Appendix: Stochastic Variance Reduced Optimization for Nonconvex Sparse Learning

A. Proof of Lemma 3.4

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

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

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

$$
\frac{\|Av\|^2_{nb}}{nb} \geq \psi_1 \|v\|^2_2 - \varphi_1 \frac{s \log d}{nb} \|v\|^2_2, \quad \text{and} \quad \frac{\|A_{S_i}v\|^2_{nb}}{b} \leq \psi_2 \|v\|^2_2 + \varphi_2 \frac{s \log d}{b} \|v\|^2_2, \quad \forall i \in [n],
$$

which further imply

$$
\rho_s^- = \inf_{\|v\|_0 \leq s} \frac{\|Av\|^2_{nb}}{\|v\|^2_2} \geq \psi_1 - \varphi_1 \frac{s \log d}{nb}, \quad \text{and} \quad \rho_s^+ = \sup_{\|v\|_0 \leq s} \frac{\|A_{S_i}v\|^2_{nb}}{\|v\|^2_2} \leq \psi_2 + \varphi_2 \frac{s \log d}{b}. \quad (A.1)
$$

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

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

By the definition of $\kappa$, this indicates $\kappa_s = \frac{\rho_s^+}{\rho_s^-} \leq \frac{4 \rho_1}{\rho_1^2}$. Then for some $C_5 \geq \frac{16 C_1 \psi_1^3}{\psi_1^2}$, 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 $\nabla F(w^*) = A^\top z/(nb)$. Since $z$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, then $A^\top j z/(nb) \sim \mathcal{N}(0, \sigma^2 \|A_{S_j}\|^2_2/(nb)^2)$ for any $j \in [d]$. Using the Mill’s inequality for tail bounds of Normal distribution, we have

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

This implies, using union bound and the assumption $\max_j \frac{\|A_{S_j}\|^2_2}{nb} \leq 1$,

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

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

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

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

$$
\|\nabla^2 F(w^*)\|^2_s \leq s \|\nabla F(w^*)\|_{\infty}^2 \leq \frac{4 \sigma^2 s \log d}{nb}. \quad (B.2)
$$
We have from Lemma 3.4 that \( s = 2k^* + (2C_5 + 1)k^* \) for some constant \( C_5 \) when \( n \) and \( b \) are large enough. For a given \( \varepsilon > 0 \) and \( \delta \in (0, 1) \), if
\[
 r \geq 4 \log \left( \frac{F(\tilde{w}^{(0)}) - F(w^*)}{\varepsilon \delta} \right),
\]
then with probability at least \( 1 - \delta - \frac{1}{\sqrt{n} \log d} \cdot d^{-3} \), we have from (3.4), (B.1) and (B.2) that
\[
\| \tilde{w}^{(r)} - w^* \|_2 \leq c_3 \sigma \sqrt{\frac{k^* \log d}{nb}},
\]
for some constant \( c_3 \), which completes the proof.

**C. Proof of Lemma 4.1**

For notational convenience, define \( w' = H_1(w) \). Let \( \text{supp}(w^*) = \mathcal{I}^* \), \( \text{supp}(w) = \mathcal{I} \), \( \text{supp}(w') = \mathcal{I}' \), and \( w'' = w - w' \) with \( \text{supp}(w'') = \mathcal{I}''. \) Clearly we have \( \mathcal{I}' \cup \mathcal{I}'' = \mathcal{I}, \mathcal{I}' \cap \mathcal{I}'' = \emptyset \), and \( \| w' \|_2^2 = \| w' \|_2^2 + \| w'' \|_2^2 \). Then we have that
\[
\| w' - w^* \|_2^2 - \| w - w^* \|_2^2 = \| w' \|_2^2 - 2 \langle w', w^* \rangle - \| w^* \|_2^2 + 2 \langle w, w^* \rangle = 2 \langle w', w^* \rangle - \| w'' \|_2^2. \tag{C.1}
\]
If \( 2 \langle w'', w^* \rangle - \| w'' \|_2^2 \leq 0 \), then (4.1) holds naturally. From this point on, we will discuss the situation when \( 2 \langle w'', w^* \rangle - \| w'' \|_2^2 > 0 \).

Let \( \mathcal{I}' \cap \mathcal{I} = \mathcal{I}' \) and \( \mathcal{I}' \cap \mathcal{I}'' = \mathcal{I}'' \), and denote \( (w^*)_I = w^1, (w^*)_I = w^2, (w')_I = w^\ast, \) and \( (w'')_I = w^{**} \). Then we have that
\[
2 \langle w'', w^* \rangle - \| w'' \|_2^2 = 2 \langle w^2, w^{**} \rangle - \| w'' \|_2^2 \leq 2 \langle w^2, w^{**} \rangle - \| w^{**} \|_2^2 \leq 2 \| w^2 \|_2 \| w^{**} \|_2 - \| w^{**} \|_2. \tag{C.2}
\]
Let \( \| \text{supp}(w^{**}) \| = |\mathcal{I}'| = k^* \) and \( w_{2, \text{max}} = \| w^{**} \|_{\infty} \), then consequently we have \( \| w^2 \|_2 = m \cdot w_{2, \text{max}} \) 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 \( \| w^2 \|_2 \), the RHS of (C.2) is maximized when:

- **Case 1:** \( m = 1 \), if \( \| w^2 \|_2 \leq w_{2, \text{max}} \);
- **Case 2:** \( m = \frac{\| w^2 \|_2}{w_{2, \text{max}}} \), if \( w_{2, \text{max}} < \| w^2 \|_2 < \sqrt{k^*} w_{2, \text{max}} \);
- **Case 3:** \( m = \sqrt{k^*} \), if \( \| w^2 \|_2 \geq \sqrt{k^*} w_{2, \text{max}} \).

**Case 1:** If \( \| w^2 \|_2 \leq w_{2, \text{max}} \), then the RHS of (C.2) is maximized when \( m = 1 \), i.e. \( w^2 \) has only one nonzero element \( w_{2, \text{max}} \). By (C.2), we have
\[
2 \langle w'', w^* \rangle - \| w'' \|_2^2 \leq 2w_{2, \text{max}} \| w^2 \|_2 - w_{2, \text{max}}^2 \leq 2w_{2, \text{max}}^2 - w_{2, \text{max}}^2 = w_{2, \text{max}}^2. \tag{C.3}
\]
Denote \( w_{1, \text{min}} \) as the smallest element of \( w^1 \) (in magnitude), which indicates that \( |w_{1, \text{min}}| \geq |w_{2, \text{max}}| \) as \( w' \) contains the largest \( k \) entries and \( w'' \) contains the smallest \( d - k \) entries of \( w \). For \( \| w - w^* \|_2^2 \), we have that
\[
\| w - w^* \|_2^2 = \| w' - w^1 \|_2^2 + \| w'' - w^{**} \|_2^2 \\
= \| w_{(\mathcal{I}^+)C} \|_2^2 + |w_{\mathcal{I}^* - \mathcal{I}^{**}}|_2^2 + \| w^{**} \|_2^2 - 2 \langle w'', w^* \rangle - \| w'' \|_2^2 \tag{C.4}
\]
where the last inequality follows from the fact that \( w_{(\mathcal{I}^+)C} \) has \( k - k^* + k^* \) entries larger than \( w_{1, \text{min}} \) (in magnitude). Combining (C.1), (C.3) and (C.5), we have that
\[
\frac{\| w' - w^* \|_2^2 - \| w - w^* \|_2^2}{\| w - w^* \|_2^2} \leq \frac{w_{2, \text{max}}}{(k - k^* + k^*)w_{1, \text{min}}^2 - w_{2, \text{max}}^2} \leq \frac{w_{2, \text{max}}}{(k - k^* + k^*)w_{2, \text{max}}^2 - w_{2, \text{max}}^2} \leq \frac{1}{k - k^*}. \tag{C.6}
\]
Combining (C.1), (C.10) and (C.11), we have

\[ 2\langle w'', w^* \rangle - ||w''||_2^2 \leq 2\sqrt{k^{**}w_{2,\text{max}}} \cdot mw_{2,\text{max}} - w_{2,\text{max}}^2 \leq k^{**}w_{2,\text{max}}^2. \]  

(C.7)

By (C.4), we have that

\[ ||w - w^*||_2^2 \geq (k - k^* + k^{**})w_{1,\text{min}}^2 + m^2w_{2,\text{max}}^2 - w_{2,\text{max}}^2 \geq (k - k^* + k^{**})w_{1,\text{min}}^2. \]  

(C.8)

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

\[ \frac{||w' - w^*||_2^2 - ||w - w^*||_2^2}{||w - w^*||_2^2} \leq \frac{k^{**}w_{2,\text{max}}^2}{(k - k^* + k^{**})w_{1,\text{min}}^2} \leq \frac{k^{**}}{k - k^* + k^{**}}. \]  

(C.9)

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

\[ 2\langle w'', w^* \rangle - ||w''||_2^2 \leq 2\sqrt{k^{**}w_{2,\text{max}}} \cdot mw_{2,\text{max}} - w_{2,\text{max}}^2 \leq k^{**}w_{2,\text{max}}^2. \]  

By (C.4), we have

\[ ||w - w^*||_2^2 \geq (k - k^* + k^{**})w_{1,\text{min}}^2 + \gamma^2w_{2,\text{max}}^2 - \gamma \sqrt{k^{**}}w_{2,\text{max}}^2 + k^{**}w_{2,\text{max}}^2. \]  

(C.10)

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

\[ \frac{||w' - w^*||_2^2 - ||w - w^*||_2^2}{||w - w^*||_2^2} \leq \frac{2\gamma \sqrt{k^{**}}w_{2,\text{max}}^2 - k^{**}w_{2,\text{max}}^2}{(k - k^* + k^{**})w_{1,\text{min}}^2 + \gamma^2w_{2,\text{max}}^2 - \gamma \sqrt{k^{**}}w_{2,\text{max}}^2 + k^{**}w_{2,\text{max}}^2} \leq \frac{2\sqrt{k^{**}} - k^{**}}{k - k^* + 2k^{**} + 2\gamma \sqrt{k^{**}}}. \]  

(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 \( \gamma \) 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

\[ \frac{||w' - w^*||_2^2 - ||w - w^*||_2^2}{||w - w^*||_2^2} \leq \max \left\{ \frac{k^{**}}{k - k^* + k^{**}}, \frac{2\sqrt{k^{**}}}{2\sqrt{k - k^* + \frac{5}{4}k^{**}} - \sqrt{k^{**}}} \right\}. \]  

(C.13)

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

\[ \max \left\{ \frac{1}{k - k^*}, \frac{k^{**}}{k - k^* + k^{**}}, \frac{2\sqrt{k^{**}}}{2\sqrt{k - k^* + \frac{5}{4}k^{**}} - \sqrt{k^{**}}} \right\} \leq \frac{2\sqrt{k^{**}}}{2\sqrt{k - k^* + \frac{5}{4}k^{**}} - \sqrt{k^{**}}} \leq \frac{2\sqrt{k^{**}}}{2\sqrt{k - k^*} - \sqrt{k^*}} \leq \frac{2\sqrt{k^{**}}}{\sqrt{k - k^*}}, \]  

which proves the result.

**D. Proof of Lemma 4.3**

Remind that the stochastic variance reduced gradient is

\[ g^{(t)}(w^{(t)}) = \nabla f_i(w^{(t)}) - \nabla f_i(\bar{w}) + \bar{\mu}, \]  

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

It is straightforward that the stochastic variance reduced gradient (D.1) satisfies
\[
\mathbb{E}g^{(t)}_I(w^{(t)}) = \mathbb{E}\nabla f_i(w^{(t)}) - \mathbb{E}\nabla f_i(\tilde{w}) + \mu = \nabla \mathcal{F}(w^{(t)}),
\]

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

Next, we bound \( \mathbb{E}\|\nabla \mathcal{F}(w^{(t)})\|^2 \). For any \( i \in [n] \) and \( w \) with \( \text{supp}(w) \subseteq I \), consider
\[
\phi_i(w) = f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^* \rangle.
\]

Since \( \nabla \phi_i(w^*) = \nabla f_i(w^*) - \nabla f_i(w^*) = 0 \), we have that \( \phi_i(w^*) = \min_w \phi_i(w) \), which implies
\[
0 = \phi_i(w^*) \leq \min_{\eta} \phi_i(w - \eta \nabla \phi_i(w)) \leq \min_{\eta} \phi_i(w) - \eta \|\nabla \phi_i(w)\|^2 + \frac{\rho_s^+ \eta^2}{2} \|\nabla \phi_i(w)\|^2
\]
\[
= \phi_i(w) - \frac{1}{2 \rho_s} \|\nabla \phi_i(w)\|^2,
\]

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
\[
\|\nabla I f_i(w) - \nabla I f_i(w^*)\|^2 \leq \|\nabla f_i(w) - \nabla f_i(w^*)\|^2
\]
\[
\leq 2 \rho_s^+ [f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^* \rangle]
\]
\[
= 2 \rho_s^+ [f_i(w) - f_i(w^*) - \langle \nabla I f_i(w^*), w - w^* \rangle].
\]

Since the sampling of \( i \) from \( [n] \) is uniform sampling, we have from (D.3)
\[
\mathbb{E}\|\nabla I f_i(w) - \nabla I f_i(w^*)\|^2 = \frac{1}{n} \sum_{i=1}^{n} \|\nabla I f_i(w) - \nabla I f_i(w^*)\|^2
\]
\[
\leq 2 \rho_s^+ [\mathcal{F}(w) - \mathcal{F}(w^*) - \langle \nabla I \mathcal{F}(w^*), w - w^* \rangle]
\]
\[
\leq 2 \rho_s^+ [\mathcal{F}(w) - \mathcal{F}(w^*) + \|\nabla I \mathcal{F}(w^*)\|_2]\]
\[
\leq 4 \rho_s^+ [\mathcal{F}(w) - \mathcal{F}(w^*)],
\]

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

By the definition of \( g^{(t)}_I \) in (D.1), we can verify the second claim as
\[
\mathbb{E}\|g^{(t)}_I(w^{(t)})\|^2 \leq 3 \mathbb{E}\|\nabla I f_i(\tilde{w}) - \nabla I f_i(w^*)\| - \nabla I \mathcal{F}(\tilde{w}) + \nabla I \mathcal{F}(w^*)\|^2
\]
\[
\leq 3 \mathbb{E}\|\nabla I f_i(\tilde{w}) - \nabla I f_i(w^*)\|^2 + 3 \mathbb{E}\|\nabla I f_i(w^*)\|^2 + 3 \mathbb{E}\|\nabla I \mathcal{F}(w^*)\|^2
\]
\[
\leq 12 \rho_s^+ \left[ \mathcal{F}(w^{(t)}) - \mathcal{F}(w^*) + \mathcal{F}(\tilde{w}) - \mathcal{F}(w^*) \right] + 3 \|\nabla I \mathcal{F}(w^*)\|^2,
\]

where the first inequality follows from \( \|a + b + c\|_2 \leq 3\|a\|_2 + 3\|b\|_2 + 3\|c\|_2 \), the second inequality follows from \( \mathbb{E}\|x - \mathbb{E}x\|^2_2 \leq \mathbb{E}\|x\|^2_2 \) with \( \mathbb{E}\|\nabla I f_i(\tilde{w}) - \nabla I f_i(w^*)\| = \nabla I \mathcal{F}(\tilde{w}) - \nabla I \mathcal{F}(w^*) \), and the last inequality follows from (D.4).