Supplementary Materials

A Proof of Convergence

A.1 Lemmas

In this subsection, we introduce two useful lemmas, which will be used in the proof of convergence.

Lemma 8 (Nesterov and Polyak (2006), Lemma 1)**.** *Let the Hessian $\nabla^2 f(\cdot)$ of the function $f(\cdot)$ be L-Lipschitz continuous with $L > 0$. Then, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have*

$$\left\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})\right\| \leqslant \frac{L}{2}\left\|\mathbf{y}-\mathbf{x}\right\|^2, \tag{29}$$

$$\left| f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) - \frac{1}{2}(\mathbf{y}-\mathbf{x})^T\nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})\right| \leqslant \frac{L}{6}\left\|\mathbf{y}-\mathbf{x}\right\|^3. \tag{30}$$

Lemma 9 (Wang et al. (2019), Lemma 3)**.** *Let $M \in \mathbb{R}, \mathbf{g} \in \mathbb{R}^d, \mathbf{H} \in \mathbb{S}^{d\times d}$, and*

$$\mathbf{s} = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d} \mathbf{g}^\top\mathbf{u} + \frac{1}{2}\mathbf{u}^\top\mathbf{H}\mathbf{u} + \frac{M}{6}\left\|\mathbf{u}\right\|^3. \tag{31}$$

Then, the following statements hold:

$$\mathbf{g} + \mathbf{H}\mathbf{s} + \frac{M}{2}\left\|\mathbf{s}\right\|\mathbf{s} = \mathbf{0}, \tag{32}$$

$$\mathbf{H} + \frac{M}{2}\left\|\mathbf{s}\right\|\mathbf{I} \succcurlyeq \mathbf{0}, \tag{33}$$

$$\mathbf{g}^\top\mathbf{s} + \frac{1}{2}\mathbf{s}^\top\mathbf{H}\mathbf{s} + \frac{M}{6}\left\|\mathbf{s}\right\|^3 \leqslant -\frac{M}{12}\left\|\mathbf{s}\right\|^3. \tag{34}$$

A.2 Proof of Theorem 1

Proof. Since $\nabla^2 f(\mathbf{x})$ is L_2-Lipschitz, thus we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\overset{\text{(i)}}{\leqslant} \nabla f(\mathbf{x}_k)^\top\mathbf{s}_{k+1} + \frac{1}{2}\mathbf{s}_{k+1}^\top\nabla f(\mathbf{x}_k)\mathbf{s}_{k+1} + \frac{L_2}{6}\left\|\mathbf{s}_{k+1}\right\|^3 \\ &\leqslant \mathbf{g}_k^\top\mathbf{s}_{k+1} + \frac{1}{2}\mathbf{s}_{k+1}^\top\mathbf{H}_k\mathbf{s}_{k+1} + \frac{M}{6}\left\|\mathbf{s}_{k+1}\right\|^3 + (\nabla f(\mathbf{x}_k) - \mathbf{g}_k)^\top\mathbf{s}_{k+1} \\ &\quad + \frac{L_2 - M}{6}\left\|\mathbf{s}_{k+1}\right\|^3 + \frac{1}{2}\mathbf{s}_{k+1}^\top(\nabla^2 f(\mathbf{x}_k) - \mathbf{H}_k)\mathbf{s}_{k+1} \\ &\overset{\text{(ii)}}{\leqslant} -\frac{3M - 2L_2}{12}\left\|\mathbf{s}_{k+1}\right\|^3 + (\nabla f(\mathbf{x}_k) - \mathbf{g}_k)^\top\mathbf{s}_{k+1} + \frac{1}{2}\mathbf{s}_{k+1}^\top(\nabla f(\mathbf{x}_k) - \mathbf{H}_k)\mathbf{s}_{k+1} \end{aligned} \tag{35}$$

where (i) follows from Lemma 8 with $\mathbf{y} = \mathbf{x}_{k+1}, \mathbf{x} = \mathbf{x}_k$ and $\mathbf{s}_{k+1} = \mathbf{x}_{k+1} - \mathbf{x}_k$, (ii) follows from eq. (34) in Lemma 9 with $\mathbf{g} = \mathbf{g}_k, \mathbf{H} = \mathbf{H}_k$ and $\mathbf{s} = \mathbf{s}_{k+1}$.

Next, we bound the terms $(\nabla f(\mathbf{x}_k) - \mathbf{g}_k)^\top\mathbf{s}_{k+1}$ and $\mathbf{s}_{k+1}^\top(\nabla f(\mathbf{x}_k) - \mathbf{H}_k)\mathbf{s}_{k+1}$. For the first term, we have that

$$\begin{aligned} (\nabla f(\mathbf{x}_k) - \mathbf{g}_k)^\top\mathbf{s}_{k+1} \leqslant \left\|\nabla f(\mathbf{x}_k) - \mathbf{g}_k\right\|\left\|\mathbf{s}_{k+1}\right\| &\overset{\text{(i)}}{\leqslant} \beta\left(\left\|\mathbf{s}_k\right\|^2 + \epsilon_1^2\right)\left\|\mathbf{s}_{k+1}\right\| = \beta\left(\|\mathbf{s}_k\|^2\|\mathbf{s}_{k+1}\| + \epsilon_1^2\|\mathbf{s}_{k+1}\|\right) \\ &\overset{\text{(ii)}}{\leqslant} \beta\left(\|\mathbf{s}_k\|^3 + \|\mathbf{s}_{k+1}\|^3 + \epsilon_1^3 + \|\mathbf{s}_{k+1}\|^3\right) = \beta\left(\|\mathbf{s}_k\|^3 + 2\|\mathbf{s}_{k+1}\|^3 + \epsilon_1^3\right), \end{aligned} \tag{36}$$

where (i) follows from Assumption 2, which gives that $\left\|\mathbf{g}_k - \nabla F(\mathbf{x}_k)\right\| \leqslant \beta\max\left\{\left\|\mathbf{s}_k\right\|^2, \epsilon_1^2\right\}$, and (ii) follows from the inequality that for $a, b \in \mathbb{R}^+$, $a^2 b \leqslant a^3 + b^3$, which can be verified by checking the cases with $a < b$ and $a \geqslant b$, respectively. Similarly, we obtain that

$$\mathbf{s}_{k+1}^\top(\nabla f(\mathbf{x}_k) - \mathbf{H}_k)\mathbf{s}_{k+1} \leqslant \left\|\nabla^2 f(\mathbf{x}_k) - \mathbf{H}_k\right\|\left\|\mathbf{s}_{k+1}\right\|^2 \overset{\text{(i)}}{\leqslant} \alpha\left(\left\|\mathbf{s}_k\right\| + \epsilon_1\right)\left\|\mathbf{s}_{k+1}\right\|^2 = \alpha\left(\|\mathbf{s}_k\|\|\mathbf{s}_{k+1}\|^2 + \epsilon_1\|\mathbf{s}_{k+1}\|^2\right)$$

$$\overset{\text{(ii)}}{\leqslant} \alpha\left(\|\mathbf{s}_k\|^3 + \|\mathbf{s}_{k+1}\|^3 + \epsilon_1^3 + \|\mathbf{s}_{k+1}\|^3\right) = \alpha\left(\|\mathbf{s}_k\|^3 + 2\|\mathbf{s}_{k+1}\|^3 + \epsilon_1^3\right), \tag{37}$$

where (i) follows from Assumption 2, which gives that $\left\|\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k)\right\| \leqslant \alpha \max\left\{\|\mathbf{s}_k\|, \epsilon_1\right\}$, and (ii) follows from the inequality that for $a, b \in \mathbb{R}^+$, $a^2 b \leqslant a^3 + b^3$.

Plugging eqs. (36) and (37) into eq. (35) yields

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leqslant -\frac{3M - 2L_2}{12}\|\mathbf{s}_{k+1}\|^3 + \beta\left(\|\mathbf{s}_k\|^3 + 2\|\mathbf{s}_{k+1}\|^3 + \epsilon_1^3\right) + \frac{\alpha}{2}\left(\|\mathbf{s}_k\|^3 + 2\|\mathbf{s}_{k+1}\|^3 + \epsilon_1^3\right) \\ &= -\left(\frac{3M - 2L_2}{12} - 2\beta - \alpha\right)\|\mathbf{s}_{k+1}\|^3 + \left(\beta + \frac{\alpha}{2}\right)\|\mathbf{s}_k\|^3 + \left(\beta + \frac{\alpha}{2}\right)\epsilon_1^3 \end{aligned} \tag{38}$$

Summing Equation (38) for 0 to k, we obtain

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_0) &\leqslant -\left(\frac{3M - 2L_2}{12} - 2\beta - \alpha\right)\sum_{i=1}^{k+1}\|\mathbf{s}_i\|^3 + \left(\beta + \frac{\alpha}{2}\right)\sum_{i=0}^{k}\|\mathbf{s}_i\|^3 + \left(\beta + \frac{\alpha}{2}\right)\sum_{i=0}^{k}\epsilon_1^3 \\ &\leqslant -\left(\frac{3M - 2L_2}{12} - 2\beta - \alpha\right)\sum_{i=1}^{k+1}\|\mathbf{s}_i\|^3 + \left(\beta + \frac{\alpha}{2}\right)\sum_{i=0}^{k+1}\|\mathbf{s}_i\|^3 + \left(\beta + \frac{\alpha}{2}\right)\sum_{i=0}^{k}\epsilon_1^3 \\ &\leqslant -\left(\frac{3M - 2L_2}{12} - 3\beta - \frac{3}{2}\alpha\right)\sum_{i=1}^{k+1}\|\mathbf{s}_i\|^3 + \left(\beta + \frac{\alpha}{2}\right)\|\mathbf{s}_0\|^3 + \left(\beta + \frac{\alpha}{2}\right)\sum_{i=0}^{k}\epsilon_1^3, \end{aligned} \tag{39}$$

We next note that

$$\sum_{i=1}^{k+1}\|\mathbf{s}_i\|^3 = \frac{1}{2}\left(\sum_{i=1}^{k+1}\|\mathbf{s}_i\|^3 + \sum_{i=1}^{k+1}\|\mathbf{s}_i\|^3\right) = \frac{1}{2}\left(\sum_{i=1}^{k+1}\|\mathbf{s}_i\|^3 + \sum_{i=0}^{k}\|\mathbf{s}_{i+1}\|^3\right) \geqslant \frac{1}{2}\sum_{i=1}^{k}\left(\|\mathbf{s}_i\|^3 + \|\mathbf{s}_{i+1}\|^3\right). \tag{40}$$

Plugging eq. (40) into eq. (39) yields that

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_0) &\leqslant -\sum_{i=1}^{k}\left(\frac{3M - 2L_2}{24} - \frac{3}{2}\beta - \frac{3}{4}\alpha\right)\left(\|\mathbf{s}_i\|^3 + \|\mathbf{s}_{i+1}\|^3\right) + \left(\beta + \frac{\alpha}{2}\right)\|\mathbf{s}_0\|^3 + \left(\beta + \frac{\alpha}{2}\right)\sum_{i=0}^{k}\epsilon_1^3 \\ &\overset{\text{(i)}}{\leqslant} -\sum_{i=1}^{k}\left(\frac{3M - 2L_2}{24} - \frac{5}{2}\beta - \frac{5}{4}\alpha\right)\left(\|\mathbf{s}_i\|^3 + \|\mathbf{s}_{i+1}\|^3\right) + \left(\beta + \frac{\alpha}{2}\right)\|\mathbf{s}_0\|^3 + \left(\beta + \frac{\alpha}{2}\right)\epsilon_1^3, \end{aligned}$$

where (i) follows from the fact that before the algorithm terminates we always have that $\|s_i\| \geqslant \epsilon_1$ or $\|s_{i+1}\| \geqslant \epsilon_1$, which gives that $\|s_i\|^3 + \|s_{i+1}\|^3 \geqslant \epsilon_1^3$. Therefore, we have

$$\begin{aligned} \sum_{i=1}^{k}\left(\frac{3M - 2L_2}{24} - \frac{5}{2}\beta - \frac{5}{4}\alpha\right)\left(\|\mathbf{s}_i\|^3 + \|\mathbf{s}_{i+1}\|^3\right) &\leqslant f(\mathbf{x}_0) - f^* + \left(\beta + \frac{\alpha}{2}\right)\|\mathbf{s}_0\|^3 + \left(\beta + \frac{\alpha}{2}\right)\epsilon_1^3 \\ &\overset{\text{(i)}}{=} f(\mathbf{x}_0) - f^* + (2\beta + \alpha)\,\epsilon_1^3 \end{aligned} \tag{41}$$

where (i) follows from the fact that $\|\mathbf{s}_0\| = \epsilon_1$. Thus, if the algorithm never terminates, then we always have that $\|s_i\| \geqslant \epsilon_1$ or $\|s_{i+1}\| \geqslant \epsilon_1$, which gives $\|s_i\|^3 + \|s_{i+1}\|^3 \geqslant \epsilon_1^3$. Following from Equation (41), we obtain that

$$k \times \gamma\epsilon_1^3 \leqslant f(\mathbf{x}_0) - f^* + (2\beta + \alpha)\,\epsilon_1^3, \tag{42}$$

where $\gamma \triangleq \left(\frac{3M-2L_2}{24} - \frac{5}{2}\beta - \frac{5}{4}\alpha\right)$. Therefore, we obtain

$$k \leqslant \frac{f(\mathbf{x}_0) - f^* + (2\beta + \alpha)\,\epsilon_1^3}{\gamma\epsilon_1^3}, \tag{43}$$

which shows that the algorithm must terminates if the total number of iterations exceeds $O(\epsilon_1^{-3})$. With the choice of ϵ_1 in Theorem 1 , we obtain that the algorithm terminates at most with total iteration $k = O(\epsilon^{-3/2})$.

Suppose that the algorithm terminates at iteration k, then according to the analysis in eq. (41), we have that

$$\sum_{i=1}^{k-1} \gamma \left(\|\mathbf{s}_i\|^3 + \|\mathbf{s}_{i+1}\|^3 \right) \leqslant f(\mathbf{x}_0) - f^* + (2\beta + \alpha)\,\epsilon_1^3. \tag{44}$$

On the other hand, according to eq. (44) and the terminal condition that $\|s_i\| \leqslant \epsilon_1$ and $\|s_{i+1}\| \leqslant \epsilon_1$, we obtain

$$\sum_{i=1}^{k} \gamma \left(\|\mathbf{s}_i\|^3 + \|\mathbf{s}_{i+1}\|^3 \right) \leqslant f(\mathbf{x}_0) - f^* + (2\beta + \alpha + 2\gamma)\,\epsilon_1^3,$$

which gives that

$$\sum_{i=1}^{k+1} \|\mathbf{s}_i\|^3 \leqslant \frac{f(\mathbf{x}_0) - f^* + (2\beta + \alpha + 2\gamma)\,\epsilon_1^3}{\gamma}. \tag{45}$$

We next consider the convergence of $\|\nabla f(x_k)\|$ and $\left\|\nabla^2 f(x_k)\right\|$. Next, we prove the convergence rate of $\nabla f(\cdot)$ and $\nabla^2 f(\cdot)$. We first derive

$$\begin{aligned}
\|\nabla f(\mathbf{x}_{k+1})\| &\overset{\text{(i)}}{=} \left\| \nabla f(\mathbf{x}_{k+1}) - \left(\mathbf{g}_k + \mathbf{H}_k \mathbf{s}_{k+1} + \frac{M}{2} \|\mathbf{s}_{k+1}\| \, \mathbf{s}_{k+1} \right) \right\| \\
&\leqslant \|\nabla f(\mathbf{x}_{k+1}) - (\mathbf{g}_k + \mathbf{H}_k \mathbf{s}_{k+1})\| + \frac{M}{2} \|\mathbf{s}_{k+1}\|^2 \\
&\leqslant \left\| \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) - \nabla^2 f(\mathbf{x}_k) \mathbf{s}_{k+1} \right\| + \|\nabla f(\mathbf{x}_k) - \mathbf{g}_k\| + \left\| (\nabla^2 f(\mathbf{x}_k) - \mathbf{H}_k) \mathbf{s}_{k+1} \right\| + \frac{M}{2} \|\mathbf{s}_{k+1}\|^2 \\
&\overset{\text{(ii)}}{\leqslant} \frac{L_2}{2} \|\mathbf{s}_{k+1}\|^2 + \beta(\|\mathbf{s}_k\|^2 + \epsilon_1^2) + \alpha(\|\mathbf{s}_k\| + \epsilon_1) \|\mathbf{s}_{k+1}\| + \frac{M}{2} \|\mathbf{s}_{k+1}\|^2 \\
&\overset{\text{(iii)}}{\leqslant} \left(\frac{L+M}{2} + 2\beta + 2\alpha \right) \epsilon_1^2 \overset{\text{(iv)}}{\leqslant} \epsilon,
\end{aligned}$$

where (i) follows from eq. (32) with $\mathbf{g} = \mathbf{g}_k, \mathbf{H} = \mathbf{H}_k$ and $\mathbf{s} = \mathbf{s}_{k+1}$, (ii) follows from eq. (29) in Lemma 8 and Assumption 2, (iii) follows from the terminal condition of the algorithm, and (iv) follows from eq. (10).

Similarly, we have

$$\begin{aligned}
\nabla^2 f(\mathbf{x}_{k+1}) &\overset{(i)}{\succcurlyeq} \mathbf{H}_k - \left\| \mathbf{H}_k - \nabla^2 f(\mathbf{x}_{k+1}) \right\| \mathbf{I} \\
&\overset{(ii)}{\succcurlyeq} -\frac{M}{2} \|\mathbf{s}_{k+1}\| \mathbf{I} - \left\| \mathbf{H}_k - \nabla^2 f(\mathbf{x}_{k+1}) \right\| \mathbf{I} \\
&\succcurlyeq -\frac{M}{2} \|\mathbf{s}_{k+1}\| \mathbf{I} - \left\| \mathbf{H}_k - \nabla^2 f(\mathbf{x}_k) \right\| \mathbf{I} - \left\| \nabla^2 f(\mathbf{x}_k) - \nabla^2 f(\mathbf{x}_{m+1}) \right\| \mathbf{I} \\
&\overset{(iii)}{\succcurlyeq} -\frac{M}{2} \|\mathbf{s}_{k+1}\| \mathbf{I} - \alpha(\|\mathbf{s}_k\| + \epsilon_1)\mathbf{I} - L_2 \|\mathbf{s}_{k+1}\| \mathbf{I} \\
&\overset{(iv)}{\succcurlyeq} -\left(\frac{M + 2L_2}{2} + 2\alpha \right) \epsilon_1 \mathbf{I} \overset{(v)}{\succcurlyeq} \epsilon \mathbf{I},
\end{aligned}$$

where (i) follows from Weyl's inequality, (ii) follows from eq. (33) with $\mathbf{H} = \mathbf{H}_m$ and $\mathbf{s} = \mathbf{s}_{m+1}$, (iii) follows from Assumption 2 and the fact that $\nabla^2 f(\cdot)$ is L_2-Lipschitz, (iv) follows from the terminal condition of the algorithm, and (v) follows from eq. (10). □

B Proofs for SVRC under Sampling with Replacement

B.1 Proof of Theorem 2

The idea of the proof is to apply the following matrix Bernstein inequality Tropp (2012) for sampling with replacement to characterize the sample complexity in order to satisfy the inexactness condition $\left\| \mathbf{H}_k - \nabla^2 F(\mathbf{x}_k) \right\| \leqslant \alpha \max\{\|\mathbf{s}_k\|, \epsilon_1\}$ with the probability at least $1 - \zeta$.

Lemma 10 (Matrix Bernstein Inequality)**.** *Consider a finite sequence* $\{\mathbf{X}_k\}$ *of independent, random matrices with dimensions* $d_1 \times d_2$*. Assume that each random matrix satisfies*

$$\mathbb{E}\mathbf{X}_k = \mathbf{0} \quad \textit{and} \quad \|\mathbf{X}_k\| \leqslant R \quad \textit{almost surely.}$$

Define

$$\sigma^2 \triangleq \max\left(\left\|\sum\nolimits_k \mathbb{E}(\mathbf{X}_k\mathbf{X}_k^*)\right\|, \left\|\sum\nolimits_k \mathbb{E}(\mathbf{X}_k^*\mathbf{X}_k)\right\|\right). \tag{46}$$

Then, for all $\epsilon \geqslant 0$*,*

$$P\left(\left\|\sum\nolimits_k \mathbf{X}_k\right\| \geqslant \epsilon\right) \leqslant 2(d_1 + d_2)\exp\left(-\frac{\epsilon^2/2}{\sigma^2 + R\epsilon/3}\right).$$

Let $\xi_H(k)$ be the collection of index that uniformly picked from $1, \cdots, N$ with replacement, and $\mathbf{X}_i$ be

$$\mathbf{X}_i = \frac{1}{|\xi_H(k)|}\left(\nabla^2 f_i(\mathbf{x}_k) - \nabla^2 f_i(\tilde{\mathbf{x}}) + \nabla^2 F(\tilde{\mathbf{x}}) - \nabla^2 F(\mathbf{x}_k)\right),$$

then we have

$$\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k) = \sum_{i \in \xi_H(k)} \mathbf{X}_i. \tag{47}$$

Moreover, we have $\mathbb{E}\mathbf{X}_i = \mathbf{0}$, and

$$\begin{aligned} R \triangleq \|\mathbf{X}_i\| &= \frac{1}{|\xi_H(k)|}\left\|\nabla^2 f_{\xi_i}(\mathbf{x}_k) - \nabla^2 f_{\xi_i}(\tilde{\mathbf{x}}) + \nabla^2 F(\tilde{\mathbf{x}}) - \nabla^2 F(\mathbf{x}_k)\right\| \\ &\overset{\text{(i)}}{\leqslant} \frac{2L_2}{|\xi_H(k)|}\left\|\mathbf{x}_k - \tilde{\mathbf{x}}\right\|, \end{aligned} \tag{48}$$

where (i) follows because $\nabla^2 f_i(\cdot)$ is L_2 Lipschitz, for $1 \leqslant i \leqslant N$.

The variance also can be bounded by

$$\begin{aligned} \sigma^2 &\triangleq \max\left(\left\|\sum\nolimits_{k \in \xi_H(k)} \mathbb{E}(\mathbf{X}_k\mathbf{X}_k^*)\right\|, \left\|\sum\nolimits_{k \in \xi_H(k)} \mathbb{E}(\mathbf{X}_k^*\mathbf{X}_k)\right\|\right) \\ &\overset{\text{(i)}}{\leqslant} \left\|\sum\nolimits_{k \in \xi_H(k)} \mathbb{E}(\mathbf{X}_k^2)\right\| \overset{\text{(ii)}}{\leqslant} \sum\nolimits_{k \in \xi_H(k)} \mathbb{E}\left\|\mathbf{X}_k^2\right\| \leqslant \sum\nolimits_{k \in \xi_H(k)} \mathbb{E}\left\|\mathbf{X}_k\right\|^2 \\ &\overset{\text{(ii)}}{\leqslant} \frac{4L_2^2}{|\xi_H(k)|}\left\|\mathbf{x}_k - \tilde{\mathbf{x}}\right\|^2 \end{aligned} \tag{49}$$

where (i) follows from the fact that $\mathbf{X}_k$ is real and symmetric, (ii) follows from Jensen's inequality, and (iii) follows from eq. (48).

Therefore, in order to satisfy $\left\|\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k)\right\| \leqslant \alpha \max\{\|\mathbf{s}_k\|, \epsilon_1\}$ with probability at least $1-\zeta$, by eq. (47), it is equivalent to require $\left\|\sum_{i \in \xi_H(k)} \mathbf{X}_i\right\| \leqslant \alpha \max\{\|\mathbf{s}_k\|, \epsilon_1\}$ with probability at least $1-\zeta$. We now apply Lemma 10 for $\mathbf{X}_i$, and it is sufficient to have:

$$2(d_1 + d_2)\exp\left(\frac{-\epsilon^2/2}{\sigma^2 + R\epsilon/3}\right) \leqslant \zeta$$

which is equivalent to have

$$\frac{1}{\sigma^2 + R\epsilon/3} \geqslant \frac{2}{\epsilon^2}\log\left(\frac{2(d_1 + d_2)}{\zeta}\right). \tag{50}$$

Plugging eqs. (48) and (49) into eq. (50) yields

$$\frac{1}{\frac{4L_2^2}{|\xi_H(k)|}\left\|\mathbf{x}_k-\tilde{\mathbf{x}}\right\|^2+\frac{2L_2}{|\xi_H(k)|}\left\|\mathbf{x}_k-\tilde{\mathbf{x}}\right\|\epsilon/3}\geqslant\frac{2}{\epsilon^2}\log\left(\frac{4d}{\zeta}\right),$$

which gives

$$|\xi_H(k)|\geqslant\left(\frac{8L_2^2}{\epsilon^2}\left\|\mathbf{x}_k-\tilde{\mathbf{x}}\right\|^2+\frac{4L_2}{3\epsilon}\left\|\mathbf{x}_k-\tilde{\mathbf{x}}\right\|\right)\log\left(\frac{4d}{\zeta}\right). \tag{51}$$

Substituting $\epsilon=\alpha\max\{\|\mathbf{s}_k\|,\epsilon_1\}$, we obtain the required sample size to be bounded by

$$|\xi_H(k)|\geqslant\left(\frac{8L_2^2}{\alpha^2\max\{\|\mathbf{s}_k\|^2,\epsilon_1^2\}}\left\|\mathbf{x}_k-\tilde{\mathbf{x}}\right\|^2+\frac{4L_2}{3\alpha\max\{\|\mathbf{s}_k\|,\epsilon_1\}}\left\|\mathbf{x}_k-\tilde{\mathbf{x}}\right\|\right)\log\left(\frac{4d}{\zeta}\right). \tag{52}$$

We next bound the sample size $|\xi_g(k)|$ for the gradient in the similar procedure. We first define $\mathbf{X}_i\in\mathbb{R}^{d\times 1}$ as

$$\mathbf{X}_i=\frac{1}{|\xi_g(k)|}\left(\nabla f_{\xi_i}(\mathbf{x}_k)-\nabla f_{\xi_i}(\tilde{\mathbf{x}})+\nabla F(\tilde{\mathbf{x}})-\nabla F(\mathbf{x}_k)\right), \tag{53}$$

then we have

$$\mathbf{g}_k-\nabla f(\mathbf{x}_k)=\sum_{i\in\xi_g(k)}\mathbf{X}_i \tag{54}$$

Furthermore,

$$R=\|\mathbf{X}_i\|=\frac{1}{|\xi_g(k)|}\left\|\nabla f_{\xi_i}(\mathbf{x}_k)-\nabla f_{\xi_i}(\tilde{\mathbf{x}})+\nabla F(\tilde{\mathbf{x}})-\nabla F(\mathbf{x}_k)\right\|\overset{\text{(i)}}{\leqslant}\frac{2L_1}{|S_{g,k}|}\left\|\mathbf{x}_k-\tilde{\mathbf{x}}\right\|, \tag{55}$$

where (i) follows because $\nabla f_i(\cdot)$ is L_1 Lipschitz, for $i=1,\dots,N$, and

$$\begin{aligned}\sigma^2\triangleq\max\left(\left\|\sum\nolimits_{k\in\xi_g(k)}\mathbb{E}(\mathbf{X}_k\mathbf{X}_k^*)\right\|,\left\|\sum\nolimits_{k\in\xi_g(k)}\mathbb{E}(\mathbf{X}_k^*\mathbf{X}_k)\right\|\right)\leqslant\sum\nolimits_{k\in\xi_H(k)}\mathbb{E}\left\|\mathbf{X}_k\right\|^2\\ \overset{\text{(ii)}}{\leqslant}\frac{4L_1^2}{|\xi_g(k)|}\left\|\mathbf{x}_k-\tilde{\mathbf{x}}\right\|^2\end{aligned}$$

In order to satisfy $\|\mathbf{g}_k-\nabla F(\mathbf{x}_k)\|\leqslant\beta\max\left\{\|\mathbf{s}_k\|^2,\epsilon_1^2\right\}$ with the probability at least $1-\zeta$, by eq. (54), it is equivalent to require $\left\|\sum_{i\in\xi_g(k)|}\mathbf{X}_i\right\|\leqslant\beta\max\left\{\|\mathbf{s}_k\|^2,\epsilon_1^2\right\}$ with the probability at least $1-\zeta$. We then apply Lemma 10 for $\mathbf{X}_i$ in the way similar to that for bounding the sample size for Hessian, with $R=\frac{2L_1}{|S_{g,k}|}\|\mathbf{x}_k-\tilde{\mathbf{x}}\|$, $\epsilon=\beta\max\left\{\|\mathbf{s}_k\|^2,\epsilon_1^2\right\}$, and $\sigma^2=\frac{4L_1^2}{|\xi_g(k)|}\|\mathbf{x}_k-\tilde{\mathbf{x}}\|^2$, and obtain the required sample size to satisfy

$$|\xi_g(k)|\geqslant\left(\frac{8L_1^2}{\beta^2\max\{\|\mathbf{s}_k\|^4,\epsilon_1^4\}}\left\|\mathbf{x}_k-\tilde{\mathbf{x}}\right\|^2+\frac{4L_1}{3\beta\max\{\|\mathbf{s}_k\|^2,\epsilon_1^2\}}\left\|\mathbf{x}_k-\tilde{\mathbf{x}}\right\|\right)\log\left(\frac{2(d+1)}{\zeta}\right). \tag{56}$$

B.2 Proof of Theorem 3

First, by eq. (13), we have

$$\sum_{i=1}^{k+1}\left\|\mathbf{x}_i-\mathbf{x}_{i-1}\right\|^3\leqslant C. \tag{57}$$

We then derive

$$\sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\left\|\mathbf{x}_{i\cdot m+j}-\mathbf{x}_{i\cdot m}\right\|^2$$

$$
\begin{aligned}
&\leqslant \sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\Big(\|\mathbf{x}_{i\cdot m+j}-\mathbf{x}_{i\cdot m+j-1}\|+\cdots+\|\mathbf{x}_{i\cdot m+1}-\mathbf{x}_{i\cdot m}\|\Big)^2\\
&\leqslant \sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\Big(\|\mathbf{x}_{i\cdot m+m-1}-\mathbf{x}_{i\cdot m+m-2}\|+\cdots+\|\mathbf{x}_{i\cdot m+1}-\mathbf{x}_{i\cdot m}\|\Big)^2\\
&= \sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\Big(\sum_{l=1}^{m-1}\|\mathbf{x}_{i\cdot m+l}-\mathbf{x}_{i\cdot m+l-1}\|\Big)^2 \overset{\text{(i)}}{\leqslant} \sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1} m\sum_{l=1}^{m-1}\|\mathbf{x}_{i\cdot m+l}-\mathbf{x}_{i\cdot m+l-1}\|^2\\
&\overset{\text{(ii)}}{\leqslant} m^2\sum_{i=0}^{k/m-1}\sum_{l=1}^{m-1}\|\mathbf{x}_{i\cdot m+l}-\mathbf{x}_{i\cdot m+l-1}\|^2 \leqslant m^2\sum_{i=1}^{k}\|\mathbf{x}_i-\mathbf{x}_{i-1}\|^2\\
&\overset{\text{(iii)}}{\leqslant} m^2k^{1/3}\big(\sum_{i=1}^{k}\|\mathbf{x}_i-\mathbf{x}_{i-1}\|^3\big)^{2/3} \overset{\text{(iv)}}{\leqslant} m^2k^{1/3}C^{2/3},
\end{aligned}
\tag{58}
$$

where (i) follows from the Cauthy-Schwaz inequality (ii) follows because j is not a variable in the inner summation, (iii) follows from Holder's inequality, and (iv) follows from eq. (57).

Similarly, we have that

$$
\begin{aligned}
\sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\|\mathbf{x}_{i\cdot m+j}-\mathbf{x}_{i\cdot m}\| &\leqslant \sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\Big(\|\mathbf{x}_{i\cdot m+j}-\mathbf{x}_{i\cdot m+j-1}\|+\cdots+\|\mathbf{x}_{i\cdot m+1}-\mathbf{x}_{i\cdot m}\|\Big)\\
&\leqslant \sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\Big(\|\mathbf{x}_{i\cdot m+m-1}-\mathbf{x}_{i\cdot m+m-2}\|+\cdots+\|\mathbf{x}_{i\cdot m+1}-\mathbf{x}_{i\cdot m}\|\Big)\\
&= \sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\Big(\sum_{l=1}^{m-1}\|\mathbf{x}_{i\cdot m+l}-\mathbf{x}_{i\cdot m+l-1}\|\Big) \overset{\text{(i)}}{\leqslant} m\sum_{i=0}^{k/m-1}\sum_{l=1}^{m-1}\|\mathbf{x}_{i\cdot m+l}-\mathbf{x}_{i\cdot m+l-1}\|\\
&\leqslant m\sum_{i=1}^{k}\|\mathbf{x}_i-\mathbf{x}_{i-1}\| \overset{\text{(ii)}}{\leqslant} mk^{2/3}\Big(\sum_{i=1}^{k}\|\mathbf{x}_i-\mathbf{x}_{i-1}\|^3\Big)^{1/3} \overset{\text{(iii)}}{\leqslant} mk^{2/3}C^{1/3},
\end{aligned}
\tag{59}
$$

where (i) follows because j is not a variable in the inner summation, (ii) follows from Holder's inequality, and (iii) follows from eq. (57).

Thus, the total sample size for Hessian is given by

$$
\begin{aligned}
&m+\frac{kN}{m}+\sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}|\xi_H(k)|\\
&\quad\overset{\text{(i)}}{\leqslant} \frac{CkN}{m}+\sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\left(\frac{8L_2^2}{\alpha^2\max\{\|\mathbf{s}_k\|^2,\epsilon_1^2\}}\|\mathbf{x}_{i\cdot m+j}-\mathbf{x}_{i\cdot m}\|^2+\frac{4L_2}{3\alpha\max\{\|\mathbf{s}_k\|,\epsilon_1\}}\|\mathbf{x}_{i\cdot m+j}-\mathbf{x}_{i\cdot m}\|\right)\log\left(\frac{4d}{\zeta}\right)\\
&\quad\leqslant \frac{CkN}{m}+\sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\left(\frac{8L_2^2}{\alpha^2\epsilon_1^2}\|\mathbf{x}_{i\cdot m+j}-\mathbf{x}_{i\cdot m}\|^2+\frac{4L_2}{3\alpha\epsilon_1}\|\mathbf{x}_{i\cdot m+j}-\mathbf{x}_{i\cdot m}\|\right)\log\left(\frac{4d}{\zeta}\right)\\
&\quad\overset{\text{(ii)}}{\leqslant} \frac{CkN}{m}+\left(\frac{8L_2^2}{\alpha^2\epsilon_1^2}m^2k^{1/3}C^{2/3}+\frac{4L_2}{3\alpha\epsilon_1}mk^{2/3}C^{1/3}\right)\log\left(\frac{4d}{\zeta}\right)\\
&\quad\overset{\text{(iii)}}{\leqslant} \log\left(\frac{4d}{\zeta}\right)\left(\frac{N}{m\epsilon^{3/2}}+\frac{C}{\epsilon^{3/2}}m^2+\frac{C}{\epsilon^{3/2}}m\right)=\log\left(\frac{4d}{\zeta}\right)\frac{C}{\epsilon^{3/2}}\left(\frac{N}{m}+m^2\right)
\end{aligned}
$$

where (i) follows form Theorem 2, and (ii) follows form eqs. (58) and (59), (iii) follows from the fact that $\zeta \leqslant 1$ and $d \geqslant 1$ which gives $\log\left(\frac{4d}{\zeta}\right) > 1$, and $\epsilon_1 = O(\epsilon^{1/2})$ such that $k = O(\epsilon^{-3/2})$ according to Theorem 1

We minimize the above bound over m, substitute the minimizer $m^\star = N^{1/3}$, and obtain

$$\sum_{i=0}^{k} |\xi_H(k)| \leqslant \frac{CN^{2/3}}{\epsilon^{3/2}} \log\left(\frac{4d}{\zeta}\right).$$

Next, according to Theorem 2, Assumption 2 is satisfies with probability at least $1-\zeta$ for gradient and $1-\zeta$ for Hessian . Thus, according to the union bound, the probability of a failure satisfaction per iteration is at most 2ζ. Then, for k iteration, the probability of failure satisfaction of Assumption 2 is at most $2k\zeta$ according to the union bound. To obtain Assumption 2 holds for the total k iteration with probability least $1-\delta$, we require

$$1 - 2k\zeta \geqslant 1 - \delta,$$

which yields

$$\zeta \leqslant \frac{\delta}{2k}.$$

Thus, with probability $1-\delta$, the algorithms successfully outputs an ϵ approximated second-order stationary point, with the total Hessian sample complexity is bounded by

$$\sum_{i=0}^{k} |\xi_H(k)| \leqslant \frac{CN^{2/3}}{\epsilon^{3/2}} \log\left(\frac{8d}{\epsilon^{3/2}\delta}\right) \leqslant \frac{CN^{2/3}}{\epsilon^{3/2}} \log\left(\frac{8d}{\epsilon\delta}\right). \tag{60}$$

which gives

$$\sum_{i=0}^{k} |\xi_H(k)| = \tilde{O}\left(\frac{N^{2/3}}{\epsilon^{3/2}}\right). \tag{61}$$

C Proof of Concentration Inequality for Sampling without replacement

The proof generalizes the Hoeffding-Serfling inequality for scalar random variables in Bardenet and Maillard (2015) to that for random matrices. We also apply various properties for handling random matrices in Tropp (2012).

C.1 Definitions and Useful Lemmas

We first introduce the definition of the matrix function following Tropp (2012), and then introduce a number of Lemmas that are useful in the proof.

Given a symmetric matrix $\mathbf{A}$, suppose its eigenvalue decomposition is given by $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \in \mathbb{R}^{d\times d}$, where $\mathbf{\Lambda} = diag(\lambda_1, \cdots, \lambda_d)$. Then a function $f : \mathbb{R} \to \mathbb{R}$ of $\mathbf{A}$ is defined as:

$$f(\mathbf{A}) \triangleq \mathbf{U}f(\mathbf{\Lambda})\mathbf{U}^T, \tag{62}$$

where $f(\mathbf{\Lambda}) = diag(f(\lambda_1), \cdots, f(\lambda_d))$, i.e., $f(\mathbf{\Lambda})$ applies the function $f(\cdot)$ to each diagonal entry of the matrix $\mathbf{\Lambda}$.

The trace exponential function tr exp : $\mathbf{A} \to \mathrm{tr}e^{\mathbf{A}}$, i.e., tr exp$(\mathbf{A})$, is defined to first apply the exponential matrix function exp$(\mathbf{A})$, and then take the trace of exp$(\mathbf{A})$. Such a function is monotone with respect to the semidefinite order:

$$\mathbf{A} \preccurlyeq \mathbf{H} \quad \Longrightarrow \quad \mathrm{tr}\ \exp(\mathbf{A}) \preccurlyeq \mathrm{tr}\ \exp(\mathbf{H}), \tag{63}$$

which follows because for two symmetric matrices $\mathbf{A}$ and $\mathbf{H}$, if $\mathbf{A} \preccurlyeq \mathbf{H}$, then $\lambda_i(\mathbf{A}) \leqslant \lambda_i(\mathbf{H})$ for every i, where $\lambda_i(\mathbf{A})$ is the i-th largest eigenvalue of $\mathbf{A}$. Furthermore, the matrix function log$(\cdot)$ is monotone with respect to the semidefinite order (see the exercise 4.2.5 in Bhatia (2007)):

$$\mathbf{0} \prec \mathbf{A} \preccurlyeq \mathbf{H} \quad \Longrightarrow \quad \log(\mathbf{A}) \preccurlyeq \log(\mathbf{H}). \tag{64}$$

The next three lemmas follow directly from Bardenet and Maillard (2015) because the proofs are applicable for matrices.

Lemma 11. *[Bardenet and Maillard (2015)] Let $\mathbf{Z}_k \triangleq \frac{1}{k}\sum_{i=1}^{k}\mathbf{X}_i$. The following reverse martingale structure holds for $\{\mathbf{Z}_k\}_{k\leqslant N}$:*

$$\mathbb{E}[\mathbf{Z}_k|\mathbf{Z}_{k+1},\cdots\mathbf{Z}_{N-1}] = \mathbf{Z}_{k+1}. \tag{65}$$

Lemma 12. *[Bardenet and Maillard (2015)] Let $\mathbf{Y}_k \triangleq \mathbf{Z}_{N-k}$ for $1\leqslant k\leqslant N-1$. For any $\lambda>0$, the following equality holds for $2\leqslant k\leqslant n$,*

$$\lambda\mathbf{Y}_k = \lambda\mathbf{Y}_{k-1} - \lambda\frac{\mathbf{X}_{N-k+1}-\mu-\mathbf{Y}_{k-1}}{N-k}. \tag{66}$$

Lemma 13. *[Bardenet and Maillard (2015)] Let $\mathbf{Y}_k \triangleq \mathbf{Z}_{N-k}$ for $1\leqslant k\leqslant N-1$. For $2\leqslant k\leqslant N$, the following equality holds*

$$\mathbb{E}[\mathbf{X}_{N-k+1}-\mu-\mathbf{Y}_{k-1}|Y_1,\cdots,\mathbf{Y}_{k-1}] = 0, \tag{67}$$

where $\mu = \frac{1}{N}\sum_{t=1}^{N}\mathbf{X}_t$.

The following lemma is an extension of Hoeffding's inequality for scalars to matrices. We include a brief proof for completeness.

Lemma 14 (Hoeffding's Inequality for Matrix). *For a random symmetric matrix $\mathbf{X}\in\mathbb{R}^{d\times d}$, suppose*

$$\mathbb{E}[\mathbf{X}] = 0 \quad \textit{and} \quad a\mathbf{I}\preccurlyeq\mathbf{X}\preccurlyeq b\mathbf{I}.$$

where a and b are real constants. Then for any $\lambda>0$, the following inequality holds

$$\mathbb{E}[e^{\lambda\mathbf{X}}] \preccurlyeq \exp\left(\frac{1}{8}\lambda^2(b-a)^2\mathbf{I}\right). \tag{68}$$

Proof. The proof follows from the standard reasoning for scalar version. We emphasize only the difference in handling matrices. Suppose the eigenvalue decomposition of the symmetric random matrix $\mathbf{X}$ can be written as $\mathbf{X}=\mathbf{U}\Lambda\mathbf{U}^T$, where $\mathbf{U}=[\mathbf{u}_1,\cdots,\mathbf{u}_d]$ and $\mathbf{\Lambda}=diag(\lambda_1,\cdots,\lambda_d)$. Therefore, we obtain $e^{\lambda\mathbf{X}}=\sum_{i=1}^{d}e^{\lambda\lambda_i}\mathbf{u}_i\mathbf{u}_i^T$. Since scalar function $e^{\lambda x}$ is convex for any $\lambda>0$, for $1\leqslant i\leqslant d$, we have

$$e^{\lambda\lambda_i} \leqslant \left(\frac{b-\lambda_i}{b-a}e^{\lambda a}+\frac{\lambda_i-a}{b-a}e^{\lambda b}\right), \tag{69}$$

which implies that

$$e^{\lambda\lambda_i}\mathbf{u}_i\mathbf{u}_i^T \preccurlyeq \left(\frac{b-\lambda_i}{b-a}e^{\lambda a}+\frac{\lambda_i-a}{b-a}e^{\lambda b}\right)\mathbf{u}_i\mathbf{u}_i^T. \tag{70}$$

Then,

$$\begin{aligned}
\mathbb{E}[e^{\lambda\mathbf{X}}] &= \mathbb{E}\left[\sum_{i=1}^{d}e^{\lambda\lambda_i}\mathbf{u}_i\mathbf{u}_i^T\right] \overset{\text{(i)}}{\preccurlyeq} \mathbb{E}\left[\sum_{i=1}^{d}\left(\frac{b-\lambda_i}{b-a}e^{\lambda a}+\frac{\lambda_i-a}{b-a}e^{\lambda b}\right)\mathbf{u}_i\mathbf{u}_i^T\right]\\
&= \mathbb{E}\left[\sum_{i=1}^{d}\frac{b}{b-a}e^{\lambda a}\mathbf{u}_i\mathbf{u}_i^T - \sum_{i=1}^{d}\frac{\lambda_i}{b-a}e^{\lambda a}\mathbf{u}_i\mathbf{u}_i^T + \sum_{i=1}^{d}\frac{\lambda_i}{b-a}e^{\lambda b}\mathbf{u}_i\mathbf{u}_i^T - \sum_{i=1}^{d}\frac{a}{b-a}e^{\lambda b}\mathbf{u}_i\mathbf{u}_i^T\right]\\
&\overset{\text{(ii)}}{=} \mathbb{E}\left[\sum_{i=1}^{d}\frac{b}{b-a}e^{\lambda a}\mathbf{u}_i\mathbf{u}_i^T - \frac{e^{\lambda a}}{b-a}\mathbf{X} + \frac{e^{\lambda b}}{b-a}\mathbf{X} - \sum_{i=1}^{d}\frac{a}{b-a}e^{\lambda b}\mathbf{u}_i\mathbf{u}_i^T\right]\\
&\overset{\text{(iii)}}{=} \mathbb{E}\left[\sum_{i=1}^{d}\frac{b}{b-a}e^{\lambda a}\mathbf{u}_i\mathbf{u}_i^T - \sum_{i=1}^{d}\frac{a}{b-a}e^{\lambda b}\mathbf{u}_i\mathbf{u}_i^T\right]\\
&\overset{\text{(iv)}}{=} \mathbb{E}\left[\frac{b}{b-a}e^{\lambda a}\mathbf{I} - \frac{a}{b-a}e^{\lambda b}\mathbf{I}\right] = \left(\frac{b}{b-a}e^{\lambda a}-\frac{a}{b-a}e^{\lambda b}\right)\mathbf{I}
\end{aligned}$$

$$\preccurlyeq \exp\left(\frac{1}{8}\lambda^2(b-a)^2\right)\mathbf{I} \overset{\text{(v)}}{=} \exp\left(\frac{1}{8}\lambda^2(b-a)^2\mathbf{I}\right), \tag{71}$$

where (i) follows from eq. (70) and the fact that the expectation of random matrix preserves the semi-definite order, (ii) follows from $\mathbf{X} = \sum_{i=1}^d \lambda_i \mathbf{u}_i\mathbf{u}_i^T$, (iii) follows because $\mathbb{E}[\mathbf{X}] = 0$, (iv) follows because $\mathbf{I} = \mathbf{U}\mathbf{U}^T = \sum_{i=1}^d \mathbf{u}_i\mathbf{u}_i^T$, and (v) follows from the standard steps in the proof of the scalar version of Hoeffding's inequality. □

Lemma 15. *Tropp (2012)[Corollary 3.3] Let* $\mathbf{H}$ *be a fixed self-adjoint matrix, and let* $\mathbf{X}$ *be a random self-adjoint matrix. The following inequality holds*

$$\mathbb{E}\ \mathrm{tr}\exp(\mathbf{H}+\mathbf{X}) \leqslant \mathrm{tr}\exp(\mathbf{H}+\log(\mathbb{E}e^{\mathbf{X}})). \tag{72}$$

Lemma 16. *Bardenet and Maillard (2015) For integer* $n \leqslant N$*, the following inequality holds*

$$\sum_{t=1}^{n}(\frac{1}{N-t})^2 \leqslant \frac{n}{(N-n)^2}(1-\frac{n-1}{N})$$

C.2 Proof of Theorem 4

First, it suffices to show the theorem only for symmetric matrices, due to the technique of *dilations* in Tropp (2012) that transforms the asymmetric matrix to a symmetric matrix while keeping the spectral norm to be the same.

Second, it also suffices to show that for $1 \leqslant i \leqslant N$, $\mathbf{X}_i$ are symmetric and bounded, i.e., $a\mathbf{I} \preccurlyeq \mathbf{X}_i \preccurlyeq b\mathbf{I}$, and $1 \leqslant n \leqslant N-1$, the following inequality holds

$$P\left(\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i - \mu\right) \geqslant \epsilon\right) \leqslant d\exp\left(-\frac{n\epsilon^2}{2(b-a)^2(1+1/n)(1-n/N)}\right).$$

This is because the above result, with $\mathbf{X}_i$ being replaced with $-\mathbf{X}_i$, implies

$$P\left(\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i - \mu\right) \leqslant -\epsilon\right) \leqslant d\exp\left(-\frac{n\epsilon^2}{2(b-a)^2(1+1/n)(1-n/N)}\right). \tag{73}$$

Then the combination of the two results completes the desired theorem.

We start the proof by applying the matrix version of Chernoff inequality as follows. Let $\mathbf{Z}_k \triangleq \frac{1}{k}\sum_{i=1}^k \mathbf{X}_i$, for any $\lambda > 0$, we obtain

$$\begin{aligned}
P\left(\lambda_{\max}(\mathbf{Z}_n) \geqslant \epsilon\right) &= P\left(\exp(\lambda\lambda_{\max}(\mathbf{Z}_n)) \geqslant \exp(\lambda\epsilon)\right) \\
&\overset{\text{(i)}}{\leqslant} \exp(-\lambda\epsilon)\mathbb{E}\exp\left(\lambda\lambda_{\max}(\mathbf{Z}_n)\right) \\
&\overset{\text{(ii)}}{\leqslant} \exp(-\lambda\epsilon)\mathbb{E}\ \lambda_{\max}\big(\exp(\lambda\mathbf{Z}_n)\big) \\
&\overset{\text{(iii)}}{\leqslant} \exp(-\lambda\epsilon)\mathbb{E}\ \mathrm{tr}\exp(\lambda\mathbf{Z}_n) \\
&\overset{\text{(iv)}}{\leqslant} \exp(-\lambda\epsilon)\ \mathrm{tr}\exp\left(\frac{\lambda^2}{2}(b-a)^2\frac{(n+1)}{n^2}\left(1-\frac{n}{N}\right)I\right) \\
&\overset{\text{(v)}}{\leqslant} d\exp\left(\frac{\lambda^2}{2}(b-a)^2\frac{(n+1)}{n^2}\left(1-\frac{n}{N}\right)\right)\exp(-\lambda\epsilon) \\
&= d\exp\left(\frac{\lambda^2}{2}(b-a)^2\frac{(n+1)}{n^2}\left(1-\frac{n}{N}\right)-\lambda\epsilon\right)
\end{aligned} \tag{74}$$

where (i) follows from the matrix version of Chernoff inequality, (ii) follows from the fact that $\exp(\cdot)$ is an increasing function, thus $\exp\left(\lambda\lambda_{\max}(\mathbf{Z}_n)\right) = \lambda_{\max}\big(\exp(\lambda\mathbf{Z}_n)$, and (iii) follows from the fact that $\lambda_{\max}(\mathbf{A}) \leqslant \mathrm{tr}(\mathbf{A})$, with $\mathbf{A} = \exp(\lambda\mathbf{Z}_n)$, we get the desire result.

We next bound $\mathbb{E}\,\mathrm{tr}\,\exp(\lambda \mathbf{Z}_n)$. Let $Y_k \triangleq Z_{N-k}$ for $1 \leqslant k \leqslant N-1$, and $\mathbb{E}_k[\,\cdot\,] \triangleq \mathbb{E}[\,\cdot\,|\mathbf{Y}_1, \cdots, \mathbf{Y}_k]$. Thus,

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}\,\exp(\lambda \mathbf{Y}_n) &\overset{\text{(i)}}{=} \mathbb{E}\,\mathrm{tr}\,\exp\left(\lambda \mathbf{Y}_{n-1} - \lambda \frac{\mathbf{X}_{N-n+1} - \mu - \mathbf{Y}_{n-1}}{N-n}\right)\\ &\overset{\text{(ii)}}{=} \mathbb{E}\,\mathbb{E}_{n-1}\,\mathrm{tr}\,\exp\left(\lambda \mathbf{Y}_{n-1} - \lambda \frac{\mathbf{X}_{N-n+1} - \mu - \mathbf{Y}_{n-1}}{N-n}\right)\\ &\overset{\text{(iii)}}{\leqslant} \mathbb{E}\,\mathrm{tr}\,\exp\left(\lambda \mathbf{Y}_{n-1} + \log \mathbb{E}_{n-1} \exp\left(-\lambda \frac{\mathbf{X}_{N-n+1} - \mu - \mathbf{Y}_{n-1}}{N-n}\right)\right),\end{aligned} \tag{75}$$

where (i) follows from Lemma 12, (ii) follows from the tower property of expectation, (iii) follows by applying Lemma 15, where $\lambda \mathbf{Y}_{n-1}$ is deterministic given $\mathbf{Y}_1, \cdots, \mathbf{Y}_k$, and $-\lambda\big(\mathbf{X}_{N-n+1} - \mu - \mathbf{Y}_{n-1}\big)/(N-n)$ is a random variable matrix.

In order to apply Lemma 14 to bound $\mathbb{E}_{n-1} \exp(-\lambda\big(\mathbf{X}_{N-n+1} - \mu - \mathbf{Y}_{n-1}\big)/(N-n))$, we first bound $\mathbf{X}_{N-n+1} - \mu - \mathbf{Y}_{n-1}$ as follows:

$$\begin{aligned}\mathbf{X}_{N-n+1} - \mu - \mathbf{Y}_{n-1} &\overset{\text{(i)}}{=} \mathbf{X}_{N-n+1} - \mu - \mathbf{Z}_{N-n+1}\\ &\overset{\text{(ii)}}{=} \mathbf{X}_{N-n+1} - \mu - \frac{1}{N-n+1} \sum_{i=1}^{N-n+1} \Big(\mathbf{X}_i - \mu\Big)\\ &= \mathbf{X}_{N-n+1} - \frac{1}{N-n+1} \sum_{i=1}^{N-n+1} \mathbf{X}_i,\end{aligned} \tag{76}$$

where (i) follows from the definition of $\mathbf{Y}_{n-1}$ and (ii) follows from the definition of $\mathbf{Z}_{N-n+1}$. Since $a\mathbf{I} \preccurlyeq \mathbf{X}_i \preccurlyeq b\mathbf{I}$, the above equality implies

$$-\frac{(b-a)}{N-n}\mathbf{I} \preccurlyeq \frac{\mathbf{X}_{N-n+1} - \mu - \mathbf{Y}_{n-1}}{N-n} \preccurlyeq \frac{(b-a)}{N-n}\mathbf{I}. \tag{77}$$

By applying Lemma 14, and the fact $\mathbb{E}_{n-1}[\mathbf{X}_{N-n+1} - \mu - \mathbf{Y}_{n-1}] = 0$ due to Lemma 13, we obtain

$$\mathbb{E}_{n-1} \exp\left(\mathbf{X}_{N-n+1} - \mu - \mathbf{Y}_{n-1}\right) \preccurlyeq \exp\left(\frac{1}{8}\lambda^2 \left(\frac{2(b-a)}{N-n}\right)^2 \mathbf{I}\right) = \exp\left(\frac{1}{2}\lambda^2 \left(\frac{b-a}{N-n}\right)^2 \mathbf{I}\right), \tag{78}$$

Substituting eq. (78) into eq. (75), we obtain

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}\,\exp(\lambda \mathbf{Y}_n) &\overset{\text{(i)}}{\leqslant} \mathbb{E}\,\mathrm{tr}\,\exp\left(\lambda \mathbf{Y}_{n-1} + \log \exp\left(\frac{1}{2}\lambda^2 \left(\frac{b-a}{N-n}\right)^2 \mathbf{I}\right)\right)\\ &= \mathbb{E}\,\mathrm{tr}\,\exp\left(\lambda \mathbf{Y}_{n-1} + \frac{\lambda^2}{2} \left(\frac{b-a}{N-n}\right)^2 \mathbf{I}\right)\\ &\cdots\cdots\\ &\overset{\text{(ii)}}{\leqslant} \mathrm{tr}\,\exp\Bigg(\log \mathbb{E}[e^{\lambda \mathbf{Y}_1}] + \sum_{t=2}^{n} \frac{\lambda^2}{2} \left(\frac{b-a}{N-t}\right)^2 \mathbf{I}\Bigg).\end{aligned} \tag{79}$$

where (i) follows from eqs. (63) and (64), and (ii) follows by applying the steps similar to obtain eq. (78) for $n-2$ times.

To bound $\mathbb{E}[e^{\lambda \mathbf{Y}_1}]$, we first note that

$$\mathbf{Y}_1 = \mathbf{Z}_{N-1} = \frac{1}{N-1} \sum_{i=1}^{N-1} \Big(\mathbf{X}_i - \mu\Big) \overset{\text{(i)}}{=} \frac{1}{N-1} \Big(N\mu - \mathbf{X}_N - (N-1)\mu\Big) = \frac{1}{N-1} \Big(\mu - \mathbf{X}_N\Big),$$

where (i) follows because $N\mu = \sum_{i=1}^{N} \mathbf{X}_i$. Thus with $a\mathbf{I} \preccurlyeq \mathbf{X}_i \preccurlyeq b\mathbf{I}$ and $a\mathbf{I} \preccurlyeq \mu \preccurlyeq b\mathbf{I}$, we obtain

$$-\frac{(b-a)}{N-1}\mathbf{I} \preccurlyeq \mathbf{Y}_1 \preccurlyeq \frac{(b-a)}{N-1}\mathbf{I}. \tag{80}$$

Applying the matrix Hoeffding lemma with eq. (80) and $\mathbb{E}[Y_1] = \mathbb{E}[Z_{N-1}] = 0$, we obtain

$$\mathbb{E}[e^{\lambda \mathbf{Y}_1}] \preccurlyeq \exp\left(\frac{1}{2}\lambda^2\left(\frac{b-1}{N-1}\right)^2\mathbf{I}\right). \tag{81}$$

Substituting eq. (81) into eq. (79), we obtain

$$\begin{aligned}\mathbb{E}\ \mathrm{tr}\ \exp(\lambda \mathbf{Y}_n) &\leqslant \mathrm{tr}\ \exp\left(\sum_{t=1}^{n}\frac{\lambda^2}{2}\left(\frac{b-a}{N-t}\right)^2\mathbf{I}\right)\\ &= \mathrm{tr}\ \exp\left(\frac{\lambda^2}{2}(b-a)^2\sum_{t=1}^{n}\left(\frac{1}{N-t}\right)^2\mathbf{I}\right)\\ &\overset{\text{(i)}}{\leqslant} \mathrm{tr}\ \exp\left(\frac{\lambda^2}{2}(b-a)^2\frac{n}{(N-n)^2}\left(1-\frac{n-1}{N}\right)\mathbf{I}\right),\end{aligned} \tag{82}$$

where (i) follows from lemma 16.

Now let $m = N - n$, where $1 \leqslant m \leqslant N-1$, and hence $\mathbf{Y}_n = \mathbf{Z}_{N-n}$. Thus, eq. (82) implies

$$\mathbb{E}\ \mathrm{tr}\ \exp(\lambda \mathbf{Z}_m) \leqslant \mathrm{tr}\ \exp\left(\frac{\lambda^2}{2}(b-a)^2\frac{(m+1)}{m^2}\left(1-\frac{m}{N}\right)\mathbf{I}\right).$$

Substituting the above bound into eq. (74), we obtain

$$\begin{aligned}P\Big(\lambda_{\max}(\mathbf{Z}_n) \geqslant \epsilon\Big) &\leqslant \exp(-\lambda\epsilon)\ \mathrm{tr}\ \exp\left(\frac{\lambda^2}{2}(b-a)^2\frac{(n+1)}{n^2}\left(1-\frac{n}{N}\right)\mathbf{I}\right)\\ &= d\exp\left(\frac{\lambda^2}{2}(b-a)^2\frac{(n+1)}{n^2}\left(1-\frac{n}{N}\right)-\lambda\epsilon\right),\end{aligned} \tag{83}$$

where the last step follows form the equation $\mathrm{tr}(a\mathbf{I}) = da$ for $\mathbf{I} \in \mathbb{R}^{d\times d}$. The proof is completed by minimizing the above bound with respect to $\lambda > 0$, and then substituting the minimizer $\lambda^\star = \frac{n\epsilon}{(b-a)^2(1+\frac{1}{n})(1-\frac{n}{N})}$.

D Proofs for SVRC under Sampling without Replacement

D.1 Proof of Theorem 5

Proof. The idea of the proof is to apply the matrix concentration inequality for sampling without replacement that we developed in Theorem 4 to characterize the sample complexity in order to satisfy the inexactness condition $\left\|\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k)\right\| \leqslant \alpha \max\{\|\mathbf{s}_k\|, \epsilon_1\}$ with the probability at least $1-\zeta$.

We first note that

$$\begin{aligned}\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k) &\overset{\text{(i)}}{=} \tfrac{1}{|\xi_H(k)|}\big[\textstyle\sum_{i\in\xi_H(k)}(\nabla^2 f_i(\mathbf{x}_k)-\nabla^2 f_i(\tilde{\mathbf{x}}))\big]+\nabla^2 F(\tilde{\mathbf{x}}_k) - \nabla^2 F(\mathbf{x}_k)\\ &= \frac{1}{|\xi_H(k)|}\sum_{i\in\xi_H(k)}\big(\nabla^2 f_i(\mathbf{x}_k) - \nabla^2 f_i(\tilde{\mathbf{x}}) + \nabla^2 F(\tilde{\mathbf{x}}) - \nabla^2 F(\mathbf{x}_k)\big)\end{aligned}$$

where (i) follows from the definition of $\mathbf{H}_k$ in Algorithm 1. In order to apply the concentration inequality (Theorem 4) to bound $\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k)$, we define, for $1 \leqslant i \leqslant N$,

$$\mathbf{X}_i = \nabla^2 f_i(\mathbf{x}_k) - \nabla^2 f_i(\tilde{\mathbf{x}}) + \nabla^2 F(\tilde{\mathbf{x}}) - \nabla^2 F(\mathbf{x}_k),$$

which gives

$$\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k) = \frac{1}{|\xi_H(k)|}\sum_{i\in\xi_H(k)}\mathbf{X}_i. \tag{84}$$

Moreover, we have $\mu \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i = \mathbf{0}$, and

$$\sigma \triangleq \|\mathbf{A}_i\| = \left\|\nabla^2 f_i(\mathbf{x}_k) - \nabla^2 f_i(\tilde{\mathbf{x}}) + \nabla^2 F(\tilde{\mathbf{x}}) - \nabla^2 F(\mathbf{x}_k)\right\| \overset{\text{(i)}}{\leqslant} 2L_2\left\|\mathbf{x}_k - \tilde{\mathbf{x}}\right\|,$$

where (i) follows because $\nabla^2 f_i(\cdot)$ is L_2 Lipschitz, for $1 \leqslant i \leqslant N$.

Thus, in order to satisfy $\left\|\mathbf{H}_k - \nabla^2 F(\mathbf{x}_k)\right\| \leqslant \alpha\max\{\|\mathbf{s}_k\|, \epsilon_1\}$ with probability at least $1-\zeta$, by eq. (84), it is equivalent to satisfy $\left\|\frac{1}{|\xi_H(k)|}\sum_{i\in\xi_H(k)}\mathbf{X}_i - \mu\right\| \leqslant\leqslant \alpha\max\{\|\mathbf{s}_k\|, \epsilon_1\}$ with probability at least $1-\zeta$. We now apply Theorem 4 for $\mathbf{X}_i$, and it is sufficient to have:

$$2(d_1+d_2)\exp\left(-\frac{n\epsilon^2}{8\sigma^2(1+1/n)(1-n/N)}\right) \leqslant \zeta,$$

which implies

$$\frac{n\epsilon^2}{8\sigma^2(1+1/n)(1-n/N)} \geqslant \log(\frac{2(d_1+d_2)}{\zeta}).$$

Using $(1+1/n) \leqslant 2$, it is sufficient to have:

$$\frac{n\epsilon^2}{16\sigma^2(1-n/N)} \geqslant \log(\frac{2(d_1+d_2)}{\zeta}),$$

which implies

$$n \geqslant \frac{1}{\frac{1}{N} + \frac{\epsilon^2}{16\sigma^2\log(2(d_1+d_2)/\zeta)}}. \tag{85}$$

We then substitute $\sigma = 2L_2\|\mathbf{x}_k - \tilde{\mathbf{x}}\|$, $\epsilon = \alpha\max\{\|\mathbf{s}_k\|, \epsilon_1\}$, and $n = |\xi_H(k)|$, and obtain the required sample size to satisfy

$$|\xi_H(k)| \geqslant \frac{1}{\frac{1}{N} + \frac{\alpha^2\max\{\|\mathbf{s}_k\|^2, \epsilon_1^2\}}{64L_2^2\|\mathbf{x}_k - \tilde{\mathbf{x}}\|^2\log(4d/\zeta)}}. \tag{86}$$

We next bound the sample size $|\xi_g(k)|$ for the gradient, the proof follows the same procedure. We first define $\mathbf{X}_i \in \mathbb{R}^{d\times 1}$ as

$$\mathbf{X}_i = \nabla f_i(\mathbf{x}_k) - \nabla f_i(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}}) - \nabla F(\mathbf{x}_k), \tag{87}$$

and hence

$$\mathbf{g}_k - \nabla F(\mathbf{x}_k) = \frac{1}{|\xi_g(k)|}\sum_{i\in\xi_g(k)}\mathbf{X}_i. \tag{88}$$

Moreover, we have $\mu = \frac{1}{N}\sum_{i\in\xi_g(k)}\mathbf{A}_i = \mathbf{0}$, and

$$\sigma \triangleq \|\mathbf{A}_i\| = \|\nabla f_i(\mathbf{x}_k) - \nabla f_i(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}}) - \nabla F(\mathbf{x}_k)\| \overset{\text{(i)}}{\leqslant} 2L_1\left\|\mathbf{x}_k - \tilde{\mathbf{x}}\right\|,$$

where (i) follows because $\nabla f_i(\cdot)$ is L_1 Lipschitz, for $1 \leqslant i \leqslant N$.

In order to satisfy $\|\mathbf{g}_k - \nabla F(\mathbf{x}_k)\| \leqslant \beta\max\{\|\mathbf{s}_k\|^2, \epsilon_1^2\}$ with probability at least $1-\zeta$, by eq. (88), it is equivalent to satisfy $\left\|\frac{1}{|\xi_g(k)|}\sum_{i\in\xi_g(k)}\mathbf{X}_i - \mu\right\| \leqslant \beta\max\{\|\mathbf{s}_k\|^2, \epsilon_1^2\}$ with probability at least $1-\zeta$. We then apply Theorem 4 for $\mathbf{X}_i$ in the way similar to that for bounding the sample size for Hessian, with $\sigma = 2L_1\|\mathbf{x}_k - \tilde{\mathbf{x}}\|$, $\mu = 0$, $\epsilon = \beta\max\{\|\mathbf{s}_k\|^2, \epsilon_1^2\}$, and $n = |\xi_g(k)|$, and obtain the required sample size to satisfy

$$|\xi_g(k)| \geqslant \frac{1}{\frac{1}{N} + \frac{\beta^2\max\{\|\mathbf{s}_k\|^4, \epsilon_1^4\}}{64L_1^2\|\mathbf{x}_k - \tilde{\mathbf{x}}\|^2\log(2(d+1)/\zeta)}}. \tag{89}$$

□

D.2 Proof of Proposition 6

Proof. The proof of Proposition 6 is similar to the proof of Theorem 5. We first define $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ as

$$\mathbf{A}_i = \nabla^2 f_i(\mathbf{x}_k) - \nabla^2 F(\mathbf{x}_k), \tag{90}$$

and hence $\mu = \frac{1}{N}\sum_{i\in\xi_H(k)} \mathbf{A}_i = \mathbf{0}$. Furthermore,

$$\sigma \triangleq \|\mathbf{A}_i\| = \left\|\nabla^2 f_i(\mathbf{x}_k) - \nabla^2 F(\mathbf{x}_k)\right\| \overset{\text{(i)}}{\leqslant} 2L_1,$$

where (i) follows from Assumption 1.

Let $\{\mathbf{X}_i\}_{i=1}^{|\xi_g(k)|} = \{\mathbf{A}_i : i \in \xi_H(k)\}$, and we have

$$\frac{1}{|\xi_H(k)|}\sum_{i\in\xi_H(k)} \mathbf{X}_i - \mu \overset{\text{(i)}}{=} \frac{1}{|\xi_H(k)|}\sum_{i\in\xi_H(k)} \mathbf{A}_i \overset{\text{(ii)}}{=} \mathbf{H}_k - \nabla^2 F(\mathbf{x}_k), \tag{91}$$

where (i) follows from the fact that $\mu = 0$ and (ii) follows from the definition of $\mathbf{H}_k$ in Algorithm 1.

We then apply Theorem 4 for $\mathbf{X}_i$ with $\sigma = 2L_1$, $\mu = 0$, $\epsilon = C_2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$, and $n = |\xi_H(k)|$, and obtain the require sampled size to satisfy

$$|\xi_H(k)| \geqslant \frac{1}{\frac{1}{N} + \frac{C_2^2\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2}{64L_1^2\log(4d/\zeta)}}. \tag{92}$$

To bound the sample size of gradient, i.e., $|\xi_g(k)|$, we follow the similar proof by constructing

$$\mathbf{A}_i = \nabla f_i(\mathbf{x}_k) - \nabla F(\mathbf{x}_k), \tag{93}$$

and applying Theorem 4 with $\sigma = 2L_0$, $\mu = 0$, $\epsilon = C_1\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$, and $n = |\xi_g(k)|$, and obtain the required sample size to satisfy

$$|\xi_g(k)| \geqslant \frac{1}{\frac{1}{N} + \frac{C_1^2\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^4}{64L_0^2\log(2(d+1)/\zeta)}}. \tag{94}$$

□

D.3 Proof of Theorem 7

Proof. Assume the algorithm terminates at iteration k, then the total Hessian complexity is given by

$$\begin{aligned}
m + \frac{kN}{m} + \sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}|\xi_H(k)| &\overset{\text{(i)}}{\leqslant} \frac{CkN}{m} + \sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\frac{1}{\frac{1}{N} + \frac{\alpha^2\max\{\|\mathbf{s}_k\|^2,\epsilon_1^2\}}{64L_2^2\|\mathbf{x}_{i\cdot m+j}-\mathbf{x}_{i\cdot m}\|^2\log(4d/\zeta)}} \\
&\leqslant \frac{CkN}{m} + \sum_{i=0}^{k/m-1}\sum_{j=1}^{m-1}\frac{64L_2^2\left\|\mathbf{x}_{i\cdot m+j} - \mathbf{x}_{i\cdot m}\right\|^2\log(4d/\zeta)}{\alpha^2\epsilon_1^2} \\
&\overset{\text{(ii)}}{\leqslant} \frac{CkN}{m} + \frac{64L_2^2}{\alpha^2\epsilon_1^2}\left(m^2k^{1/3}C^{2/3}\right)\log\left(\frac{4d}{\zeta}\right) \\
&\overset{\text{(iii)}}{\leqslant} C\log\left(\frac{4d}{\zeta}\right)\left(\frac{N}{m\epsilon^{3/2}} + \frac{m^2}{\epsilon^{3/2}}\right) = \frac{C}{\epsilon^{3/2}}\log\left(\frac{4d}{\zeta}\right)\left(\frac{N}{m} + m^2\right)
\end{aligned}$$

where (i) follows form Theorem 5, and (ii) follows form eq. (58), (iii) follows from the fact that $\zeta < 1$ and $d \geqslant 1$ which gives $\log\left(\frac{4d}{\zeta}\right) > 1$, and the fact that $\epsilon_1 = O(\epsilon^{1/2})$ such that $k = O(\epsilon^{-3/2})$ according to Theorem 1.

We minimize the above bound over m, substitute the minimizer $m^{\star} = N^{1/3}$, and follows the similar procedure in the proof of eq. (13) to ensure a successful event overall iteration with at least $1 - \delta$, which gives that

$$\sum_{i=0}^{k} |\xi_H(k)| \leqslant \frac{CN^{2/3}}{\epsilon^{3/2}} \log\left(\frac{8d}{\epsilon\delta}\right). \tag{95}$$

Thus, we have

$$\sum_{i=0}^{k} |\xi_H(k)| = \tilde{O}\left(\frac{N^{3/2}}{\epsilon^{3/2}}\right). \tag{96}$$

□