathbf{Y}_i\bigg\|_2^q = \mathbb{E}\bigg\|\sum_{i=1}^{N}\mathbb{E}_{\mathbf{Y}'}(\mathbf{Y}_i - \mathbf{Y}_i{}')\bigg\|_2^q \leq \mathbb{E}\bigg[\mathbb{E}_{\mathbf{Y}'}\bigg\|\sum_{i=1}^{N}(\mathbf{Y}_i - \mathbf{Y}_i{}')\bigg\|_2^q\bigg] = \mathbb{E}\bigg\|\sum_{i=1}^{N}(\mathbf{Y}_i - \mathbf{Y}_i{}')\bigg\|_2^q. \tag{C.2}$$

The first equality holds due to $\mathbb{E}_{\mathbf{Y}'}\mathbf{Y}_i{}' = 0$, the first inequality holds because $\|\cdot\|_2^q$ is a convex function, and the second equality holds because we combine the iterated expectation into a single expectation.

Note that $\mathbf{Y}_i - \mathbf{Y}_i{}'$ has the same distribution as $\mathbf{Y}_i{}' - \mathbf{Y}_i$, thus the independent sequence $\{\xi_i(\mathbf{Y}_i - \mathbf{Y}_i{}') : 1 \leq i \leq n\}$ has the same distribution as $\{\mathbf{Y}_i - \mathbf{Y}_i{}' : 1 \leq i \leq N\}$, where ξ_i are independent Rademacher random variables, also independent with $\mathbf{Y}_i, \mathbf{Y}_i{}'$. Therefore,

$$\mathbb{E}\bigg\|\sum_{i=1}^{N}(\mathbf{Y}_i - \mathbf{Y}_i{}')\bigg\|_2^q = \mathbb{E}\bigg\|\sum_{i=1}^{N}\xi_i(\mathbf{Y}_i - \mathbf{Y}_i{}')\bigg\|_2^q. \tag{C.3}$$

Furthermore, we have

$$\mathbb{E}\bigg\|\sum_{i=1}^{N}\xi_i(\mathbf{Y}_i - \mathbf{Y}_i{}')\bigg\|_2^q \leq \mathbb{E}\bigg[2^{q-1}\bigg(\bigg\|\sum_{i=1}^{N}\xi_i\mathbf{Y}_i\bigg\|_2^q + \bigg\|\sum_{i=1}^{N}\xi_i\mathbf{Y}_i{}'\bigg\|_2^q\bigg)\bigg] = 2^q\cdot\mathbb{E}\bigg\|\sum_{i=1}^{N}\xi_i\mathbf{Y}_i\bigg\|_2^q. \tag{C.4}$$

The first inequality holds due to $\|\mathbf{A}-\mathbf{B}\|_2^q \leq (\|\mathbf{A}\|_2+\|\mathbf{B}\|_2)^q \leq 2^{q-1}(\|\mathbf{A}\|_2^q+\|\mathbf{B}\|_2^q)$, where we let $\mathbf{A} = \sum_{i=1}^N \xi_i\mathbf{Y}_i, \mathbf{B} = \sum_{i=1}^N \xi_i\mathbf{Y}_i{}'$; the equality holds due to the identical distribution of $\{\xi\mathbf{Y}_i\}$ and $\{\xi\mathbf{Y}_i{}'\}$. Submitting (C.3), (C.4) into (C.2) yields

$$\mathbb{E}\bigg\|\sum_{i=1}^{N}\mathbf{Y}_i\bigg\|_2^q \leq 2^q\cdot\mathbb{E}\bigg\|\sum_{i=1}^{N}\xi_i\mathbf{Y}_i\bigg\|_2^q \tag{C.5}$$

Note that both sides of (C.5) are greater than 0, then we take q-th root for both sides, we have

$$\bigg[\mathbb{E}\bigg\|\sum_{i=1}^{N}\mathbf{Y}_i\bigg\|_2^q\bigg]^{1/q} \leq 2\bigg[\mathbb{E}\bigg\|\sum_{i=1}^{N}\xi_i\mathbf{Y}_i\bigg\|_2^q\bigg]^{1/q}. \tag{C.6}$$

Next, we have the inequality chain:

$$\begin{aligned} 2\bigg[\mathbb{E}\bigg\|\sum_{i=1}^{N}\xi_i\mathbf{Y}_i\bigg\|_2^q\bigg]^{1/q} &\leq 2\bigg[\mathbb{E}\bigg\|\sum_{i=1}^{N}\xi_i\mathbf{Y}_i\bigg\|_{S_r}^q\bigg]^{1/q} \\ &= 2\bigg[\mathbb{E}_{\mathbf{Y}_i}\bigg(\mathbb{E}_{\xi_i}\bigg\|\sum_{i=1}^{N}\xi_i\mathbf{Y}_i\bigg\|_{S_r}^q\bigg)\bigg]^{1/q} \\ &\leq 2\bigg[\mathbb{E}_{\mathbf{Y}_i}\bigg(\mathbb{E}_{\xi_i}\bigg\|\sum_{i=1}^{N}\xi_i\mathbf{Y}_i\bigg\|_{S_r}^r\bigg)^{q/r}\bigg]^{1/q}, \end{aligned} \tag{C.7}$$

where the first inequality holds due to $\|\cdot\|_2 \leq \|\cdot\|_{S_r}$, the second inequality holds due to Lyapunov's inequality D.1, where we set s, t in D.1 as $s = q, t = r$. Since $q < r$, then the second inequality holds. Note we have

$$\begin{aligned} 2\bigg[\mathbb{E}_{\mathbf{Y}_i}\bigg(\mathbb{E}_{\xi_i}\bigg\|\sum_{i=1}^N \xi_i \mathbf{Y}_i\bigg\|_{S_r}^r\bigg)^{q/r}\bigg]^{1/q} &\le 2\sqrt{r}\bigg[\mathbb{E}\bigg\|\bigg(\sum_{i=1}^N \mathbf{Y}_i^2\bigg)^{1/2}\bigg\|_{S_r}^q\bigg]^{1/q} \\ &\le 2\sqrt{r}\bigg[\mathbb{E}\bigg(p^{1/r}\bigg\|\bigg(\sum_{i=1}^N \mathbf{Y}_i^2\bigg)^{1/2}\bigg\|_2\bigg)^q\bigg]^{1/q} \\ &\le 2\sqrt{er}\bigg[\mathbb{E}\bigg\|\bigg(\sum_{i=1}^N \mathbf{Y}_i^2\bigg)^{1/2}\bigg\|_2^q\bigg]^{1/q} \\ &= 2\sqrt{er}\bigg[\mathbb{E}\bigg\|\sum_{i=1}^N \mathbf{Y}_i^2\bigg\|_2^{q/2}\bigg]^{1/q}, \end{aligned} \tag{C.8}$$

where the first inequality holds due to Proposition D.2; the second inequality holds because $\|\mathbf{A}\|_{S_r} \le p^{1/r}\|\mathbf{A}\|_2$, where we set $\mathbf{A} = \big(\sum_{i=1}^N \mathbf{Y}_i^2\big)^{1/2}$ and p is the dimension of $\mathbf{A}$; the third inequality holds because $p^{1/r} \le p^{1/(2\log p)} = \sqrt{e}$.

Finally, we use Proposition D.3 to bound (C.8). Since $\mathbf{Y}_i^2$ are independent, random, positive-semidefinite matrices, we can set $\mathbf{W}_i$ in Proposition D.3 as $\mathbf{W}_i = \mathbf{Y}_i^2$. Meanwhile, $q/2 \ge 1$, so we have

$$\bigg[\mathbb{E}\bigg\|\sum_{i=1}^N \mathbf{Y}_i^2\bigg\|_2^{q/2}\bigg]^{2/q} \le \bigg[\bigg\|\sum_{i=1}^N \mathbb{E}\mathbf{Y}_i^2\bigg\|_2^{1/2} + 2\sqrt{er}\big(\mathbb{E}\max_i \|\mathbf{Y}_i\|_2^q\big)^{1/q}\bigg]^2,$$

which immediately implies

$$\bigg[\mathbb{E}\bigg\|\sum_{i=1}^N \mathbf{Y}_i^2\bigg\|_2^{q/2}\bigg]^{1/q} \le \bigg\|\sum_{i=1}^N \mathbb{E}\mathbf{Y}_i^2\bigg\|_2^{1/2} + 2\sqrt{er}\big(\mathbb{E}\max_i \|\mathbf{Y}_i\|_2^q\big)^{1/q}. \tag{C.9}$$

Submitting (C.7), (C.8),(C.9) into (C.6), we have the proof completed. □

C.3 Proof of Lemma B.3

Proof. We have

$$\begin{aligned} \|\nabla F(\mathbf{x}_t^s + \mathbf{h})\|_2 &= \Big\|\nabla F(\mathbf{x}_t^s + \mathbf{h}) - \nabla F(\mathbf{x}_t^s) - \nabla^2 F(\mathbf{x}_t^s)\mathbf{h} + \mathbf{v}_t^s + \mathbf{U}_t^s\mathbf{h} \\ &\quad + \big(\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\big) + \Big(\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\Big)\mathbf{h}\Big\|_2 \\ &\le \big\|\nabla F(\mathbf{x}_t^s + \mathbf{h}) - \nabla F(\mathbf{x}_t^s) - \nabla^2 F(\mathbf{x}_t^s)\mathbf{h}\big\|_2 + \Big\|\mathbf{v}_t^s + \mathbf{U}_t^s\mathbf{h} + \frac{M}{2}\|\mathbf{h}\|_2\mathbf{h}\Big\|_2 \\ &\quad + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2 + \big\|\big(\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big)\mathbf{h}\big\|_2 + \frac{M}{2}\|\mathbf{h}\|_2^2, \end{aligned} \tag{C.10}$$

where the inequality holds due to triangle inequality. In the following, we are going to bound the right hand side of (C.10). For the first term in the right hand side of (C.10), it can be bounded as

$$\big\|\nabla F(\mathbf{x}_t^s + \mathbf{h}) - \nabla F(\mathbf{x}_t^s) - \nabla^2 F(\mathbf{x}_t^s)\mathbf{h}\big\|_2 \le \frac{\rho}{2}\|\mathbf{h}\|_2^2 \le \frac{M}{4}\|\mathbf{h}\|_2^2,$$

where the first inequality holds due to Assumption 4.1 and the second inequality holds due to $2\rho \le M$. For the second term in the the right hand side of (C.10), it equals to

$$\Big\|\mathbf{v}_t^s + \mathbf{U}_t^s\mathbf{h} + \frac{M}{2}\|\mathbf{h}\|_2\mathbf{h}\Big\|_2 = \big\|\nabla m_t^s(\mathbf{h})\big\|_2.$$

And the final term can be bounded as

$$\Big\|\Big(\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\Big)\mathbf{h}\Big\|_2 \le \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2 \cdot \|\mathbf{h}\|_2 \le \frac{1}{M}\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_2^2 + \frac{M}{4}\|\mathbf{h}\|_2^2,$$

where the last inequality is due to Young's inequality. Putting all these bounds together and submit them into (C.10), we have

$$\big\|\nabla F(\mathbf{x}_t^s+\mathbf{h})\big\|_2 \le M\|\mathbf{h}\|_2^2 + \|\nabla F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|_2 + \frac{1}{M}\big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2^2 + \big\|\nabla m_t^s(\mathbf{h})\big\|_2.$$

□

C.4 Proof of Lemma B.4

Proof. We have

$$\begin{aligned}\nabla^2 F(\mathbf{x}_t^s+\mathbf{h}) &\succeq \nabla^2 F(\mathbf{x}_t^s) - \rho\|\mathbf{h}\|_2\mathbf{I}\\ &\succeq \mathbf{U}_t^s - \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2\mathbf{I} - \rho\|\mathbf{h}\|_2\mathbf{I}\\ &\succeq -\frac{M}{2}\|\mathbf{h}_t^s\|_2\mathbf{I} - \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2\mathbf{I} - \rho\|\mathbf{h}\|_2\mathbf{I},\end{aligned}$$

where the first inequality holds because $\nabla^2 F$ is ρ-Hessian Lipschitz, the last inequality holds due to (A.2) in Lemma A.1. Thus we have

$$\begin{aligned}-\lambda_{\min}\big(\nabla^2 F(\mathbf{x}_t^s+\mathbf{h})\big) &\le \frac{M}{2}\|\mathbf{h}_t^s\|_2 + \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2 + \rho\|\mathbf{h}\|_2\\ &= \frac{M}{2}(\|\mathbf{h}_t^s\|_2 - \|\mathbf{h}\|_2) + \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2 + (\rho + M/2)\|\mathbf{h}\|_2\\ &\le M\|\mathbf{h}\|_2 + \big\|\nabla^2 F(\mathbf{x}_t^s) - \mathbf{U}_t^s\big\|_2 + M\big|\|\mathbf{h}_t^s\|_2 - \|\mathbf{h}\|_2\big|,\end{aligned}$$

where the last inequality holds because $\rho \le M/2$. □

D Additional Lemmas and Propositions

Lemma D.1 (Lyapunov's inequality)**.** (Durrett, 2010) For a random variable X, when $0 < s < t$, it holds that

$$(\mathbb{E}|X|^s)^{1/s} \le (\mathbb{E}|X|^t)^{1/t}.$$

We list two propositions about matrix concentration inequality below. As they play key roles in our next analysis, we use them without proof:

Proposition D.2 (Matrix Khintchine inequality)**.** (Mackey et al., 2014) Suppose that $r > 2$. Consider a finite sequence $\{\mathbf{A}_i, 1 \le i \le N\}$ of deterministic, self-adjoint matrices. Then

$$\bigg[\mathbb{E}\bigg\|\sum_{i=1}^N \xi_i\mathbf{A}_i\bigg\|_{S_r}^r\bigg]^{1/r} \le \sqrt{r}\bigg\|\bigg[\sum_{i=1}^N \mathbf{A}_i^2\bigg]^{1/2}\bigg\|_{S_r},$$

where sequence ξ_i consists of independent Rademacher random variables.

Proposition D.3. (Chen et al., 2012) Suppose that $q \ge 1$, and fix $r \ge \max\{q, 2\log p\}$. Consider $\mathbf{W}_1, ..., \mathbf{W}_N$ of independent, random, positive-definite matrices with dimension $p \times p$. Then

$$\bigg[\mathbb{E}\bigg\|\sum_{i=1}^N \mathbf{W}_i\bigg\|_2^q\bigg]^{1/q} \le \bigg[\bigg\|\sum_{i=1}^N \mathbb{E}\mathbf{W}_i\bigg\|_2^{1/2} + 2\sqrt{er}\big(\mathbb{E}\max_i\|\mathbf{W}_i\|_2^q\big)^{1/(2q)}\bigg]^2.$$