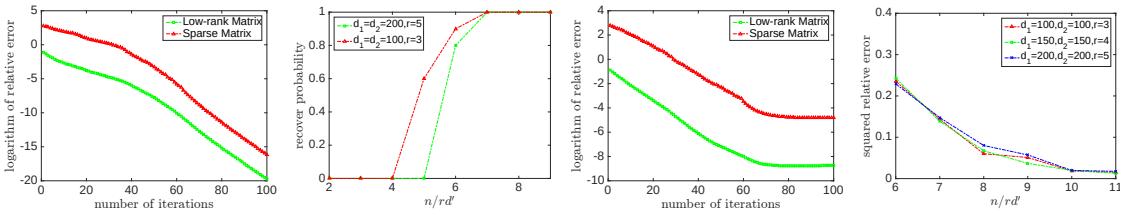
A Additional Experiments

In this section, we present additional experimental results to verify the linear convergence rate, sample complexity, and statistical rate of our proposed algorithm.

A.1 Robust Matrix Sensing

Our data are generated from the same procedure as described before. In addition, we study the same experimental setting as before except we choose $\alpha = r, \nu = 1, \beta = 0.1$. Figure 2 summarized the experimental results for robust matrix sensing. Figure 2(a) and 2(c) illustrate the relative error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/\|\mathbf{X}^*\|_F^2$ in log scale versus number of iterations. Note that, we only lay out results under setting $d_1 = d_2 = 100, r = 3$ with number of observations $n = 0.2 * d_1 d_2$ to avoid redundancy. These plots demonstrate the linear rate of convergence of our algorithm. Figure 2(b) demonstrates the sample complexity requirement to achieve exact recovery for low-rank structure in the noiseless setting. Note that we say $\widehat{\mathbf{X}}$ achieves exact recovery if $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F/\|\mathbf{X}^*\|_F \leq 10^{-3}$. It confirms our theoretical results regarding the sample complexity. The statistical error for the low-rank matrix is demonstrated in Figure 2(d), which is consistent with our result $O(rd/n)$.

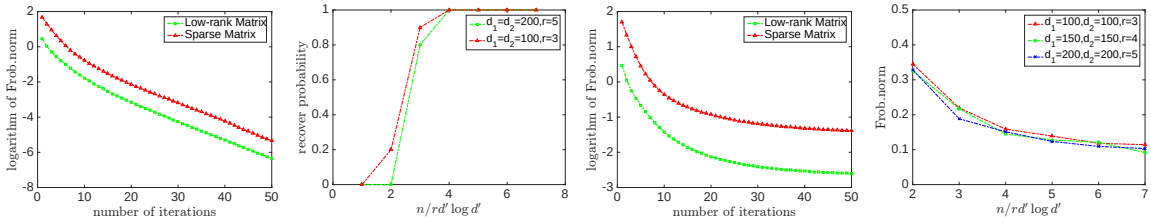


(a) Noiseless Case (b) Sample Complexity (c) Noisy Case (d) Statistical Error

Figure 2: Experimental results for robust matrix sensing. (a),(c) Relative error in log scale vs. number of iterations in the noiseless and noisy settings respectively. (b) Recovering probability of low-rank matrix vs. scaled sample size in the noiseless setting. (d) Relative error vs. scaled sample size in the noisy setting.

A.2 Robust PCA

We generate the data according to the same procedure as before. Furthermore, we consider the same experimental settings as robust matrix sensing except. The experimental results for robust PCA are summarized in Figure 3. In detail, Figures 3(a) and 3(c) report the squared estimation error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/(d_1 d_2)$ in log scale versus number of iterations. Note that we only lay out the results under fully observed model with setting $d_1 = d_2 = 200, r = 5$, because other settings will give us similar plots, and we leave them out for simplicity. The results verify the linear convergence rate of our algorithm. In the noiseless setting, the sample complexity for achieving exactly recovery of the low-rank matrix is illustrated in Figure 3(b). The result of recovery probability indicates the sample complexity requirement $n = O(rd \log d)$ for robust PCA. Finally, Figure 3(d) demonstrates the statistical error for the low-rank matrix, which is at the order $O(rd \log d/n)$. Although our theoretical results suggest $O(r^2 d \log d)$ sample complexity and $O(r^2 d \log d)$ statistical error, the simulation results indicate that both the sample complexity and the statistical error scale linearly with rd.



(a) Noiseless Case (b) Sample Complexity (c) Noisy Case (d) Statistical Error

Figure 3: Experimental results for robust PCA. (a),(c) Squared estimation error in log scale vs. number of iterations in the noiseless and noisy settings respectively. (b) Recovering probability of low-rank matrix vs. scaled sample size in the noiseless setting. (d) Squared relative error vs. scaled sample size in the noisy setting.

B Proof of the Main Theory

In this section, we establish the proof of our main theory. Before proceeding any further, we introduce the following notations. For any index set $\Omega \subseteq [d_1] \times [d_2]$, let $\Omega_{i,*}$ and $\Omega_{*,j}$ be the i-th row and j-th column of Ω respectively. Denote the column and row space of $\mathbf{A}$ by $\mathrm{col}(\mathbf{A})$ and $\mathrm{row}(\mathbf{A})$ respectively. Let the top $d_1 \times r$ and bottom $d_2 \times r$ matrices of any matrix $\mathbf{A} \in \mathbb{R}^{(d_1+d_2)\times r}$ be $\mathbf{A}_U$ and $\mathbf{A}_V$ respectively. Let the nuclear norm of any matrix $\mathbf{A}$ be $\|\mathbf{A}\|_*$. Denote $\mathbf{Z} = [\mathbf{U}; \mathbf{V}] \in \mathbb{R}^{(d_1+d_2)\times r}$, then according to (3.3), we reformulate the regularized objective function as follows

$$\widetilde{F}_n(\mathbf{Z}, \mathbf{S}) = F_n(\mathbf{U}, \mathbf{V}, \mathbf{S}) = \mathcal{L}_n(\mathbf{U}\mathbf{V}^\top + \mathbf{S}) + \frac{1}{8}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2. \tag{B.1}$$

Therefore, the corresponding gradient regarding to $\mathbf{Z}$ is as follows

$$\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}, \mathbf{S}) = \begin{bmatrix} \nabla_{\mathbf{U}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top + \mathbf{S}) + \frac{1}{2}\mathbf{U}(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}) \\ \nabla_{\mathbf{V}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top + \mathbf{S}) + \frac{1}{2}\mathbf{V}(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}) \end{bmatrix}. \tag{B.2}$$

B.1 Proof of Theorem 4.6

In order to prove Theorem 4.6, we need to make use of the following lemmas. Since both low-rank and sparse structures exist in our model, it is necessary to derive the convergence results for both structures. Lemma B.1, proved in Section C.1 characterizes the convergence of the low rank structure, while Lemma B.2, proved in Section C.2 corresponds to the convergence of the sparse structure.

Lemma B.1 (Convergence for Low-Rank Structure)**.** Suppose the sample loss function $\mathcal{L}_n$ satisfies Conditions 4.2 and 4.4. Recall that $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$ is the unknown rank-r matrix that satisfies (3.1), $\mathbf{S}^*$ is the unknown s-sparse matrix with at most β-fraction nonzero entries per row and column. There exist constants c_1, c_2 and c_3 such that if $\mathbf{Z}^t \in \mathbb{B}(c_2\sqrt{\sigma_r})$ with $c_2 \leq \min\{1/4, \sqrt{\mu_1'/[10(L_1 + 1 + 8/\mu_2)]}\}$, and we set the step size $\eta = c_1/\sigma_1$ with $c_1 \leq \min\{1/32, \mu_1/(192L_1^2)\}$, then the output of Algorithm 1 $\mathbf{Z}^t = [\mathbf{U}^t; \mathbf{V}^t]$ satisfies

$$d^2(\mathbf{Z}^{t+1}, \mathbf{Z}^*) \leq \rho_1 d^2(\mathbf{Z}^t, \mathbf{Z}^*) - \frac{\eta\mu_1}{4}\|\mathbf{X}^t - \mathbf{X}^*\|_F^2 + \Gamma_1\|\mathbf{S}^t - \mathbf{S}^*\|_F^2 + \Gamma_2\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_2^2,$$

provided that $\beta \leq 1/(c_3\alpha r\kappa)$ with $c_3 \geq 720(\gamma+1)\mu_2/\mu_1'$, where contraction parameter $\rho_1 = 1 - \eta\mu_1'\sigma_r/40$, $\mu_1' = \min\{\mu_1, 2\}$, $\Gamma_1 = 48\eta^2(1+K)^2\sigma_1 + \eta(\mu_2 + 4K^2/\mu_1)$, and $\Gamma_2 = 48\eta^2 r\sigma_1 + 2\eta(8r/\mu_1 + r/L_1)$.

Lemma B.2 (Convergence for Sparse Structure)**.** Suppose the sample loss function $\mathcal{L}_n$ satisfies Conditions 4.3 and 4.4. Recall that $\mathbf{X}^*$ is the unknown rank-r matrix, $\mathbf{S}^*$ is the unknown s-sparse matrix. If we set the step size $\tau \leq 1/(3L_2)$ and choose appropriate parameters γ, γ', then the output of Algorithm 1 satisfies

$$\|\mathbf{S}^{t+1} - \mathbf{S}^*\|_F^2 \leq \rho_2\|\mathbf{S}^t - \mathbf{S}^*\|_F^2 + \Gamma_3\|\mathbf{X}^t - \mathbf{X}^*\|_F^2 + \Gamma_4\|\mathbf{H}^t\|_F^2 + \Gamma_5\|\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_{\infty,\infty}^2.$$

Here, ρ_2 is the contraction parameter satisfying $\rho_2 = C(\gamma, \gamma') \cdot (1 - \mu_2\tau/4) < 1$, where $C(\gamma, \gamma')$ is defined in Theorem 4.6, and Γ_3, Γ_4 and Γ_5 are constants satisfying

$$\Gamma_3 = C(\gamma, \gamma') \cdot \left(\frac{4\tau K^2}{\mu_2} + 3\tau^2(1+K)^2\right), \quad \Gamma_4 = C(\gamma, \gamma') \cdot \frac{\tau(\gamma+1)\beta\alpha r\sigma_1}{\mu_2},$$
$$\Gamma_5 = C(\gamma, \gamma') \cdot \left(\frac{4\tau(\gamma'+1)s}{\mu_2} + 3\tau^2(2\gamma'+1)s\right).$$

Now we are ready to prove Theorem 4.6.

Proof of Theorem 4.6. Given a fixed step size τ, we set γ, γ' such that $\gamma' \geq 1 + 256/(\mu_2^2\tau^2)$ and $\gamma \geq \max\{5, 1 + 64^2/(\mu_2\tau)^2\}$, then we obtain

$$\rho_2 = \left(1 + \sqrt{\frac{2}{\gamma - 1}}\right)^2 \cdot \left(1 + \frac{2}{\sqrt{\gamma' - 1}}\right) \cdot \left(1 - \frac{\mu_2\tau}{4}\right) \leq 1 - \frac{\mu_2\tau}{16}.$$

Consider iteration stage t. According to Lemmas B.1 and B.2, we have

$$d^2(\mathbf{Z}^{t+1},\mathbf{Z}^*)+\frac{1}{\sigma_1}\|\mathbf{S}^{t+1}-\mathbf{S}^*\|_F^2\le\bigg(\rho_1+\frac{\Gamma_4}{\sigma_1}\bigg)\cdot d^2(\mathbf{Z}^t,\mathbf{Z}^*)+\frac{1}{\sigma_1}(\rho_2+\Gamma_1\sigma_1)\cdot\|\mathbf{S}^t-\mathbf{S}^*\|_F^2$$
$$+\bigg(-\frac{\eta\mu_1}{4}+\frac{\Gamma_3}{\sigma_1}\bigg)\cdot\|\mathbf{X}^t-\mathbf{X}^*\|_F^2+\Gamma_2\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2+\frac{\Gamma_5}{\sigma_1}\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_{\infty,\infty}^2.$$

Recall the formula of Γ_1 and Γ_3,Γ_4 from Lemmas B.1 and B.2 respectively. Note that under condition $\eta=c_1/\sigma_1$ and $\beta=c_3/(\alpha r\kappa)$, we can set c_3 to be sufficiently small such that

$$\Gamma_4=C(\gamma,\gamma')\cdot\frac{\tau(\gamma+1)\beta\alpha r\sigma_1}{\mu_2}\le\frac{c_1\mu_1'\sigma_r}{80},$$

where $\mu_1'=\min\{\mu_1,2\}$, which implies that $\rho_1+\Gamma_4/\sigma_1\le 1-\eta\mu_1'\sigma_r/80$. Besides, under condition that K is sufficiently small, we can set $c_1\le\min\{\mu_2/50,\tau/96\}$ such that the following inequality holds

$$\Gamma_1\sigma_1=48c_1^2(1+K)^2+c_1\bigg(\frac{4K^2}{\mu_1}+\mu_2\bigg)\le 50c_1^2+2c_1\mu_2\le 3\mu_2c_1\le\frac{\mu_2\tau}{32}.\tag{B.3}$$

Finally, consider the formula of Γ_3. Note that similarly we can set K to be small enough such that

$$\Gamma_3=C(\gamma,\gamma')\cdot\bigg(\frac{4\tau K^2}{\mu_2}+3\tau^2(1+K)^2\bigg)\le 4\tau^2,$$

thus as long as τ is sufficiently small, there exist c_1 such that $16\tau^2/\mu_1\le c_1\le\min\{\mu_2/50,\tau/96\}$, which implies $\Gamma_3\le c_1\mu_1/4$ while ensuring (B.3) holds as well. Therefore, we obtain

$$d^2(\mathbf{Z}^{t+1},\mathbf{Z}^*)+\frac{1}{\sigma_1}\|\mathbf{S}^{t+1}-\mathbf{S}^*\|_F^2\le\bigg(1-\frac{\eta\mu_1'\sigma_r}{80}\bigg)\cdot d^2(\mathbf{Z}^t,\mathbf{Z}^*)+\frac{1}{\sigma_1}\bigg(1-\frac{\mu_2\tau}{32}\bigg)\cdot\|\mathbf{S}^t-\mathbf{S}^*\|_F^2$$
$$+\Gamma_2\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2+\frac{\Gamma_5}{\sigma_1}\|\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_{\infty,\infty}^2.$$

For simplicity, we denote $D(\mathbf{Z}^t,\mathbf{S}^t)=d^2(\mathbf{Z}^t,\mathbf{Z}^*)+\|\mathbf{S}^t-\mathbf{S}^*\|_F^2/\sigma_1$, and $\rho=\max\big\{1-\eta\mu_1'\sigma_r/80,1-\mu_2\tau/32\big\}\in(0,1)$, then we have

$$D(\mathbf{Z}^{t+1},\mathbf{S}^{t+1})\le\rho D(\mathbf{Z}^t,\mathbf{S}^t)+\Gamma_2\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2+\frac{\Gamma_5}{\sigma_1}\|\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_{\infty,\infty}^2.$$

Recall the formula of Γ_2 and Γ_5 in Lemmas B.1 and B.2 respectively. Under Condition 4.5, we can always set the sample size n to be large enough such that

$$\Gamma_2\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2+\frac{\Gamma_5}{\sigma_1}\|\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_{\infty,\infty}^2\le\Gamma_2\epsilon_1^2(n,\delta)+\frac{\Gamma_5}{\sigma_1}\epsilon_2^2(n,\delta)\le(1-\rho)c_2^2\sigma_r$$

holds with probability at least $1-\delta$. Thus as long as $D(\mathbf{Z}^0,\mathbf{S}^0)\le c_2^2\sigma_r$, we have by induction $D(\mathbf{Z}^t,\mathbf{S}^t)\le c_2^2\sigma_r$ for any $t\ge 0$, which implies $\mathbf{Z}^t\in\mathbb{B}(c_2\sqrt{\sigma_r})$, for any $t\ge 0$. Hence, we obtain

$$D(\mathbf{Z}^t,\mathbf{S}^t)\le\rho^tD(\mathbf{Z}^0,\mathbf{S}^0)+\frac{\Gamma_2}{1-\rho}\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2+\frac{\Gamma_5}{(1-\rho)\sigma_1}\|\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_{\infty,\infty}^2,$$

which completes the proof. □

B.2 Proof of Theorem 4.8

In order to prove Theorem 4.8, we need to make use of the following lemma. Lemma B.3 characterizes a variation of regularity condition for the sample loss function $\mathcal{L}_n$ with respect to the sparse structure, which is proved in Section C.3.

Lemma B.3. Suppose the sample loss function $\mathcal{L}_n$ satisfies Condition 4.3. Given a fixed rank-r matrix $\mathbf{X}$, for any sparse matrices $\mathbf{S}_1, \mathbf{S}_2 \in \mathbb{R}^{d_1 \times d_2}$ with cardinality at most $\gamma' s$, we have

$$\langle \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}+\mathbf{S}_1) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}+\mathbf{S}_2), \mathbf{S}_1 - \mathbf{S}_2 \rangle \geq \frac{\mu_2}{2}\|\mathbf{S}_1 - \mathbf{S}_2\|_F^2 + \frac{1}{2L_2}\big\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}+\mathbf{S}_1) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}+\mathbf{S}_2)\big)\big\|_F^2,$$

where $\Omega \subseteq [d_1] \times [d_2]$ is an index set with cardinality at most $\widetilde{s}$ such that $\mathrm{supp}(\mathbf{S}_1) \subseteq \Omega$ and $\mathcal{P}_\Omega$ is the projection operator onto Ω.

Proof of Theorem 4.8. Consider a fixed iteration ℓ in Algorithm 2. As for the sparse structure, we have

$$\mathbf{S}_{\ell+1} = \mathcal{H}_{\lambda s}(\mathbf{S}_\ell - \tau' \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}_\ell + \mathbf{S}_\ell)).$$

Denote $\Omega' = \mathrm{supp}(\mathbf{S}^*) \cup \mathrm{supp}(\mathbf{S}_\ell) \cup \mathrm{supp}(\mathbf{S}_{\ell+1})$, then we have $\lambda s \leq |\Omega'| \leq (2\lambda+1)s$. We further denote $\widetilde{\mathbf{S}}_{\ell+1} = \mathcal{P}_{\Omega'}\big(\mathbf{S}_\ell - \tau' \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}_\ell + \mathbf{S}_\ell)\big)$, then we obtain $\mathbf{S}_{\ell+1} = \mathcal{H}_{\lambda s}(\widetilde{\mathbf{S}}_{\ell+1})$. Thus, according to Lemma 3.3 in [35], we have

$$\|\mathbf{S}_{\ell+1} - \mathbf{S}^*\|_F^2 \leq \left(1 + \frac{2}{\sqrt{\lambda' - 1}}\right) \cdot \|\widetilde{\mathbf{S}}_{\ell+1} - \mathbf{S}^*\|_F^2. \tag{B.4}$$

Therefore, it is sufficient to upper bound $\|\widetilde{\mathbf{S}}_{\ell+1} - \mathbf{S}^*\|_F$ for the sparse structure. We have

$$\begin{aligned}
\|\widetilde{\mathbf{S}}_{\ell+1} - \mathbf{S}^*\|_F &= \big\|\mathbf{S}_\ell - \mathbf{S}^* - \tau' \mathcal{P}_{\Omega'}\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}_\ell + \mathbf{S}_\ell)\big)\big\|_F \\
&\leq \underbrace{\big\|\mathbf{S}_\ell - \mathbf{S}^* - \tau' \mathcal{P}_{\Omega'}\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}_\ell + \mathbf{S}_\ell) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}_\ell + \mathbf{S}^*)\big)\big\|_F}_{I_1} \\
&\quad + \tau' \underbrace{\|\mathcal{P}_{\Omega'}\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}_\ell + \mathbf{S}^*) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\big)\|_F}_{I_2} + \tau' \underbrace{\|\mathcal{P}_{\Omega'}\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\big)\|_F}_{I_3},
\end{aligned} \tag{B.5}$$

where the second inequality follows from the triangle inequality. As for the first term I_1 in (B.5), according to Lemma B.3, we have

$$\begin{aligned}
I_1^2 &\leq (1 - \mu_2 \tau') \cdot \|\mathbf{S}_\ell - \mathbf{S}^*\|_F^2 - \left(\frac{\tau'}{L_2} - \tau'^2\right) \cdot \|\mathcal{P}_{\Omega'}\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}_\ell + \mathbf{S}_\ell) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}_\ell + \mathbf{S}^*)\big)\|_F^2 \\
&\leq (1 - \mu_2 \tau') \cdot \|\mathbf{S}_\ell - \mathbf{S}^*\|_F^2,
\end{aligned} \tag{B.6}$$

provided that $\tau' \leq 1/L_2$. Consider the second term I_2 in (B.5). Note that $|\Omega'| \leq (2\lambda+1)s$, thus according to the definition of Frobenius norm, we have

$$\begin{aligned}
I_2 &= \sup_{\|\mathbf{W}\|_F \leq 1} \big\langle \mathcal{P}_{\Omega'}\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}_\ell + \mathbf{S}^*) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\big), \mathbf{W} \big\rangle \\
&\leq \sup_{\|\mathbf{W}\|_F \leq 1} \big\{ |\langle \mathbf{X}_\ell - \mathbf{X}^*, \mathcal{P}_{\Omega'}(\mathbf{W}) \rangle| + K \|\mathbf{X}_\ell - \mathbf{X}^*\|_F \cdot \|\mathcal{P}_{\Omega'}(\mathbf{W})\|_F \big\} \\
&\leq \|\mathbf{X}_\ell - \mathbf{X}^*\|_{\infty,\infty} \cdot \|\mathcal{P}_{\Omega'}(\mathbf{W})\|_{1,1} + K \|\mathbf{X}_\ell - \mathbf{X}^*\|_F \leq 4\zeta^* \sqrt{\lambda s} + K \|\mathbf{X}_\ell - \mathbf{X}^*\|_F,
\end{aligned} \tag{B.7}$$

where the first inequality follows from Condition 4.4, the second inequality holds because $|\langle \mathbf{A}, \mathbf{B} \rangle| \leq \|\mathbf{A}\|_{1,1} \cdot \|\mathbf{B}\|_{\infty,\infty}$ and $\|\mathcal{P}_{\Omega'}(\mathbf{W})\|_F \leq \|\mathbf{W}\|_F \leq 1$, and the last inequality is due to the fact that $\|\mathbf{X}_\ell\|_{\infty,\infty} \leq \zeta^*$, $\|\mathbf{X}^*\|_{\infty,\infty} \leq \zeta^*$ and the triangle inequality. And for the third term I_3, we have

$$I_3 \leq \sqrt{(2\lambda+1)s} \cdot \|\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_{\infty,\infty}. \tag{B.8}$$

Therefore, plugging (B.6), (B.7) and (B.8) into (B.5), we obtain

$$\begin{aligned}
\|\widetilde{\mathbf{S}}_{\ell+1} - \mathbf{S}^*\|_F &\leq \sqrt{1 - \mu_2 \tau'} \cdot \|\mathbf{S}_\ell - \mathbf{S}^*\|_F + \tau' K \|\mathbf{X}_\ell - \mathbf{X}^*\|_F + 2\tau' \zeta^* \sqrt{s} \\
&\quad + \tau' \sqrt{(2\lambda+1)s} \cdot \|\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_{\infty,\infty}.
\end{aligned} \tag{B.9}$$

Hence, combining (B.4) and (B.9), we obtain the following result for sparse structure

$$\|\mathbf{S}_{\ell+1}-\mathbf{S}^*\|_F \le \bigg(1+\frac{2}{\sqrt{\lambda}-1}\bigg)\cdot\big(\sqrt{1-\mu_2\tau'}\cdot\|\mathbf{S}_\ell-\mathbf{S}^*\|_F+\tau' K\|\mathbf{X}_\ell-\mathbf{X}^*\|_F\big) + \tau'\bigg(1+\frac{2}{\sqrt{\lambda}-1}\bigg)\cdot(4\zeta^*\sqrt{\lambda s}+\sqrt{3\lambda s}\cdot\|\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_{\infty,\infty}). \tag{B.10}$$

Next, let us consider the low-rank structure. According to Algorithm 2, we have

$$\mathbf{X}_{\ell+1}=\mathcal{P}_{\lambda',\zeta^*}(\mathbf{X}_\ell-\eta'\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}_\ell+\mathbf{S}_\ell)),$$

where the projection operator $\mathcal{P}_{\lambda',\zeta^*}$ is defined as

$$\mathcal{P}_{\lambda',\zeta^*}(\mathbf{X})=\underset{\mathrm{rank}(\mathbf{Y})\le\lambda' r,\,\|\mathbf{Y}\|_{\infty,\infty}\le\zeta^*}{\mathrm{argmin}}\|\mathbf{Y}-\mathbf{X}\|_F,\ \text{for any } \mathbf{X}\in\mathbb{R}^{d_1\times d_2}.$$

Let the singular value decomposition of $\mathbf{X}_\ell,\mathbf{X}_{\ell+1}$ be $\mathbf{X}_\ell=\overline{\mathbf{U}}^\ell\mathbf{\Sigma}^\ell\overline{\mathbf{V}}^{\ell\top}$ and $\mathbf{X}_{\ell+1}=\overline{\mathbf{U}}^{\ell+1}\mathbf{\Sigma}^{\ell+1}\overline{\mathbf{V}}^{\ell+1\top}$ respectively. Define the following subspace spanned by the column vectors of $\overline{\mathbf{U}}^*,\overline{\mathbf{U}}^\ell$ and $\overline{\mathbf{U}}^{\ell+1}$ as

$$\mathrm{span}(\widetilde{\mathbf{U}})=\mathrm{span}\big\{\overline{\mathbf{U}}^*,\overline{\mathbf{U}}^\ell,\overline{\mathbf{U}}^{\ell+1}\big\}=\mathrm{col}(\overline{\mathbf{U}}^*)+\mathrm{col}(\overline{\mathbf{U}}^\ell)+\mathrm{col}(\overline{\mathbf{U}}^{\ell+1}),$$

where each column vector of $\widetilde{\mathbf{U}}$ is a basis vector of the above subspace. Similarly, we define the subspace spanned by the column vectors of $\overline{\mathbf{V}}^*,\overline{\mathbf{V}}^\ell$ and $\overline{\mathbf{V}}^{\ell+1}$ as

$$\mathrm{span}(\widetilde{\mathbf{V}})=\mathrm{span}\big\{\overline{\mathbf{V}}^*,\overline{\mathbf{V}}^\ell,\overline{\mathbf{V}}^{\ell+1}\big\}=\mathrm{col}(\overline{\mathbf{V}}^*)+\mathrm{col}(\overline{\mathbf{V}}^\ell)+\mathrm{col}(\overline{\mathbf{V}}^{\ell+1}),$$

Note that $\mathbf{X}^*$ has rank r, $\mathbf{X}_\ell$ and $\mathbf{X}_{\ell+1}$ has rank at most $\lambda' r$, thus both $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ have at most $(2\lambda'+1)r$ columns. Moreover, we further define the following subspace

$$\mathcal{A}=\big\{\mathbf{\Delta}\in\mathbb{R}^{d_1\times d_2}\,\big|\,\mathrm{row}(\mathbf{\Delta})\subseteq\mathrm{span}(\widetilde{\mathbf{V}})\text{ and }\mathrm{col}(\mathbf{\Delta})\subseteq\mathrm{span}(\widetilde{\mathbf{U}})\big\}.$$

Let $\Pi_{\mathcal{A}}$ be the projection operator onto $\mathcal{A}$, then for any $\mathbf{X}\in\mathbb{R}^{d_1\times d_2}$, we have $\Pi_{\mathcal{A}}(\mathbf{X})=\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\mathbf{X}\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top$. Note that for any $\mathbf{X}\in\mathbb{R}^{d_1\times d_2}$, we have $\mathrm{rank}\big(\Pi_{\mathcal{A}}(\mathbf{X})\big)\le(2\lambda'+1)r$, since $\mathrm{rank}(\mathbf{AB})\le\min\{\mathrm{rank}(\mathbf{A}),\mathrm{rank}(\mathbf{B})\}$. Besides, we denote

$$\widetilde{\mathbf{X}}_{\ell+1}=\mathbf{X}_\ell-\eta'\Pi_{\mathcal{A}}\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}_\ell+\mathbf{S}_\ell)\big).$$

Similar to the proof of Theorem 5.9 in [49], we have $\mathbf{X}_{\ell+1}$ is actually the best rank-$\lambda' r$ approximation of $\widetilde{\mathbf{X}}_{\ell+1}$ satisfying the infinity norm constraint, or in other words, $\mathbf{X}_{\ell+1}=\mathcal{P}_{\lambda',\zeta^*}(\widetilde{\mathbf{X}}_{\ell+1})$. Note that $\mathcal{P}_{\lambda',\zeta^*}(\mathbf{X}^*)=\mathbf{X}^*$, thus according to Lemma 3.18 in [35], we obtain

$$\|\mathbf{X}_{\ell+1}-\mathbf{X}^*\|_F^2=\|\mathcal{P}_{\lambda',\zeta^*}(\widetilde{\mathbf{X}}_{\ell+1})-\mathbf{X}^*\|_F^2\le\bigg(1+\frac{2}{\sqrt{\lambda'}-1}\bigg)\cdot\|\widetilde{\mathbf{X}}_{\ell+1}-\mathbf{X}^*\|_F^2. \tag{B.11}$$

Thus, it suffices to bound the term $\|\widetilde{\mathbf{X}}_{\ell+1}-\mathbf{X}^*\|_F$. Note that $\mathbf{X}^*\in\mathcal{A}$, thus according to the triangle inequality, we have

$$\|\widetilde{\mathbf{X}}_{\ell+1}-\mathbf{X}^*\|_F \le \underbrace{\|\mathbf{X}_\ell-\mathbf{X}^*-\eta'\Pi_{\mathcal{A}}(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}_\ell+\mathbf{S}_\ell)-\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}_\ell))\|_F}_{I_1'} + \eta'\underbrace{\|\Pi_{\mathcal{A}}(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}_\ell)-\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*))\|_F}_{I_2'}+\eta'\underbrace{\|\Pi_{\mathcal{A}}(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*))\|_F}_{I_3'}. \tag{B.12}$$

Consider I_1' in (B.12) first. According to Lemma B.2 in [50], we have

$$I_1'^2\le(1-\eta'\mu_1)\cdot\|\mathbf{X}_\ell-\mathbf{X}^*\|_F^2, \tag{B.13}$$

provided that $\eta' \leq 1/L_1$. As for the second term I_2' in (B.12), by the definition of Frobenius norm, we have

$$I_2' = \sup_{\|\mathbf{W}\|_F \leq 1} \langle \Pi_{\mathcal{A}_{3r}}(\nabla_{\mathbf{X}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}_\ell) - \nabla_{\mathbf{X}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)), \mathbf{W} \rangle$$
$$\leq (1+K) \cdot \|\mathbf{S}_\ell - \mathbf{S}^*\|_F \cdot \|\Pi_{\mathcal{A}}(\mathbf{W})\|_F \leq (1+K) \cdot \|\mathbf{S}_\ell - \mathbf{S}^*\|_F, \tag{B.14}$$

where the first inequality holds because of Condition 4.4. As for I_3', we have

$$I_3' \leq \sqrt{(2\lambda'+1)r} \cdot \|\Pi_{\mathcal{A}}(\nabla_{\mathbf{X}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*))\|_2 \leq \sqrt{(2\lambda'+1)r} \cdot \|\nabla_{\mathbf{X}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_2. \tag{B.15}$$

Therefore, plugging (B.13), (B.14) and (B.15) into (B.12), we obtain

$$\|\widetilde{\mathbf{X}}_{\ell+1} - \mathbf{X}^*\|_F \leq \sqrt{1-\eta'\mu_1} \cdot \|\mathbf{X}_\ell - \mathbf{X}^*\|_F + \eta'(1+K) \cdot \|\mathbf{S}_\ell - \mathbf{S}^*\|_F + \eta'\sqrt{3\lambda' r'} \|\nabla_{\mathbf{X}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_2. \tag{B.16}$$

Finally, combining (B.11) and (B.16), we obtain the following result for low rank structure

$$\|\mathbf{X}_{\ell+1} - \mathbf{X}^*\|_F \leq \left(1 + \frac{2}{\sqrt{\lambda'-1}}\right) \cdot \left(\sqrt{1-\eta'\mu_1} \cdot \|\mathbf{X}_\ell - \mathbf{X}^*\|_F + \eta'(1+K) \cdot \|\mathbf{S}_\ell - \mathbf{S}^*\|_F\right)$$
$$+ \eta' \left(1 + \frac{2}{\sqrt{\lambda'-1}}\right) \cdot \sqrt{3\lambda' r} \|\nabla_{\mathbf{X}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_2. \tag{B.17}$$

Hence, combining (B.10) and (B.17), we obtain

$$\|\mathbf{X}_{\ell+1} - \mathbf{X}^*\|_F + \|\mathbf{S}_{\ell+1} - \mathbf{S}^*\|_F \leq \rho_1' \|\mathbf{X}_\ell - \mathbf{X}^*\|_F + \rho_2' \|\mathbf{S}_\ell - \mathbf{S}^*\|_F + 4\sqrt{\lambda}\tau' \left(1 + \frac{2}{\sqrt{\lambda-1}}\right) \cdot \zeta^* \sqrt{s}$$
$$+ \Gamma_1 \|\nabla_{\mathbf{X}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_2 + \Gamma_2 \|\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_{\infty,\infty}, \tag{B.18}$$

where $\Gamma_1 = \eta'(1 + 2/\sqrt{\lambda'-1})\sqrt{3\lambda' r}$, $\Gamma_2 = \tau'(1 + 2/\sqrt{\lambda-1})\sqrt{3\lambda s}$, and contraction parameter ρ_1', ρ_2' are defined in Theorem 4.8. Note that we set $\eta' = 1/(6\mu_1) \leq 1/L_1$, $\tau' = 3/(4\mu_2) \leq 1/L_2$, and we assume $\mu_1 \geq 1/3$. Then with sufficient large λ and λ' and structural Lipschitz gradient parameter K small enough, we could guarantee $\rho_1', \rho_2' \in (0, 19/20)$. Plugging in the definition of $\zeta^* = c_0 \alpha r \kappa / \sqrt{d_1 d_2}$, we complete the proof by induction. □

B.3 Proof of Theorem 4.9

Proof. To prove Theorem 4.9, it is sufficient to verify the assumption $D(\mathbf{Z}^0, \mathbf{S}^0) \leq c_4^2 \sigma_r$ in Theorem 4.6. Thus, according to Theorem 4.8, it is sufficient to make sure the right hand side of (4.2) is small enough.

As for the optimization error, i.e., the first term on the R.H.S. of (4.2), we can perform $L \geq \log\{c\sigma_r/(2\|\mathbf{X}^*\|_F + 2\|\mathbf{S}^*\|_F\}/\log(\rho')$ iterations in Algorithm 2 to make sure the optimization error is sufficiently small such that $\rho'^L \cdot (\|\mathbf{X}^*\|_F + \|\mathbf{S}^*\|_F) \leq c\sigma_r/2$, where $c = \min\{1/2, c_4/4\}$.

On the other hand, for the statistical error, i.e., the last three terms on the R.H.S. of (4.2), we assume $s \leq cd_1d_2/(\alpha^2 r^2 \kappa^2)$, where c is a small enough constant, and sample size n is sufficiently large such that $\Gamma_1\sqrt{r}\epsilon_1(n,\delta) + \Gamma_2\sqrt{s}\epsilon_2(n,\delta) \leq c\sigma_r/4$. Putting pieces together, we arrive at $\|\mathbf{X}^0 - \mathbf{X}^*\|_F + \|\mathbf{S}^0 - \mathbf{S}^*\|_F \leq c \cdot \sigma_r$. Finally, based on Lemma 5.14 in [46], the initial assumption that $D(\mathbf{Z}^0, \mathbf{S}^0) \leq c_4^2 \sigma_r$ in Theorem 4.6 is satisfied, which completes the proof. □

C Proofs of Technical Lemmas

C.1 Proof of Lemma B.1

In order to prove Theorem B.1, we need to make use of the following lemmas. Lemma C.1 characterizes a local curvature property of the low-rank structure, which gives us the lower bound of the inner product term. We provide its proof in Section D.1. Lemma C.2, proved in Section D.2, characterizes a local smoothness property of the low-rank structure and gives us an upper bound of the Frobenius term.

Lemma C.1 (Local Curvature Property for Low-Rank Structure)**.** Suppose the sample loss function $\mathcal{L}_n$ satisfies Conditions 4.2 and 4.4. Recall that $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$ is the unknown rank-r matrix that satisfies (3.1), and $\mathbf{S}^*$ is the unknown s-sparse matrix. Let $\mathbf{Z} \in \mathbb{R}^{(d_1+d_2)\times r}$ be any matrix with $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$, where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$, $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$

satisfy $\|\mathbf{U}\|_{2,\infty} \leq 2\sqrt{\alpha r\sigma_1/d_1}$ and $\|\mathbf{V}\|_{2,\infty} \leq 2\sqrt{\alpha r\sigma_1/d_2}$. Let $\mathbf{S} \in \mathbb{R}^{d_1\times d_2}$ be any matrix with at most β'-fraction nonzero entries per row and column and satisfying $\|\mathbf{S}\|_0 \leq s' \leq \widetilde{s}$. Denote the optimal rotation with respect to $\mathbf{Z}$ by $\mathbf{R} = \mathrm{argmin}_{\widetilde{\mathbf{R}}\in\mathbb{Q}_r} \|\mathbf{Z} - \mathbf{Z}^*\widetilde{\mathbf{R}}\|_F$, and $\mathbf{H} = \mathbf{Z} - \mathbf{Z}^*\mathbf{R}$, then we have

$$\begin{aligned}\langle \nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z},\mathbf{S}),\mathbf{H}\rangle \geq{}& \frac{\mu_1}{4}\|\mathbf{X}-\mathbf{X}^*\|_F^2 + \frac{1}{16}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2 + \left(\frac{\mu_1'}{20}\sigma_r - C\right)\cdot\|\mathbf{H}\|_F^2 - \left(\frac{L_1+1}{8}+\frac{1}{\mu_2}\right)\cdot\|\mathbf{H}\|_F^4\\ &- \left(\frac{\mu_2}{2}+\frac{2K^2}{\mu_1}\right)\cdot\|\mathbf{S}-\mathbf{S}^*\|_F^2 - \left(\frac{8r}{\mu_1}+\frac{r}{L_1}\right)\cdot\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2,\end{aligned}$$

where $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, $\mu_1' = \min\{\mu_1, 2\}$, and $C = 18(\beta'+\beta)\alpha r\sigma_1/\mu_2$.

Lemma C.2 (Local Smoothness Property for Low-Rank Structure)**.** Suppose the sample loss function $\mathcal{L}_n$ satisfies Conditions 4.2 and 4.4. Recall that $\mathbf{X}^*$ is the unknown rank-r matrix and $\mathbf{S}^*$ is the unknown s-sparse matrix. For any matrix $\mathbf{Z} = [\mathbf{U};\mathbf{V}] \in \mathbb{R}^{(d_1+d_2)\times r}$ and $\mathbf{S} \in \mathbb{R}^{d_1\times d_2}$ with at most s' nonzero entries satisfying $s' \leq \widetilde{s}$, we have

$$\begin{aligned}\|\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z},\mathbf{S})\|_F^2 \leq{}& \left(12L_1^2\|\mathbf{X}-\mathbf{X}^*\|_F^2 + 12(1+K)^2\cdot\|\mathbf{S}-\mathbf{S}^*\|_F^2 + \|\mathbf{U}^\top\mathbf{U}-\mathbf{V}^\top\mathbf{V}\|_F^2\right)\cdot\|\mathbf{Z}\|_2^2\\ &+ 12r\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2\cdot\|\mathbf{Z}\|_2^2,\end{aligned}$$

where $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$.

Proof of Lemma B.1. Recall $\mathbf{Z}^* = [\mathbf{U}^*;\mathbf{V}^*]$ and $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$, where $\mathbf{U}^* = \overline{\mathbf{U}}^*(\mathbf{\Sigma}^*)^{1/2}, \mathbf{V}^* = \overline{\mathbf{V}}^*(\mathbf{\Sigma}^*)^{1/2}$, we have $\|\mathbf{Z}^*\|_2 = \sqrt{2\sigma_1}$. According to our initial ball assumption $\mathbf{Z}^0 \in \mathbb{B}(\sqrt{\sigma_r}/4)$, there exists an orthogonal matrix $\mathbf{R} \in \mathbb{R}^{r\times r}$ such that $\|\mathbf{Z}^0 - \mathbf{Z}^*\mathbf{R}\|_F \leq \sqrt{\sigma_r}/4$, thus we obtain

$$\sqrt{\sigma_1} \leq \|\mathbf{Z}^*\|_2 - \|\mathbf{Z}^0 - \mathbf{Z}^*\mathbf{R}\|_2 \leq \|\mathbf{Z}^0\|_2 \leq \|\mathbf{Z}^*\|_2 + \|\mathbf{Z}^0 - \mathbf{Z}^*\mathbf{R}\|_F \leq 2\sqrt{\sigma_1}.$$

Recall (3.1) and the definition of $\mathcal{C}_1, \mathcal{C}_2$ in (3.2), then it is obvious that $\mathbf{U}^* \in \mathcal{C}_1$ and $\mathbf{V}^* \in \mathcal{C}_2$. Consider a fixed iteration stage t, we denote

$$\begin{aligned}\widetilde{\mathbf{U}}^{t+1} &= \mathbf{U}^t - \eta\nabla_{\mathbf{U}}\mathcal{L}_n(\mathbf{U}^t\mathbf{V}^{t\top} + \mathbf{S}^t) - \frac{1}{2}\eta\mathbf{U}^t(\mathbf{U}^{t\top}\mathbf{U}^t - \mathbf{V}^{t\top}\mathbf{V}^t),\\ \widetilde{\mathbf{V}}^{t+1} &= \mathbf{V}^t - \eta\nabla_{\mathbf{V}}\mathcal{L}_n(\mathbf{U}^t\mathbf{V}^{t\top} + \mathbf{S}^t) - \frac{1}{2}\eta\mathbf{V}^t(\mathbf{V}^{t\top}\mathbf{V}^t - \mathbf{U}^{t\top}\mathbf{U}^t).\end{aligned}$$

Denote $\widetilde{\mathbf{Z}}^{t+1} = [\widetilde{\mathbf{U}}^{t+1};\widetilde{\mathbf{V}}^{t+1}]$, and $\mathbf{Z}^t = [\mathbf{U}^t;\mathbf{V}^t]$, for any iteration stage t, then according to (B.2), we have $\widetilde{\mathbf{Z}}^{t+1} = \mathbf{Z}^t - \eta\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}^t,\mathbf{S}^t)$. Besides, according to Algorithm 1, we obtain

$$\mathbf{U}^{t+1} = \mathcal{P}_{\mathcal{C}_1}(\widetilde{\mathbf{U}}^{t+1}) \quad \text{and} \quad \mathbf{V}^{t+1} = \mathcal{P}_{\mathcal{C}_2}(\widetilde{\mathbf{V}}^{t+1}).$$

Recall $\mathbf{Z}^* = [\mathbf{U}^*;\mathbf{V}^*]$, and $\mathbf{R}^t = \mathrm{argmin}_{\mathbf{R}\in\mathbb{Q}_r}\|\mathbf{Z}^t - \mathbf{Z}^*\mathbf{R}\|_F$, for any t. Denote $\mathbf{H}^t = \mathbf{Z} - \mathbf{Z}^*\mathbf{R}^t$. Since $\mathcal{C}_1$, $\mathcal{C}_2$ are both rotation-invariant constraint sets, and $\mathbf{U}^* \in \mathcal{C}_1$, $\mathbf{V}^* \in \mathcal{C}_2$, we have

$$\begin{aligned}d^2(\mathbf{Z}^{t+1},\mathbf{Z}^*) &\leq \|\mathbf{Z}^{t+1} - \mathbf{Z}^*\mathbf{R}^t\|_F^2\\ &\leq \|\mathbf{Z}^t - \eta\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}^t,\mathbf{S}^t) - \mathbf{Z}^*\mathbf{R}^t\|_F^2\\ &= d^2(\mathbf{Z}^t,\mathbf{Z}^*) - 2\eta\langle\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}^t,\mathbf{S}^t),\mathbf{H}^t\rangle + \eta^2\|\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}^t,\mathbf{S}^t)\|_F^2,\end{aligned} \tag{C.1}$$

where the first inequality follows from Definition 4.1, and the second inequality is due to the nonexpansive property of projection $\mathcal{P}_{\mathcal{C}_i}$ onto $\mathcal{C}_i$, where $i \in \{1,2\}$. Therefore, it suffices to lower bound the inner product term $\langle\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}^t,\mathbf{S}^t),\mathbf{H}^t\rangle$ and upper bound the term $\|\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}^t,\mathbf{S}^t)\|_F^2$, respectively. According to Algorithm 1, we have $(\mathbf{U}^t,\mathbf{V}^t)$ satisfies the condition of $(\mathbf{U},\mathbf{V})$ in Lemma C.1, and $\mathbf{S}^t$ has at most $\gamma\beta$-fraction nonzero entries per row and column with $\|\mathbf{S}\|_0 \leq \gamma' s \leq \widetilde{s}$. Denote $\mathbf{X}^t = \mathbf{U}^t\mathbf{V}^{t\top}$, then according to Lemma C.1, we obtain

$$\begin{aligned}\langle\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}^t,\mathbf{S}^t),\mathbf{H}^t\rangle \geq{}& \frac{\mu_1}{4}\|\mathbf{X}^t-\mathbf{X}^*\|_F^2 + \frac{1}{16}\|\mathbf{U}^{t\top}\mathbf{U}^t - \mathbf{V}^{t\top}\mathbf{V}^t\|_F^2 + \left(\frac{\mu_1'}{20}\sigma_r - C\right)\cdot\|\mathbf{H}^t\|_F^2\\ &- \left(\frac{L_1+1}{8}+\frac{1}{\mu_2}\right)\cdot\|\mathbf{H}^t\|_F^4 - \left(\frac{\mu_2}{2}+\frac{2K^2}{\mu_1}\right)\cdot\|\mathbf{S}^t-\mathbf{S}^*\|_F^2 - \left(\frac{8r}{\mu_1}+\frac{r}{L_1}\right)\cdot\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2,\end{aligned}$$

where $\mu_1' = \min\{\mu_1, 2\}$, and $C = 18(\gamma+1)\beta\alpha r\sigma_1/\mu_2$. Besides, according to Lemma C.2, we have

$$\|\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}^t,\mathbf{S}^t)\|_F^2 \le \Big(12L_1^2\|\mathbf{X}^t-\mathbf{X}^*\|_F^2 + 12(1+K)^2\cdot\|\mathbf{S}^t-\mathbf{S}^*\|_F^2 + \|\mathbf{U}^{t\top}\mathbf{U}^t-\mathbf{V}^{t\top}\mathbf{V}^t\|_F^2\Big)\cdot\|\mathbf{Z}^t\|_2^2 + 12r\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2\cdot\|\mathbf{Z}^t\|_2^2.$$

Note that under the assumption of $\mathbf{Z}^t \in \mathbb{B}(c_2\sqrt{\sigma_r})$, where $c_2 \le 1/4$, we have $\|\mathbf{Z}^t\|_2 \le \|\mathbf{Z}^*\mathbf{R}^t\|_2 + \|\mathbf{Z}^t - \mathbf{Z}^*\mathbf{R}^t\|_2 \le 2\sqrt{\sigma_1}$, since $\|\mathbf{Z}^*\|_2^2 = 2\sigma_1$. Thus, if we set the step size $\eta = c_1/\sigma_1$, where $c_1 \le \min\{1/32, \mu_1/(192L_1^2)\}$, and we assume $\beta \le 1/(c_3\alpha r\kappa)$ with c_3 large enough such that $c_3 \ge 720(\gamma+1)\mu_2/\mu_1'$, we have

$$-2\eta\langle\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}^t,\mathbf{S}^t),\mathbf{H}^t\rangle + \eta^2\|\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}^t,\mathbf{S}^t)\|_F^2 \le -\frac{\eta\mu_1}{4}\|\mathbf{X}^t-\mathbf{X}^*\|_F^2 - \frac{\eta\mu_1'\sigma_r}{20}\|\mathbf{H}^t\|_F^2 + \eta\bigg(\frac{L_1+1}{4}+\frac{2}{\mu_2}\bigg)\cdot\|\mathbf{H}^t\|_F^4 + C_1\|\mathbf{S}^t-\mathbf{S}^*\|_F^2 + C_2\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2,$$

where $C_1 = 48\eta^2(1+K)^2\sigma_1 + \eta(\mu_2 + 4K^2/\mu_1)$, and $C_2 = 48\eta^2 r\sigma_1 + 2\eta(8r/\mu_1 + r/L_1)$. Note that according to our assumption, $\|\mathbf{H}^t\|_F^2 \le c_2^2\sigma_r$ with $c_2^2 \le \mu_1'/[10(L_1+1+8/\mu_2)]$, thus by (C.1), we obtain

$$d^2(\mathbf{Z}^{t+1},\mathbf{Z}^*) \le \bigg(1-\frac{\eta\mu_1'\sigma_r}{40}\bigg)d^2(\mathbf{Z}^t,\mathbf{Z}^*) - \frac{\eta\mu_1}{4}\|\mathbf{X}^t-\mathbf{X}^*\|_F^2 + C_1\|\mathbf{S}^t-\mathbf{S}^*\|_F^2 + C_2\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2,$$

which completes the proof. □

C.2 Proof of Lemma B.2

In order to prove Lemma B.2, we need to utilize the following lemma. Inspired by [56], we present Lemma C.3, which characterizes a nearly non-expansiveness property of the truncation operator $\mathcal{T}_\theta$, as long as θ is large enough. We provides its proof in Section D.3 for completeness.

Lemma C.3. Suppose $\mathbf{S}^* \in \mathbb{R}^{d_1\times d_2}$ is the unknown sparse matrix with at most β-fraction nonzero entries per row and column. For any matrix $\mathbf{S} \in \mathbb{R}^{d_1\times d_2}$, we have

$$\|\mathcal{T}_{\gamma\beta}(\mathbf{S}) - \mathbf{S}^*\|_F^2 \le \bigg(1+\sqrt{\frac{2}{\gamma-1}}\bigg)^2\cdot\|\mathbf{S}-\mathbf{S}^*\|_F^2,$$

where $\gamma > 1$ is a parameter.

Now we are ready to prove Lemma B.2.

Proof of Lemma B.2. Consider a fixed iteration stage t. For the sparse structure, according to Algorithm 1, we have

$$\mathbf{S}^{t+1} = \mathcal{T}_{\gamma\beta}\circ\mathcal{H}_{\gamma' s}\big(\mathbf{S}^t - \tau\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{U}^t\mathbf{V}^{t\top}+\mathbf{S}^t)\big).$$

Denote $\bar{\mathbf{S}}^{t+1} = \mathcal{H}_{\gamma' s}\big(\mathbf{S}^t - \tau\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{U}^t\mathbf{V}^{t\top}+\mathbf{S}^t)\big)$, then we have $\mathbf{S}^{t+1} = \mathcal{T}_{\gamma\beta}(\bar{\mathbf{S}}^{t+1})$. To begin with according to Lemma C.3, we have

$$\|\mathbf{S}^{t+1}-\mathbf{S}^*\|_F^2 = \|\mathcal{T}_{\gamma\beta}(\bar{\mathbf{S}}^{t+1}) - \mathbf{S}^*\|_F^2 \le \bigg(1+\sqrt{\frac{2}{\gamma-1}}\bigg)^2\cdot\|\bar{\mathbf{S}}^{t+1}-\mathbf{S}^*\|_F^2. \tag{C.2}$$

Moreover, denote $\Omega = \Omega^*\cup\Omega^t\cup\Omega^{t+1}$, where $\Omega^* = \mathrm{supp}(\mathbf{S}^*)$, $\Omega^t = \mathrm{supp}(\mathbf{S}^t)$ and $\Omega^{t+1} = \mathrm{supp}(\bar{\mathbf{S}}^{t+1})$. Obviously, the cardinality of Ω satisfies $\gamma' s \le |\Omega| \le (2\gamma'+1)s$. Based on Ω, we define $\widetilde{\mathbf{S}}^{t+1}$ as follows

$$\widetilde{\mathbf{S}}^{t+1} = \mathcal{P}_\Omega\big(\mathbf{S}^t - \tau\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{U}^t\mathbf{V}^{t\top}+\mathbf{S}^t)\big) = \mathbf{S}^t - \tau\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{U}^t\mathbf{V}^{t\top}+\mathbf{S}^t)\big), \tag{C.3}$$

where $\mathcal{P}_\Omega$ is the projection operator onto the index set Ω. Note that $\Omega^{t+1}\subseteq\Omega$, thus we have $\bar{\mathbf{S}}^{t+1} = \mathcal{H}_{\gamma s}(\widetilde{\mathbf{S}}^{t+1})$. According to Lemma 3.3 in [35], we have

$$\|\bar{\mathbf{S}}^{t+1}-\mathbf{S}^*\|_F^2 \le \bigg(1+\frac{2}{\sqrt{\gamma'-1}}\bigg)\cdot\|\widetilde{\mathbf{S}}^{t+1}-\mathbf{S}^*\|_F^2. \tag{C.4}$$

Therefore, it is sufficient to upper bound $\|\widetilde{\mathbf{S}}^{t+1} - \mathbf{S}^*\|_F^2$. By (C.3), we have

$$\|\widetilde{\mathbf{S}}^{t+1} - \mathbf{S}^*\|_F^2 = \|\mathbf{S}^t - \mathbf{S}^*\|_F^2 - 2\tau \underbrace{\langle \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^t), \mathbf{S}^t - \mathbf{S}^* \rangle}_{I_1} + \tau^2 \underbrace{\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^t)\big)\|_F^2}_{I_2}, \tag{C.5}$$

where the equality holds because $\langle \mathcal{P}_\Omega(\mathbf{A}), \mathbf{B} \rangle = \langle \mathbf{A}, \mathcal{P}_\Omega(\mathbf{B}) \rangle$. In the following discussions, we are going to bound I_1 and I_2 respectively. Consider the term I_1 first, we have

$$\begin{aligned} I_1 = &\underbrace{\langle \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^t) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^*), \mathbf{S}^t - \mathbf{S}^* \rangle}_{I_{11}} + \underbrace{\langle \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^*) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*), \mathbf{S}^t - \mathbf{S}^* \rangle}_{I_{12}} \\ &+ \underbrace{\langle \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*), \mathbf{S}^t - \mathbf{S}^* \rangle}_{I_{13}}. \end{aligned} \tag{C.6}$$

As for the first term I_{11} in (C.6), according to Lemma B.3, we have

$$I_{11} \geq \frac{\mu_2}{2} \|\mathbf{S}^t - \mathbf{S}^*\|_F^2 + \frac{1}{2L_2} \big\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^t) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^*)\big)\big\|_F^2. \tag{C.7}$$

Note that we have $\mathrm{supp}(\mathbf{S}^t - \mathbf{S}^*) \subseteq \Omega^t \cup \Omega^*$, where $\Omega^t \cup \Omega^*$ has at most $(\gamma+1)\beta$-fraction nonzero entries per row and column. Denote $\mathbf{R}^t$ as the optimal rotation with respect to $\mathbf{Z}^t = [\mathbf{U}^t; \mathbf{V}^t]$, and $\mathbf{H}^t = \mathbf{Z}^t - \mathbf{Z}^* \mathbf{R}^t$. According to Condition 4.4, we obtain the bound of I_{12} in (C.6)

$$\begin{aligned} |I_{12}| &\leq |\langle \mathbf{X}^t - \mathbf{X}^*, \mathbf{S}^t - \mathbf{S}^* \rangle| + K \|\mathbf{X}^t - \mathbf{X}^*\|_F \cdot \|\mathbf{S}^t - \mathbf{S}^*\|_F \\ &\leq \|\mathcal{P}_{\Omega^t \cup \Omega^*}(\mathbf{X}^t - \mathbf{X}^*)\|_F \cdot \|\mathbf{S}^t - \mathbf{S}^*\|_F + K \|\mathbf{X}^t - \mathbf{X}^*\|_F \cdot \|\mathbf{S}^t - \mathbf{S}^*\|_F \\ &\leq \sqrt{18(\gamma+1)\beta\alpha r \sigma_1} \|\mathbf{H}^t\|_F \cdot \|\mathbf{S}^t - \mathbf{S}^*\|_F + K \|\mathbf{X}^t - \mathbf{X}^*\|_F \cdot \|\mathbf{S}^t - \mathbf{S}^*\|_F, \end{aligned} \tag{C.8}$$

where the second inequality holds because $|\langle \mathbf{A}, \mathbf{B} \rangle| \leq \|\mathbf{A}\|_F \cdot \|\mathbf{B}\|_F$, and the last inequality follows from Lemma 14 in [56]. As for the last term I_{13} in (C.6), we have

$$|I_{13}| \leq \|\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_{\infty,\infty} \cdot \|\mathbf{S}^t - \mathbf{S}^*\|_{1,1} \leq \sqrt{(\gamma'+1)s} \cdot \|\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_{\infty,\infty} \cdot \|\mathbf{S}^t - \mathbf{S}^*\|_F, \tag{C.9}$$

where the first inequality holds because $|\langle \mathbf{A}, \mathbf{B} \rangle| \leq \|\mathbf{A}\|_{\infty,\infty} \cdot \|\mathbf{B}\|_{1,1}$, and the second inequality follows from the fact that $\mathbf{S}^t - \mathbf{S}^*$ has at most $(\gamma'+1)s$ nonzero entries. Therefore, plugging (C.7), (C.8) and (C.9) into (C.6), we obtain the lower bound of I_1

$$\begin{aligned} I_1 \geq &\frac{\mu_2}{8} \|\mathbf{S}^t - \mathbf{S}^*\|_F^2 + \frac{1}{2L_2} \big\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^t) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^*)\big)\big\|_F^2 - \frac{2K^2}{\mu_2} \|\mathbf{X}^t - \mathbf{X}^*\|_F^2 \\ &- \frac{36(\gamma+1)\beta\alpha r \sigma_1}{\mu_2} \|\mathbf{H}^t\|_F^2 - \frac{2(\gamma'+1)s}{\mu_2} \|\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_{\infty,\infty}^2. \end{aligned} \tag{C.10}$$

Next, consider the term I_2 in (C.5). We have

$$\begin{aligned} I_2 \leq &3\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^t) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^*)\big\|_F^2 + 3\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^*) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\big\|_F^2 \\ &+ 3\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_F^2 \end{aligned} \tag{C.11}$$

As for the second term in (C.11), according to the definition of Frobenius norm, we have

$$\begin{aligned} \|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^*) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\big)\|_F &= \sup_{\|\mathbf{W}\| \leq 1} \langle \mathcal{P}_\Omega\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^*) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\big), \mathbf{W} \rangle \\ &= \sup_{\|\mathbf{W}\| \leq 1} \langle \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^t + \mathbf{S}^*) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*), \mathcal{P}_\Omega(\mathbf{W}) \rangle \\ &\leq (1+K) \cdot \|\mathbf{X}^t - \mathbf{X}^*\|_F \cdot \|\mathcal{P}_\Omega(\mathbf{W})\|_F \\ &\leq (1+K) \cdot \|\mathbf{X}^t - \mathbf{X}^*\|_F, \end{aligned} \tag{C.12}$$

where the second equality holds because $\langle \mathcal{P}_\Omega(\mathbf{A}), \mathbf{B} \rangle = \langle \mathbf{A}, \mathcal{P}_\Omega(\mathbf{B}) \rangle$, and the first inequality holds because of Condition 4.4. As for the last term in (C.11), note that $|\Omega| \leq (2\gamma'+1)s$, thus we have

$$\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_F^2 \leq (2\gamma'+1)s \|\nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_{\infty,\infty}^2. \tag{C.13}$$

Therefore, plugging (C.12) and (C.13) into (C.11), we obtain the upper bound of I_2

$$I_2 \leq 3\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^t+\mathbf{S}^t)-\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^t+\mathbf{S}^*)\big)\|_F^2+3(1+K)^2\cdot\|\mathbf{X}^t-\mathbf{X}^*\|_F^2 \\ +3(2\gamma'+1)s\|\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_{\infty,\infty}^2. \tag{C.14}$$

If we set the step size $\tau \leq 1/(3L_2)$, then by plugging (C.10) and (C.14) into (C.5), we have

$$\|\widetilde{\mathbf{S}}^{t+1}-\mathbf{S}^*\|_F^2 \leq \left(1-\frac{\mu_2\tau}{4}\right)\cdot\|\mathbf{S}^t-\mathbf{S}^*\|_F^2+C_3\|\mathbf{X}^t-\mathbf{X}^*\|_F^2+C_4\|\mathbf{H}^t\|_F^2+C_5\|\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_{\infty,\infty}^2, \tag{C.15}$$

where $C_3=4\tau K^2/\mu_2+3\tau^2(1+K)^2$, $C_4=72\tau(\gamma+1)\beta\alpha r\sigma_1/\mu_2$ and $C_5=4\tau(\gamma'+1)s/\mu_2+3\tau^2(2\gamma'+1)s$. Thus combining (C.2), (C.4) and (C.15), we obtain

$$\|\mathbf{S}^{t+1}-\mathbf{S}^*\|_F^2 \leq \rho\|\mathbf{S}^t-\mathbf{S}^*\|_F^2+C(\gamma,\gamma')\cdot\big(C_3\|\mathbf{X}^t-\mathbf{X}^*\|_F^2+C_4\|\mathbf{H}^t\|_F^2+C_5\|\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_{\infty,\infty}^2\big),$$

which completes the proof. □

C.3 Proof of Lemma B.3

In order to proof Lemma B.3, we need to make use of the following lemma, which can be derived following the standard proof of Lipschitz continuous gradient property [40].

Lemma C.4. Suppose the sample loss function $\mathcal{L}_n$ satisfies Conditions 4.3. Given a fixed rank-r matrix $\mathbf{X}\in\mathbb{R}^{d_1\times d_2}$, then for any sparse matrices $\mathbf{S}_1,\mathbf{S}_2\in\mathbb{R}^{d_1\times d_2}$ with cardinality at most $\widetilde{s}$, we have

$$\mathcal{L}_n(\mathbf{X}+\mathbf{S}_1)\geq\mathcal{L}_n(\mathbf{X}+\mathbf{S}_2)+\big\langle\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_2),\mathbf{S}_1-\mathbf{S}_2\big\rangle \\ +\frac{1}{2L_2}\big\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_1)-\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_2)\big)\big\|_F^2,$$

where $\Omega\subseteq[d_1]\times[d_2]$ is an index set with cardinality at most $\widetilde{s}$ such that $\mathrm{supp}(\mathbf{S}_1)\subseteq\Omega$ and $\mathcal{P}_\Omega$ is the projection operator onto Ω.

Now we are ready to prove Lemma B.3.

Proof of Lemma B.3. Since the sample loss function $\mathcal{L}_n$ satisfies the restricted strong convexity Condition 4.3, we have

$$\mathcal{L}_n(\mathbf{X}+\mathbf{S}_2)\geq\mathcal{L}_n(\mathbf{X}+\mathbf{S}_1)+\langle\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_1),\mathbf{S}_2-\mathbf{S}_1\rangle+\frac{\mu_2}{2}\|\mathbf{S}_2-\mathbf{S}_1\|_F^2. \tag{C.16}$$

According to Lemma C.4, we have

$$\mathcal{L}_n(\mathbf{X}+\mathbf{S}_1)\geq\mathcal{L}_n(\mathbf{X}+\mathbf{S}_2)+\big\langle\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_2),\mathbf{S}_1-\mathbf{S}_2\big\rangle \\ +\frac{1}{2L_2}\big\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_1)-\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_2)\big)\big\|_F^2. \tag{C.17}$$

Therefore, combining (C.16) and (C.17), we obtain

$$\langle\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_1)-\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_2),\mathbf{S}_1-\mathbf{S}_2\rangle\geq\frac{\mu_2}{2}\|\mathbf{S}_1-\mathbf{S}_2\|_F^2 \\ +\frac{1}{2L_2}\big\|\mathcal{P}_\Omega\big(\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_1)-\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_2)\big)\big\|_F^2,$$

which completes the proof. □

D Proofs of Auxiliary Lemmas in Appendix C

To begin with, we introduce the following notations for simplicity. Consider $\mathbf{Z} \in \mathbb{R}^{(d_1+d_2)\times r}$, for $\mathbf{U} \in \mathbb{R}^{d_1\times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2\times r}$, and $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, we let $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$. Let $\mathbf{R} = \operatorname{argmin}_{\widetilde{\mathbf{R}}\in\mathbb{Q}_r} \|\mathbf{Z} - \mathbf{Z}^*\widetilde{\mathbf{R}}\|_F$ be the optimal rotation regarding to $\mathbf{Z}$, and $\mathbf{H} = \mathbf{Z} - \mathbf{Z}^*\mathbf{R} = [\mathbf{H}_U; \mathbf{H}_V]$ with $\mathbf{H}_U \in \mathbb{R}^{d_1\times r}$ and $\mathbf{H}_V \in \mathbb{R}^{d_2\times r}$.

Besides, we introduce the following projection metrics, which are essential for proving the following lemmas. Denote by $\overline{\mathbf{U}}_1, \overline{\mathbf{U}}_2, \overline{\mathbf{U}}_3$ the left singular matrices of $\mathbf{X}, \mathbf{U}, \mathbf{H}_U$ respectively. Let $\widetilde{\mathbf{U}}$ be the matrix spanned by the column vectors of $\overline{\mathbf{U}}_1, \overline{\mathbf{U}}_2$ and $\overline{\mathbf{U}}_3$, i.e.,

$$\operatorname{col}(\widetilde{\mathbf{U}}) = \operatorname{span}\big\{\overline{\mathbf{U}}_1, \overline{\mathbf{U}}_2, \overline{\mathbf{U}}_3\big\} = \operatorname{col}(\overline{\mathbf{U}}_1) + \operatorname{col}(\overline{\mathbf{U}}_2) + \operatorname{col}(\overline{\mathbf{U}}_3). \tag{D.1}$$

It is easy to show that $\widetilde{\mathbf{U}}$ is an orthonormal matrix with at most $3r$ columns. Here, the sum of two subspaces is defined as $\mathbf{U}_1 + \mathbf{U}_2 = \{\mathbf{u}_1 + \mathbf{u}_2 \,|\, \mathbf{u}_1 \in \mathbf{U}_1, \mathbf{u}_2 \in \mathbf{U}_2\}$. Similarly, denote by $\overline{\mathbf{V}}_1, \overline{\mathbf{V}}_2, \overline{\mathbf{V}}_3$ the right singular matrices of $\mathbf{X}, \mathbf{V}, \mathbf{H}_V$ respectively. Again, let $\widetilde{\mathbf{V}}$ be the matrix spanned by the column of $\overline{\mathbf{V}}_1, \overline{\mathbf{V}}_2$ and $\overline{\mathbf{V}}_3$, i.e.,

$$\operatorname{col}(\widetilde{\mathbf{V}}) = \operatorname{span}\big\{\overline{\mathbf{V}}_1, \overline{\mathbf{V}}_2, \overline{\mathbf{V}}_3\big\} = \operatorname{col}(\overline{\mathbf{V}}_1) + \operatorname{col}(\overline{\mathbf{V}}_2) + \operatorname{col}(\overline{\mathbf{V}}_3), \tag{D.2}$$

where the rank of $\widetilde{\mathbf{V}}$ is at most $3r$.

D.1 Proof of Lemma C.1

Proof. Recall $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$. We denote $\widetilde{\mathbf{Z}} = [\mathbf{U}; -\mathbf{V}] \in \mathbb{R}^{(d_1+d_2)\times r}$, then we can rewrite the regularization term $\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2$ as $\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2$ and its gradient with respect to $\mathbf{Z}$ as $\nabla_{\mathbf{Z}}(\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2) = 4\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top\mathbf{Z}$. According to the formula of $\nabla\widetilde{F}_n(\mathbf{Z}, \mathbf{S})$ in (B.2), we have

$$\langle\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z}, \mathbf{S}), \mathbf{H}\rangle = \underbrace{\langle\nabla_{\mathbf{U}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top + \mathbf{S}), \mathbf{H}_U\rangle + \langle\nabla_{\mathbf{V}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top + \mathbf{S}), \mathbf{H}_V\rangle}_{I_1} + \frac{1}{2}\underbrace{\langle\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top\mathbf{Z}, \mathbf{H}\rangle}_{I_2}, \tag{D.3}$$

where $\widetilde{\mathbf{Z}} = [\mathbf{U}; -\mathbf{V}]$, and $\mathbf{H}_U, \mathbf{H}_V$ denote the top $d_1 \times r$ and bottom $d_2 \times r$ submatrices of $\mathbf{H}$ respectively. Note that $\nabla_{\mathbf{U}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top + \mathbf{S}) = \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X} + \mathbf{S})\mathbf{V}$, and $\nabla_{\mathbf{V}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top + \mathbf{S}) = [\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X} + \mathbf{S})]^\top\mathbf{U}$. Consider the term I_1 in (D.3) first, we have

$$\begin{aligned}
I_1 &= \langle\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X} + \mathbf{S}), \mathbf{U}\mathbf{V}^\top - \mathbf{U}^*\mathbf{V}^{*\top} + \mathbf{H}_U\mathbf{H}_V^\top\rangle \\
&= \underbrace{\langle\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*), \mathbf{X} - \mathbf{X}^* + \mathbf{H}_U\mathbf{H}_V^\top\rangle}_{I_{11}} + \underbrace{\langle\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*), \mathbf{X} - \mathbf{X}^* + \mathbf{H}_U\mathbf{H}_V^\top\rangle}_{I_{12}} \\
&+ \underbrace{\langle\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X} + \mathbf{S}) - \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}), \mathbf{X} - \mathbf{X}^* + \mathbf{H}_U\mathbf{H}_V^\top\rangle}_{I_{13}}.
\end{aligned} \tag{D.4}$$

In the following discussions, we are going to bound I_{11}, I_{12} and I_{13} respectively. For the first term I_{11} in (D.4), we have

$$\begin{aligned}
|I_{11}| &\le \|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_2 \cdot \big(\|\mathbf{X} - \mathbf{X}^*\|_* + \|\mathbf{H}_U\mathbf{H}_V^\top\|_*\big) \\
&\le \|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_2 \cdot \big(\sqrt{2r}\|\mathbf{X} - \mathbf{X}^*\|_F + \sqrt{r}\|\mathbf{H}_U\mathbf{H}_V\|_F\big) \\
&\le \frac{\mu_1}{16}\|\mathbf{X} - \mathbf{X}^*\|_F^2 + \frac{L_1}{16}\|\mathbf{H}\|_F^4 + \left(\frac{8r}{\mu_1} + \frac{r}{L_1}\right) \cdot \|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*)\|_2^2,
\end{aligned} \tag{D.5}$$

where the first inequality holds because of Von Neumann trace inequality, the second inequality is due to $\mathbf{X} - \mathbf{X}^*$ has rank at most $2r$ and $\mathbf{H}_U\mathbf{H}_V^\top$ has rank at most r, and the last inequality holds because $\|\mathbf{H}_U\mathbf{H}_V^\top\|_F \le \|\mathbf{H}_U\| \cdot \|\mathbf{H}_V\|_F \le \|\mathbf{H}\|_F^2/2$ and $2ab \le ta^2 + b^2/t$, for any $t > 0$. As for the second term I_{12} in (D.4), note that $\mathbf{X} - \mathbf{X}^* + \mathbf{H}_U\mathbf{H}_V^\top$ has rank at most $3r$, thus according to the structural Lipschitz gradient Condition 4.4, we have

$$\begin{aligned}
|I_{12}| &\le |\langle\mathbf{S} - \mathbf{S}^*, \mathbf{X} - \mathbf{X}^* + \mathbf{H}_U\mathbf{H}_V^\top\rangle| + K\|\mathbf{X} - \mathbf{X}^* + \mathbf{H}_U\mathbf{H}_V^\top\|_F \cdot \|\mathbf{S} - \mathbf{S}^*\|_F \\
&\le |\langle\mathbf{S} - \mathbf{S}^*, \mathbf{X} - \mathbf{X}^*\rangle| + \|\mathbf{S} - \mathbf{S}^*\|_F \cdot \|\mathbf{H}_U\mathbf{H}_V^\top\|_F + K\|\mathbf{X} - \mathbf{X}^* + \mathbf{H}_U\mathbf{H}_V^\top\|_F \cdot \|\mathbf{S} - \mathbf{S}^*\|_F \\
&\le |\langle\mathbf{S} - \mathbf{S}^*, \mathbf{X} - \mathbf{X}^*\rangle| + \frac{1 + K}{2}\|\mathbf{S} - \mathbf{S}^*\|_F \cdot \|\mathbf{H}\|_F^2 + K\|\mathbf{X} - \mathbf{X}^*\|_F \cdot \|\mathbf{S} - \mathbf{S}^*\|_F,
\end{aligned} \tag{D.6}$$

where the second inequality follows from triangle inequality and the fact that $|\langle \mathbf{A}, \mathbf{B}\rangle| \leq \|\mathbf{A}\|_F \cdot \|\mathbf{B}\|_F$, and the last inequality is due to triangle inequality and the fact that $\|\mathbf{H}_U\mathbf{H}_V^\top\|_F \leq \|\mathbf{H}_U\|_F \cdot \|\mathbf{H}_V\|_F \leq \|\mathbf{H}\|_F^2/2$. Therefore, it suffices to bound the first term $|\langle \mathbf{S}-\mathbf{S}^*, \mathbf{X}-\mathbf{X}^*\rangle|$. Denote the support of $\mathbf{S}-\mathbf{S}^*$ by Ω, then according to our assumption, Ω has at most $\beta'+\beta$ fraction nonzero entries per row and column. By Lemma 14 in [56], we further obtain

$$
\begin{aligned}
|I_{12}| &\leq \|\mathbf{S}-\mathbf{S}^*\|_F \cdot \|\mathcal{P}_\Omega(\mathbf{U}\mathbf{V}^\top - \mathbf{U}\mathbf{V}^*)\|_F + \frac{1+K}{2}\|\mathbf{S}-\mathbf{S}^*\|_F \cdot \|\mathbf{H}\|_F^2 + K\|\mathbf{X}-\mathbf{X}^*\|_F \cdot \|\mathbf{S}-\mathbf{S}^*\|_F \\
&\leq \sqrt{18(\beta'+\beta)\alpha r\sigma_1}\|\mathbf{H}\|_F \cdot \|\mathbf{S}-\mathbf{S}^*\|_F + \frac{1+K}{2}\|\mathbf{S}-\mathbf{S}^*\|_F \cdot \|\mathbf{H}\|_F^2 + K\|\mathbf{X}-\mathbf{X}^*\|_F \cdot \|\mathbf{S}-\mathbf{S}^*\|_F \\
&\leq \frac{\mu_1}{8}\|\mathbf{X}-\mathbf{X}^*\|_F^2 + \bigg(\frac{\mu_2}{2} + \frac{2K^2}{\mu_1}\bigg) \cdot \|\mathbf{S}-\mathbf{S}^*\|_F^2 + \frac{18(\beta'+\beta)\alpha r\sigma_1}{\mu_2}\|\mathbf{H}\|_F^2 + \frac{(1+K)^2}{4\mu_2}\|\mathbf{H}\|_F^4, \qquad \text{(D.7)}
\end{aligned}
$$

where the first inequality holds because $|\langle \mathbf{A}, \mathcal{P}_\Omega(\mathbf{B})\rangle| \leq \|\mathcal{P}_\Omega(\mathbf{A})\|_F \cdot \|\mathbf{B}\|_F$, and the second inequality is due to Lemma 14 in [56], and the last inequality holds because $2ab \leq ta^2 + b^2/t$, for any $t > 0$. Finally, we consider the last term I_{13} in (D.4). Recall the orthonormal projection matrices $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ in (D.1) and (D.2). According to Lemma B.2 in [50], we have

$$
\begin{aligned}
\langle \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}), \mathbf{X}-\mathbf{X}^*\rangle &\geq \frac{1}{4L_1}\|\widetilde{\mathbf{U}}^\top(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}))\|_F^2 \\
&\quad + \frac{1}{4L_1}\|(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}))\widetilde{\mathbf{V}}\|_F^2 + \frac{\mu_1}{2}\|\mathbf{X}-\mathbf{X}^*\|_F^2. \qquad \text{(D.8)}
\end{aligned}
$$

As for the remaining term in I_{13}, we have

$$
\begin{aligned}
|\langle \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}), \mathbf{H}_U\mathbf{H}_V^\top\rangle| &= |\langle \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}), \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\mathbf{H}_U\mathbf{H}_V^\top\rangle| \\
&\leq \frac{1}{2}\big\|\widetilde{\mathbf{U}}^\top\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})\big)\|_F \cdot \|\mathbf{H}\|_F^2 \\
&\leq \frac{1}{2L_1}\big\|\widetilde{\mathbf{U}}^\top\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})\big\|_F^2 + \frac{L_1}{8}\|\mathbf{H}\|_F^4, \qquad \text{(D.9)}
\end{aligned}
$$

where the equality is due to the fact that $\mathrm{col}(\overline{\mathbf{U}}_3) \subseteq \mathrm{col}(\widetilde{\mathbf{U}})$, where $\overline{\mathbf{U}}_3$ is the left singular matrix of $\mathbf{H}_U$, which implies that $\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\mathbf{H}_U = \mathbf{H}_U$, the first inequality holds because $|\langle \mathbf{A}, \mathbf{B}\mathbf{C}\rangle| \leq \|\mathbf{A}\|_F \cdot \|\mathbf{B}\mathbf{C}\|_F \leq \|\mathbf{A}\|_F \cdot \|\mathbf{B}\|_2 \cdot \|\mathbf{C}\|_F$ and $\|\widetilde{\mathbf{U}}\|_2 = 1$, and the last inequality holds because $2ab \leq ta^2 + b^2/t$, for any $t > 0$. Similarly, we have

$$
\begin{aligned}
|\langle \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}), \mathbf{H}_U\mathbf{H}_V^\top\rangle| &\leq \frac{1}{2}\big\|\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})\big)\widetilde{\mathbf{V}}\|_F \cdot \|\mathbf{H}\|_F^2 \\
&\leq \frac{1}{2L_1}\big\|\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})\big)\widetilde{\mathbf{V}}\|_F^2 + \frac{L_1}{8}\|\mathbf{H}\|_F^4. \qquad \text{(D.10)}
\end{aligned}
$$

Therefore, combining (D.8), (D.9) and (D.10), we obtain the lower bound of I_{13}

$$
I_{13} \geq \frac{\mu_1}{2}\|\mathbf{X}-\mathbf{X}^*\|_F^2 - \frac{L_1}{8}\|\mathbf{H}\|_F^4. \qquad \text{(D.11)}
$$

Hence, combining (D.5), (D.7) and (D.11), we further obtain the lower bound of I_1 in (D.3)

$$
\begin{aligned}
I_1 \geq \frac{3\mu_1}{8}\|\mathbf{X}-\mathbf{X}^*\|_F^2 - \frac{18(\beta'+\beta)\alpha r\sigma_1}{\mu_2}\|\mathbf{H}\|_F^2 - \bigg(\frac{L_1}{8} + \frac{(1+K)^2}{4\mu_2}\bigg) \cdot \|\mathbf{H}\|_F^4 \\
- \bigg(\frac{\mu_2}{2} + \frac{2K^2}{\mu_1}\bigg) \cdot \|\mathbf{S}-\mathbf{S}^*\|_F^2 - \bigg(\frac{8r}{\mu_1} + \frac{r}{L_1}\bigg) \cdot \|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2. \qquad \text{(D.12)}
\end{aligned}
$$

Besides, according to Lemma C.1 in [49], we obtain the following lower bound regarding I_2 in (D.3)

$$
I_2 \geq \frac{1}{2}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2 - \frac{1}{2}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F \cdot \|\mathbf{H}\|_F^2 \geq \frac{1}{4}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2 - \frac{1}{4}\|\mathbf{H}\|_F^4. \qquad \text{(D.13)}
$$

Note that $K \in (0,1)$, by plugging (D.12) and (D.13) into (D.3), we have

$$\langle \nabla_{\mathbf{Z}} \widetilde{F}_n(\mathbf{Z},\mathbf{S}), \mathbf{H}\rangle \geq \frac{3\mu_1}{8}\|\mathbf{X}-\mathbf{X}^*\|_F^2 + \frac{1}{8}\|\widetilde{\mathbf{Z}}^\top \mathbf{Z}\|_F^2 - \frac{18(\beta'+\beta)\alpha r\sigma_1}{\mu_2}\|\mathbf{H}\|_F^2 - \left(\frac{L_1+1}{8}+\frac{1}{\mu_2}\right)\cdot\|\mathbf{H}\|_F^4 - \left(\frac{\mu_2}{2}+\frac{2K^2}{\mu_1}\right)\cdot\|\mathbf{S}-\mathbf{S}^*\|_F^2 - \left(\frac{8r}{\mu_1}+\frac{r}{L_1}\right)\cdot\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2. \tag{D.14}$$

Furthermore, denote $\widetilde{\mathbf{Z}}^* = [\mathbf{U}^*; -\mathbf{V}^*]$, then we obtain the following result

$$\begin{aligned}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2 &= \langle \mathbf{Z}\mathbf{Z}^\top - \mathbf{Z}^*\mathbf{Z}^{*\top}, \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top - \widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top}\rangle + \langle \mathbf{Z}^*\mathbf{Z}^{*\top}, \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top\rangle + \langle \mathbf{Z}\mathbf{Z}^\top, \widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top}\rangle \\ &\geq \langle \mathbf{Z}\mathbf{Z}^\top - \mathbf{Z}^*\mathbf{Z}^{*\top}, \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top - \widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top}\rangle \\ &= \|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^*\mathbf{U}^{*\top}\|_F^2 + \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^*\mathbf{V}^{*\top}\|_F^2 - 2\|\mathbf{X}-\mathbf{X}^*\|_F^2,\end{aligned} \tag{D.15}$$

where the first equality follows from $\widetilde{\mathbf{Z}}^{*\top}\mathbf{Z}^* = 0$, and the inequality is due to the fact that $\langle \mathbf{A}\mathbf{A}^\top, \mathbf{B}\mathbf{B}^\top\rangle = \|\mathbf{A}^\top\mathbf{B}\|_F^2 \geq 0$. Therefore, by (D.15), we obtain

$$4\|\mathbf{X}-\mathbf{X}^*\|_F^2 + \|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2 = \|\mathbf{Z}\mathbf{Z}^\top - \mathbf{Z}^*\mathbf{Z}^{*\top}\|_F^2 \geq 4(\sqrt{2}-1)\sigma_r\|\mathbf{H}\|_F^2, \tag{D.16}$$

where the inequality follows from Lemma 5.4 in [46] and the fact that $\sigma_r^2(\mathbf{Z}^*) = 2\sigma_r(\mathbf{X}^*) = 2\sigma_r$. Denote $\mu_1' = \min\{\mu_1, 2\}$, then by plugging (D.16) into (D.14), we obtain

$$\begin{aligned}\langle \nabla_{\mathbf{Z}} \widetilde{F}_n(\mathbf{Z},\mathbf{S}), \mathbf{H}\rangle \geq{}& \frac{\mu_1}{4}\|\mathbf{X}-\mathbf{X}^*\|_F^2 + \frac{1}{16}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2 + \left(\frac{\mu_1'}{20}\sigma_r - \frac{18(\beta'+\beta)\alpha r\sigma_1}{\mu_2}\right)\cdot\|\mathbf{H}\|_F^2 \\ &- \left(\frac{L_1+1}{8}+\frac{1}{\mu_2}\right)\cdot\|\mathbf{H}\|_F^4 - \left(\frac{\mu_2}{2}+\frac{2K^2}{\mu_1}\right)\cdot\|\mathbf{S}-\mathbf{S}^*\|_F^2 - \left(\frac{8r}{\mu_1}+\frac{r}{L_1}\right)\cdot\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2,\end{aligned}$$

which completes the proof. □

D.2 Proof of Lemma C.2

Proof. According to the formula of $\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z},\mathbf{S})$ in (B.2), we have

$$\|\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z},\mathbf{S})\|_F^2 \leq 2\|\nabla_{\mathbf{U}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top+\mathbf{S})\|_F^2 + 2\|\nabla_{\mathbf{V}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top+\mathbf{S})\|_F^2 + \|\mathbf{U}^\top\mathbf{U}-\mathbf{V}^\top\mathbf{V}\|_F^2\cdot\|\mathbf{Z}\|_2^2, \tag{D.17}$$

where the inequality follows from the fact that $\|\mathbf{A}+\mathbf{B}\|_F^2 \leq 2\|\mathbf{A}\|_F^2 + 2\|\mathbf{B}\|_F^2$, $\|\mathbf{A}\mathbf{B}\|_F \leq \|\mathbf{A}\|_2\cdot\|\mathbf{B}\|_F$, and $\max\{\|\mathbf{U}\|_2, \|\mathbf{V}\|_2\} \leq \|\mathbf{Z}\|_2$. Consider the first term $\|\nabla_{\mathbf{U}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top+\mathbf{S})\|_F^2$. Denote $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, then we have

$$\begin{aligned}\|\nabla_{\mathbf{U}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top+\mathbf{S})\|_F^2 \leq{}& 3\underbrace{\big\|\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})\big)\mathbf{V}\big\|_F^2}_{I_1} \\ &+ 3\underbrace{\big\|\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\big)\mathbf{V}\big\|_F^2}_{I_2} + 3\underbrace{\big\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\mathbf{V}\big\|_F^2}_{I_3},\end{aligned} \tag{D.18}$$

where the inequality holds because $\nabla_{\mathbf{U}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top+\mathbf{S}) = \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S})\mathbf{V}$ and $\|\mathbf{A}+\mathbf{B}+\mathbf{C}\|_F^2 \leq 3(\|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2+\|\mathbf{C}\|_F^2)$. In the following discussion, we are going to upper bound I_1,I_2 and I_3 separately. As for I_1, according to the orthonormal projection matrix $\widetilde{\mathbf{V}}$ defined in (D.2), we have

$$\begin{aligned}I_1 &= \big\|\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})\big)\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\mathbf{V}\big\|_F^2 \\ &\leq \big\|\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})\big)\widetilde{\mathbf{V}}\big\|_F^2\cdot\|\widetilde{\mathbf{V}}^\top\mathbf{V}\|_2^2 \\ &\leq \big\|\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})\big)\widetilde{\mathbf{V}}\big\|_F^2\cdot\|\mathbf{V}\|_2^2,\end{aligned} \tag{D.19}$$

where the equality holds because $\mathrm{col}(\mathbf{V}) \subseteq \mathrm{col}(\widetilde{\mathbf{V}})$, which implies that $\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\mathbf{V} = \mathbf{V}$, the first inequality is due to the fact that $\|\mathbf{A}\mathbf{B}\|_F \leq \|\mathbf{A}\|_F\cdot\|\mathbf{B}\|_2$, and the last inequality holds because $\|\mathbf{A}\mathbf{B}\|_2 \leq \|\mathbf{A}\|_2\cdot\|\mathbf{B}\|_2$ and the

fact that $\widetilde{\mathbf{V}}$ is orthonormal. Moreover, consider the second term I_2 in (D.18). According to the definition of Frobenius norm, we have

$$\begin{aligned}\big\|\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})-\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\big)\mathbf{V}\big\|_F &= \sup_{\|\mathbf{W}\|_F\le 1}\langle\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})-\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\big)\mathbf{V},\mathbf{W}\rangle\\ &\le \sup_{\|\mathbf{W}\|_F\le 1}(1+K)\cdot\|\mathbf{S}-\mathbf{S}^*\|_F\cdot\|\mathbf{W}\mathbf{V}^\top\|_F\\ &\le (1+K)\cdot\|\mathbf{S}-\mathbf{S}^*\|_F\cdot\|\mathbf{V}\|_2, \end{aligned}\tag{D.20}$$

where the first inequality follows from the structural Lipschitz gradient Condition 4.4 and the fact that $|\langle\mathbf{A},\mathbf{B}\rangle|\le\|\mathbf{A}\|_F\cdot\|\mathbf{B}\|_F$, and the second one holds because $\|\mathbf{AB}\|_F\le\|\mathbf{A}\|_F\cdot\|\mathbf{B}\|_2$ and $\|\mathbf{W}\|_F\le 1$. Finally, consider the last term I_3 in (D.18), we have

$$I_3\le\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2\cdot\|\mathbf{V}\|_F^2\le r\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2\cdot\|\mathbf{V}\|_2^2.\tag{D.21}$$

Thus, combining (D.19), (D.20) and (D.21), we obtain

$$\begin{aligned}\|\nabla_{\mathbf{U}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top+\mathbf{S})\|_F^2 \le{}& 3\big\|\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S})-\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})\big)\widetilde{\mathbf{V}}\big\|_F^2\cdot\|\mathbf{V}\|_2^2\\ &+3(1+K)^2\cdot\|\mathbf{S}-\mathbf{S}^*\|_F^2\cdot\|\mathbf{V}\|_2^2+3r\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2\cdot\|\mathbf{V}\|_2^2.\end{aligned}\tag{D.22}$$

As for the second term $\|\nabla_{\mathbf{V}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top+\mathbf{S})\|_F^2$ in (D.17), based on similar techniques, we obtain

$$\begin{aligned}\|\nabla_{\mathbf{V}}\mathcal{L}_n(\mathbf{U}\mathbf{V}^\top+\mathbf{S})\|_F^2 \le{}& 3\big\|\widetilde{\mathbf{U}}^\top\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S})-\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})\big)\big\|_F^2\cdot\|\mathbf{U}\|_2^2\\ &+3(1+K)^2\cdot\|\mathbf{S}-\mathbf{S}^*\|_F^2\cdot\|\mathbf{U}\|_2^2+3r\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2\cdot\|\mathbf{U}\|_2^2,\end{aligned}\tag{D.23}$$

where $\widetilde{\mathbf{U}}$ is an orthonormal matrix defined in (D.1). According to Lemma C.1 in [50] and Condition 4.2, we have

$$\begin{aligned}I &= \big\|\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})-\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S})\big)\widetilde{\mathbf{V}}\big\|_F^2+\big\|\widetilde{\mathbf{U}}^\top\big(\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})-\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S})\big)\big\|_F^2\\ &\le 4L_1\big(\mathcal{L}_n(\mathbf{X}^*+\mathbf{S})-\mathcal{L}_n(\mathbf{X}+\mathbf{S})-\langle\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}),\mathbf{X}^*-\mathbf{X}\rangle\big)\le 2L_1^2\cdot\|\mathbf{X}-\mathbf{X}^*\|_F^2.\end{aligned}\tag{D.24}$$

Therefore, plugging (D.22), (D.23) and (D.24) into (D.17), we obtain

$$\begin{aligned}\|\nabla_{\mathbf{Z}}\widetilde{F}_n(\mathbf{Z},\mathbf{S})\|_F^2 \le{}& \Big(12L_1^2\|\mathbf{X}-\mathbf{X}^*\|_F^2+12(1+K)^2\cdot\|\mathbf{S}-\mathbf{S}^*\|_F^2+\|\mathbf{U}^\top\mathbf{U}-\mathbf{V}^\top\mathbf{V}\|_F^2\Big)\cdot\|\mathbf{Z}\|_2^2\\ &+12r\|\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^*+\mathbf{S}^*)\|_2^2\cdot\|\mathbf{Z}\|_2^2,\end{aligned}$$

where the inequality holds because $\max\{\|\mathbf{U}\|_2,\|\mathbf{V}\|_2\}\le\|\mathbf{Z}\|_2$. Thus, we finish the proof. □

D.3 Proof of Lemma C.3

Proof. Denote the support of $\mathbf{S}^*$ and $\mathcal{T}_{\gamma\beta}(\mathbf{S})$ by Ω^* and Ω respectively. According to the definition of the truncation operator $\mathcal{T}_\alpha$, we have

$$\begin{aligned}\|\mathcal{T}_{\gamma\beta}(\mathbf{S})-\mathbf{S}^*\|_F^2 &= \|\mathcal{P}_\Omega(\mathcal{T}_{\gamma\beta}(\mathbf{S})-\mathbf{S}^*)\|_F^2+\|\mathcal{P}_{\Omega^*\setminus\Omega}(\mathcal{T}_{\gamma\beta}(\mathbf{S})-\mathbf{S}^*)\|_F^2\\ &= \|\mathcal{P}_\Omega(\mathbf{S}-\mathbf{S}^*)\|_F^2+\|\mathcal{P}_{\Omega^*\setminus\Omega}(-\mathbf{S}^*)\|_F^2,\end{aligned}\tag{D.25}$$

where the second inequality holds because $[\mathcal{T}_{\gamma\beta}(\mathbf{S})]_{i,j}=S_{i,j}$ if $(i,j)\in\Omega$, and $[\mathcal{T}_{\gamma\beta}(\mathbf{S})]_{i,j}=0$ otherwise. For any $(i,j)\in\Omega^*\setminus\Omega$, we claim

$$\big|(\mathbf{S}-\mathbf{S}^*+\mathbf{S}^*)_{i,j}\big|\le\max\Big\{\underbrace{\big|(\mathbf{S}-\mathbf{S}^*)_{i,*}^{(\gamma\beta d_2-\beta d_2)}\big|}_{I_1},\underbrace{\big|(\mathbf{S}-\mathbf{S}^*)_{*,j}^{(\gamma\beta d_1-\beta d_1)}\big|}_{I_2}\Big\},\tag{D.26}$$

where we denote the k-th largest element in magnitude of $(\mathbf{S}-\mathbf{S}^*)_{i,*}$ by $(\mathbf{S}-\mathbf{S}^*)_{i,*}^{(k)}$, and the k-th largest element in magnitude of $(\mathbf{S}-\mathbf{S}^*)_{*,j}$ by $(\mathbf{S}-\mathbf{S}^*)_{*,j}^{(k)}$. In the following discussion, we are going to prove claim (D.26) by

contradiction. Suppose $\big|(\mathbf{S}-\mathbf{S}^*+\mathbf{S}^*)_{i,j}\big| = |S_{i,j}| > \max\{I_1, I_2\}$, where $(i,j) \in \Omega^* \setminus \Omega$. Noticing $\mathbf{S}^*$ has at most β-fraction nonzero entries per row and column, we have

$$I_1 \geq \big|\mathbf{S}^{(\gamma\beta d_2 - \beta d_2)}_{i,*}\big| \quad \text{and} \quad I_2 \geq \big|\mathbf{S}^{(\gamma\beta d_1 - \beta d_1)}_{*,j}\big|.$$

Thus we have $|S_{i,j}| \geq \max\big\{\big|\mathbf{S}^{(\gamma\beta d_2 - \beta d_2)}_{i,*}\big|, \big|\mathbf{S}^{(\gamma\beta d_1 - \beta d_1)}_{*,j}\big|\big\}$, which contradicts with the fact that $(i,j) \in \Omega^* \setminus \Omega$. Therefore, based on (D.26), we obtain

$$\begin{aligned}
\|\mathcal{P}_{\Omega^*\setminus\Omega}(\mathbf{S}-\mathbf{S}^*+\mathbf{S}^*)\|_F^2 &= \sum_{(i,j)\in\Omega^*\setminus\Omega} \big|(\mathbf{S}-\mathbf{S}^*+\mathbf{S}^*)_{i,j}\big|^2 \\
&\leq \sum_{(i,j)\in\Omega^*\setminus\Omega} \frac{\|(\mathbf{S}-\mathbf{S}^*)_{i,*}\|_2^2}{(\gamma-1)\beta d_2} + \sum_{(i,j)\in\Omega^*\setminus\Omega} \frac{\|(\mathbf{S}-\mathbf{S}^*)_{*,j}\|_2^2}{(\gamma-1)\beta d_1} \\
&\leq \frac{2}{\gamma-1}\|\mathbf{S}-\mathbf{S}^*\|_F^2, \qquad \text{(D.27)}
\end{aligned}$$

where the first inequality is due to (D.26), and the second inequality holds because for each row and column of Ω^*, it has at most β-fraction nonzero elements. Thus we obtain

$$\begin{aligned}
\|\mathcal{P}_{\Omega^*\setminus\Omega}(-\mathbf{S}^*)\|_F^2 &= \|\mathcal{P}_{\Omega^*\setminus\Omega}(\mathbf{S}-\mathbf{S}^*) - \mathcal{P}_{\Omega^*\setminus\Omega}(\mathbf{S}-\mathbf{S}^*+\mathbf{S}^*)\|_F^2 \\
&\leq (1+c)\cdot\|\mathcal{P}_{\Omega^*\setminus\Omega}(\mathbf{S}-\mathbf{S}^*)\|_F^2 + \bigg(1+\frac{1}{c}\bigg)\cdot\|\mathcal{P}_{\Omega^*\setminus\Omega}(\mathbf{S}-\mathbf{S}^*+\mathbf{S}^*)\|_F^2 \\
&\leq (1+c)\cdot\|\mathcal{P}_{\Omega^*\setminus\Omega}(\mathbf{S}-\mathbf{S}^*)\|_F^2 + \frac{c+1}{c}\cdot\frac{2}{\gamma-1}\|\mathbf{S}-\mathbf{S}^*\|_F^2, \qquad \text{(D.28)}
\end{aligned}$$

where the second inequality holds because $\|\mathbf{A}+\mathbf{B}\|_F^2 \leq (1+c)\cdot\|\mathbf{A}\|_F^2 + (1+1/c)\cdot\|\mathbf{B}\|_F^2$, for any $c>0$, and the second inequality is due to (D.27). Therefore, plugging in (D.28) into (D.25), we have

$$\begin{aligned}
\|\mathcal{T}_{\gamma\beta}(\mathbf{S})-\mathbf{S}^*\|_F^2 &\leq \|\mathcal{P}_{\Omega}(\mathbf{S}-\mathbf{S}^*)\|_F^2 + (1+c)\cdot\|\mathcal{P}_{\Omega^*\setminus\Omega}(\mathbf{S}-\mathbf{S}^*)\|_F^2 + \frac{c+1}{c}\cdot\frac{2}{\gamma-1}\|\mathbf{S}-\mathbf{S}^*\|_F^2 \\
&\leq \bigg(1+c+\frac{2(c+1)}{c(\gamma-1)}\bigg)\cdot\|\mathbf{S}-\mathbf{S}^*\|_F^2 \\
&= \bigg(1+\frac{2}{\gamma-1}+2\sqrt{\frac{2}{\gamma-1}}\bigg)\cdot\|\mathbf{S}-\mathbf{S}^*\|_F^2,
\end{aligned}$$

where we set $c=\sqrt{(\gamma-1)/2}$ in the last step. Thus we complete the proof. □

E Proofs of Specific Models

In this section, we provide proofs for specific models. In the following discussions, we let $d = \max\{d_1, d_2\}$.

E.1 Proofs of Robust Matrix Sensing

For matrix sensing, recall that we have the linear measurement operator $\mathcal{A}$ with each sensing matrix $\mathbf{A}_i$ sampled independently from $\mathbf{\Sigma}$-Gaussian ensemble, where $\mathrm{vec}(\mathbf{A}_i) \sim N(0, \mathbf{\Sigma})$. In particular, we consider $\mathbf{\Sigma} = \mathbf{I}$ and here $\mathrm{vec}(\mathbf{A}_i)$ denotes the vectorization of matrix $\mathbf{A}_i$. In order to prove the results for matrix sensing, we first lay out several lemmas, which are essential to prove the results for robust matrix sensing. The first lemma is useful to verify the restricted strong convexity and smoothness conditions in Condition 4.2.

Lemma E.1. [37] Suppose we have the linear measurement operator $\mathcal{A}$ with each sensing matrix $\mathbf{A}_i$ sampled independently from $\mathbf{I}$-Gaussian ensemble, then there exists constants c_0, c_1 such that for all $\mathbf{\Delta} \in \mathbb{R}^{d_1\times d_2}$ with rank at most $2\widetilde{r}$, it holds with probability at least $1-\exp(-c_0 n)$ that

$$\bigg|\frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{n} - \frac{1}{2}\big\|\mathrm{vec}(\mathbf{\Delta})\big\|_2^2\bigg| \leq c_1\frac{\widetilde{r}d}{n}\big\|\mathbf{\Delta}\big\|_F^2. \qquad \text{(E.1)}$$

The second lemma is useful to verify the restricted strong convexity and smoothness conditions in Condition 4.3.

Lemma E.2. [43] For any random matrix $\mathbf{A} \in \mathbb{R}^{n \times d_1 d_2}$, which is drawn from the $\mathbf{\Sigma}$-Gaussian ensemble, and the cardinalities of all vector $\mathbf{s} \in \mathbb{R}^{d_1 d_2}$ satisfy $|\mathbf{s}| \leq \widetilde{s}$. If we have sample size $n \geq c_2 \widetilde{s} \log d$, then the following inequality holds with probability at least $1 - c_3 \exp(-c_4 n)$

$$c_5 \|\mathbf{\Sigma}^{1/2}\mathbf{s}\|_2^2 - c_6 \frac{\log d}{n} \|\mathbf{s}\|_1 \leq \frac{\|\mathbf{A}\mathbf{s}\|_2^2}{n} \leq c_7 \|\mathbf{\Sigma}^{1/2}\mathbf{s}\|_2^2 + c_8 \frac{\log d}{n} \|\mathbf{s}\|_1,$$

where $\{c_i\}_{i=2}^8$ are universal constants.

The next lemma verifies the structural Lipschitz gradient condition in Condition 4.4.

Lemma E.3. Consider robust matrix sensing with objective loss function defined in section 4.2. There exist constants C_0, C_1 such that the following inequality holds with probability at least $1 - \exp(-C_0 d)$

$$|\langle \nabla_{\mathbf{X}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}) - \nabla_{\mathbf{X}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*), \mathbf{X} \rangle - \langle \mathbf{S} - \mathbf{S}^*, \mathbf{X} \rangle| \leq K \|\mathbf{X}\|_F \cdot \|\mathbf{S} - \mathbf{S}^*\|_F,$$
$$|\langle \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X} + \mathbf{S}^*) - \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*), \mathbf{S} \rangle - \langle \mathbf{X} - \mathbf{X}^*, \mathbf{S} \rangle| \leq K \|\mathbf{X} - \mathbf{X}^*\|_F \cdot \|\mathbf{S}\|_F,$$

for all low-rank matrices $\mathbf{X}, \mathbf{X}^*$ with rank at most $\widetilde{r}$ and all sparse matrices $\mathbf{S}, \mathbf{S}^*$ with sparsity at most $\widetilde{s}$, where $\widetilde{r}, \widetilde{s}$ are defined in Condition 4.2, and the structural Lipschitz gradient parameter $K = C_1 \sqrt{(rd + s) \log d / n}$.

The last lemma verifies the condition in Condition 4.5 for robust matrix sensing.

Lemma E.4. Consider robust matrix sensing, suppose each sensing matrix $\mathbf{A}_i$ is sampled independently from $\mathbf{I}$-Gaussian ensemble and each element of noise vector $\boldsymbol{\epsilon}$ follows i.i.d. sub-Gaussian distribution with parameter ν. Then we have the following inequalities hold with probability at least $1 - C_2/d$ in terms of spectral norm and infinity norm respectively

$$\left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbf{A}_i \right\|_2 \leq C_3 \nu \sqrt{\frac{d}{n}} \quad \text{and} \quad \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbf{A}_i \right\|_{\infty,\infty} \leq C_4 \nu \sqrt{\frac{\log d}{n}},$$

where C_2, C_3, C_4 are universal constants.

Now, we are ready to prove Corollary 4.11.

Proof of Corollary 4.11. In order to prove Corollary 4.11, we only need to verify the restricted strong convex and smoothness conditions in Conditions 4.2 and 4.3, the structural Lipschitz gradient condition in Condition 4.4, and the condition in Condition 4.5.

Recall that we have the sample loss function for robust matrix sensing as $\mathcal{L}_n(\mathbf{X} + \mathbf{S}) := \|\mathbf{y} - \mathcal{A}_n(\mathbf{X} + \mathbf{S})\|_2^2/(2n)$. Therefore, for all given sparse matrices $\mathbf{S}$, we have the following holds for all matrices $\mathbf{X}_1, \mathbf{X}_2$ with rank at most $\widetilde{r}$

$$\mathcal{L}_n(\mathbf{X}_1 + \mathbf{S}) - \mathcal{L}_n(\mathbf{X}_2 + \mathbf{S}) - \langle \nabla_{\mathbf{X}} \mathcal{L}_n(\mathbf{X}_2 + \mathbf{S}), \mathbf{X}_2 - \mathbf{X}_1 \rangle = \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{n},$$

where $\mathbf{\Delta} = \mathbf{X}_2 - \mathbf{X}_1$. According to Lemma E.1, if we have $n > c_1' \widetilde{r} d$, where c' is some constants. Then, with probability at least $1 - \exp(-c_0 n)$, we have the restricted strong convexity and smoothness conditions in Condition 4.2 hold with parameter $\mu_1 = 4/9$ and $L_1 = 5/9$. In addition, for all given low-rank matrices $\mathbf{X}$, we have the following holds for all matrices $\mathbf{S}_1, \mathbf{S}_2$ with sparsity at most $\widetilde{s}$

$$\mathcal{L}_n(\mathbf{X} + \mathbf{S}_1) - \mathcal{L}_n(\mathbf{X} + \mathbf{S}_2) - \langle \nabla_{\mathbf{S}} \mathcal{L}_n(\mathbf{X} + \mathbf{S}_2), \mathbf{S}_2 - \mathbf{S}_1 \rangle = \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{n},$$

where $\mathbf{\Delta} = \mathbf{S}_2 - \mathbf{S}_1$. Furthermore, we can obtain $\|\mathcal{A}(\mathbf{\Delta})\|_2^2 = \|\mathbf{A}\boldsymbol{\delta}\|_2^2$, where we have $\mathbf{A} \in \mathbb{R}^{n \times d_1 d_2}$ with each row $\mathbf{A}_{i*} = \mathrm{vec}(\mathbf{A}_i)$, and $\boldsymbol{\delta} = \mathrm{vec}(\mathbf{\Delta})$. Therefore, according to Lemma E.2, we have

$$c_1 \|\boldsymbol{\delta}\|_2^2 - c_2 \frac{\log d}{n} \|\boldsymbol{\delta}\|_1 \leq \frac{\|\mathbf{A}\boldsymbol{\delta}\|_2^2}{n} \leq c_3 \|\boldsymbol{\delta}\|_2^2 + c_4 \frac{\log d}{n} \|\boldsymbol{\delta}\|_1.$$

Thus provided that $n > c_5\widetilde{s}\log d$, with probability at least $1 - c_3\exp(-c_4 n)$, the restricted strong convexity and smoothness conditions in Condition 4.3 hold with parameters $\mu_2 = 4/9$ and $L_2 = 5/9$.

Next, according to Lemma E.3, with probability at least $1-\exp(-C_0 d)$, we can establish the structural Lipschitz gradient condition in Condition 4.5 with parameter $K = C_1\sqrt{(rd+s)\log d/n}$.

Finally, we will verify the condition in Condition 4.5. By the definition of the objective loss function for robust matrix sensing, we have $\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*) = \sum_{i=1}^n \epsilon_i \mathbf{A}_i/n$ and $\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*) = \sum_{i=1}^n \epsilon_i \mathbf{A}_i/n$. Therefore, according to Lemma E.4, with probability at least $1 - C_2/d$, we can establish the condition in Condition 4.5 with parameters $\epsilon_1 = C_3\nu\sqrt{d/n}$ and $\epsilon_2 = C_4\nu\sqrt{\log d/n}$. This completes the proof. □

E.2 Proofs of Robust PCA

Note that since robust PCA under fully observed model is a special case of robust PCA under partially observed model, thus we just lay out the proofs of robust PCA under partially observed model. To prove the results of partially observed robust PCA, we need the following lemmas, which are essential to establish the restricted strong convexity and smoothness conditions in Conditions 4.2 and 4.3. Note that the following lemmas only work for robust PCA under noisy observation model.

Lemma E.5. [38] There exist universal constants $\{c_i\}_{i=1}^4$ such that if the number of observations $n \geq c_1 rd\log d$, and the following condition is satisfied for all $\mathbf{\Delta} \in \mathbb{R}^{d_1\times d_2}$

$$\sqrt{\frac{d_1 d_2}{r}}\frac{\|\mathbf{\Delta}\|_{\infty,\infty}}{\|\mathbf{\Delta}\|_F}\cdot\frac{\|\mathbf{\Delta}\|_*}{\|\mathbf{\Delta}\|_F} \leq \frac{1}{c_2}\sqrt{n/(d\log d)}, \tag{E.2}$$

we have, with probability at least $1 - c_3/d$, that the following holds

$$\left|\frac{\|\mathcal{A}(\mathbf{\Delta})\|_2}{\sqrt{n}} - \frac{\|\mathbf{\Delta}\|_F}{\sqrt{d_1 d_2}}\right| \leq \frac{1}{10}\frac{\|\mathbf{\Delta}\|_F}{\sqrt{d_1 d_2}}\left(1 + \frac{c_4\sqrt{d_1 d_2}\|\mathbf{\Delta}\|_{\infty,\infty}}{\sqrt{n}\|\mathbf{\Delta}\|_F}\right).$$

Lemma E.6. There exist universal constants $\{c_i\}_{i=1}^5$ such that as long as $n \geq c_1\log d$, we have with probability at least $1 - c_2\exp(-c_3\log d)$ that

$$\left|\frac{\|\mathcal{A}(\mathbf{\Delta})\|_2}{\sqrt{n}} - \frac{\|\mathbf{\Delta}\|_F}{\sqrt{d_1 d_2}}\right| \leq \frac{1}{2}\frac{\|\mathbf{\Delta}\|_F}{\sqrt{d_1 d_2}} + \frac{c_5\|\mathbf{\Delta}\|_{\infty,\infty}}{\sqrt{n}} \quad \text{for all} \quad \mathbf{\Delta} \in \mathcal{C}(n), \tag{E.3}$$

where we have the set $\mathcal{C}(n)$ as follows

$$\mathcal{C}(n) = \left\{\mathbf{\Delta} \in \mathbb{R}^{d_1\times d_2} \mid \frac{\|\mathbf{\Delta}\|_{1,1}}{\|\mathbf{\Delta}\|_F}\cdot\frac{\|\mathbf{\Delta}\|_{\infty,\infty}}{\|\mathbf{\Delta}\|_F} \leq c_4\sqrt{\frac{n}{d_1 d_2\log d}}\right\}.$$

The next lemma verifies the structural Lipschitz gradient condition in Condition 4.4.

Lemma E.7. Consider partially observed robust PCA with objective loss function defined in section4.2. There exist constants C_0, C_1 such that the following inequality holds with probability at least $1 - \exp(-C_0 d)$

$$|\langle\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*), \mathbf{X}\rangle - \langle\mathbf{S} - \mathbf{S}^*, \mathbf{X}\rangle| \leq K\|\mathbf{X}\|_F\cdot\|\mathbf{S} - \mathbf{S}^*\|_F,$$
$$|\langle\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X} + \mathbf{S}^*) - \nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*), \mathbf{S}\rangle - \langle\mathbf{X} - \mathbf{X}^*, \mathbf{S}\rangle| \leq K\|\mathbf{X} - \mathbf{X}^*\|_F\cdot\|\mathbf{S}\|_F,$$

for all low-rank matrices $\mathbf{X}, \mathbf{X}^*$ with rank at most $\widetilde{r}$ and all sparse matrices $\mathbf{S}, \mathbf{S}^*$ with sparsity at most $\widetilde{s}$, where $\widetilde{r}, \widetilde{s}$ are defined in Condition 4.2, and the structural Lipschitz gradient parameter $K = C_1\sqrt{(rd+s)\log d/n}$.

The last lemma verifies the condition in Condition 4.5 for partially observed robust PCA.

Lemma E.8. Consider partially observed robust PCA. If $\mathbf{A}_{jk} = \mathbf{e}_j\mathbf{e}_k^\top$ is uniformly distributed on Ω, then for i.i.d. zero mean random variables ϵ_{jk} with variance ν^2, we have the following inequalities hold with probability at least $1 - C_2/d$ in terms of spectral norm and infinity norm respectively

$$\left\|\frac{1}{p}\sum_{j,k\in\Omega}\epsilon_{jk}\mathbf{A}_{jk}\right\|_2 \leq C_3\nu\sqrt{\frac{d\log d}{p}} \quad \text{and} \quad \left\|\frac{1}{p}\sum_{jk\in\Omega}\epsilon_{jk}\mathbf{A}_{jk}\right\|_{\infty,\infty} \leq C_4\nu\sqrt{\frac{\log d}{p}},$$

where C_2, C_3, C_4 are universal constants, and $p = n/(d_1 d_2)$.

Now, we are ready to prove Corollary 4.15.

Proof of Corollary 4.15. To prove Corollary 4.15, we need to verify the restricted strong convexity and smoothness conditions in Conditions 4.2 and 4.3, the structural Lipschitz gradient condition in Condition 4.4, and the condition in Condition 4.5.

In the following discussion, we let $\mathbf{A}_{jk} = \mathbf{e}_j\mathbf{e}_k^\top$, where $\mathbf{e}_i, \mathbf{e}_j$ are basis vectors with d_1 and d_2 dimensions, and we let $\mathcal{A}$ be the corresponding transformation operator. In addition, let the number of observations to be $|\Omega| = n$. Therefore, the objective loss function for robust PCA in 4.2 can be rewritten as

$$\mathcal{L}_n(\mathbf{X}+\mathbf{S}) := \frac{1}{2p}\sum_{(j,k)\in\Omega}\big(\langle\mathbf{A}_{jk}, \mathbf{X}+\mathbf{S}\rangle - Y_{jk}\big)^2.$$

Therefore, for all given sparse matrices $\mathbf{S}$, we have the following holds for all matrices $\mathbf{X}_1, \mathbf{X}_2$ satisfying incoherence condition with rank at most $\widetilde{r}$

$$\mathcal{L}_n(\mathbf{X}_1+\mathbf{S}) - \mathcal{L}_n(\mathbf{X}_2+\mathbf{S}) - \langle\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}_2+\mathbf{S}), \mathbf{X}_2-\mathbf{X}_1\rangle = \frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2^2}{p},$$

where $\boldsymbol{\Delta} = \mathbf{X}_1 - \mathbf{X}_2$, and $p = n/(d_1d_2)$. Now, we are ready to prove the restricted strong convexity and smoothness conditions in Condition 4.2.

Case 1: If $\boldsymbol{\Delta}$ not satisfies condition (E.2), then we have

$$\begin{aligned}\|\boldsymbol{\Delta}\|_F^2 &\le C_0\big(\sqrt{d_1d_2}\|\boldsymbol{\Delta}\|_\infty\big)\|\boldsymbol{\Delta}\|_*\sqrt{\frac{d\log d}{nr}}\\ &\le 2C_0\alpha_1\sqrt{d_1d_2}\|\boldsymbol{\Delta}\|_*\sqrt{\frac{d\log d}{nr}}\\ &\le 2C_0\alpha_1\sqrt{2\widetilde{r}d_1d_2}\|\boldsymbol{\Delta}\|_F\sqrt{\frac{d\log d}{nr}},\end{aligned}$$

where $\widetilde{\alpha} = \alpha r/\sqrt{d_1d_2}$ due to the incoherence condition of low rank matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, and the last inequality comes from $\mathrm{rank}(\boldsymbol{\Delta}) \le 2\widetilde{r}$. Thus, by the definition of $\widetilde{r}$, we can obtain

$$\|\boldsymbol{\Delta}\|_F^2 \le C_1\alpha^2\sigma_1^2\frac{r^2d\log d}{n}. \tag{E.4}$$

Case 2: If $\boldsymbol{\Delta}$ satisfies condition (E.2), then according to Lemma E.5, we have

$$\left|\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2}{\sqrt{p}} - \|\boldsymbol{\Delta}\|_F\right| \le \frac{\|\boldsymbol{\Delta}\|_F}{10}\left(1 + \frac{C_2\sqrt{d_1d_2}\|\boldsymbol{\Delta}\|_{\infty,\infty}}{\sqrt{n}\|\boldsymbol{\Delta}\|_F}\right).$$

Thus if $C_2\sqrt{d_1d_2}\|\boldsymbol{\Delta}\|_{\infty,\infty}/(\sqrt{n}\|\boldsymbol{\Delta}\|_F) \ge C_3$, we have

$$\|\boldsymbol{\Delta}\|_F^2 \le C_4\frac{\widetilde{\alpha}^2}{p}.$$

Otherwise, if $C_2\sqrt{d_1d_2}\|\boldsymbol{\Delta}\|_{\infty,\infty}/(\sqrt{n}\|\boldsymbol{\Delta}\|_F) \le C_3$, we have

$$\frac{8}{9}\|\boldsymbol{\Delta}\|_F^2 \le \frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2^2}{p} \le \frac{10}{9}\|\boldsymbol{\Delta}\|_F^2,$$

which gives us the restricted strong convexity and smoothness conditions in Condition 4.2 with parameters $\mu_1 = 8/9, L_1 = 10/9$.

Next, we prove the restricted strong convexity and smoothness conditions in Condition 4.3. For all given low-rank matrices $\mathbf{X}$, we have the following holds for all matrices $\mathbf{S}_1, \mathbf{S}_2$ with at most $\widetilde{s}$ nonzero entries and infinity norm bound $\alpha_1/\sqrt{d_1d_2}$

$$\mathcal{L}_n(\mathbf{X}+\mathbf{S}_1) - \mathcal{L}_n(\mathbf{X}+\mathbf{S}_2) - \langle\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}+\mathbf{S}_2), \mathbf{S}_2-\mathbf{S}_1\rangle = \frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2^2}{p},$$

where $\boldsymbol{\Delta} = \mathbf{S}_1 - \mathbf{S}_2$, and $p = n/(d_1 d_2)$.

Case 1: If $\boldsymbol{\Delta} \notin \mathcal{C}(n)$, then we can get

$$\|\boldsymbol{\Delta}\|_F^2 \leq C_5\big(\sqrt{d_1 d_2}\|\boldsymbol{\Delta}\|_{\infty,\infty}\big) \cdot \|\boldsymbol{\Delta}\|_{1,1}\sqrt{\frac{\log d}{n}} \leq 2C_5\alpha_1\|\boldsymbol{\Delta}\|_{1,1}\sqrt{\frac{\log d}{n}},$$

where the last inequality is due to the fact that $\|\boldsymbol{\Delta}\|_\infty = \|\mathbf{S}_1 - \mathbf{S}_2\|_\infty \leq 2\alpha_1/\sqrt{d_1 d_2}$. Therefore, we can obtain

$$\|\boldsymbol{\Delta}\|_F^2 \leq 2C_5\sqrt{2\widetilde{s}}\alpha_1\|\boldsymbol{\Delta}\|_F\sqrt{\frac{\log d}{n}},$$

where the inequality holds because $\boldsymbol{\Delta}$ has at most $2\widetilde{s}$ nonzero entries. Therefore, by the definition of $\widetilde{s}$, we have

$$\|\boldsymbol{\Delta}\|_F^2 \leq C_6\alpha_1^2\frac{s\log d}{n}. \tag{E.5}$$

Case 2: If $\boldsymbol{\Delta} \in \mathcal{C}(n)$, we have

$$\left|\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2}{\sqrt{p}} - \|\boldsymbol{\Delta}\|_F\right| \leq \frac{1}{2}\|\boldsymbol{\Delta}\|_F + \frac{c_5\sqrt{d_1 d_2}\|\boldsymbol{\Delta}\|_{\infty,\infty}}{\sqrt{n}\|\boldsymbol{\Delta}\|_F},$$

If $\sqrt{n}\|\boldsymbol{\Delta}\|_F \leq C_7\sqrt{d_1 d_2}\|\boldsymbol{\Delta}\|_{\infty,\infty}$, we can obtain $\|\boldsymbol{\Delta}\|_F^2 \leq C_7'\alpha_1^2/n$. Otherwise, if we have $\sqrt{n}\|\boldsymbol{\Delta}\|_F \geq C_7\sqrt{d_1 d_2}\|\boldsymbol{\Delta}\|_{\infty,\infty}$, according to Lemma E.6, we obtain

$$\frac{8}{9}\|\boldsymbol{\Delta}\|_F^2 \leq \frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2^2}{p} \leq \frac{10}{9}\|\boldsymbol{\Delta}\|_F^2,$$

which implies the restricted strong convexity and smoothness conditions in Condition 4.3 hold with parameters $\mu_2 = 8/9, L_2 = 10/9$.

Next, according to Lemma E.7, with probability at least $1-\exp(-C_0 d)$, we can establish the structural Lipschitz gradient condition in Condition 4.5 with parameter $K = C_1\sqrt{(rd+s)\log d/n}$.

Finally, we verify the condition in Condition 4.5. By the definition of the objective loss function for robust PCA, we have $\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*) = \sum_{j,k\in\Omega}\epsilon_{jk}\mathbf{A}_{jk}/p$ and $\nabla_{\mathbf{S}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*) = \sum_{j,k\in\Omega}\epsilon_{jk}\mathbf{A}_{jk}/p$, where ϵ_{jk} are i.i.d. Gaussian variables with variance $\nu^2/(d_1 d_2)$. Therefore, according to Lemma E.8, with probability at least $1 - C_8'/d$, we have $\|\sum_{j,k\in\Omega}\epsilon_{jk}\mathbf{A}_{jk}/p\|_2^2 \leq C_8\nu^2 d\log d/n$. In addition, we have $\|\sum_{jk\in\Omega}\epsilon_{jk}\mathbf{A}_{jk}/p\|_{\infty,\infty}^2 \leq C_9\nu^2\log d/n$. Furthermore, we have additional estimation error bounds (E.4) and (E.5) when we derive the restricted strong convexity and smoothness conditions. Therefore, we can establish the condition in Condition 4.5 with parameters $\epsilon_1^2 = C_8\max\{\alpha_1^2,\nu^2\}d/n$ and $\epsilon_2^2 = C_9\max\{\alpha_1^2,\nu^2\}\log d/n$. This completes the proof. □

F Proofs of Technical Lemmas in Appendix E

F.1 Proof of Lemma E.3

Proof. In order to verify the structural Lipschitz gradient condition, we need to make use of the Bernstein-type inequality for sub-exponential random variables in [47] as well as the corresponding covering arguments for low-rank and sparse structures, respectively.

By the definition of the objective loss function of matrix sensing, we have

$$\langle\nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}) - \nabla_{\mathbf{X}}\mathcal{L}_n(\mathbf{X}^* + \mathbf{S}^*), \mathbf{X}\rangle = \frac{1}{n}\sum_{i=1}^n\langle\mathbf{A}_i, \mathbf{S} - \mathbf{S}^*\rangle\langle\mathbf{A}_i, \mathbf{X}\rangle = \frac{1}{n}\sum_{i=1}^n Y_i,$$

where $Y_i = \langle\mathbf{A}_i, \mathbf{S}-\mathbf{S}^*\rangle\langle\mathbf{A}_i, \mathbf{X}\rangle$. Note that $\langle\mathbf{A}_i, \mathbf{S}-\mathbf{S}^*\rangle$, $\langle\mathbf{A}_i, \mathbf{X}\rangle$ follow i.i.d. normal distribution $N(0, \|\mathbf{S}-\mathbf{S}^*\|_F^2)$ and $N(0, \|\mathbf{X}\|_F^2)$ respectively. Thus Y_i follows i.i.d. chi-square distribution which is also sub-exponential. Besides, $\mathbb{E}(Y_i) = \langle\mathbf{S} - \mathbf{S}^*, \mathbf{X}\rangle$, and we have

$$\big\|Y_i - \mathbb{E}[Y_i]\big\|_{\psi_1} \leq 2\big\|Y_i\big\|_{\psi_1} \leq 2\|\langle\mathbf{A}_i, \mathbf{S} - \mathbf{S}^*\rangle\|_{\psi_2} \cdot \|\langle\mathbf{A}_i, \mathbf{X}\rangle\|_{\psi_2} \leq 2C^2\|\mathbf{S} - \mathbf{S}^*\|_F \cdot \|\mathbf{X}\|_F = \lambda,$$

where C is a universal constant. Thus, by applying Proposition 5.16 in [47], for $Y_i - \mathbb{E}[Y_i]$, we obtain

$$\mathbb{P}\bigg\{\bigg|\sum_{i=1}^n \frac{1}{n}(Y_i - \mathbb{E}[Y_i])\bigg| \geq t\bigg\} \leq 2\exp\bigg[-c\min\bigg(\frac{nt^2}{\lambda^2}, \frac{nt}{\lambda}\bigg)\bigg].$$

According the covering argument for low-rank matrices Lemma 3.1 in [8] and covering number for sparse matrices in [48], we have

$$\begin{aligned}
&\mathbb{P}\bigg\{\sup_{(\mathbf{S}-\mathbf{S}^*)\in\mathcal{N}_\epsilon^{cs},\mathbf{X}\in\mathcal{N}_\epsilon^{3r}}\bigg|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{A}_i,\mathbf{S}-\mathbf{S}^*\rangle\langle\mathbf{A}_i,\mathbf{X}\rangle - \langle\mathbf{S}-\mathbf{S}^*,\mathbf{X}\rangle\bigg| \geq t\bigg\}\\
&\leq 2|\mathcal{N}_\epsilon^{cs}||\mathcal{N}_\epsilon^{3r}|\exp\bigg[-c_1\min\bigg(\frac{nt^2}{\lambda^2},\frac{nt}{\lambda}\bigg)\bigg]\\
&\leq 2\bigg(\frac{9}{\epsilon}\bigg)^{(d_1+d_2+1)3r}\cdot\bigg(\frac{c_2d_1d_2}{cs\epsilon}\bigg)^{cs}\cdot\exp\bigg[-c_1\min\bigg(\frac{nt^2}{\lambda^2},\frac{nt}{\lambda}\bigg)\bigg]\\
&\leq \exp\bigg[c_3\big(rd\log(1/\epsilon)+s\max\{\log d,\log(1/\epsilon)\}\big)-c_1\min\bigg(\frac{nt^2}{\lambda^2},\frac{nt}{\lambda}\bigg)\bigg]\leq\exp(-c'd),
\end{aligned}\tag{F.1}$$

where c_1, c_2, c_3 are constants, $\lambda = 2C^2$, and the first inequality follows from union bound, the second inequality is due to the covering arguments, and the last inequality holds by setting $t = c_4\sqrt{(rd+s)\log d/n}$. Besides, note that for any $\mathbf{X}\in\mathcal{M}_{3r}$, $\mathbf{S}\in\mathbf{S}^*+\mathcal{M}_{cs}$, there exists $\mathbf{X}_1\in\mathcal{N}_\epsilon^{3r}$, $\mathbf{S}_1\in\mathbf{S}^*+\mathcal{N}_\epsilon^{cs}$ such that $\|\mathbf{X}-\mathbf{X}_1\|_F\leq\epsilon$ and $\|\mathbf{S}-\mathbf{S}_1\|_F\leq\epsilon$. Thus, we have

$$\begin{aligned}
&\bigg|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{A}_i,\mathbf{S}-\mathbf{S}^*\rangle\langle\mathbf{A}_i,\mathbf{X}\rangle - \frac{1}{n}\sum_{i=1}^n\langle\mathbf{A}_i,\mathbf{S}_1-\mathbf{S}^*\rangle\langle\mathbf{A}_i,\mathbf{X}_1\rangle\bigg|\\
&\leq\bigg|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{A}_i,\mathbf{S}-\mathbf{S}^*\rangle\langle\mathbf{A}_i,\mathbf{X}-\mathbf{X}_1\rangle\bigg| + \bigg|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{A}_i,\mathbf{S}-\mathbf{S}_1\rangle\langle\mathbf{A}_i,\mathbf{X}_1\rangle\bigg|\\
&\leq\sqrt{L_1L_2}\|\mathbf{S}-\mathbf{S}^*\|_F\cdot\|\mathbf{X}-\mathbf{X}_1\|_F+\sqrt{L_1L_2}\|\mathbf{S}-\mathbf{S}_1\|_F\cdot\|\mathbf{X}_1\|_F\leq 2\epsilon\sqrt{L_1L_2},
\end{aligned}\tag{F.2}$$

where the first inequality holds because of triangle inequality, and the second inequality follows from the restricted strong smoothness condition for both low-rank and sparse structures. Similarly, we have

$$|\langle\mathbf{S}-\mathbf{S}^*,\mathbf{X}\rangle - \langle\mathbf{S}_1-\mathbf{S}^*,\mathbf{X}_1\rangle| \leq \|\mathbf{S}-\mathbf{S}^*\|_F\cdot\|\mathbf{X}-\mathbf{X}_1\|_F+\|\mathbf{S}-\mathbf{S}_1\|_F\cdot\|\mathbf{X}_1\|_F\leq 2\epsilon,\tag{F.3}$$

Therefore, combining (F.1), (F.2) and (F.3), by triangle inequality, we obtain

$$\sup_{(\mathbf{S}-\mathbf{S}^*)\in\mathcal{M}_{cs},\mathbf{X}\in\mathcal{M}_{3r}}\bigg|\frac{1}{n}\sum_{i=1}^n\langle\mathbf{A}_i,\mathbf{S}-\mathbf{S}^*\rangle\langle\mathbf{A}_i,\mathbf{X}\rangle - \langle\mathbf{S}-\mathbf{S}^*,\mathbf{X}\rangle\bigg| \leq t+2\epsilon\sqrt{L_1L_2}+2\epsilon,$$

with probability at least $1-\exp(-c'd)$. We establish the incoherence condition by setting $\epsilon = t/(2\sqrt{L_1L_2}+2)$ in (F.3). By similar techniques, we can prove the second inequality in Lemma E.3. Note that we obtain $K = C\sqrt{(rd+s)\log d/n}$ in Lemma E.3. □

F.2 Proof of Lemma E.4

Proof. The first inequality in Lemma E.4 has been established in [37] Lemma 6. We provided the second inequality using Bernstein-type inequality and Union Bound. Recall that, we have

$$\bigg\|\frac{1}{n}\sum_{i=1}^n\epsilon_i\mathbf{A}_i\bigg\|_{\infty,\infty} = \max_{j,k}\bigg|\frac{1}{n}\sum_{i=1}^n\epsilon_iA_{jk}^i\bigg| = \max_{j,k}\bigg|\frac{1}{n}\sum_{i=1}^nZ_{jk}^i\bigg|,$$

where we let $Z_{jk}^i = \epsilon_iA_{jk}^i$. Since Z_{jk}^i are independent centered sub-exponential random variables for $i=1,\ldots,n$ with $\max_i\|Z_{jk}^i\|_{\psi_1}\leq 2\max_i\|\epsilon_i\|_{\psi_2}\cdot\|A_{jk}^i\|_{\psi_2}\leq 2\nu$, according to Proposition 5.16 in [47], we have

$$\mathbb{P}\bigg\{\bigg|\frac{1}{n}\sum_{i=1}^nZ_{jk}^i\bigg|\geq t\bigg\}\leq 2\exp\bigg(-C'\frac{nt^2}{\nu^2}\bigg).$$

Thus by union bound, we have

$$\mathbb{P}\bigg\{ \max_{j,k} \bigg| \frac{1}{n} \sum_{i=1}^{n} Z_{jk}^{i} \bigg| \geq t \bigg\} \leq 2d_1 d_2 \exp\bigg(- C' \frac{nt^2}{\nu^2} \bigg).$$

Let $t = C_2 \nu \sqrt{\log d / n}$, we have the following inequality holds with probability at least $1 - C/d$

$$\bigg\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathbf{A}_i \bigg\|_{\infty,\infty} \leq C_2 \nu \sqrt{\frac{\log d}{n}}.$$

□

Thus, we complete the proof.

F.3 Proof of Lemma E.6

The proof of this lemma is inspired by the proof of Theorem 1 in [38], and we extended it to the sparse case. In order to prove Lemma E.6, we only need to prove the inequality (E.3) holds with high probability. Specifically, we consider the following event

$$E = \bigg\{ \exists\ \mathbf{S} \in \mathcal{C}(n) \mid \bigg| \frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \bigg| \leq \frac{1}{2} \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} + \frac{32 \|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}} \bigg\}.$$

Therefore, we want to establish the probability for event E, and we need the following lemmas.

Lemma F.1. Consider the robust PCA under observation model in section 4.2, for $\ell = 1, 2, \ldots$, we have

$$\mathbb{P}(E_\ell) \leq \exp(-c_1 n \alpha^{2\ell} \mu^2),$$

where we have

$$E_\ell := \bigg\{ \exists\ \mathbf{S} \in \mathcal{B}'(\alpha^\ell \mu) \mid \bigg| \frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \bigg| \geq \frac{5}{12} \frac{\alpha^\ell \mu}{\sqrt{d_1 d_2}} + \frac{32 \|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}} \bigg\},$$

and

$$\mathcal{B}'(\alpha^\ell \mu) = \bigg\{ \mathbf{S} \in \mathcal{C}(n, s) \mid \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \leq \frac{\alpha^\ell \mu}{\sqrt{d_1 d_2}} \bigg\}.$$

Proof of Lemma E.6. The reminder of this proof is to derive the probability of the event E. In order to establish the probability of the event E, we make use of the peeling argument of the Frobenius norm $\|\mathbf{S}\|_F$. Let $\mu = c\sqrt{\log d / n}$, and $\alpha = 6/5$. For $\ell = 1, 2, \ldots$, we define the sets

$$\mathcal{S}_\ell := \bigg\{ \mathbf{S} \in \mathcal{C}(n, s) \ \bigg|\ \frac{\alpha^{\ell-1} \mu}{\sqrt{d_1 d_2}} \leq \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \leq \frac{\alpha^\ell \mu}{\sqrt{d_1 d_2}} \bigg\}.$$

Therefore, if the event E holds, there exist a matrix $\mathbf{S}$ that must belongs to $\mathcal{S}_\ell$ for some $\ell = 1, 2, \ldots$ such that

$$\bigg| \frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \bigg| \geq \frac{1}{2} \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} + \frac{32 \|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}} \geq \frac{1}{2} \frac{\alpha^{\ell-1} \mu}{\sqrt{d_1 d_2}} + \frac{32 \|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}} = \frac{5}{12} \frac{\alpha^\ell \mu}{\sqrt{d_1 d_2}} + \frac{32 \|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}},$$

where the last equality is due to the fact that $\alpha = 6/5$.

Next, consider following events E_ℓ, for $\ell = 1, 2, \ldots$

$$E_\ell := \bigg\{ \exists\ \mathbf{S} \in \mathcal{B}'(\alpha^\ell \mu) \mid \bigg| \frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \bigg| \geq \frac{5}{12} \frac{\alpha^\ell \mu}{\sqrt{d_1 d_2}} + \frac{32 \|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}} \bigg\},$$

where we have the constraint set

$$\mathcal{B}'(\alpha^\ell \mu) = \bigg\{ \mathbf{S} \in \mathcal{C}(n, s) \mid \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \leq \frac{\alpha^\ell \mu}{\sqrt{d_1 d_2}} \bigg\}.$$

Since $\mathbf{S} \in \mathcal{S}_\ell$ implies that $\mathbf{S} \in \mathcal{B}(\alpha^\ell \mu)$, we can get $E \subset \bigcup_{\ell=1}^{\infty} E_\ell$. Therefore, we only need to upper bound the probability $\mathbb{P}(\bigcup_{\ell=1}^{\infty} E_\ell)$. In order to do so, we need upper bound the probability $\mathbb{P}(E_\ell)$. According to Lemma F.1, we have $\mathbb{P}(E_\ell) \leq \exp(-c_1 n \alpha^{2\ell} \mu^2)$. Therefore, we can obtain

$$\mathbb{P}(E) \leq \mathbb{P}\Big(\bigcup_{\ell=1}^{\infty} E_\ell \Big) \leq \sum_{\ell=1}^{\infty} \mathbb{P}(E_\ell) \leq \sum_{\ell=1}^{\infty} \exp(-c_1 n \alpha^{2\ell} \mu^2).$$

Thus according to the inequality $a \leq e^a$, we can obtain

$$\mathbb{P}(E) \leq \sum_{\ell=1}^{\infty} \exp\big(-2\ell c_1 n \mu^2 \log \alpha \big) \leq \frac{\exp\big(-2c_1 n \mu^2 \log \alpha \big)}{1 - \exp\big(-2c_1 n \mu^2 \log \alpha \big)} = \frac{\exp(-c_2 \log d)}{1 - \exp(-c_2 \log d)},$$

where the last equality comes from the definition $\mu = c\sqrt{\log d / n}$, and this implies $\mathbb{P}(E) \leq c_3 \exp(-c_2 \log d)$. □

F.4 Proof of Lemma E.7

Proof. The proof of this Lemma is similar to the proof of Lemma E.3, using Proposition 5.16 in [47] and covering number argument, with probability at least $1 - \exp(-c_1 d)$, we can obtain the restricted Lipschitz gradient condition in Condition 4.4 with parameter $K = c_2\sqrt{(rd + s)\log d / n}$. □

F.5 Proof of Lemma E.8

Proof. For the first inequality in Lemma E.8, it has been established in [38] Proposition 1. For the second inequality in Lemma E.8, we use the similar proof as in the proof of Lemma E.4. By proposition 5.16 in [47] and union bound, with probability at least $1 - C/d$, we can obtain the required inequality. □

G Proof of Auxiliary Lemmas in Appendix F

In order to prove Lemma F.1, we need the following lemmas.

Lemma G.1. We have the following holds with probability at least $1 - C\exp(-C_1 n D^2)$

$$\max_{k=1\ldots,N(D/8)} \bigg| \frac{\|\mathcal{A}(\mathbf{S}^k)\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}^k\|_F}{\sqrt{d_1 d_2}} \bigg| \leq \frac{D}{8\sqrt{d_1 d_2}} + \frac{32\|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}}.$$

Lemma G.2. We have the following holds

$$\sup_{\mathbf{\Delta} \in \mathcal{D}(\delta)} \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2}{\sqrt{n}} \leq \frac{D}{2\sqrt{d_1 d_2}},$$

where we have

$$\mathcal{D}(\delta) := \{\mathbf{\Delta} \in \mathbb{R}^{d_1 \times d_2} \mid \|\mathbf{\Delta}\|_F \leq \delta,\ \|\mathbf{\Delta}\|_{1,1} \leq 2\rho(D),\ \|\mathbf{\Delta}\|_0 \leq 2\widetilde{s}\},$$

and $\rho(D) \leq D^2 / \big(c\sqrt{\log d / n}\big)$.

Proof of Lemma F.1. The proof of this lemma is inspired by the proof of Lemma 3 in [38]. Note that since the definition of the constraint set $\mathcal{C}(n)$ and E is invariant to rescaling of $\mathbf{S}$, we can assume w.l.o.g. that $\|\mathbf{S}\|_{\infty,\infty} = 1/\sqrt{d_1 d_2}$. Therefore, it is equivalent to consider following events

$$E_\ell := \bigg\{ \exists\ \mathbf{S} \in \mathcal{B}(\alpha^\ell \mu) \mid \bigg| \frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \bigg| \geq \frac{3\alpha^\ell \mu}{4\sqrt{d_1 d_2}} + \frac{32}{\sqrt{n d_1 d_2}} \bigg\}$$

where we have the constraint set

$$\mathcal{B}(\alpha^\ell \mu) = \bigg\{ \mathbf{S} \in \mathcal{C}(n, s) \mid \|\mathbf{S}\|_{\infty,\infty} \leq \frac{1}{\sqrt{d_1 d_2}},\ \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \leq \frac{\alpha^\ell \mu}{\sqrt{d_1 d_2}},\ \|\mathbf{S}\|_{1,1} \leq \rho(\alpha^\ell \mu) \bigg\},$$

where $\rho(\alpha^\ell\mu) \leq (\alpha^\ell\mu)^2/\big(c\sqrt{\log d/n}\big)$. Define

$$Z_n(\alpha^\ell\mu) := \sup_{\mathbf{S}\in\mathcal{B}(\alpha^\ell\mu)} \left| \frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \right|.$$

For simplicity, we use D to denote $\alpha^\ell\mu$ in the following discussion. Therefore, we just need to prove the following probability bound

$$\mathbb{P}\bigg(Z_n(D) \geq \frac{3D}{4\sqrt{d_1 d_2}} + \frac{32}{\sqrt{n d_1 d_2}} \bigg) \leq c_3 \exp(-c_4 n D^2).$$

Suppose $\mathbf{S}^1, \ldots, \mathbf{S}^{N(\delta)}$ are a $\delta-$covering of $\mathcal{B}(D)$ in terms of Frobenius norm. Therefore, for any $\mathbf{S} \in \mathcal{B}(D)$, there exist a matrix $\mathbf{\Delta} \in \mathbb{R}^{d_1\times d_2}$ and some index $k \in \{1, \ldots, N(\delta)\}$ satisfying $\mathbf{S} = \mathbf{S}^k + \mathbf{\Delta}$, where $\|\mathbf{\Delta}\|_F \leq \delta$. Thus we can obtain

$$\begin{aligned}
\frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} &= \frac{\|\mathcal{A}(\mathbf{S}^k + \mathbf{\Delta})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}^k + \mathbf{\Delta}\|_F}{\sqrt{d_1 d_2}} \\
&\leq \frac{\|\mathcal{A}(\mathbf{S}^k)\|_2}{\sqrt{n}} + \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}^k\|_F}{\sqrt{d_1 d_2}} + \frac{\|\mathbf{\Delta}\|_F}{\sqrt{d_1 d_2}} \\
&\leq \left| \frac{\|\mathcal{A}(\mathbf{S}^k)\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}^k\|_F}{\sqrt{d_1 d_2}} \right| + \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2}{\sqrt{n}} + \frac{\delta}{\sqrt{d_1 d_2}}.
\end{aligned}$$

In addition we can get

$$\left| \frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} \right| \leq \left| \frac{\|\mathcal{A}(\mathbf{S}^k)\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}^k\|_F}{\sqrt{d_1 d_2}} \right| + \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2}{\sqrt{n}} + \frac{\delta}{\sqrt{d_1 d_2}}.$$

Therefore, we have

$$Z_n(D) \leq \frac{\delta}{\sqrt{d_1 d_2}} + \max_{k=1\ldots,N(\delta)} \left| \frac{\|\mathcal{A}(\mathbf{S}^k)\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}^k\|_F}{\sqrt{d_1 d_2}} \right| + \sup_{\mathbf{\Delta}\in\mathcal{D}(\delta)} \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2}{\sqrt{n}}, \tag{G.1}$$

where we have $\mathcal{D}(\delta) := \{\mathbf{\Delta} \in \mathbb{R}^{d_1\times d_2} \mid \|\mathbf{\Delta}\|_F \leq \delta,\ \|\mathbf{\Delta}\|_{1,1} \leq 2\rho(D),\ \|\mathbf{\Delta}\|_0 \leq 2cs\}$. We establish the high probability bound of (G.1) with $\delta = D/8$. First, according to Lemma G.1, we have

$$\max_{k=1\ldots,N(D/8)} \left| \frac{\|\mathcal{A}(\mathbf{S}^k)\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}^k\|_F}{\sqrt{d_1 d_2}} \right| \leq \frac{D}{8\sqrt{d_1 d_2}} + \frac{32\|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}}, \tag{G.2}$$

holds with probability at least $1 - c\exp(-c_1 n D^2)$.

Next, according to Lemma G.2, we have

$$\sup_{\mathbf{\Delta}\in\mathcal{D}(\delta)} \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2}{\sqrt{n}} \leq \frac{D}{2\sqrt{d_1 d_2}}, \tag{G.3}$$

holds with probability at least $1 - c_2\exp(-c_3 n D^2)$.

Therefore, combining (G.2) and (G.3), we can get

$$Z_n(D) \leq \frac{D}{8\sqrt{d_1 d_2}} + \frac{D}{8\sqrt{d_1 d_2}} + \frac{D}{2\sqrt{d_1 d_2}} + \frac{32\|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}} \leq \frac{3D}{4\sqrt{d_1 d_2}} + \frac{32}{\sqrt{n d_1 d_2}},$$

holds with probability at least $1 - c_3\exp(-c_4 n D^2)$, and the last inequality comes from that $\|\mathbf{S}\|_{\infty,\infty} \leq 1/\sqrt{d_1 d_2}$. □

H Proofs of Auxiliary Lemmas in Appendix G

H.1 Proof of Lemma G.1

Proof. First, we prove that for a fixed matrix $\mathbf{S}$, we have the following inequality holds

$$\mathbb{P}\bigg(\bigg|\frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}}\bigg| \geq \frac{\delta}{\sqrt{d_1 d_2}} + \frac{32\|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}}\bigg) \leq C\exp(-C_1 n\delta^2).$$

Since we have

$$\frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sqrt{\sum_{j,k\in\Omega}\langle\mathbf{A}_{jk},\mathbf{S}\rangle^2} = \frac{1}{\sqrt{n}}\sup_{\|\mathbf{w}\|_2=1}\sum_{j,k\in\Omega} w_i\langle\mathbf{A}_{jk},\mathbf{S}\rangle,$$

we consider

$$\begin{aligned}\sqrt{d_1 d_2}\frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} = \frac{\sqrt{d_1 d_2}}{\sqrt{n}}\sqrt{\sum_{j,k\in\Omega}\langle\mathbf{A}_{jk},\mathbf{S}\rangle^2} &= \frac{1}{\sqrt{n}}\sup_{\|\mathbf{w}\|_2=1}\sum_{j,k\in\Omega} w_{jk}\langle\sqrt{d_1 d_2}\mathbf{A}_{jk},\mathbf{S}\rangle \\ &= \frac{1}{\sqrt{n}}\sup_{\|\mathbf{w}\|_2=1}\sum_{j,k\in\Omega} w_{jk}Y_{jk},\end{aligned}$$

where we have the random variables Y_{jk} satisfying $|Y_{jk}| = |\langle\sqrt{d_1 d_2}\mathbf{A}_{jk},\mathbf{S}\rangle| \leq \sqrt{d_1 d_2}\|\mathbf{S}\|_{\infty,\infty} = 1$. Therefore, according to lemma I.1, we have

$$\mathbb{P}\bigg(\bigg|\frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \mathbb{E}\bigg[\frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}}\bigg]\bigg| \geq \frac{\delta}{\sqrt{d_1 d_2}} + \frac{16}{\sqrt{n d_1 d_2}}\bigg) \leq C\exp(-C_1 n\delta^2). \tag{H.1}$$

In addition, we have

$$\begin{aligned}\bigg|\frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}} - \mathbb{E}\bigg[\frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}}\bigg]\bigg| &= \bigg|\sqrt{\mathbb{E}\bigg[\frac{\|\mathcal{A}(\mathbf{S})\|_2^2}{n}\bigg]} - \mathbb{E}\bigg[\frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}}\bigg]\bigg| \\ &\leq \sqrt{\mathbb{E}\bigg[\frac{\|\mathcal{A}(\mathbf{S})\|_2^2}{n}\bigg] - \mathbb{E}\bigg(\bigg[\frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}}\bigg]\bigg)^2} \leq \frac{16}{\sqrt{n d_1 d_2}}.\end{aligned} \tag{H.2}$$

Therefore, combining (H.1) and (H.2), we can obtain

$$\mathbb{P}\bigg(\bigg|\frac{\|\mathcal{A}(\mathbf{S})\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}\|_F}{\sqrt{d_1 d_2}}\bigg]\bigg| \geq \frac{\delta}{d_1 d_2} + \frac{32}{\sqrt{n d_1 d_2}}\bigg) \leq C\exp(-C_1 n\delta^2).$$

Next, according to Lemma 4 in [38], there exists a $\delta-$covering of $\mathcal{B}(D)$ such that

$$\log N(\delta) \leq C_3(\rho(D)/\delta)^2\log d.$$

Therefore, we can get

$$\begin{aligned}\mathbb{P}\bigg[\max_{k=1\ldots,N(D/8)}\bigg|\frac{\|\mathcal{A}(\mathbf{S}^k)\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}^k\|_F}{\sqrt{d_1 d_2}}\bigg| \geq \frac{\delta}{\sqrt{d_1 d_2}} + \frac{32}{\sqrt{n d_1 d_2}}\bigg] &\leq C\exp(-C_1 n\delta^2 + \log N(\delta)) \\ &\leq C\exp(-C_1 n\delta^2 + C_3(\rho(D)/\delta)^2\log d).\end{aligned}$$

Since we have $\delta = D/8$ and $\rho(D) = C_4 D^2/\sqrt{\log d/n}$, we can obtain

$$\mathbb{P}\bigg[\max_{k=1\ldots,N(D/8)}\bigg|\frac{\|\mathcal{A}(\mathbf{S}^k)\|_2}{\sqrt{n}} - \frac{\|\mathbf{S}^k\|_F}{\sqrt{d_1 d_2}}\bigg| \geq \frac{\delta}{\sqrt{d_1 d_2}} + \frac{32\|\mathbf{S}\|_{\infty,\infty}}{\sqrt{n}}\bigg] \leq C\exp(-C_2 n\delta^2),$$

which complete the proof. □

H.2 Proof of Lemma G.2

Proof. According to Lemma 5 in [38], we have following results

$$\mathbb{P}\bigg[\bigg|\sup_{\boldsymbol{\Delta}\in\mathcal{D}(\delta)}\sqrt{d_1d_2}\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2}{\sqrt{n}}-\mathbb{E}\bigg[\sup_{\boldsymbol{\Delta}\in\mathcal{D}(\delta)}\sqrt{d_1d_2}\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2}{\sqrt{n}}\bigg]\bigg|\geq\delta\bigg]\leq C\exp(-C_1n\delta^2), \tag{H.3}$$

and

$$\bigg(\mathbb{E}\bigg[\sup_{\boldsymbol{\Delta}\in\mathcal{D}(\delta)}\sqrt{d_1d_2}\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2}{\sqrt{n}}\bigg]\bigg)^2\leq 16\sqrt{d_1d_2}\|\boldsymbol{\Delta}\|_\infty\mathbb{E}\bigg[\sup_{\boldsymbol{\Delta}\in\mathcal{D}(\delta)}\frac{1}{n}\sum_{j,k\in\Omega}\xi_{jk}\langle\mathbf{A}_{jk},\boldsymbol{\Delta}\rangle\bigg]+\delta^2,$$

where ξ_{jk} are independent Rademacher variables. Furthermore, by the duality between norms, we can obtain

$$\frac{1}{n}\sum_{j,k\in\Omega}\xi_{jk}\langle\mathbf{A}_{jk},\boldsymbol{\Delta}\rangle\leq\bigg\|\frac{1}{n}\sum_{j,k\in\Omega}\xi_{jk}\mathbf{A}_{jk}\bigg\|_\infty\cdot\|\boldsymbol{\Delta}\|_{1,1}\leq\rho(D)\bigg\|\frac{1}{n}\sum_{j,k\in\Omega}\xi_{jk}\mathbf{A}_{jk}\bigg\|_{\infty,\infty},$$

where the last inequality is due to the fact that $\boldsymbol{\Delta}\in\mathcal{D}(\delta)$. Finally, we have

$$\bigg\|\frac{1}{n}\sum_{j,k\in\Omega}\xi_{jk}\mathbf{A}_{jk}\bigg\|_{\infty,\infty}\leq C\sqrt{\frac{\log d}{n}}. \tag{H.4}$$

To prove this, we use Hoeffding's inequality and Union Bound. By the definition of $\mathbf{A}_i$, we can obtain

$$\bigg\|\frac{1}{n}\sum_{i=1}^n\xi_i\mathbf{A}_i\bigg\|_{\infty,\infty}=\max_{j,k}\bigg|\frac{1}{n}\sum_{i=1}^n\xi_i\mathbf{e}_j^i\mathbf{e}_k^i\bigg|=\max_{j,k}\bigg|\frac{1}{n}\sum_{i=1}^nZ_{jk}^i\bigg|,$$

where we have $Z_{jk}^i=\xi_iA_{jk}$. Thus we can get $|Z_{jk}^i|\leq|\xi_i|=1$, and we conclude that Z_{jk}^i are independent centered sub-Gaussian random variables for $i=1,\ldots,n$. Therefore, following the same procedure as in the proof of Lemma E.4, we can obtain inequality (H.4). Therefore, we can obtain

$$\bigg(\mathbb{E}\bigg[\sup_{\boldsymbol{\Delta}\in\mathcal{D}(\delta)}\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2}{\sqrt{n}}\bigg]\bigg)^2\leq C\frac{\|\boldsymbol{\Delta}\|_{\infty,\infty}}{\sqrt{d_1d_2}}\rho(D)\sqrt{\frac{\log d}{n}}+\frac{\delta^2}{d_1d_2}\leq C'\frac{D^2}{d_1d_2},$$

where the last inequality comes from the definition of $\rho(D)$, δ and $\|\boldsymbol{\Delta}\|_{\infty,\infty}\leq 2/\sqrt{d_1d_2}$. It implies that

$$\mathbb{E}\bigg[\sup_{\boldsymbol{\Delta}\in\mathcal{D}(\delta)}\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2}{\sqrt{n}}\bigg]\leq C''\frac{D}{\sqrt{d_1d_2}}. \tag{H.5}$$

Thus combining (H.3) and (H.5), we have

$$\sup_{\boldsymbol{\Delta}\in\mathcal{D}(\delta)}\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2}{\sqrt{n}}\leq\frac{D}{2\sqrt{d_1d_2}}$$

holds with probability at least $1-C\exp(-C_1nD^2)$. □

I Other Auxiliary Lemmas

Lemma I.1. [31] Consider independent random variables $Y_1,\ldots,Y_n$ such that $a_i\leq Y_i\leq b_i$ for $i=1,\ldots,n$. Let

$$Z:=\sup_{\mathbf{t}\in\mathcal{T}}\sum_{i=1}^nt_iY_i,$$

where $\mathcal{T}$ is a family of vectors $\mathbf{t}\in\mathbb{R}^n$ such that $\sigma=\sup_{t\in\mathcal{T}}\big(\sum_{i=1}^nt_i^2(b_i-a_i)^2\big)^{1/2}\leq\infty$. Then, for any $r\geq 0$, we have

$$\mathbb{P}\big(|Z-m_Z|\geq r\big)\leq 4\exp\Big(-\frac{r^2}{4\sigma^2}\Big),$$

where m_Z is a median of Z. Furthermore, we have

$$|\mathbb{E}(Z)-m_Z|\leq 4\sqrt{\pi}\sigma\quad\text{and}\quad\mathrm{Var}(Z)\leq 16\sigma^2.$$