
Efficient Stochastic Gradient Hard Thresholding

Pan Zhou* Xiao-Tong Yuan† Jiashi Feng*

* Learning & Vision Lab, National University of Singapore, Singapore

† B-DAT Lab, Nanjing University of Information Science & Technology, Nanjing, China
pzhou@u.nus.edu xtyuan@nuist.edu.cn elefjia@nus.edu.sg

Abstract

This supplementary document contains the technical proofs of convergence results and some additional numerical results of the NeurIPS’18 paper entitled “Efficient Stochastic Gradient Hard Thresholding”. It is structured as follows. The proof of the results in Section 3, including Theorems 1 and 2 and Corollary 1, is presented in Appendix B. Then Appendix C provides the proof of the results in Section 4, including Theorems 3 and Corollary 2. Next, Appendix D gives the proof of auxiliary lemmas. Finally, the detailed descriptions of datasets and more experimental results are provided in Appendix E.

A Auxiliary Lemmas

In this section, we introduce auxiliary lemmas which will be used for proving the results in the manuscript. For readability, we defer the proofs of some lemmas into Appendix D.

Lemma 1. [1] When $\Phi_k(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a vector hard thresholding operator that keeps the largest (in absolute value) k entries in \mathbf{x} and sets other entries to zero, we have that for any \mathbf{x}^* with $\|\mathbf{x}^*\|_0 = k^* < k$,

$$\|\Phi_k(\mathbf{x}) - \mathbf{x}^*\|^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}\right) \|\mathbf{x} - \mathbf{x}^*\|^2. \quad (3)$$

On the other hand, when $\Phi_k(\mathbf{x}) : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^{d_1 \times d_2}$ is a matrix hard thresholding operator that keeps the largest k top singular values of matrix \mathbf{x} and sets other singular values to zero, we have that the property (3) still holds for any \mathbf{x}^* with $\text{rank}(\mathbf{x}^*) = k^* < k$.

Lemma 2. [2] Let $\Phi_k(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a vector hard thresholding operator and $\mathbf{y} = \Phi_k(\mathbf{x})$. Then for any \mathbf{y}^* with $\|\mathbf{y}^*\|_0 = k^* < k$, we have

$$\|\mathbf{y} - \mathbf{x}\|^2 \leq \frac{d - k}{d - k^*} \|\mathbf{y}^* - \mathbf{x}\|^2.$$

On the other hand, when $\Phi_k(\mathbf{x}) : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^{d_1 \times d_2}$ is a matrix hard thresholding operator, we have that for any \mathbf{y}^* with $\text{rank}(\mathbf{y}^*) = k^* < k$,

$$\|\mathbf{y} - \mathbf{x}\|^2 \leq \frac{r - k}{r - k^*} \|\mathbf{y}^* - \mathbf{x}\|^2.$$

where $r = \text{rank}(\mathbf{x})$.

Lemma 3. [3] Assume that \mathbf{g}^t is the sampled gradient in Algorithm 1 for sparsity- or rank-constrained problem. Then the gradient variance of the gradient estimation \mathbf{g}^t can be bounded as follows

$$\mathbb{E}\|\mathbf{g}^t - \nabla f(\mathbf{x}^t)\|^2 \leq \frac{n - s_t}{n} \frac{1}{s_t} B_t,$$

where $B_t = \frac{1}{n-1} \sum_{i=1}^n \|\nabla f_i(\mathbf{x}^t) - \nabla f(\mathbf{x}^t)\|^2$ and \mathbf{x}^t denotes the variable at the t -th iteration.

Lemma 4. Assume that \mathbf{g}^t is the sampled gradient in Algorithm 1 for the sparsity- or rank-constrained problem. Then we can bound $\mathbb{E}\|\mathbf{g}_{\mathcal{I}}^t\|^2$ as follows

$$\mathbb{E}\|\mathbf{g}_{\mathcal{I}}^t\|^2 \leq \frac{3}{s_t} B_t + 6\ell_s (f(\mathbf{x}^t) - f(\mathbf{x}^*) + \langle \nabla_{\mathcal{I}} f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle) + 3\|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2,$$

where $B_t = \frac{1}{n-1} \sum_{i=1}^n \|\nabla f_i(\mathbf{x}^t) - \nabla f(\mathbf{x}^t)\|^2$, $\mathcal{I} = \text{supp}(\mathbf{x}^*) \cup \text{supp}(\mathbf{x}^t) \cup \text{supp}(\mathbf{x}^{t+1})$ and ℓ_s denotes the smooth parameter with $s = 2k + k^*$. Here k^* denotes the cardinality of the support set $\text{supp}(\mathbf{x}^*)$.

Proof. We defer the proof of Lemma 4 to Appendix D.1. \square

Lemma 5. For both vector variable \mathbf{x} in sparsity-constrained problem or matrix variable \mathbf{x} in rank-constrained problem, we have

$$\|\mathbf{x}^t - \mathbf{x}^*\|^2 \leq \frac{4}{\rho_s} (f(\mathbf{x}^*) - f(\mathbf{x}^t)) + \frac{8}{\rho_s^2} \|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2 + \frac{8}{\rho_s^2} \|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\|^2,$$

where $\mathcal{I}^t = \text{supp}(\mathbf{x}^t)$ and $\mathcal{I}^* = \text{supp}(\mathbf{x}^*)$.

Proof. See the proof of Lemma 5 in Appendix D.2. \square

Let us denote $\rho(\mathbf{A})$ the spectral radius of a square matrix \mathbf{A} , i.e., the largest (in magnitude) eigenvalue of \mathbf{A} .

Lemma 6. Let $\mathbf{A} \in \mathbb{R}^{d \times d}$. Assume that $\mu \mathbf{I} \preceq \mathbf{A} \preceq \ell \mathbf{I}$ for some $0 < \mu < \ell$. Then the following inequality holds

$$\rho \left(\begin{bmatrix} (1+\nu)\mathbf{I} - \eta\mathbf{A} & -\nu\mathbf{I} \\ \mathbf{I} & 0 \end{bmatrix} \right) \leq \max \left\{ |1 - \sqrt{\eta\mu}|, |1 - \sqrt{\eta\ell}| \right\}$$

for $\nu = \max \left\{ |1 - \sqrt{\eta\mu}|^2, |1 - \sqrt{\eta\ell}|^2 \right\}$.

Proof. We defer the proof of Lemma 6 to Appendix D.3. \square

B Proofs for Section 3

B.1 Proof of Theorem 1

Proof. Here we also first consider the sparsity-constrained problem. Assume $\mathbf{v} = \mathbf{x}^t - \eta\mathbf{g}_{\mathcal{I}}^t$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^t \cup \mathcal{I}^{t+1}$, where $\mathcal{I}^* = \text{supp}(\mathbf{x}^*)$, $\mathcal{I}^t = \text{supp}(\mathbf{x}^t)$ and $\mathcal{I}^{t+1} = \text{supp}(\mathbf{x}^{t+1})$. Then we have

$$\begin{aligned} & \mathbb{E}\|\mathbf{v} - \mathbf{x}^*\|^2 \\ &= \mathbb{E}\|\mathbf{x}^t - \eta\mathbf{g}_{\mathcal{I}}^t - \mathbf{x}^*\|^2 \\ &= \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + \eta^2 \mathbb{E}\|\mathbf{g}_{\mathcal{I}}^t\|^2 - 2\eta \mathbb{E}\langle \mathbf{x}^t - \mathbf{x}^*, \mathbf{g}_{\mathcal{I}}^t \rangle \\ &\stackrel{\textcircled{1}}{\leq} \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + \eta^2 \mathbb{E}\|\mathbf{g}_{\mathcal{I}}^t\|^2 - 2\eta \mathbb{E}(f(\mathbf{x}^t) - f(\mathbf{x}^*)) \\ &\stackrel{\textcircled{2}}{\leq} \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + \eta^2 \left[\frac{3}{s_t} B_t + 6\ell_s (f(\mathbf{x}^t) - f(\mathbf{x}^*) + \langle \nabla_{\mathcal{I}} f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle) + 3\|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2 \right] \\ &\quad - 2\eta \mathbb{E}[f(\mathbf{x}^t) - f(\mathbf{x}^*)] \\ &\stackrel{\textcircled{3}}{=} \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + 2\eta(3\eta\ell_s - 1) [f(\mathbf{x}^t) - f(\mathbf{x}^*)] + 6\eta^2 \ell_s \langle \nabla_{\mathcal{I}} f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle + \frac{3\eta^2}{s_t} B_t + 3\eta^2 \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2, \end{aligned} \tag{4}$$

where $\textcircled{1}$ holds by using $\|(\mathbf{x}^t - \mathbf{x}^*)_{\mathcal{I}^c}\| = 0$ and the convexity of $f(\mathbf{x})$, namely, $f(\mathbf{x}^*) - f(\mathbf{x}^t) \geq \langle \mathbf{x}^* - \mathbf{x}^t, \nabla f(\mathbf{x}^t) \rangle = \langle \mathbf{x}^* - \mathbf{x}^t, \nabla_{\mathcal{I}} f(\mathbf{x}^t) \rangle$, and $\textcircled{2}$ holds by using Lemma 4, and $\textcircled{3}$ holds by using $\|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|_2^2 \leq \|\nabla_{\tilde{\mathcal{I}}} f(\mathbf{x}^*)\|_2^2$ where $\tilde{\mathcal{I}} = \text{supp}(\Phi_{2k}(\nabla f(\mathbf{x}^*))) \cup \text{supp}(\mathbf{x}^*)$.

Next, we apply Lemma 1 and obtain

$$\begin{aligned} \mathbb{E}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 &\leq \alpha \mathbb{E}\|\mathbf{v} - \mathbf{x}^*\|^2 \\ &\leq \alpha \left[\mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + 2\eta(3\eta\ell - 1) [f(\mathbf{x}^t) - f(\mathbf{x}^*)] + 6\eta^2\ell_s \langle \nabla_{\mathcal{I}} f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle \right] \\ &\quad + \alpha \left[\frac{3\eta^2}{s_t} B_t + 3\eta^2 \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2 \right], \end{aligned} \quad (5)$$

where $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. On the other hand, we have

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \geq \langle \nabla f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle + \frac{\rho_s}{2} \|\mathbf{x}^t - \mathbf{x}^*\|^2. \quad (6)$$

By setting $\eta \leq \frac{1}{3\ell_s}$ and plugging Eqn. (6) into Eqn. (5), we can obtain

$$\begin{aligned} &\mathbb{E}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 \\ &\leq \alpha [1 + \rho\eta(3\eta\ell_s - 1)] \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + 2\alpha\eta(6\eta\ell_s - 1) \langle \nabla f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle + \frac{3\alpha\eta^2}{s_t} B_t + 3\alpha\eta^2 \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2 \\ &\stackrel{\textcircled{1}}{\leq} \alpha [1 + \rho\eta(3\eta\ell_s - 1)] \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + \frac{3\alpha\eta^2}{s_t} B + 3\alpha\eta^2 \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2, \end{aligned} \quad (7)$$

where $\textcircled{1}$ holds since we set $\eta = \frac{1}{6\ell_s}$ and $B = \max_t B_t$. Then we let

$$\beta := \alpha [1 + \rho\eta(3\eta\ell_s - 1)] = \alpha \left(1 - \frac{1}{12\kappa_s} \right) \in (0, 1) \quad \text{where} \quad \kappa_s = \frac{\ell_s}{\rho_s}.$$

We further set $s_t = \tau/\omega^t$ and assume that τ is large enough such that

$$\gamma := \frac{3\alpha\eta^2 B}{\tau} = \frac{\alpha B}{12\tau\ell_s^2} \leq \delta \|\mathbf{x}^0 - \mathbf{x}^*\|^2, \quad (8)$$

where δ is a positive constant and will be discussed later. Now we use mathematical induction to prove

$$\mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 \leq \theta^t \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2} \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2, \quad (\forall t), \quad (9)$$

where $\theta < 1$ is a constant and will be given below.

Obviously, when $t = 0$, Eqn. (9) holds. Now assume that for all $k \leq t$, Eqn. (9) holds. Then for $k = t + 1$, we have

$$\begin{aligned} \mathbb{E}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 &\leq \beta \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + \delta \omega^t \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{\alpha}{12\ell_s^2} \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2 \\ &\leq (\beta\theta^t + \delta\omega^t) \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{\alpha}{12\ell_s^2} \left[\frac{\beta}{1-\beta} + 1 \right] \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2 \\ &\stackrel{\textcircled{1}}{\leq} \theta^{t+1} \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2} \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2 \end{aligned}$$

where $\textcircled{1}$ holds since we let

$$\omega = \theta = \beta + \delta. \quad (10)$$

This means that if Eqn. (8) holds, then Eqn. (9) always holds. So the conclusion holds.

Finally, we discuss the value of k such that $\beta \in (0, 1)$. It is easily to check that if $k \geq \left(\frac{6400}{9}\kappa_s^2 + 1\right)k^*$, where $\kappa_s = \ell_s/\rho_s$, then we have $\beta = \alpha \left(1 - \frac{1}{12\kappa_s}\right) < 1 - \frac{1}{120\kappa_s} - \frac{1}{160\kappa_s^2}$.

So we just let

$$k \geq (712\kappa_s^2 + 1)k^*, \quad \text{and} \quad \beta = \alpha \left(1 - \frac{1}{12\kappa_s}\right) \leq 1 - \frac{1}{120\kappa_s}.$$

Then let $\tau \geq \frac{40\alpha B}{3\ell_s \rho_s \|\mathbf{x}^0 - \mathbf{x}^*\|^2}$. We have

$$\omega = \theta = \beta + \delta \leq 1 - \frac{1}{120\kappa_s} + \frac{3}{480\kappa_s} = 1 - \frac{1}{480\kappa_s}.$$

Therefore, we have

$$\mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 \leq \theta^t \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2} \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2, \quad (\forall t), \quad (11)$$

where $\beta = \alpha \left(1 - \frac{1}{12\kappa_s}\right) \leq 1 - \frac{1}{120\kappa_s}$ and $\theta \leq 1 - \frac{1}{480\kappa_s}$. Then we can further derive

$$\begin{aligned} \mathbb{E}\|\mathbf{x}^T - \mathbf{x}^*\| &\leq \sqrt{\mathbb{E}\|\mathbf{x}^T - \mathbf{x}^*\|^2} \leq \sqrt{\theta^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2} \|\nabla_{\tilde{\mathcal{I}}} f(\mathbf{x}^*)\|^2} \\ &\leq \theta^{\frac{T}{2}} \|\mathbf{x}^0 - \mathbf{x}^*\| + \frac{\sqrt{\alpha}}{\ell_s \sqrt{12(1-\beta)}} \|\nabla_{\tilde{\mathcal{I}}} f(\mathbf{x}^*)\|, \end{aligned}$$

where $\theta \leq 1 - \frac{1}{480\kappa_s}$. So we obtain the result on sparsity-constrained problem.

For rank-constrained problem, its proof is very similar to the proof for sparsity-constrained problem. Firstly, assume that the skinny SVDs of \mathbf{x}^t , \mathbf{x}^{t+1} , \mathbf{x}^* are respectively $\mathbf{x}^t = \mathbf{U}_t \boldsymbol{\Sigma}_t \mathbf{V}_t^T$, $\mathbf{x}^{t+1} = \mathbf{U}_{t+1} \boldsymbol{\Sigma}_{t+1} \mathbf{V}_{t+1}^T$ and $\mathbf{x}^* = \mathbf{U}_* \boldsymbol{\Sigma}_* \mathbf{V}_*^T$. Then we have $\mathbf{x}^t = \mathbf{U}_t \mathbf{U}_t^T \mathbf{x}^t$, $\mathbf{x}^{t+1} = \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{x}^{t+1}$ and $\mathbf{x}^* = \mathbf{U}_* \mathbf{U}_*^T \mathbf{x}^*$. We further define the projection operation $\mathcal{P}_{\mathcal{U}}(\mathbf{x}) = \mathbf{U} \mathbf{U}^T \mathbf{x}$ which projects \mathbf{x} in the subspace spanned by \mathbf{U} . Then let $\mathbf{v} = \mathbf{x}^t - \eta \mathbf{g}_{\mathcal{I}}^t$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^t \cup \mathcal{I}^{t+1}$, where $\mathcal{I}^* = \text{supp}(\mathbf{x}^*)$, $\mathcal{I}^t = \text{supp}(\mathbf{x}^t)$ and $\mathcal{I}^{t+1} = \text{supp}(\mathbf{x}^{t+1})$. The notation $\mathbf{x}_{\mathcal{I}^t}^t = \mathbf{U}_t \mathbf{U}_t^T \mathbf{x}^t = \mathcal{P}_{\mathcal{U}_t}(\mathbf{x}^t)$. For the specifical notation $\mathbf{g}_{\mathcal{I}}^t$, it denotes $\mathbf{g}_{\mathcal{I}}^t = \mathcal{P}_{\mathcal{I}}(\mathbf{g}^t) = (\mathbf{U}_t \mathbf{U}_t^T + \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T + \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T + \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T) \mathbf{g}^t$ where $\mathcal{P}_{\mathcal{I}}$ is a projection.

Then we prove the result similar to Eqn. (4) as follows:

$$\begin{aligned} &\mathbb{E}\|\mathbf{v} - \mathbf{x}^*\|^2 \\ &= \mathbb{E}\|\mathbf{x}^t - \eta \mathbf{g}_{\mathcal{I}}^t - \mathbf{x}^*\|^2 \\ &= \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + \eta^2 \mathbb{E}\|\mathbf{g}_{\mathcal{I}}^t\|^2 - 2\eta \mathbb{E}\langle \mathbf{x}^t - \mathbf{x}^*, \mathbf{g}_{\mathcal{I}}^t \rangle \\ &\stackrel{\textcircled{1}}{\leq} \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + \eta^2 \mathbb{E}\|\mathbf{g}_{\mathcal{I}}^t\|^2 - 2\eta \mathbb{E}(f(\mathbf{x}^t) - f(\mathbf{x}^*)) \\ &\stackrel{\textcircled{2}}{\leq} \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + \eta^2 \left[\frac{3}{s_t} B_t + 6\ell_s (f(\mathbf{x}^t) - f(\mathbf{x}^*) + \langle \nabla_{\mathcal{I}} f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle) + 3\|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2 \right] \\ &\quad - 2\eta \mathbb{E}[f(\mathbf{x}^t) - f(\mathbf{x}^*)] \\ &\stackrel{\textcircled{3}}{=} \mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + 2\eta(3\eta\ell_s - 1)[f(\mathbf{x}^t) - f(\mathbf{x}^*)] + 6\eta^2 \ell_s \langle \nabla_{\mathcal{I}} f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle + \frac{3\eta^2}{s_t} B_t + 3\eta^2 \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2, \end{aligned}$$

where $\textcircled{1}$ holds since we have $\mathcal{P}_{\mathcal{I}}(\mathbf{x}^t - \mathbf{x}^*) = (\mathbf{U}_t \mathbf{U}_t^T + \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T + \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T + \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T)(\mathbf{U}_t \mathbf{U}_t^T \mathbf{x}^t - \mathbf{U}_* \mathbf{U}_*^T \mathbf{x}^*) = \mathbf{x}^t - \mathbf{x}^*$ which gives $\|(\mathbf{x}^t - \mathbf{x}^*)_{\mathcal{I}^c}\| = \|(\mathbf{I} - \mathcal{P}_{\mathcal{I}})(\mathbf{x}^t - \mathbf{x}^*)\| = 0$ and $\mathbb{E}\langle \mathbf{x}^t - \mathbf{x}^*, \mathbf{g}_{\mathcal{I}}^t \rangle = \mathbb{E}\langle \mathbf{x}^t - \mathbf{x}^*, \mathcal{P}_{\mathcal{I}}(\mathbf{g}^t) \rangle = \mathbb{E}\langle \mathcal{P}_{\mathcal{I}}(\mathbf{x}^t - \mathbf{x}^*), \mathcal{P}_{\mathcal{I}}(\mathbf{g}^t) \rangle = \mathbb{E}\langle \mathbf{x}^t - \mathbf{x}^*, \mathbf{g}^t \rangle = \mathbb{E}\langle \mathbf{x}^t - \mathbf{x}^*, \nabla f(\mathbf{x}^t) \rangle \leq f(\mathbf{x}^*) - f(\mathbf{x}^t)$. $\textcircled{2}$ still uses the results in Lemma 4. $\textcircled{3}$ holds by using $\|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|_2^2 \leq \|\nabla_{\tilde{\mathcal{I}}} f(\mathbf{x}^*)\|_2^2$ where $\tilde{\mathcal{I}} = \text{supp}(\Phi_{2k}(\nabla f(\mathbf{x}^*))) \cup \text{supp}(\mathbf{x}^*)$.

Before we apply Lemma 1 to obtain similar result in Eqn. (5). We first establish

$$\begin{aligned} \Phi_k(\mathbf{v}) &= \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T (\mathbf{x}^t - \eta \mathbf{g}_{\mathcal{I}}^t) = \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{x}^t - \eta \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T [(\mathbf{U}_t \mathbf{U}_t^T + \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T + \mathbf{U}_* \mathbf{U}_*^T \\ &\quad - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T + \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T)] \mathbf{g}^t \\ &= \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{x}^t - \eta \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{g}^t = \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T (\mathbf{x}^t - \eta \mathbf{g}^t) = \Phi_k(\mathbf{x}^t - \eta \mathbf{g}^t) = \mathbf{x}^{t+1}. \end{aligned} \quad (12)$$

Then we apply Lemma 1 to establish

$$\begin{aligned}\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 &= \|\Phi_k(\mathbf{v}) - \mathbf{x}^*\|^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}\right) \|\mathbf{v} - \mathbf{x}^*\|^2 \\ &\leq \alpha \left[\mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 + 2\eta(3\eta\ell - 1) [f(\mathbf{x}^t) - f(\mathbf{x}^*)] + 6\eta^2\ell_s \langle \nabla_{\mathcal{I}} f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle \right] \\ &\quad + \alpha \left[\frac{3\eta^2}{s_t} B_t + 3\eta^2 \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2 \right],\end{aligned}$$

where $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$.

Next by using Assumption 1 in which we have $f(\mathbf{x}^t) - f(\mathbf{x}^*) \geq \langle \nabla f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle + \frac{\rho_s}{2} \|\mathbf{x}^t - \mathbf{x}^*\|^2$, we can establish the result in Eqn. (7). Finally, since the deduction after Eqn. (7) does not rely on the rank-constrained problem, we can just follow the above proof and obtain the desired result on rank-constrained problem. The proof is completed. \square

B.2 Proof of Corollary 1

Proof. Here we use the result in Eqn. (11) in Section B.1:

$$\mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 \leq \theta^t \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2} \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2, \quad (\forall t),$$

where $\theta \leq 1 - \frac{1}{480\kappa_s}$ and $\beta = \alpha \left(1 - \frac{1}{12\kappa_s}\right)$. Then to achieve $\theta^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 \leq \epsilon$, we have

$$T \geq \log_{1/\theta} \left(\frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\epsilon} \right)$$

In this way, we have

$$\begin{aligned}\mathbb{E}\|\mathbf{x}^T - \mathbf{x}^*\|_2 &\leq \sqrt{\mathbb{E}\|\mathbf{x}^T - \mathbf{x}^*\|_2^2} \leq \sqrt{\theta^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \frac{\alpha}{12(1-\beta)\ell_s^2} \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2} \\ &\leq \sqrt{\epsilon + \frac{\alpha}{12(1-\beta)\ell_s^2} \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2} \leq \sqrt{\epsilon} + \frac{\sqrt{\alpha}}{\ell_s \sqrt{12(1-\beta)}} \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|.\end{aligned}$$

Therefore, the IFO complexity is

$$\begin{aligned}\tau \left[1 + \frac{1}{\omega} + \dots + \frac{1}{\omega^{T-1}} \right] &= \tau \frac{(1/\omega)^{\log_{1/\theta} \left(\frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\epsilon} \right)} - 1}{1/\omega - 1} \stackrel{\textcircled{1}}{=} \frac{\tau}{1/\omega - 1} \left[\frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\epsilon} - 1 \right] \\ &\leq \frac{6400\alpha B}{\rho_s \ell_s} \cdot \frac{\kappa_s}{\epsilon} \stackrel{\textcircled{3}}{=} \mathcal{O} \left(\frac{\kappa_s}{\epsilon} \right),\end{aligned}$$

where $\textcircled{1}$ uses $\omega = \theta = 1 - \frac{1}{480\kappa_s}$; $\textcircled{2}$ uses $\tau \geq \frac{40\alpha B}{3\rho_s \ell_s \|\mathbf{x}^0 - \mathbf{x}^*\|^2}$; $\textcircled{3}$ holds since (1) the parameter ρ_s is the strong convex parameter at sparsity/low-rank level s and thus is not very small since s is much smaller than the feature dimension, (2) we can always scale the problem such that ρ_s is not small. Notice, such a scale does not affect the ratio B/ℓ_s since they always scale at the same order. Thus, we have the IFO complexity $\mathcal{O} \left(\frac{\kappa_s}{\epsilon} \right)$.

On the other hand, we have

$$\log_{1/\theta} \left(\frac{1}{\epsilon} \right) = \frac{\log \left(\frac{1}{\epsilon} \right)}{\log \left(\frac{1}{\theta} \right)} = \frac{\log \left(\frac{1}{\epsilon} \right)}{\log \left(1 + \frac{1}{480\kappa_s - 1} \right)} = \frac{\log \left(\frac{1}{\epsilon} \right)}{\log \left(1 + \frac{1}{480\kappa_s - 1} \right)} \stackrel{\textcircled{1}}{\leq} \mathcal{O} \left(\kappa_s \log \left(\frac{1}{\epsilon} \right) \right),$$

where $\textcircled{1}$ holds since we have $\log(1+x) \geq \log(2) \cdot x$ for $x \in [0, 1]$. The proof is completed. \square

B.3 Proof of Theorem 2

Proof. We first prove the result for sparsity-constrained problem. In this case, the variable \mathbf{x} is vector. Let $\mathcal{I} = \mathcal{I}^{t+1} \cup \mathcal{I}^t \cup \mathcal{I}^*$, where $\mathcal{I}^* = \text{supp}(\mathbf{x}^*)$, $\mathcal{I}^t = \text{supp}(\mathbf{x}^t)$ and $\mathcal{I}^{t+1} = \text{supp}(\mathbf{x}^{t+1})$. Recall

that $\mathbf{x}^{t+1} = \Phi_k(\mathbf{x}^t - \eta \mathbf{g}^t)$. Then we have

$$\begin{aligned}
& f(\mathbf{x}^{t+1}) \\
& \leq f(\mathbf{x}^t) + \langle \nabla f(\mathbf{x}^t), \mathbf{x}^{t+1} - \mathbf{x}^t \rangle + \frac{\ell_s}{2} \|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \\
& \leq f(\mathbf{x}^t) + \langle \mathbf{g}^t, \mathbf{x}^{t+1} - \mathbf{x}^t \rangle + \frac{\ell_s}{2} \|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + \|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\| \|\mathbf{x}^{t+1} - \mathbf{x}^t\| \\
& = f(\mathbf{x}^t) + \frac{1}{2\eta} \|\mathbf{x}_{\mathcal{I}}^{t+1} - \mathbf{x}_{\mathcal{I}}^t + \eta \mathbf{g}_{\mathcal{I}}^t\|^2 - \frac{\eta \|\mathbf{g}_{\mathcal{I}}^t\|^2}{2} - \frac{1-\eta\ell_s}{2\eta} \|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + \|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\| \|\mathbf{x}^{t+1} - \mathbf{x}^t\| \\
& = f(\mathbf{x}^t) + \frac{1}{2\eta} \|\mathbf{x}_{\mathcal{I}}^{t+1} - \mathbf{x}_{\mathcal{I}}^t + \eta \mathbf{g}_{\mathcal{I}}^t\|^2 - \frac{\eta \|\mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2}{2} - \frac{\eta \|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2}{2} - \frac{1-\eta\ell_s}{2\eta} \|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \\
& \quad + \|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\| \|\mathbf{x}^{t+1} - \mathbf{x}^t\| \\
& = f(\mathbf{x}^t) + \frac{1}{2\eta} \left(\|\mathbf{x}_{\mathcal{I}}^{t+1} - \mathbf{x}_{\mathcal{I}}^t + \eta \mathbf{g}_{\mathcal{I}}^t\|^2 - \eta^2 \|\mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 \right) - \frac{\eta \|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2}{2} - \frac{1-\eta\ell_s}{2\eta} \|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \\
& \quad + \|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\| \|\mathbf{x}^{t+1} - \mathbf{x}^t\|.
\end{aligned} \tag{13}$$

Now we bound the second term $\|\mathbf{x}_{\mathcal{I}}^{t+1} - \mathbf{x}_{\mathcal{I}}^t + \eta \mathbf{g}_{\mathcal{I}}^t\|^2 - \eta^2 \|\mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2$. Here we adopt similar strategy in [2]. Since we have $\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*) = \mathcal{I}^{t+1} \setminus (\mathcal{I}^t \cup \mathcal{I}^*) \subseteq \mathcal{I}^{t+1}$, then we can establish $\mathbf{x}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^{t+1} = \mathbf{x}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t - \eta \mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t$. On the other hand, we have $\mathbf{x}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t = 0$ which further yields $\mathbf{x}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^{t+1} = -\eta \mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t$. Next, we choose a set $\mathcal{R} \subseteq \mathcal{I}^t \setminus \mathcal{I}^{t+1}$ such that $|\mathcal{R}| = |\mathcal{I}^{t+1} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)|$. We can find such a set \mathcal{R} because we have $|\mathcal{I}^{t+1} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)| = |\mathcal{I}^t \setminus \mathcal{I}^{t+1}| - |(\mathcal{I}^{t+1} \cap \mathcal{I}^*) \setminus \mathcal{I}^t|$. Besides, since $\mathbf{x}^{t+1} = \Phi_k(\mathbf{x}^t - \eta \mathbf{g}^t)$, we can establish:

$$\eta^2 \|\mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 = \|\mathbf{x}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^{t+1}\|^2 \geq \|\mathbf{x}_{\mathcal{R}}^t - \eta \mathbf{g}_{\mathcal{R}}^t\|^2. \tag{14}$$

Then combining Eqn. (14) and the fact that $\mathbf{x}_{\mathcal{R}}^{t+1} = 0$, we have

$$\begin{aligned}
\|\mathbf{x}_{\mathcal{I}}^{t+1} - \mathbf{x}_{\mathcal{I}}^t + \eta \mathbf{g}_{\mathcal{I}}^t\|^2 - \eta^2 \|\mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 & \leq \|\mathbf{x}_{\mathcal{I}}^{t+1} - \mathbf{x}_{\mathcal{I}}^t + \eta \mathbf{g}_{\mathcal{I}}^t\|^2 - \|\mathbf{x}_{\mathcal{R}}^{t+1} - \mathbf{x}_{\mathcal{R}}^t + \eta \mathbf{g}_{\mathcal{R}}^t\|^2 \\
& = \|\mathbf{x}_{\mathcal{I} \setminus \mathcal{R}}^{t+1} - \mathbf{x}_{\mathcal{I} \setminus \mathcal{R}}^t + \eta \mathbf{g}_{\mathcal{I} \setminus \mathcal{R}}^t\|^2.
\end{aligned} \tag{15}$$

Next, we bound the size of $\mathcal{I} \setminus \mathcal{R}$ as $|\mathcal{I} \setminus \mathcal{R}| \leq |\mathcal{I}^{t+1}| + |(\mathcal{I}^t \setminus \mathcal{I}^{t+1}) \setminus \mathcal{R}| + |\mathcal{I}^*| \leq k + |(\mathcal{I}^{t+1} \cap \mathcal{I}^*) \setminus \mathcal{I}^t| + k^* \leq k + 2k^*$. Also, since $\mathcal{I}^{t+1} \subseteq (\mathcal{I} \setminus \mathcal{R})$, we have $\mathbf{x}_{\mathcal{I} \setminus \mathcal{R}}^{t+1} = \Phi_k(\mathbf{x}_{\mathcal{I} \setminus \mathcal{R}}^t - \eta \mathbf{g}_{\mathcal{I} \setminus \mathcal{R}}^t)$. By combining Eqn. (15) and Lemma 2, we can obtain

$$\begin{aligned}
& \|\mathbf{x}_{\mathcal{I}}^{t+1} - \mathbf{x}_{\mathcal{I}}^t + \eta \mathbf{g}_{\mathcal{I}}^t\|^2 - \eta^2 \|\mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 \\
& \leq \frac{2k^*}{k+k^*} \|\mathbf{x}_{\mathcal{I} \setminus \mathcal{R}}^* - \mathbf{x}_{\mathcal{I} \setminus \mathcal{R}}^t + \eta \mathbf{g}_{\mathcal{I} \setminus \mathcal{R}}^t\|^2 \\
& \leq \frac{2k^*}{k+k^*} \|\mathbf{x}_{\mathcal{I}}^* - \mathbf{x}_{\mathcal{I}}^t + \eta \mathbf{g}_{\mathcal{I}}^t\|^2 \\
& = \frac{2k^*}{k+k^*} (\|\mathbf{x}^* - \mathbf{x}^t\|^2 + 2\eta \langle \mathbf{g}^t, \mathbf{x}^* - \mathbf{x}^t \rangle + \eta^2 \|\mathbf{g}_{\mathcal{I}}^t\|^2) \\
& \stackrel{\textcircled{1}}{=} \frac{2k^*}{k+k^*} (\|\mathbf{x}^* - \mathbf{x}^t\|^2 + 2\eta \langle \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle + \eta^2 \|\mathbf{g}_{\mathcal{I}}^t\|^2) + \xi \langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle \\
& \leq \frac{2k^*}{k+k^*} \left[\|\mathbf{x}^* - \mathbf{x}^t\|^2 + 2\eta \left(f(\mathbf{x}^*) - f(\mathbf{x}^t) - \frac{\rho_s}{2} \|\mathbf{x}^* - \mathbf{x}^t\|^2 \right) + \eta^2 \|\mathbf{g}_{\mathcal{I}}^t\|^2 \right] + \xi \langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle \\
& = \frac{4\eta k^*}{k+k^*} (f(\mathbf{x}^*) - f(\mathbf{x}^t)) + \frac{2(1-\eta\rho_s)k^*}{k+k^*} \|\mathbf{x}^* - \mathbf{x}^t\|^2 + \frac{2\eta^2 k^*}{k+k^*} \|\mathbf{g}_{\mathcal{I}}^t\|^2 + \xi \langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle \\
& = \frac{4\eta k^*}{k+k^*} (f(\mathbf{x}^*) - f(\mathbf{x}^t)) + \frac{2(1-\eta\rho_s)k^*}{k+k^*} \|\mathbf{x}^* - \mathbf{x}^t\|^2 + \frac{2\eta^2 k^*}{k+k^*} \|\mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 + \frac{2\eta^2 k^*}{k+k^*} \|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2 \\
& \quad + \xi \langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle,
\end{aligned}$$

where $\xi = \frac{4\eta k^*}{k+k^*}$ in ①. By substituting the above inequality into Eqn. (13), we can further obtain

$$\begin{aligned}
& f(\mathbf{x}^{t+1}) \\
& \leq f(\mathbf{x}^t) + \frac{2k^*}{k+k^*}(f(\mathbf{x}^*) - f(\mathbf{x}^t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|\mathbf{x}^* - \mathbf{x}^t\|^2 + \frac{\eta k^*}{k+k^*}\|\mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|^2 \\
& \quad + \left(\frac{\eta k^*}{k+k^*} - \frac{\eta}{2}\right)\|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2 - \frac{1-\eta\ell_s}{2\eta}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + \frac{\xi}{2\eta}\langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle \\
& \quad + \|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\|\|\mathbf{x}^{t+1} - \mathbf{x}^t\| \\
& \stackrel{\textcircled{1}}{\leq} f(\mathbf{x}^t) + \frac{2k^*}{k+k^*}(f(\mathbf{x}^*) - f(\mathbf{x}^t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|\mathbf{x}^* - \mathbf{x}^t\|^2 - \left[\frac{1-\eta\ell_s}{2\eta} - \frac{k^*}{\eta(k+k^*)}\right]\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \\
& \quad + \left(\frac{\eta k^*}{k+k^*} - \frac{\eta}{2}\right)\|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2 + \frac{\xi}{2\eta}\langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle + \|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\|\|\mathbf{x}^{t+1} - \mathbf{x}^t\| \\
& \stackrel{\textcircled{2}}{\leq} f(\mathbf{x}^t) + \frac{2k^*}{k+k^*}(f(\mathbf{x}^*) - f(\mathbf{x}^t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|\mathbf{x}^* - \mathbf{x}^t\|^2 - \left(\frac{\eta}{2} - \frac{\eta k^*}{k+k^*}\right)\|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2 \\
& \quad + \frac{\xi}{2\eta}\langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle + \frac{\eta(k+k^*)}{2((1-\eta\ell_s)k - (1+\eta\ell_s)k^*)}\|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\|^2,
\end{aligned}$$

where in ① we used $\|\mathbf{x}^{t+1} - \mathbf{x}^t\| \geq \eta\|\mathbf{g}_{\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}^t\|$, and in ② we have used the basic inequality $ab \leq \frac{a^2}{4c} + cb^2, \forall c > 0$. By invoking Lemma 5 in the above we further get

$$\begin{aligned}
& f(\mathbf{x}^{t+1}) \\
& \leq f(\mathbf{x}^t) + \frac{2k^*}{k+k^*} \left(1 + \frac{2(1-\eta\rho_s)}{\eta\rho_s}\right) (f(\mathbf{x}^*) - f(\mathbf{x}^t)) - \left(\frac{\eta}{2} - \frac{(\eta^2\rho_s^2 + 8(1-\eta\rho_s))k^*}{\eta\rho_s^2(k+k^*)}\right)\|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2 \\
& \quad + \frac{\xi}{2\eta}\langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle + \left(\frac{\eta(k+k^*)}{2((1-\eta\ell_s)k - (1+\eta\ell_s)k^*)} + \frac{8(1-\eta\rho_s)k^*}{\eta\rho_s^2(k+k^*)}\right)\|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\|^2.
\end{aligned}$$

Let us now consider $\eta = \frac{1}{2\ell_s}$ in the above inequality, which leads to

$$\begin{aligned}
f(\mathbf{x}^{t+1}) & \leq f(\mathbf{x}^t) + \frac{2(4\ell_s - \rho_s)k^*}{\rho_s(k+k^*)}(f(\mathbf{x}^*) - f(\mathbf{x}^t)) - \left(\frac{1}{4\ell_s} - \frac{(\rho_s^2 - 16\rho_s\ell_s + 32\ell_s^2)k^*}{2\ell_s\rho_s^2(k+k^*)}\right)\|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2 \\
& \quad + \ell_s\xi\langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle + \left(\frac{k+k^*}{2\ell_s(k-3k^*)} + \frac{8(2\ell_s - \rho_s)k^*}{\rho_s^2(k+k^*)}\right)\|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\|^2.
\end{aligned}$$

Since $k \geq \left(1 + \frac{64\ell_s^2}{\rho_s^2}\right)k^*$, with algebra manipulation we can further show that

$$\begin{aligned}
f(\mathbf{x}^{t+1}) & \leq f(\mathbf{x}^t) + \frac{\rho_s}{8\ell_s}(f(\mathbf{x}^*) - f(\mathbf{x}^t)) + \ell_s\xi\langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle + \left(\frac{5}{4\ell_s} + \frac{8\ell_s}{\rho_s^2}\right)\|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\|^2 \\
& \leq f(\mathbf{x}^t) + \frac{\rho_s}{8\ell_s}(f(\mathbf{x}^*) - f(\mathbf{x}^t)) + \ell_s\xi\langle \mathbf{g}^t - \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle + \frac{37\ell_s}{4\rho_s^2}\|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\|^2.
\end{aligned}$$

Taking expectation (conditioned on \mathbf{x}^t) on both sides of the above we arrive at

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}^{t+1}) \mid \mathbf{x}^t] & \leq f(\mathbf{x}^t) + \frac{\rho_s}{8\ell_s}(f(\mathbf{x}^*) - f(\mathbf{x}^t)) + \frac{37\ell_s}{4\rho_s^2}\mathbb{E}[\|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\|^2 \mid \mathbf{x}^t] \\
& \stackrel{\textcircled{1}}{\leq} f(\mathbf{x}^t) + \frac{\rho_s}{8\ell_s}(f(\mathbf{x}^*) - f(\mathbf{x}^t)) + \frac{37\ell_s}{4\rho_s^2 s_t} B,
\end{aligned}$$

where ① uses Lemma 3, in which $B = \max_t B_t$ and $B_t = \frac{1}{n-1} \sum_{i=1}^n \|\nabla f_i(\mathbf{x}^t) - \nabla f(\mathbf{x}^t)\|^2$. By further taking expectation on \mathbf{x}^t we obtain

$$\mathbb{E}[f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*)] \leq \left(1 - \frac{\rho_s}{8\ell_s}\right)\mathbb{E}[f(\mathbf{x}^t) - f(\mathbf{x}^*)] + \frac{37\ell_s}{4\rho_s^2 s_t} B.$$

We further set $\beta = 1 - \frac{1}{8\kappa_s}$ and $s_t = \tau/\omega^t$, and then assume that τ is large enough such that

$$\gamma := \frac{37\ell_s B}{4\tau\rho_s^2} \leq \delta[f(\mathbf{x}^0) - f(\mathbf{x}^*)], \quad (16)$$

where δ is a positive constant and will be discussed later. Now we use mathematical induction to prove

$$\mathbb{E}[f(\mathbf{x}^t) - f(\mathbf{x}^*)] \leq \theta^t[f(\mathbf{x}^0) - f(\mathbf{x}^*)], \quad (\forall t), \quad (17)$$

where $\theta < 1$ is a constant and will be given below.

Obviously, when $t = 0$, Eqn. (17) holds. Now assume that for all $k \leq t$, Eqn. (17) holds. Then for $k = t + 1$, we have

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*)] &\leq \beta \mathbb{E}[f(\mathbf{x}^t) - f(\mathbf{x}^*)] + \delta \omega^t [f(\mathbf{x}^0) - f(\mathbf{x}^*)] \\ &\leq (\beta \theta^t + \delta \omega^t) [f(\mathbf{x}^0) - f(\mathbf{x}^*)] \stackrel{\textcircled{1}}{\leq} \theta^{t+1} [f(\mathbf{x}^0) - f(\mathbf{x}^*)], \end{aligned}$$

where $\textcircled{1}$ holds since we let

$$\omega = \theta = \beta + \delta. \quad (18)$$

This means that if Eqn. (16) holds, then Eqn. (17) always holds. So the conclusion holds.

Then let $\tau \geq \frac{148B\kappa_s^2}{\rho_s[f(\mathbf{x}^0) - f(\mathbf{x}^*)]}$ which gives $\delta \leq \frac{1}{16\kappa_s}$. We have

$$\omega = \theta = \beta + \delta \leq 1 - \frac{1}{8\kappa_s} + \frac{1}{16\kappa_s} = 1 - \frac{1}{16\kappa_s}.$$

Therefore, we have

$$\mathbb{E}[f(\mathbf{x}^t) - f(\mathbf{x}^*)] \leq \left(1 - \frac{1}{16\kappa_s}\right)^t [f(\mathbf{x}^0) - f(\mathbf{x}^*)].$$

The proof is completed. \square

C Proofs for Section 4

C.1 Proofs of Theorem 3

Proof. Here we also first consider the sparsity-constrained problem. Let us consider $\mathcal{I} = \mathcal{I}^{t+1} \cup \mathcal{I}^t \cup \mathcal{I}^{t-1} \cup \mathcal{I}^*$ and $\mathbf{y}^{t+1} = \mathbf{x}^t - \eta \mathbf{g}_{\mathcal{I}}^t + \nu(\mathbf{x}^t - \mathbf{x}^{t-1})$. Since $f(\mathbf{x})$ is quadratic with Hessian \mathbf{H} , we know that

$$\begin{aligned} \mathbf{y}_{\mathcal{I}}^{t+1} - \mathbf{x}^* &= \mathbf{x}^t - \mathbf{x}^* - \eta \nabla_{\mathcal{I}} f(\mathbf{x}^t) + \eta \nabla_{\mathcal{I}} f(\mathbf{x}^*) + \nu(\mathbf{x}^t - \mathbf{x}^{t-1}) + \eta(\nabla_{\mathcal{I}} f(\mathbf{x}^t) - \mathbf{g}_{\mathcal{I}}^t) - \eta \nabla_{\mathcal{I}} f(\mathbf{x}^*) \\ &= ((1 + \nu)\mathbf{I} - \eta \mathbf{H}_{\mathcal{I}\mathcal{I}})(\mathbf{x}^t - \mathbf{x}^*) - \nu(\mathbf{x}^{t-1} - \mathbf{x}^*) + \eta(\nabla_{\mathcal{I}} f(\mathbf{x}^t) - \mathbf{g}_{\mathcal{I}}^t) - \eta \nabla_{\mathcal{I}} f(\mathbf{x}^*). \end{aligned} \quad (19)$$

Since the above iterate depends on the previous two iterates, we consider the following three-term recurrence in matrix form:

$$\begin{bmatrix} \mathbf{y}_{\mathcal{I}}^{t+1} - \mathbf{x}^* \\ \mathbf{x}^t - \mathbf{x}^* \end{bmatrix} = \begin{bmatrix} (1 + \nu)\mathbf{I} - \eta \mathbf{H}_{\mathcal{I}\mathcal{I}} & -\nu \mathbf{I} \\ \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}^t - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} + \eta \begin{bmatrix} \nabla_{\mathcal{I}} f(\mathbf{x}^t) - \mathbf{g}_{\mathcal{I}}^t - \nabla_{\mathcal{I}} f(\mathbf{x}^*) \\ 0 \end{bmatrix}.$$

Since $\mathbf{x}^{t+1} = \Phi_k(\mathbf{y}^{t+1}) = \Phi_k(\mathbf{y}_{\mathcal{I}}^{t+1})$, based on Lemma 1 we get $\|\mathbf{x}^{t+1} - \mathbf{x}^*\| \leq \sqrt{\alpha} \|\mathbf{y}_{\mathcal{I}}^{t+1} - \mathbf{x}^*\| \leq \alpha \|\mathbf{y}_{\mathcal{I}}^{t+1} - \mathbf{x}^*\|$, where $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k-k^*}}$. Let $\widehat{\mathcal{I}} = \text{supp}(\Phi_{3k}(\nabla f(\mathbf{x}^*))) \cup \text{supp}(\mathbf{x}^*)$ and $\widehat{s} = 3k + k^*$. Then

$$\begin{aligned}
& \mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^{t+1} - \mathbf{x}^* \\ \mathbf{x}^t - \mathbf{x}^* \end{bmatrix} \right\| \\
& \leq \alpha \mathbb{E} \left\| \begin{bmatrix} \mathbf{y}_{\mathcal{I}}^{t+1} - \mathbf{x}^* \\ \mathbf{x}^t - \mathbf{x}^* \end{bmatrix} \right\| \\
& \leq \alpha \mathbb{E} \left\| \begin{bmatrix} (1+\nu)\mathbf{I} - \eta\mathbf{H}_{\mathcal{I}\mathcal{I}} & -\nu\mathbf{I} \\ \mathbf{I} & 0 \end{bmatrix} \right\| \left\| \begin{bmatrix} \mathbf{x}^t - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} \right\| + \alpha \eta \mathbb{E} \left\| \begin{bmatrix} \nabla_{\mathcal{I}} f(\mathbf{x}^t) - \mathbf{g}_{\mathcal{I}}^t - \nabla_{\mathcal{I}} f(\mathbf{x}^*) \\ 0 \end{bmatrix} \right\| \\
& \leq \alpha \mathbb{E} \left\| \begin{bmatrix} (1+\nu)\mathbf{I}_{\mathcal{I}\mathcal{I}} - \eta\mathbf{H}_{\mathcal{I}\mathcal{I}} & -\nu\mathbf{I}_{\mathcal{I}\mathcal{I}} \\ \mathbf{I}_{\mathcal{I}\mathcal{I}} & 0 \end{bmatrix} \right\| \left\| \begin{bmatrix} \mathbf{x}^t - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} \right\| + \alpha \eta \mathbb{E} \|\nabla_{\mathcal{I}} f(\mathbf{x}^t) - \mathbf{g}_{\mathcal{I}}^t\| + \eta \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\| \\
& \stackrel{\textcircled{1}}{\leq} \alpha \frac{\sqrt{\ell_{\widehat{s}}} - \sqrt{\rho_{\widehat{s}}}}{\sqrt{\ell_{\widehat{s}}} + \sqrt{\rho_{\widehat{s}}}} \mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^t - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} \right\| + \alpha \eta \mathbb{E} \|\nabla_{\mathcal{I}} f(\mathbf{x}^t) - \mathbf{g}_{\mathcal{I}}^t\| + \alpha \eta \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\| \\
& \stackrel{\textcircled{2}}{\leq} \alpha \left(1 - \sqrt{\frac{\rho_{\widehat{s}}}{\ell_{\widehat{s}}}}\right) \mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^t - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} \right\| + \alpha \eta \mathbb{E} \|\nabla_{\mathcal{I}} f(\mathbf{x}^t) - \mathbf{g}_{\mathcal{I}}^t\| + \alpha \eta \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\| \\
& \stackrel{\textcircled{3}}{\leq} \left(1 - \frac{1}{2} \sqrt{\frac{\rho_{\widehat{s}}}{\ell_{\widehat{s}}}}\right) \mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^t - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} \right\| + \alpha \eta \mathbb{E} \|\nabla_{\mathcal{I}} f(\mathbf{x}^t) - \mathbf{g}_{\mathcal{I}}^t\| + \alpha \eta \|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\| \\
& \stackrel{\textcircled{4}}{\leq} \left(1 - \frac{1}{2} \sqrt{\frac{\rho_{\widehat{s}}}{\ell_{\widehat{s}}}}\right) \mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^t - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} \right\| + \alpha \eta \sqrt{\frac{B}{s_t}} + \alpha \eta \|\nabla_{\widehat{\mathcal{I}}} f(\mathbf{x}^*)\|, \tag{20}
\end{aligned}$$

where $\textcircled{1}$ follows from Lemma 6 with $\eta = \frac{4}{(\sqrt{\rho_{\widehat{s}}} + \sqrt{\ell_{\widehat{s}}})^2}$ and $\nu = \max\{|1 - \sqrt{\eta\rho_{\widehat{s}}}|^2, |1 - \sqrt{\eta\ell_{\widehat{s}}}|^2\}$ (For the sake of readability, we have assumed without loss of generality that $\rho \left(\begin{bmatrix} (1+\nu)\mathbf{I}_{\mathcal{I}\mathcal{I}} - \eta\mathbf{H}_{\mathcal{I}\mathcal{I}} & -\nu\mathbf{I}_{\mathcal{I}\mathcal{I}} \\ \mathbf{I}_{\mathcal{I}\mathcal{I}} & 0 \end{bmatrix} \right) = \left\| \begin{bmatrix} (1+\nu)\mathbf{I}_{\mathcal{I}\mathcal{I}} - \eta\mathbf{H}_{\mathcal{I}\mathcal{I}} & -\nu\mathbf{I}_{\mathcal{I}\mathcal{I}} \\ \mathbf{I}_{\mathcal{I}\mathcal{I}} & 0 \end{bmatrix} \right\|$. Otherwise, we may simply run sufficient rounds of heavy-ball iteration over \mathcal{I} before applying the hard thresholding operation and similar complexity bounds can be obtained with more involved arguments), $\textcircled{2}$ follows from the fact $\ell_{\widehat{s}} \geq \rho_{\widehat{s}}$, $\textcircled{3}$ follows from the condition $k \geq \left(1 + \frac{16\ell_{\widehat{s}}}{\rho_{\widehat{s}}}\right) k^*$ which implies $\alpha \leq 1 + \frac{1}{2} \sqrt{\frac{\rho_{\widehat{s}}}{\ell_{\widehat{s}}}}$, $\textcircled{4}$ uses Lemma 3 with $B = \max_t B_t$ and $\|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\| \leq \|\nabla_{\widehat{\mathcal{I}}} f(\mathbf{x}^*)\|$ where $\widehat{\mathcal{I}} = \text{supp}(\Phi_{3k}(\nabla f(\mathbf{x}^*))) \cup \text{supp}(\mathbf{x}^*)$.

Now we let

$$\beta := 1 - \frac{1}{2\sqrt{\kappa_s}} \quad \text{where} \quad \kappa_s = \frac{\ell_s}{\rho_s}.$$

We further set $s_t = \tau/\omega^t$ and assume that τ is large enough such that

$$\gamma := \frac{\alpha\eta\sqrt{B}}{\sqrt{\tau}} \leq \frac{8\sqrt{B}}{\sqrt{\tau}(\sqrt{\rho_{\widehat{s}}} + \sqrt{\ell_{\widehat{s}}})^2} \leq \delta(\|\mathbf{x}^0 - \mathbf{x}^*\| + \|\mathbf{x}^{-1} - \mathbf{x}^*\|), \tag{21}$$

where δ is a positive constant and will be discussed later. Now we use mathematical induction to prove

$$\mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^t - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} \right\| \leq \theta^t \left\| \begin{bmatrix} \mathbf{x}^0 - \mathbf{x}^* \\ \mathbf{x}^{-1} - \mathbf{x}^* \end{bmatrix} \right\| + \frac{8}{(1-\beta)(\sqrt{\rho_{\widehat{s}}} + \sqrt{\ell_{\widehat{s}}})^2} \|\nabla_{\widehat{\mathcal{I}}} f(\mathbf{x}^*)\|, \quad (\forall t), \tag{22}$$

where $\theta < 1$ is a constant and will be given below.

Obviously, when $t = 0$, Eqn. (22) holds. Now assume that for all $k \leq t$, Eqn. (22) holds. Then for $k = t + 1$, we have

$$\begin{aligned} \mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^t - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} \right\| &\leq \beta \mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^{t-1} - \mathbf{x}^* \\ \mathbf{x}^{t-2} - \mathbf{x}^* \end{bmatrix} \right\| + \delta \omega^{\frac{t}{2}} \left\| \begin{bmatrix} \mathbf{x}^0 - \mathbf{x}^* \\ \mathbf{x}^{-1} - \mathbf{x}^* \end{bmatrix} \right\| + \frac{8}{(\sqrt{\rho_s} + \sqrt{\ell_s})^2} \|\nabla_{\hat{\mathcal{I}}} f(\mathbf{x}^*)\| \\ &\leq (\beta \theta^t + \delta \omega^{\frac{t}{2}}) \mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^0 - \mathbf{x}^* \\ \mathbf{x}^{-1} - \mathbf{x}^* \end{bmatrix} \right\| + \frac{8}{(\sqrt{\rho_s} + \sqrt{\ell_s})^2} \left[\frac{\beta}{1 - \beta} + 1 \right] \|\nabla_{\hat{\mathcal{I}}} f(\mathbf{x}^*)\| \\ &\stackrel{\textcircled{1}}{\leq} \theta^t \mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^0 - \mathbf{x}^* \\ \mathbf{x}^{-1} - \mathbf{x}^* \end{bmatrix} \right\| + \frac{8}{(1 - \beta)(\sqrt{\rho_s} + \sqrt{\ell_s})^2} \|\nabla_{\hat{\mathcal{I}}} f(\mathbf{x}^*)\| \end{aligned}$$

where $\textcircled{1}$ holds since we let

$$\omega = \theta^2 \quad \text{and} \quad \theta = \beta + \delta. \quad (23)$$

This means that if Eqn. (23) holds, then Eqn. (22) always holds. So the conclusion holds.

Then let $\tau \geq \frac{81B\kappa_s}{4(\sqrt{\rho_s} + \sqrt{\ell_s})^4 \|\mathbf{x}^0 - \mathbf{x}^*\|^2}$. Recall $\mathbf{x}^{-1} = \mathbf{x}^0$, we have

$$\theta = \beta + \delta \leq 1 - \frac{1}{2\sqrt{\kappa_s}} + \frac{4}{9\sqrt{\kappa_s}} = 1 - \frac{1}{18\sqrt{\kappa_s}}.$$

Therefore, we have

$$\begin{aligned} \mathbb{E} \|\mathbf{x}^t - \mathbf{x}^*\| &\leq 2\theta^t \|\mathbf{x}^0 - \mathbf{x}^*\| + \frac{8}{(1 - \beta)(\sqrt{\rho_s} + \sqrt{\ell_s})^2} \|\nabla_{\hat{\mathcal{I}}} f(\mathbf{x}^*)\| \\ &= 2\theta^t \|\mathbf{x}^0 - \mathbf{x}^*\| + \frac{16\sqrt{\kappa_s}}{(\sqrt{\rho_s} + \sqrt{\ell_s})^2} \|\nabla_{\hat{\mathcal{I}}} f(\mathbf{x}^*)\|, \quad (\forall t) \end{aligned}$$

where $\beta = 1 - \frac{1}{2\sqrt{\kappa_s}}$ and $\theta \leq 1 - \frac{1}{18\sqrt{\kappa_s}}$ with $\omega = \theta^2$. So the result on sparsity-constrained problem holds.

Now we consider the rank-constrained problem. Since the proof accesses the Hessian, here we need to vectorize the matrix variable \mathbf{x} . For notation simplicity, we use $\tilde{\mathbf{x}} \in \mathbb{R}^{d_1 d_2}$ to denote the vectorization of $\mathbf{x} \in \mathbb{R}^{d_1 \times d_2}$. Assume that the skinny SVDs of \mathbf{x}^t , and \mathbf{x}^* are respectively $\mathbf{x}^t = \mathbf{U}_t \Sigma_t \mathbf{V}_t^T$ and $\mathbf{x}^* = \mathbf{U}_* \Sigma_* \mathbf{V}_*^T$. Then we have $\mathbf{x}^t = \mathbf{U}_t \mathbf{U}_t^T \mathbf{x}^t$ and $\mathbf{x}^* = \mathbf{U}_* \mathbf{U}_*^T \mathbf{x}^*$. We further define the projection operation $\mathcal{P}_{\mathcal{U}}(\mathbf{x}) = \mathbf{U} \mathbf{U}^T \mathbf{x}$ which projects \mathbf{x} in the subspace spanned by \mathbf{U} . Then let $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^{t-1} \cup \mathcal{I}^t \cup \mathcal{I}^{t+1}$, where $\mathcal{I}^* = \text{supp}(\mathbf{x}^*)$, $\mathcal{I}^{t-1} = \text{supp}(\mathbf{x}^{t-1})$, $\mathcal{I}^t = \text{supp}(\mathbf{x}^t)$ and $\mathcal{I}^{t+1} = \text{supp}(\mathbf{x}^{t+1})$. The notation $\mathbf{x}_{\mathcal{I}^t}^t = \mathbf{U}_t \mathbf{U}_t^T \mathbf{x}^t = \mathcal{P}_{\mathcal{U}_t}(\mathbf{x}^t)$. For the specifical notation $\mathbf{y}_{\mathcal{I}}$, it denotes $\mathbf{y}_{\mathcal{I}} = \mathcal{P}_{\mathcal{I}}(\mathbf{y}_{\mathcal{I}}) = (\mathbf{U}_{t-1} \mathbf{U}_{t-1}^T + \mathbf{U}_t \mathbf{U}_t^T + \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T + \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_{t-1} \mathbf{U}_{t-1}^T \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T - \mathbf{U}_{t-1} \mathbf{U}_{t-1}^T \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_{t-1} \mathbf{U}_{t-1}^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T + \mathbf{U}_{t-1} \mathbf{U}_{t-1}^T \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T) \mathbf{g}^t$ where $\mathcal{P}_{\mathcal{I}}$ is a projection. Here \mathbf{y} can be \mathbf{x} , \mathbf{g}^t and $\nabla f(\mathbf{x})$.

Then we also prove Eqn. (19) holds. Let $\tilde{\mathbf{y}}^t$, $\tilde{\mathbf{x}}^*$, $\tilde{\mathbf{x}}^t$ and $\nabla \tilde{f}(\mathbf{x})$ respectively denote the vectorization of \mathbf{y}^t , \mathbf{x}^* , \mathbf{x}^t and $\nabla f(\mathbf{x})$. The notation $\tilde{\mathbf{x}}_{\mathcal{I}}$ denotes the vectorization of $\mathbf{U}_{\mathcal{I}} \mathbf{U}_{\mathcal{I}}^T \tilde{\mathbf{x}}$. Then we have

$$\begin{aligned} \tilde{\mathbf{y}}_{\mathcal{I}}^{t+1} - \tilde{\mathbf{x}}^* &= \tilde{\mathbf{x}}^t - \tilde{\mathbf{x}}^* - \eta \nabla_{\mathcal{I}} \tilde{f}(\mathbf{x}^t) + \eta \nabla_{\mathcal{I}} \tilde{f}(\mathbf{x}^*) + \nu (\tilde{\mathbf{x}}^t - \tilde{\mathbf{x}}^{t-1}) + \eta (\nabla_{\mathcal{I}} \tilde{f}(\mathbf{x}^t) - \tilde{\mathbf{g}}_{\mathcal{I}}^t) - \eta \nabla_{\mathcal{I}} \tilde{f}(\mathbf{x}^*) \\ &= ((1 + \nu) \mathbf{I} - \eta \mathbf{H}) (\tilde{\mathbf{x}}^t - \tilde{\mathbf{x}}^*) - \nu (\tilde{\mathbf{x}}^{t-1} - \tilde{\mathbf{x}}^*) + \eta (\nabla_{\mathcal{I}} \tilde{f}(\tilde{\mathbf{x}}^t) - \tilde{\mathbf{g}}_{\mathcal{I}}^t) - \eta \nabla_{\mathcal{I}} \tilde{f}(\mathbf{x}^*), \end{aligned}$$

where $\mathbf{H} = \mathcal{P}_{\mathcal{I}}(\nabla^2 \tilde{f}(\tilde{\mathbf{z}}))$. Here $\nabla^2 \tilde{f}(\tilde{\mathbf{z}})$ comes from the fact that $\nabla_{\mathcal{I}} \tilde{f}(\mathbf{x}^t) - \nabla_{\mathcal{I}} \tilde{f}(\mathbf{x}^*) = \mathcal{P}_{\mathcal{I}}(\nabla \tilde{f}(\mathbf{x}^t) - \nabla \tilde{f}(\mathbf{x}^*)) \stackrel{\textcircled{1}}{=} \mathcal{P}_{\mathcal{I}} \nabla^2 \tilde{f}(\tilde{\mathbf{z}}) (\tilde{\mathbf{x}}^t - \tilde{\mathbf{x}}^*)$ in which $\textcircled{1}$ uses the second differential property of $f(\mathbf{x})$ and thus there exists a matrix $\tilde{\mathbf{z}}$ such that $\nabla \tilde{f}(\mathbf{x}^t) - \nabla \tilde{f}(\mathbf{x}^*) = \nabla^2 \tilde{f}(\tilde{\mathbf{z}}) (\tilde{\mathbf{x}}^t - \tilde{\mathbf{x}}^*)$. Then by Assumptions 1 and 2, we have $\rho_s \mathbf{I} \preceq \mathcal{P}_{\mathcal{I}}(\mathbf{H}) \preceq \ell_s \mathbf{I}$ since $\|\mathcal{P}_{\mathcal{I}}\|_2 \leq 1$. We obtain $\|\mathcal{P}_{\mathcal{I}}\|_2 \leq 1$ since we have $\mathcal{P}_{\mathcal{I}}^T \mathcal{P}_{\mathcal{I}} = \mathcal{P}_{\mathcal{I}}$.

On the other hand, we can follow Eqn. (12) in Section B.1 to prove

$$\begin{aligned} \Phi_k(\mathbf{y}_{\mathcal{I}}^{t+1}) &= \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T [\mathbf{x}^t - \eta \mathbf{g}_{\mathcal{I}}^t + \nu (\mathbf{x}^t - \mathbf{x}^{t-1})] \\ &= \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T [\mathbf{x}^t + \nu (\mathbf{x}^t - \mathbf{x}^{t-1})] + \eta \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathcal{P}_{\mathcal{I}}(\mathbf{g}^t) \\ &\stackrel{\textcircled{1}}{=} \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T [\mathbf{x}^t + \nu (\mathbf{x}^t - \mathbf{x}^{t-1})] + \eta \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{g}^t = \mathbf{x}^{t+1}, \end{aligned}$$

where ① plugs $\mathcal{P}_{\mathcal{I}}$ defined above and obtains $\mathbf{U}_{t+1}\mathbf{U}_{t+1}^T\mathcal{P}_{\mathcal{I}} = \mathbf{U}_{t+1}\mathbf{U}_{t+1}^T$. Then we apply Lemma 1 to establish

$$\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2 = \|\Phi_k(\mathbf{y}_{\mathcal{I}}^{t+1}) - \mathbf{x}^*\|^2 \leq \left(1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}\right) \|\mathbf{y}_{\mathcal{I}}^{t+1} - \mathbf{x}^*\|^2,$$

where $\alpha = 1 + \frac{2\sqrt{k^*}}{\sqrt{k - k^*}}$. Therefore, we can establish

$$\mathbb{E} \left\| \begin{bmatrix} \mathbf{x}^{t+1} - \mathbf{x}^* \\ \mathbf{x}^t - \mathbf{x}^* \end{bmatrix} \right\| \leq \alpha \mathbb{E} \left\| \begin{bmatrix} \mathbf{y}_{\mathcal{I}}^{t+1} - \mathbf{x}^* \\ \mathbf{x}^t - \mathbf{x}^* \end{bmatrix} \right\|.$$

Then we can establish the first inequality in Eqn. (20). The following proof does not depend the property on rank-constrained problem. So we can just follow the above proof sketch for sparsity-constrained problem to prove the result on the rank-constrained problem. The proof is completed. \square

C.2 Proof of Corollary 2

Proof. To achieve ϵ -accurate solution, let

$$2\theta^t \|\mathbf{x}^0 - \mathbf{x}^*\| \leq \sqrt{\epsilon}$$

where $\tilde{\beta} = 1 - \frac{1}{2\sqrt{\kappa_{\hat{s}}}}$ and $\theta \leq 1 - \frac{1}{18\sqrt{\kappa_{\hat{s}}}}$ with $\omega = \theta^2$, we have

$$T \geq \log_{1/\theta} \left(\frac{2\|\mathbf{x}^0 - \mathbf{x}^*\|}{\sqrt{\epsilon}} \right).$$

Therefore, the IFO complexity is

$$\begin{aligned} \tau \left[1 + \frac{1}{\omega} + \dots + \frac{1}{\omega^{T-1}} \right] &= \tau \frac{(1/\omega)^{\log_{1/\theta} \left(\frac{2\|\mathbf{x}^0 - \mathbf{x}^*\|}{\sqrt{\epsilon}} \right)} - 1}{1/\omega - 1} \stackrel{\text{①}}{=} \frac{\tau}{1/\omega - 1} \left[\frac{4\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\epsilon} - 1 \right] \\ &\leq \frac{\tau}{1/\omega - 1} \left[\frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\epsilon} \right] \stackrel{\text{②}}{\leq} \frac{81B\kappa_{\hat{s}}}{4(\sqrt{\rho_{\hat{s}}} + \sqrt{\ell_{\hat{s}}})^4 \epsilon} \frac{1}{1/\sqrt{1 - \frac{1}{18\sqrt{\kappa_{\hat{s}}}}} - 1}} \stackrel{\text{③}}{\leq} \frac{81B \cdot 36\sqrt{\kappa_{\hat{s}}}}{\rho_{\hat{s}}(\sqrt{\rho_{\hat{s}}} + \sqrt{\ell_{\hat{s}}})^2 \epsilon} \frac{\ell_s}{(\sqrt{\rho_{\hat{s}}} + \sqrt{\ell_{\hat{s}}})^2} \\ &\leq \frac{81B \cdot 36\sqrt{\kappa_{\hat{s}}}}{\rho_{\hat{s}}(\sqrt{\rho_{\hat{s}}} + \sqrt{\ell_{\hat{s}}})^2 \epsilon} \frac{\ell_{\hat{s}}}{(\sqrt{\rho_{\hat{s}}} + \sqrt{\ell_{\hat{s}}})^2} = \mathcal{O} \left(\frac{\sqrt{\kappa_{\hat{s}}}B}{\rho_{\hat{s}}\ell_{\hat{s}}\epsilon} \right) \stackrel{\text{④}}{=} \mathcal{O} \left(\frac{\sqrt{\kappa_{\hat{s}}}}{\epsilon} \right) \end{aligned}$$

where ① uses $\omega = \theta^2$; ② uses $\theta \leq 1 - \frac{1}{18\sqrt{\kappa_{\hat{s}}}}$ and $\tau \geq \frac{81B\kappa_{\hat{s}}}{4(\sqrt{\rho_{\hat{s}}} + \sqrt{\ell_{\hat{s}}})^4 \|\mathbf{x}^0 - \mathbf{x}^*\|^2}$; ③ uses $\frac{1}{1/\sqrt{1 - \frac{1}{18\sqrt{\kappa_{\hat{s}}}}} - 1} \leq \frac{1}{1 - \sqrt{1 - \frac{1}{18\sqrt{\kappa_{\hat{s}}}}}} \leq 36\sqrt{\kappa_{\hat{s}}}$ since we have $1 - \sqrt{1 - a} \geq \frac{1}{2}a$ for $a \in (0, 1)$; ④ holds since (1) the parameter $\rho_{\hat{s}}$ is the strong convex parameter at sparsity/low-rank level s and thus is not very small since \hat{s} is much smaller than the feature dimension, (2) we can always scale the problem such that $\rho_{\hat{s}}$ is not small. Notice, such a scale does not affect the ratio $B/\ell_{\hat{s}}$ since they always scale at the same order. Thus, we have the IFO complexity $\mathcal{O} \left(\frac{\sqrt{\kappa_{\hat{s}}}}{\epsilon} \right)$.

On the other hand, we have

$$\log_{1/\theta} \left(\frac{1}{\sqrt{\epsilon}} \right) = \frac{\log \left(\frac{1}{\sqrt{\epsilon}} \right)}{\log \left(\frac{1}{\theta} \right)} = \frac{\log \left(\frac{1}{\sqrt{\epsilon}} \right)}{\log \left(1 + \frac{1}{18\sqrt{\kappa_{\hat{s}} - 1}} \right)} = \frac{\log \left(\frac{1}{\sqrt{\epsilon}} \right)}{\log \left(1 + \frac{1}{18\sqrt{\kappa_{\hat{s}} - 1}} \right)} \stackrel{\text{①}}{\leq} \mathcal{O} \left(\sqrt{\kappa_{\hat{s}}} \log \left(\frac{1}{\sqrt{\epsilon}} \right) \right),$$

where ① holds since we have $\log(1 + x) \geq \log(2) \cdot x$ for $x \in [0, 1]$. The proof is completed. \square

D Proof of Auxiliary Lemmas

D.1 Proof of Lemma 4

Proof. Firstly, for both vector \mathbf{x} and matrix variable \mathbf{x} we can decompose $\mathbb{E}\|\mathbf{g}^t\|^2$ and bound it as follows:

$$\begin{aligned} \mathbb{E}\|\mathbf{g}_{\mathcal{I}}^t\|^2 &= \mathbb{E}\|\mathbf{g}_{\mathcal{I}}^t - \nabla_{\mathcal{I}}f(\mathbf{x}^t) + \nabla_{\mathcal{I}}f(\mathbf{x}^t) - \nabla_{\mathcal{I}}f(\mathbf{x}^*) + \nabla_{\mathcal{I}}f(\mathbf{x}^*)\|^2 \\ &\leq 3\mathbb{E}\|\mathbf{g}_{\mathcal{I}}^t - \nabla_{\mathcal{I}}f(\mathbf{x}^t)\|^2 + 3\mathbb{E}\|\nabla_{\mathcal{I}}f(\mathbf{x}^t) - \nabla_{\mathcal{I}}f(\mathbf{x}^*)\|_2^2 + 3\|\nabla_{\mathcal{I}}f(\mathbf{x}^*)\|^2 \\ &\stackrel{\text{①}}{\leq} \frac{3}{s_t} B_t + 3\mathbb{E}\|\nabla_{\mathcal{I}}f(\mathbf{x}^t) - \nabla_{\mathcal{I}}f(\mathbf{x}^*)\|^2 + 3\|\nabla_{\mathcal{I}}f(\mathbf{x}^*)\|^2, \end{aligned}$$

where ① use Lemma 3. Now we bound the second term. We define a function

$$h_i(\mathbf{x}) = f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla_{\mathcal{I}} f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle.$$

It is easy to check that $\nabla h_i(\mathbf{x}^*) = 0$, which implies $h_i(\mathbf{x}^*) = \min_{\mathbf{x}} h_i(\mathbf{x})$. In this way, for vector variable \mathbf{x} , we have

$$\begin{aligned} 0 = h_i(\mathbf{x}^*) &\leq \min_{\eta} h_i(\mathbf{x} - \eta \nabla_{\mathcal{I}} h_i(\mathbf{x})) \leq \min_{\eta} h_i(\mathbf{x}) - \eta \langle \nabla h_i(\mathbf{x}), \nabla_{\mathcal{I}} h_i(\mathbf{x}) \rangle + \frac{\eta^2 \ell_s}{2} \|\nabla_{\mathcal{I}} h_i(\mathbf{x})\|_2^2 \\ &\stackrel{\text{①}}{=} \min_{\eta} h_i(\mathbf{x}) - \eta \|\nabla_{\mathcal{I}} h_i(\mathbf{x})\|^2 + \frac{\eta^2 \ell_s}{2} \|\nabla_{\mathcal{I}} h_i(\mathbf{x})\|^2 \\ &\stackrel{\text{②}}{=} h_i(\mathbf{x}) - \frac{1}{2\ell_s} \|\nabla_{\mathcal{I}} h_i(\mathbf{x})\|_2^2, \end{aligned} \tag{24}$$

where ① holds since for vector \mathbf{x} , we have $\langle \nabla h_i(\mathbf{x}), \nabla_{\mathcal{I}} h_i(\mathbf{x}) \rangle = \|\nabla_{\mathcal{I}} h_i(\mathbf{x})\|^2$ and ② holds by optimizing $\eta = \frac{1}{\ell_s}$.

Now we consider the matrix variable \mathbf{x} . Firstly, assume that the skinny SVDs of \mathbf{x}^t , \mathbf{x}^{t+1} , \mathbf{x}^* are respectively $\mathbf{x}^t = \mathbf{U}_t \Sigma_t \mathbf{V}_t^T$, $\mathbf{x}^{t+1} = \mathbf{U}_{t+1} \Sigma_{t+1} \mathbf{V}_{t+1}^T$ and $\mathbf{x}^* = \mathbf{U}_* \Sigma_* \mathbf{V}_*^T$. Then we have $\mathbf{x}^t = \mathbf{U}_t \mathbf{U}_t^T \mathbf{x}^t$, $\mathbf{x}^{t+1} = \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{x}^{t+1}$ and $\mathbf{x}^* = \mathbf{U}_* \mathbf{U}_*^T \mathbf{x}^*$. We further define the projection operation $\mathcal{P}_{\mathcal{U}}(\mathbf{x}) = \mathbf{U} \mathbf{U}^T \mathbf{x}$ which projects \mathbf{x} in the subspace spanned by \mathbf{U} . Then let $\mathbf{v} = \mathbf{x}^t - \eta \mathbf{g}_{\mathcal{I}}^t$ and $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^t \cup \mathcal{I}^{t+1}$, where $\mathcal{I}^* = \text{supp}(\mathbf{x}^*)$, $\mathcal{I}^t = \text{supp}(\mathbf{x}^t)$ and $\mathcal{I}^{t+1} = \text{supp}(\mathbf{x}^{t+1})$. The notation $\mathbf{x}_{\mathcal{I}^t}^t = \mathbf{U}_t \mathbf{U}_t^T \mathbf{x}^t = \mathcal{P}_{\mathcal{U}_t}(\mathbf{x}^t)$. For the special notation $\mathbf{g}_{\mathcal{I}}^t$, it denotes $\mathbf{g}_{\mathcal{I}}^t = \mathcal{P}_{\mathcal{I}}(\mathbf{g}^t) = (\mathbf{U}_t \mathbf{U}_t^T + \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T + \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T + \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_{t+1} \mathbf{U}_{t+1}^T \mathbf{U}_* \mathbf{U}_*^T) \mathbf{g}^t$ where $\mathcal{P}_{\mathcal{I}}$ is a projection. Then we also have

$$\begin{aligned} \langle \nabla h_i(\mathbf{x}), \nabla_{\mathcal{I}} h_i(\mathbf{x}) \rangle &= \langle \nabla h_i(\mathbf{x}), \mathcal{P}_{\mathcal{I}}(\nabla h_i(\mathbf{x})) \rangle \stackrel{\text{①}}{=} \langle \mathcal{P}_{\mathcal{I}}(\nabla h_i(\mathbf{x})), \mathcal{P}_{\mathcal{I}}(\nabla h_i(\mathbf{x})) \rangle = \|\nabla_{\mathcal{I}} h_i(\mathbf{x})\|^2, \\ \text{where ① holds since } \langle \mathcal{P}_{\mathcal{I}}(\nabla_{\mathcal{I}} h_i(\mathbf{x})), \mathcal{P}_{\mathcal{I}}(\nabla_{\mathcal{I}} h_i(\mathbf{x})) \rangle &= \langle \nabla h_i(\mathbf{x}), \mathcal{P}_{\mathcal{I}}^T \mathcal{P}_{\mathcal{I}}(\nabla_{\mathcal{I}} h_i(\mathbf{x})) \rangle = \\ &= \langle \nabla h_i(\mathbf{x}), \mathcal{P}_{\mathcal{I}}(\nabla_{\mathcal{I}} h_i(\mathbf{x})) \rangle \text{ due to } \mathcal{P}_{\mathcal{I}}^T \mathcal{P}_{\mathcal{I}} = \mathcal{P}_{\mathcal{I}}. \end{aligned}$$

Thus, Eqn. (24) holds for both vector variable \mathbf{x} and matrix variable \mathbf{x} . It further yields

$$\|\nabla_{\mathcal{I}} f_i(\mathbf{x}) - \nabla_{\mathcal{I}} f_i(\mathbf{x}^*)\|_2^2 \leq 2\ell_s (f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla_{\mathcal{I}} f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle).$$

Then we are ready to bound the second term:

$$\begin{aligned} \mathbb{E} \|\nabla_{\mathcal{I}} f(\mathbf{x}^t) - \nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2 &= \frac{1}{n} \sum_{i=1}^n \|\nabla_{\mathcal{I}} f_i(\mathbf{x}^t) - \nabla_{\mathcal{I}} f_i(\mathbf{x}^*)\|^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n 2\ell_s (f_i(\mathbf{x}^t) - f_i(\mathbf{x}^*) - \langle \nabla_{\mathcal{I}} f_i(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle) \\ &= 2\ell_s (f(\mathbf{x}^t) - f(\mathbf{x}^*) + \langle \nabla_{\mathcal{I}} f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle). \end{aligned}$$

Therefore, for both vector variable \mathbf{x} and matrix variable \mathbf{x} we have

$$\mathbb{E} \|\mathbf{g}_{\mathcal{I}}^t\|^2 \leq \frac{3}{s_t} B_t + 6\ell_s (f(\mathbf{x}^t) - f(\mathbf{x}^*) + \langle \nabla_{\mathcal{I}} f(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle) + 3\|\nabla_{\mathcal{I}} f(\mathbf{x}^*)\|^2.$$

This completes the proof. \square

D.2 Proof of Lemma 5

Proof. We first consider vector variable \mathbf{x} . From the strong convexity we get

$$\begin{aligned} f(\mathbf{x}^*) &\geq f(\mathbf{x}^t) + \langle \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle + \frac{\rho_s}{2} \|\mathbf{x}^* - \mathbf{x}^t\|^2 \\ &\stackrel{\text{①}}{=} f(\mathbf{x}^t) + \langle \nabla_{\mathcal{I}^t \cup \mathcal{I}^*} f(\mathbf{x}^t) - \mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t + \mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t, \mathbf{x}^* - \mathbf{x}^t \rangle + \frac{\rho_s}{2} \|\mathbf{x}^* - \mathbf{x}^t\|^2 \\ &\stackrel{\text{②}}{\geq} f(\mathbf{x}^t) - \frac{2}{\rho_s} \|\nabla f(\mathbf{x}^t) - \mathbf{g}^t\|^2 - \frac{2}{\rho_s} \|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2 - \frac{\rho_s}{4} \|\mathbf{x}^* - \mathbf{x}^t\|^2 + \frac{\rho_s}{2} \|\mathbf{x}^* - \mathbf{x}^t\|^2 \\ &= f(\mathbf{x}^t) - \frac{2}{\rho_s} \|f(\mathbf{x}^t) - \mathbf{g}^t\|^2 - \frac{2}{\rho_s} \|\mathbf{g}_{\mathcal{I}^t \cup \mathcal{I}^*}^t\|^2 + \frac{\rho_s}{4} \|\mathbf{x}^* - \mathbf{x}^t\|^2, \end{aligned} \tag{25}$$

where ② holds since we use $\langle \mathbf{x}, \mathbf{y} \rangle \geq -\left(\frac{1}{2c}\|\mathbf{x}\|_2^2 + \frac{c}{2}\|\mathbf{y}\|_2^2\right)$ for arbitrary $c \geq 0$. By rearranging both sides of the above we get the desired bound.

Then we consider the matrix variable $\mathbf{x} \in \mathbb{R}^{d_1 \times d_2}$ for rank-constrained problem. Firstly, assume that the skinny SVDs of \mathbf{x}^t and \mathbf{x}^* are respectively $\mathbf{x}^t = \mathbf{U}_t \boldsymbol{\Sigma}_t \mathbf{V}_t^T$ and $\mathbf{x}^* = \mathbf{U}_* \boldsymbol{\Sigma}_* \mathbf{V}_*^T$. Then we have $\mathbf{x}^t = \mathbf{U}_t \mathbf{U}_t^T \mathbf{x}^t$ and $\mathbf{x}^* = \mathbf{U}_* \mathbf{U}_*^T \mathbf{x}^*$. We further define the projection operation $\mathcal{P}_U(\mathbf{x}) = \mathbf{U} \mathbf{U}^T \mathbf{x}$ which projects \mathbf{x} in the subspace spanned by \mathbf{U} . Then let $\mathcal{I} = \mathcal{I}^* \cup \mathcal{I}^t$, where $\mathcal{I}^* = \text{supp}(\mathbf{x}^*)$ and $\mathcal{I}^t = \text{supp}(\mathbf{x}^t)$. The notation $\mathbf{x}_{\mathcal{I}^t}^t = \mathbf{U}_t \mathbf{U}_t^T \mathbf{x}^t = \mathcal{P}_{\mathbf{U}_t}(\mathbf{x}^t)$. For the specific notation $\nabla_{\mathcal{I}} f(\mathbf{x}^t)$, it denotes $\nabla_{\mathcal{I}} f(\mathbf{x}^t) = \mathcal{P}_{\mathcal{I}}(\nabla f(\mathbf{x}^t)) = (\mathbf{U}_t \mathbf{U}_t^T + \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_* \mathbf{U}_*^T) \nabla f(\mathbf{x}^t)$ where $\mathcal{P}_{\mathcal{I}}$ is a projection. Then we have

$$\mathcal{P}_{\mathcal{I}}(\mathbf{x}^* - \mathbf{x}^t) = (\mathbf{U}_t \mathbf{U}_t^T + \mathbf{U}_* \mathbf{U}_*^T - \mathbf{U}_t \mathbf{U}_t^T \mathbf{U}_* \mathbf{U}_*^T) (\mathbf{U}_* \mathbf{U}_*^T \mathbf{x}^* - \mathbf{U}_t \mathbf{U}_t^T \mathbf{x}^t) = \mathbf{x}^* - \mathbf{x}^t,$$

which further gives

$$\begin{aligned} \langle \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle &= \langle \nabla f(\mathbf{x}^t), \mathcal{P}_{\mathcal{I}}(\mathbf{x}^* - \mathbf{x}^t) \rangle \stackrel{\textcircled{1}}{=} \langle \mathcal{P}_{\mathcal{I}}(\nabla f(\mathbf{x}^t)), \mathcal{P}_{\mathcal{I}}(\mathbf{x}^* - \mathbf{x}^t) \rangle \\ &= \langle \nabla_{\mathcal{I}} f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \rangle, \end{aligned}$$

where ① holds since $\langle \mathcal{P}_{\mathcal{I}}(\nabla f(\mathbf{x}^t)), \mathcal{P}_{\mathcal{I}}(\mathbf{x}^* - \mathbf{x}^t) \rangle = \langle \nabla f(\mathbf{x}^t), \mathcal{P}_{\mathcal{I}}^T \mathcal{P}_{\mathcal{I}}(\mathbf{x}^* - \mathbf{x}^t) \rangle = \langle \nabla f(\mathbf{x}^t), \mathcal{P}_{\mathcal{I}}(\mathbf{x}^* - \mathbf{x}^t) \rangle$ due to $\mathcal{P}_{\mathcal{I}}^T \mathcal{P}_{\mathcal{I}} = \mathcal{P}_{\mathcal{I}}$.

In this way, ① in Eqn. (25) holds. Thus, the above result also hold for matrix variable in rank-constrained problem. The proof is completed. \square

D.3 Proof of Lemma 6

Proof. Let $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d$ be the eigenvalues of \mathbf{A} and $\boldsymbol{\Lambda}$ be a diagonal matrix whose diagonal entries are $\{\lambda_i\}$ in a non-decreasing order. By proper manipulation we get

$$\rho \left(\begin{bmatrix} (1+\nu)\mathbf{I} - \eta\mathbf{A} & -\nu\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \right) = \rho \left(\begin{bmatrix} (1+\nu)\mathbf{I} - \eta\boldsymbol{\Lambda} & -\nu\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \right) = \max_{i \in [d]} \rho \left(\begin{bmatrix} 1+\nu - \eta\lambda_i & -\nu \\ 1 & 0 \end{bmatrix} \right),$$

where in the second equality we have used the fact that it is possible to permute the matrix to a block diagonal matrix with 2×2 blocks. For each $i \in [d]$, the eigenvalues of the 2×2 matrices are given by the roots of

$$\lambda^2 - (1+\nu - \eta\lambda_i)\lambda + \nu = 0.$$

Given that $\nu \geq |1 - \sqrt{\eta\lambda_i}|^2$, the roots of the above equation are imaginary and both have magnitude $\sqrt{\nu}$. Since $\nu = \max\{|1 - \sqrt{\eta\mu}|^2, |1 - \sqrt{\eta\ell}|^2\}$, the magnitude of each root is at most $\max\{|1 - \sqrt{\eta\mu}|, |1 - \sqrt{\eta\ell}|\}$. This proves the desired spectral norm bound. \square

E Additional Experimental Results

E.1 Descriptions of Testing Datasets

We briefly introduce the seven testing datasets in the manuscript. Among them, three datasets are provided in the LibSVM website¹, including rcv1, real-sim and epsilon. We also evaluate our algorithms on mnist² for handwriting recognition, news20³ for news classification, coil100⁴ and caltech256⁵ for image classification. Their detailed information is summarized in Table 2. We can observe that these datasets are different from each other in feature dimension, training samples, and class numbers, etc. It should be mentioned that for caltech256 including 256 kinds of objects and one background class, we use its OverFeat feature, while for other datasets, we all use their raw data.

¹<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

²<http://yann.lecun.com/exdb/mnist/>

³<http://qwone.com/~jason/20Newsgroups/>

⁴<http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php>

⁵<https://authors.library.caltech.edu/7694/>

Table 2: Descriptions of the ten testing datasets.

	#class	#sample	#feature		#class	#sample	#feature
rcv1	2	20,242	47,236	news20	20	62,061	15,935
real-sim	2	72,309	20,958	coil100	100	7,200	1,024
epsilon	2	100,000	2,000	caltech256	257	5,140	2,000
mnist	10	60,000	784				

E.2 More Experiments on A Single Pass over Data

Finally, we give more experimental results on a single pass over data. Following the setting in Section 5 in manuscript, here we test the considered algorithms on logistic regression with regularization parameters $\lambda = 10^{-5}$. We follow our theoretical results to exponentially expand the mini-batch size s_k in HSG-HT and AHSG-HT and set $\tau = 1$. Figure 3 summarizes the numerical results in this setting. One can observe that on these optimization problems, most algorithms still achieve high accuracy after one pass over data, while HSG-HT and AHSG-HT also converge significantly faster than the other algorithms. These observations are consistent with the results in Figure 1 in the manuscript. All these results demonstrate the high efficiency of HSG-HT and AHSG-HT and also confirm the theoretical implication of Corollary 1 and 2 that HSG-HT and AHSG-HT always have lowest hard thresholding complexity than the compared algorithms and have lower in IFO complexity than other considered variance-reduced algorithms linearly depending on the sample size n , when the desired accuracy is moderately small and data scale is large.

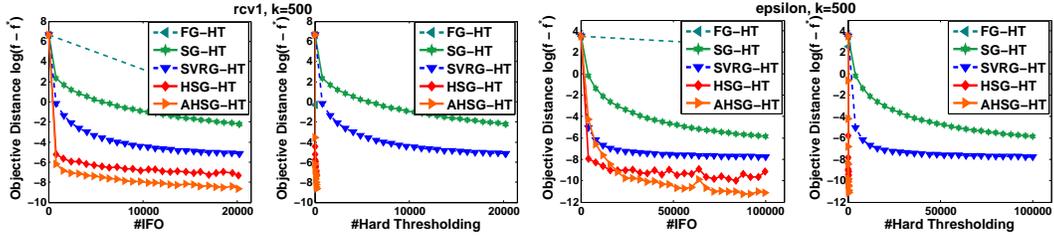


Figure 3: Single-epoch processing: comparison among hard thresholding algorithms for a single pass over data on sparse logistic regression with regularization parameter $\lambda = 10^{-5}$.

Since the objective loss decreases fast along with the hard thresholding iteration, here we magnify the subfigures in Figure 1 in the manuscript and the above Figure 3 which display the objective loss decrease along with the hard thresholding iteration. In this way, the objective loss decrease along with the hard thresholding iteration can be viewed better. From Figure 4, one can easily observe that AHSG-HT and HSG-HT converge much faster than the compared algorithms. Moreover, AHSG-HT achieves higher optimization accuracy which demonstrates that AHSG-HT is superior over HSG-HT in hard thresholding complexity. All these results confirm our theoretical implication of Corollary 1 and 2.

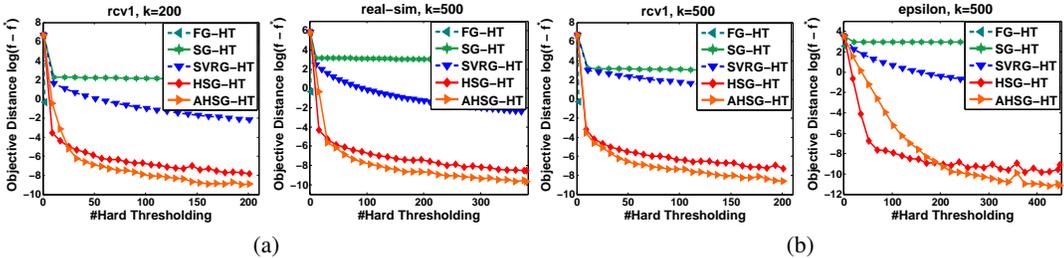


Figure 4: Comparison of hard thresholding complexity in single-epoch processing. (a) magnifies the hard thresholding iterations in Figure 1 in the manuscript. (b) magnifies the hard thresholding iterations in Figure 3 above.

References

- [1] X. Li, R. Arora, H. Liu, J. Haupt, and T. Zhao. Nonconvex sparse learning via stochastic optimization with progressive variance reduction. *Proc. Int'l Conf. Machine Learning*, 2016. [1](#)
- [2] P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional M-estimation. In *Proc. Conf. Neural Information Processing Systems*, pages 685–693, 2014. [1](#), [6](#)
- [3] M. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012. [1](#)