Appendix : Stochastic Variance Reduced Optimization for Nonconvex Sparse Learning

A. Proof of Lemma 3.4

For any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$ in sparse linear model, we have $\nabla^2 \mathcal{F}(\boldsymbol{w}) = \mathbf{A}^\top \mathbf{A}$ and

$$\mathcal{F}(\boldsymbol{w}) - \mathcal{F}(\boldsymbol{w}') - \langle \nabla \mathcal{F}(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle = \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}')^\top \nabla^2 \mathcal{F}(\boldsymbol{w}'')(\boldsymbol{w} - \boldsymbol{w}') = \frac{1}{2}\|\mathbf{A}(\boldsymbol{w} - \boldsymbol{w}')\|_2^2,$$

where $\boldsymbol{w}''$ is between $\boldsymbol{w}$ and $\boldsymbol{w}'$ and $\|\boldsymbol{w} - \boldsymbol{w}'\|_0 \leq 2k \leq s$. Let $\boldsymbol{v} = \boldsymbol{w} - \boldsymbol{w}'$, then $\|\boldsymbol{v}\|_0 \leq s$ and $\|\boldsymbol{v}\|_1^2 \leq s\|\boldsymbol{v}\|_2^2$. By (3.8), we have

$$\frac{\|\mathbf{A}\boldsymbol{v}\|_2^2}{nb} \geq \psi_1 \|\boldsymbol{v}\|_2^2 - \varphi_1 \frac{s \log d}{nb}\|\boldsymbol{v}\|_2^2, \text{ and } \frac{\|\mathbf{A}_{\mathcal{S}_i *}\boldsymbol{v}\|_2^2}{b} \leq \psi_2 \|\boldsymbol{v}\|_2^2 + \varphi_2 \frac{s \log d}{b}\|\boldsymbol{v}\|_2^2, \forall i \in [n],$$

which further imply

$$\rho_s^- = \inf_{\|\boldsymbol{v}\|_0 \leq s} \frac{\|\mathbf{A}\boldsymbol{v}\|_2^2}{nb\|\boldsymbol{v}\|_2^2} \geq \psi_1 - \varphi_1 \frac{s \log d}{nb}, \text{ and } \rho_s^+ = \sup_{\|\boldsymbol{v}\|_0 \leq s, i \in [n]} \frac{\|\mathbf{A}_{\mathcal{S}_i *}\boldsymbol{v}\|_2^2}{b\|\boldsymbol{v}\|_2^2} \leq \psi_2 + \varphi_2 \frac{s \log d}{b}. \tag{A.1}$$

If $b \geq \frac{\varphi_2 s \log d}{\psi_2}$ and $n \geq \frac{2\varphi_1 \psi_2}{\psi_1 \varphi_2}$, then we have $nb \geq \frac{2\varphi_1 s \log d}{\psi_1}$. Combining these with (A.1), we have

$$\rho_s^- \geq \frac{1}{2}\psi_1, \text{ and } \rho_s^+ \leq 2\psi_2.$$

By the definition of κ, this indicates $\kappa_s = \frac{\rho_s^+}{\rho_s^-} \leq \frac{4\psi_2}{\psi_1}$. Then for some $C_5 \geq \frac{16 C_1 \psi_2^2}{\psi_1^2}$, we have

$$k = C_5 k^* \geq C_1 \kappa_s^2 k^*.$$

B. Proof of Theorem 3.5

For sparse linear model, we have $\nabla \mathcal{F}(\boldsymbol{w}^*) = \mathbf{A}^\top \mathbf{z}/(nb)$. Since $\mathbf{z}$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, then $\mathbf{A}_{*j}^\top \mathbf{z}/(nb) \sim \mathcal{N}(0, \sigma^2 \|\mathbf{A}_{*j}\|_2^2/(nb)^2)$ for any $j \in [d]$. Using the Mill's inequality for tail bounds of Normal distribution, we have

$$\mathbb{P}\left(\left|\frac{\mathbf{A}_{*j}^\top \mathbf{z}}{nb}\right| > 2\sigma\sqrt{\frac{\log d}{nb}}\right) = \mathbb{P}\left(\left|\frac{\mathbf{A}_{*j}^\top \mathbf{z}}{\sigma\|\mathbf{A}_{*j}\|_2}\right| > 2\frac{\sqrt{nb \log d}}{\|\mathbf{A}_{*j}\|_2}\right) \leq \|\mathbf{A}_{*j}\|_2 \sqrt{\frac{1}{2\pi nb \log d}} \exp\left(-4\frac{nb \log d}{\|\mathbf{A}_{*j}\|_2^2}\right).$$

This implies, using union bound and the assumption $\frac{\max_j \|\mathbf{A}_{*j}\|_2}{\sqrt{nb}} \leq 1$,

$$\mathbb{P}\left(\left\|\frac{\mathbf{A}_{*j}^\top \mathbf{z}}{nb}\right\|_\infty > 2\sigma\sqrt{\frac{\log d}{nb}}\right) \leq \frac{d^{-3}}{\sqrt{\pi nb \log d}}.$$

Then we have the following result holds with probability at least $1 - \frac{1}{\sqrt{nb \log d}} \cdot d^{-3}$

$$\|\nabla \mathcal{F}(\boldsymbol{w}^*)\|_\infty \leq \left\|\frac{\mathbf{A}^\top \mathbf{z}}{nb}\right\|_\infty \leq 2\sigma\sqrt{\frac{\log d}{nb}}. \tag{B.1}$$

Conditioning on (B.1), it follows consequently that

$$\|\nabla_{\widetilde{\mathcal{I}}} \mathcal{F}(\boldsymbol{w}^*)\|_2^2 \leq s\|\nabla \mathcal{F}(\boldsymbol{w}^*)\|_\infty^2 \leq \frac{4\sigma^2 s \log d}{nb}. \tag{B.2}$$

We have from Lemma 3.4 that $s = 2k + k^* = (2C_5 + 1)\,k^*$ for some constant C_5 when n and b are large enough. For a given $\varepsilon > 0$ and $\delta \in (0,1)$, if

$$r \geq 4\log\left(\frac{\mathcal{F}(\widetilde{\boldsymbol{w}}^{(0)}) - \mathcal{F}(\boldsymbol{w}^*)}{\varepsilon\delta}\right),$$

then with probability at least $1 - \delta - \frac{1}{\sqrt{nb\log d}} \cdot d^{-3}$, we have from (3.4), (B.1) and (B.2) that

$$\|\widetilde{\boldsymbol{w}}^{(r)} - \boldsymbol{w}^*\|_2 \leq c_3\sigma\sqrt{\frac{k^*\log d}{nb}},$$

for some constant c_3, which completes the proof.

C. Proof of Lemma 4.1

For notational convenience, define $\boldsymbol{w}' = \mathcal{H}_k(\boldsymbol{w})$. Let $\mathrm{supp}(\boldsymbol{w}^*) = \mathcal{I}^*$, $\mathrm{supp}(\boldsymbol{w}) = \mathcal{I}$, $\mathrm{supp}(\boldsymbol{w}') = \mathcal{I}'$, and $\boldsymbol{w}'' = \boldsymbol{w} - \boldsymbol{w}'$ with $\mathrm{supp}(\boldsymbol{w}'') = \mathcal{I}''$. Clearly we have $\mathcal{I}' \cup \mathcal{I}'' = \mathcal{I}$, $\mathcal{I}' \cap \mathcal{I}'' = \emptyset$, and $\|\boldsymbol{w}\|_2^2 = \|\boldsymbol{w}'\|_2^2 + \|\boldsymbol{w}''\|_2^2$. Then we have that

$$\|\boldsymbol{w}' - \boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 = \|\boldsymbol{w}'\|_2^2 - 2\langle\boldsymbol{w}', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}\|_2^2 + 2\langle\boldsymbol{w}, \boldsymbol{w}^*\rangle = 2\langle\boldsymbol{w}'', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}''\|_2^2. \tag{C.1}$$

If $2\langle\boldsymbol{w}'', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}''\|_2^2 \leq 0$, then (4.1) holds naturally. From this point on, we will discuss the situation when $2\langle\boldsymbol{w}'', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}''\|_2^2 > 0$.

Let $\mathcal{I}^* \cap \mathcal{I}' = \mathcal{I}^{*1}$ and $\mathcal{I}^* \cap \mathcal{I}'' = \mathcal{I}^{*2}$, and denote $(\boldsymbol{w}^*)_{\mathcal{I}^{*1}} = \boldsymbol{w}^{*1}$, $(\boldsymbol{w}^*)_{\mathcal{I}^{*2}} = \boldsymbol{w}^{*2}$, $(\boldsymbol{w}')_{\mathcal{I}^{*1}} = \boldsymbol{w}^{1*}$, and $(\boldsymbol{w}'')_{\mathcal{I}^{*2}} = \boldsymbol{w}^{2*}$. Then we have that

$$2\langle\boldsymbol{w}'', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}''\|_2^2 = 2\langle\boldsymbol{w}^{2*}, \boldsymbol{w}^{*2}\rangle - \|\boldsymbol{w}''\|_2^2 \leq 2\langle\boldsymbol{w}^{2*}, \boldsymbol{w}^{*2}\rangle - \|\boldsymbol{w}^{2*}\|_2^2 \leq 2\|\boldsymbol{w}^{2*}\|_2\|\boldsymbol{w}^{*2}\|_2 - \|\boldsymbol{w}^{2*}\|_2^2. \tag{C.2}$$

Let $|\mathrm{supp}(\boldsymbol{w}^{2*})| = |\mathcal{I}^{*2}| = k^{**}$ and $w_{2,\max} = \|\boldsymbol{w}^{2*}\|_\infty$, then consequently we have $\|\boldsymbol{w}^{2*}\|_2 = m \cdot w_{2,\max}$ for some $m \in [1, \sqrt{k^{**}}]$. Notice that we are interested in $1 \leq k^{**} \leq k^*$, because (4.1) holds naturally if $k^{**} = 0$. In terms of $\|\boldsymbol{w}^{*2}\|_2$, the RHS of (C.2) is maximized when:

Case 1: $m = 1$, if $\|\boldsymbol{w}^{*2}\|_2 \leq w_{2,\max}$;

Case 2: $m = \frac{\|\boldsymbol{w}^{*2}\|_2}{w_{2,\max}}$, if $w_{2,\max} < \|\boldsymbol{w}^{*2}\|_2 < \sqrt{k^{**}}w_{2,\max}$, ;

Case 3: $m = \sqrt{k^{**}}$, if $\|\boldsymbol{w}^{*2}\|_2 \geq \sqrt{k^{**}}w_{2,\max}$.

Case 1: If $\|\boldsymbol{w}^{*2}\|_2 \leq w_{2,\max}$, then the RHS of (C.2) is maximized when $m = 1$, i.e. $\boldsymbol{w}^{2*}$ has only one nonzero element $w_{2,\max}$. By (C.2), we have

$$2\langle\boldsymbol{w}'', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}''\|_2^2 \leq 2w_{2,\max}\|\boldsymbol{w}^{*2}\|_2 - w_{2,\max}^2 \leq 2w_{2,\max}^2 - w_{2,\max}^2 = w_{2,\max}^2. \tag{C.3}$$

Denote $w_{1,\min}$ as the smallest element of $\boldsymbol{w}^{1*}$ (in magnitude), which indicates that $|w_{1,\min}| \geq |w_{2,\max}|$ as $\boldsymbol{w}'$ contains the largest k entries and $\boldsymbol{w}''$ contains the smallest $d-k$ entries of $\boldsymbol{w}$. For $\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2$, we have that

$$\begin{aligned}
\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 &= \|\boldsymbol{w}' - \boldsymbol{w}^{*1}\|_2^2 + \|\boldsymbol{w}'' - \boldsymbol{w}^{*2}\|_2^2 \\
&= \|\boldsymbol{w}_{(\mathcal{I}^{*1})^C}\|_2^2 + \|\boldsymbol{w}_{\mathcal{I}^{*1}} - \boldsymbol{w}^{*1}\|_2^2 + \|\boldsymbol{w}^{*2}\|_2^2 - (2\langle\boldsymbol{w}'', \boldsymbol{w}^*\rangle - \|\boldsymbol{w}''\|_2^2) &\text{(C.4)}\\
&\geq (k - k^* + k^{**})w_{1,\min}^2 - w_{2,\max}^2 &\text{(C.5)}
\end{aligned}$$

where the last inequality follows from the fact that $\boldsymbol{w}_{(\mathcal{I}^{*1})^C}$ has $k - k^* + k^{**}$ entries larger than $w_{1,\min}$ (in magnitude). Combining (C.1), (C.3) and (C.5), we have that

$$\begin{aligned}
\frac{\|\boldsymbol{w}' - \boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2}{\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2} &\leq \frac{w_{2,\max}^2}{(k - k^* + k^{**})w_{1,\min}^2 - w_{2,\max}^2} \\
&\leq \frac{w_{2,\max}^2}{(k - k^* + k^{**})w_{2,\max}^2 - w_{2,\max}^2} \leq \frac{1}{k - k^*}. &\text{(C.6)}
\end{aligned}$$

Case 2: If $w_{2,\max} < \|\boldsymbol{w}^{*2}\|_2 < \sqrt{k^{**}} w_{2,\max}$, then the RHS of (C.2) is maximized when $m = \frac{\|\boldsymbol{w}^{*2}\|_2}{w_{2,\max}}$. By (C.2), we have that

$$2\langle \boldsymbol{w}'', \boldsymbol{w}^* \rangle - \|\boldsymbol{w}''\|_2^2 \leq 2\sqrt{k^{**}} w_{2,\max} \cdot m w_{2,\max} - w_{2,\max}^2 \leq k^{**} w_{2,\max}^2. \tag{C.7}$$

By (C.4), we have that

$$\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 \geq (k - k^* + k^{**}) w_{1,\min}^2 + m^2 w_{2,\max}^2 - w_{2,\max}^2 \geq (k - k^* + k^{**}) w_{1,\min}^2. \tag{C.8}$$

Combining (C.1), (C.7) and (C.8), we have that

$$\frac{\|\boldsymbol{w}' - \boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2}{\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2} \leq \frac{k^{**} w_{2,\max}^2}{(k - k^* + k^{**}) w_{1,\min}^2} \leq \frac{k^{**}}{k - k^* + k^{**}}. \tag{C.9}$$

Case 3: If $\|\boldsymbol{w}^{*2}\|_2 \geq \sqrt{k^{**}} w_{2,\max}$, then the RHS of (C.2) is maximized when $m = \sqrt{k^{**}}$. Let $\|\boldsymbol{w}^{*2}\|_2 = \gamma w_{2,\max}$ for some $\gamma \geq \sqrt{k^{**}}$. We have from (C.2) that

$$2\langle \boldsymbol{w}'', \boldsymbol{w}^* \rangle - \|\boldsymbol{w}''\|_2^2 \leq 2\gamma\sqrt{k^{**}} w_{2,\max}^2 - k^{**} w_{2,\max}^2. \tag{C.10}$$

By (C.4), we have

$$\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 \geq (k - k^* + k^{**}) w_{1,\min}^2 + \gamma^2 w_{2,\max}^2 - \gamma\sqrt{k^{**}} w_{2,\max}^2 + k^{**} w_{2,\max}^2. \tag{C.11}$$

Combining (C.1), (C.10) and (C.11), we have

$$\begin{aligned} \frac{\|\boldsymbol{w}' - \boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2}{\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2} &\leq \frac{2\gamma\sqrt{k^{**}} w_{2,\max}^2 - k^{**} w_{2,\max}^2}{(k - k^* + k^{**}) w_{1,\min}^2 + \gamma^2 w_{2,\max}^2 - \gamma\sqrt{k^{**}} w_{2,\max}^2 + k^{**} w_{2,\max}^2} \\ &\leq \frac{2\gamma\sqrt{k^{**}} - k^{**}}{k - k^* + 2k^{**} + \gamma^2 - 2\gamma\sqrt{k^{**}}}. \end{aligned} \tag{C.12}$$

Inspecting the RHS of (C.12) carefully, we can see that it is either a bell shape function or a monotone decreasing function when $\gamma \geq \sqrt{k^{**}}$. Setting the first derivative of the RHS in terms of γ to zero, we have $\gamma = \frac{1}{2}\sqrt{k^{**}} + \sqrt{k - k^* + \frac{5}{4}k^{**}}$ (the other root is smaller than $\sqrt{k^{**}}$). Denoting $\gamma_* = \max\{\sqrt{k^{**}}, \frac{1}{2}\sqrt{k^{**}} + \sqrt{k - k^* + \frac{5}{4}k^{**}}\}$ and plugging it into the RHS of (C.12), we have

$$\frac{\|\boldsymbol{w}' - \boldsymbol{w}^*\|_2^2 - \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2}{\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2} \leq \max\left\{ \frac{k^{**}}{k - k^* + k^{**}}, \frac{2\sqrt{k^{**}}}{2\sqrt{k - k^* + \frac{5}{4}k^{**}} - \sqrt{k^{**}}} \right\}. \tag{C.13}$$

Combining (C.6), (C.9) and (C.13), and taking $k > k^*$ and $k^* \geq k^{**} \geq 1$ into consideration, we have

$$\begin{aligned} \max\left\{ \frac{1}{k - k^*}, \frac{k^{**}}{k - k^* + k^{**}}, \frac{2\sqrt{k^{**}}}{2\sqrt{k - k^* + \frac{5}{4}k^{**}} - \sqrt{k^{**}}} \right\} &\leq \frac{2\sqrt{k^{**}}}{2\sqrt{k - k^* + \frac{5}{4}k^{**}} - \sqrt{k^{**}}} \\ &\leq \frac{2\sqrt{k^*}}{2\sqrt{k - k^*} - \sqrt{k^*}} \leq \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}, \end{aligned}$$

which proves the result.

D. Proof of Lemma 4.3

Remind that the stochastic variance reduced gradient is

$$\mathbf{g}^{(t)}(\boldsymbol{w}^{(t)}) = \nabla f_{i_t}(\boldsymbol{w}^{(t)}) - \nabla f_{i_t}(\widetilde{\boldsymbol{w}}) + \widetilde{\boldsymbol{\mu}}, \tag{D.1}$$

where $\widetilde{\boldsymbol{\mu}} = \nabla \mathcal{F}(\widetilde{\boldsymbol{w}}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\widetilde{\boldsymbol{w}})$.

It is straightforward that the stochastic variance reduced gradient (D.1) satisfies

$$\mathbb{E}\mathbf{g}^{(t)}(\boldsymbol{w}^{(t)}) = \mathbb{E}\nabla f_{i_t}(\boldsymbol{w}^{(t)}) - \mathbb{E}\nabla f_{i_t}(\widetilde{\boldsymbol{w}}) + \widetilde{\boldsymbol{\mu}} = \nabla \mathcal{F}(\boldsymbol{w}^{(t)}),$$

Thus $\mathbf{g}^{(t)}(\boldsymbol{w}^{(t)})$ is a unbiased estimate of $\nabla\mathcal{F}(\boldsymbol{w}^{(t)})$ and the first claim is verified.

Next, we bound $\mathbb{E}\|\mathbf{g}_{\mathcal{I}}^{(t)}(\boldsymbol{w}^{(t)})\|_2^2$. For any $i \in [n]$ and $\boldsymbol{w}$ with $\mathrm{supp}(\boldsymbol{w}) \subseteq \mathcal{I}$, consider

$$\phi_i(\boldsymbol{w}) = f_i(\boldsymbol{w}) - f_i(\boldsymbol{w}^*) - \langle \nabla f_i(\boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^* \rangle.$$

Since $\nabla\phi_i(\boldsymbol{w}^*) = \nabla f_i(\boldsymbol{w}^*) - \nabla f_i(\boldsymbol{w}^*) = \mathbf{0}$, we have that $\phi_i(\boldsymbol{w}^*) = \min_{\boldsymbol{w}} \phi_i(\boldsymbol{w})$, which implies

$$\begin{aligned} 0 = \phi_i(\boldsymbol{w}^*) &\le \min_{\eta} \phi_i(\boldsymbol{w} - \eta\nabla\phi_i(\boldsymbol{w})) \le \min_{\eta} \phi_i(\boldsymbol{w}) - \eta\|\nabla\phi_i(\boldsymbol{w})\|_2^2 + \frac{\rho_s^+\eta^2}{2}\|\nabla\phi_i(\boldsymbol{w})\|_2^2 \\ &= \phi_i(\boldsymbol{w}) - \frac{1}{2\rho_s^+}\|\nabla\phi_i(\boldsymbol{w})\|_2^2, \end{aligned} \tag{D.2}$$

where the last inequality follows from the RSS condition and the last equality follows from the fact that $\eta = 1/\rho_s^+$ minimizes the function. By (D.2), we have

$$\begin{aligned} \|\nabla_{\mathcal{I}} f_i(\boldsymbol{w}) - \nabla_{\mathcal{I}} f_i(\boldsymbol{w}^*)\|_2^2 &\le \|\nabla f_i(\boldsymbol{w}) - \nabla f_i(\boldsymbol{w}^*)\|_2^2 \\ &\le 2\rho_s^+ \left[f_i(\boldsymbol{w}) - f_i(\boldsymbol{w}^*) - \langle \nabla f_i(\boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^*\rangle\right] \\ &= 2\rho_s^+ \left[f_i(\boldsymbol{w}) - f_i(\boldsymbol{w}^*) - \langle \nabla_{\mathcal{I}} f_i(\boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^*\rangle\right]. \end{aligned} \tag{D.3}$$

Since the sampling of i from $[n]$ is uniform sampling, we have from (D.3)

$$\begin{aligned} \mathbb{E}\|\nabla_{\mathcal{I}} f_i(\boldsymbol{w}) - \nabla_{\mathcal{I}} f_i(\boldsymbol{w}^*)\|_2^2 &= \frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\mathcal{I}} f_i(\boldsymbol{w}) - \nabla_{\mathcal{I}} f_i(\boldsymbol{w}^*)\|_2^2 \\ &\le 2\rho_s^+ \left[\mathcal{F}(\boldsymbol{w}) - \mathcal{F}(\boldsymbol{w}^*) - \langle \nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^*\rangle\right] \\ &\le 2\rho_s^+ \left[\mathcal{F}(\boldsymbol{w}) - \mathcal{F}(\boldsymbol{w}^*) + |\langle \nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^*\rangle|\right] \\ &\le 4\rho_s^+ \left[\mathcal{F}(\boldsymbol{w}) - \mathcal{F}(\boldsymbol{w}^*)\right], \end{aligned} \tag{D.4}$$

where the last inequality is from the RSC condition of $\mathcal{F}(\boldsymbol{w})$.

By the definition of $\mathbf{g}_{\mathcal{I}}^{(t)}$ in (D.1), we can verify the second claim as

$$\begin{aligned} \mathbb{E}\|\mathbf{g}_{\mathcal{I}}^{(t)}(\boldsymbol{w}^{(t)})\|_2^2 &\le 3\mathbb{E}\|\left[\nabla_{\mathcal{I}} f_{i_t}(\widetilde{\boldsymbol{w}}) - \nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^*)\right] - \nabla_{\mathcal{I}}\mathcal{F}(\widetilde{\boldsymbol{w}}) + \nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2 \\ &\quad + 3\mathbb{E}\|\nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^{(t)}) - \nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2 \\ &\le 3\mathbb{E}\|\nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^{(t)}) - \nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^*)\|_2^2 + 3\mathbb{E}\|\nabla_{\mathcal{I}} f_{i_t}(\widetilde{\boldsymbol{w}}) - \nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2 \\ &\le 12\rho_s^+ \left[\mathcal{F}(\boldsymbol{w}^{(t)}) - \mathcal{F}(\boldsymbol{w}^*) + \mathcal{F}(\widetilde{\boldsymbol{w}}) - \mathcal{F}(\boldsymbol{w}^*)\right] + 3\|\nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)\|_2^2, \end{aligned}$$

where the first inequality follows from $\|\mathbf{a} + \boldsymbol{b} + \mathbf{c}\|_2^2 \le 3\|\mathbf{a}\|_2^2 + 3\|\boldsymbol{b}\|_2^2 + 3\|\mathbf{c}\|_2^2$, the second inequality follows from $\mathbb{E}\|\mathbf{x} - \mathbb{E}\mathbf{x}\|_2^2 \le \mathbb{E}\|\mathbf{x}\|_2^2$ with $\mathbb{E}\left[\nabla_{\mathcal{I}} f_{i_t}(\widetilde{\boldsymbol{w}}) - \nabla_{\mathcal{I}} f_{i_t}(\boldsymbol{w}^*)\right] = \nabla_{\mathcal{I}}\mathcal{F}(\widetilde{\boldsymbol{w}}) - \nabla_{\mathcal{I}}\mathcal{F}(\boldsymbol{w}^*)$, and the last inequality follows from (D.4).