# A. Additional Applications and Experimental Results

In this section, we present the application of our generic framework to one-bit matrix completion as well as additional experimental results for matrix sensing.

## A.1. One-bit Matrix Completion

Compared with matrix completion, we only observe the sign of each noisy entries of the unknown low-rank matrix $\mathbf{X}^*$ in one-bit matrix completion (Davenport et al., 2014; Cai & Zhou, 2013). We consider the uniform sampling model, which has been studied in existing literature (Davenport et al., 2014; Cai & Zhou, 2013; Ni & Gu, 2016). More specifically, we consider the following observation model, which is based on a differentiable function $f : \mathbb{R} \to [0, 1]$

$$Y_{jk} = \begin{cases} +1, & \text{with probability } f(X^*_{jk}), \\ -1, & \text{with probability } 1 - f(X^*_{jk}), \end{cases} \tag{A.1}$$

where we use a binary matrix $\mathbf{Y}$ to denote the observation matrix in (A.1). In addition, if the function $f$ is a cumulative distribution function with respect to $-Z_{jk}$, then we can rewrite the observation model (A.1) as follows

$$Y_{jk} = \begin{cases} +1, & \text{if } X^*_{jk} + Z_{jk} > 0, \\ -1, & \text{if } X^*_{jk} + Z_{jk} < 0, \end{cases} \tag{A.2}$$

where we use $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$ to denote the noise matrix with i.i.d. elements $Z_{jk}$. A lot of functions can be applied to observation model (A.1), and we consider the broadly-used logistic function $f(X_{jk}) = e^{X_{jk}}/(1+e^{X_{jk}})$ as the observation probability function in our study, which is equivalent to the fact that each noise element $Z_{jk}$ in model (A.2) follows the standard logistic distribution. Similar to matrix completion, we use $\Omega \subseteq [d_1] \times [d_2]$ to denote the index set of the observed elements. Therefore, given the logistic function $f$ and the index set $\Omega$, we define $F_\Omega(\mathbf{U}, \mathbf{V})$ for one-bit matrix completion as follows

$$F_\Omega(\mathbf{U}, \mathbf{V}) := \mathcal{L}_\Omega(\mathbf{U}\mathbf{V}^\top) + \mathcal{R}(\mathbf{U}, \mathbf{V}) = \frac{1}{n}\sum_{i=1}^{n} F_{\Omega_{\mathcal{S}_i}}(\mathbf{U}, \mathbf{V}),$$

where $\mathcal{L}_\Omega(\mathbf{U}\mathbf{V}^\top)$ is the negative log-likelihood function such that

$$\mathcal{L}_\Omega(\mathbf{U}\mathbf{V}^\top) = -\frac{1}{N}\sum_{(j,k)\in\Omega}\Big\{\mathbb{1}\{Y_{jk} = 1\}\log\big(f(\mathbf{U}_{j*}\mathbf{V}^\top_{*k})\big) + \mathbb{1}\{Y_{jk} = -1\}\log\big(1 - f(\mathbf{U}_{j*}\mathbf{V}^\top_{*k})\big)\Big\}.$$

Therefore, for each component function, we have

$$F_{\Omega_{\mathcal{S}_i}}(\mathbf{U}, \mathbf{V}) = \mathcal{L}_{\Omega_{\mathcal{S}_i}}(\mathbf{U}\mathbf{V}^\top) + \mathcal{R}(\mathbf{U}, \mathbf{V}),$$

where $\{\Omega_{\mathcal{S}_i}\}_{i=1}^n$ denote the mutually disjoint subsets such that $\cup_{i=1}^n \Omega_{\mathcal{S}_i} = \Omega$. In addition, we have $|\Omega_{\mathcal{S}_i}| = b$ for $i = 1, \ldots, n$ such that $|\Omega| = nb$. And $\mathcal{L}_{\Omega_{\mathcal{S}_i}}(\mathbf{U}\mathbf{V}^\top)$ is defined as

$$\mathcal{L}_{\Omega_{\mathcal{S}_i}}(\mathbf{U}\mathbf{V}^\top) = \frac{1}{b}\sum_{(j,k)\in\Omega_{\mathcal{S}_i}}\Big\{\mathbb{1}\{Y_{jk} = 1\}\log\big(f(\mathbf{U}_{j*}\mathbf{V}^\top_{*k})\big) + \mathbb{1}\{Y_{jk} = -1\}\log\big(1 - f(\mathbf{U}_{j*}\mathbf{V}^\top_{*k})\big)\Big\}.$$

## A.2. Theoretical Guarantees for One-bit Matrix Completion

We establish the theoretical guarantee of our algorithm for one-bit matrix completion. We obtain the restricted strong convexity and smoothness conditions for $\mathcal{L}_N$ with parameters $\mu = C_1\mu_\alpha$ and $L = C_2 L_\alpha$. In addition, we are able to get the restricted strong smoothness condition for each component function $\mathcal{L}_i$ with parameter $L' = c_0 L_\alpha > L$. Here, $\mu_\alpha$ and $L_\alpha$ are defined as

$$\mu_\alpha \le \min\left(\inf_{|x|\le\alpha}\left\{\frac{f'^2(x)}{f^2(x)} - \frac{f''(x)}{f(x)}\right\}, \inf_{|x|\le\alpha}\left\{\frac{f'^2(x)}{(1-f(x))^2} + \frac{f''(x)}{1-f(x)}\right\}\right), \tag{A.3}$$

$$L_\alpha \ge \max\left(\sup_{|x|\le\alpha}\left\{\frac{f'^2(x)}{f^2(x)} - \frac{f''(x)}{f(x)}\right\}, \sup_{|x|\le\alpha}\left\{\frac{f'^2(x)}{(1-f(x))^2} + \frac{f''(x)}{1-f(x)}\right\}\right), \tag{A.4}$$

where $f(x)$ is the function used in (A.1), and each element $X_{jk}$ satisfies $|X_{jk}| \leq \alpha$. Note that given the function $f(x)$ and constant $\alpha$, we can calculate $\mu_\alpha$ and $L_\alpha$, which are fixed constants and do not rely on dimension of the unknown low-rank matrix. For example, if we have logistic function, we can get $\mu_\alpha = e^\alpha/(1+e^\alpha)^2$ and $L_\alpha = 1/4$. Furthermore, we define $\gamma_\alpha$ as follows, which reflects the steepness property of the sample loss function $\mathcal{L}_N(\cdot)$

$$\gamma_\alpha \geq \sup_{|x|\leq\alpha} \left\{ \frac{|f'(x)|}{f(x)\big(1-f(x)\big)} \right\}. \tag{A.5}$$

Moreover, we can derive the upper bound of the $\nabla\mathcal{L}_N(\mathbf{X}^*)$ in terms of spectral norm. If we choose the step size $\eta = c_1'/\sigma_1$, where $c_1' = \mu'/\big(c_0'\kappa\big)$, and the inner loop iterations $m \geq c_2'\kappa^2$, where $c_0', c_1'$ and $c_2'$ are some constants, then we have the following convergence result of our algorithm for the model of matrix completion.

**Corollary A.1.** Consider one-bit matrix completion under uniform sampling model with log-concave function $f$ in (A.1). Suppose $\mathbf{X}^*$ satisfies the incoherence condition. There exist constants $\{c_i\}_{i=1}^7$ such that if we choose parameters $\eta = c_1/\sigma_1$, where $c_1 = \mu'/\big(c_2\kappa\big)$, $m \geq c_3\kappa^2$, and the number of observations satisfies $N \geq c_4 r^2 d\log d$, then for any initial solution $\widetilde{\mathbf{Z}}^0 \in \mathbb{B}(c_5\sqrt{\sigma_r})$, with probability at least $1 - c_6/d$, the output of our Algorithm 1 satisfies

$$\mathbb{E}\big[d^2(\widetilde{\mathbf{Z}}^S, \mathbf{Z}^*)\big] \leq \rho^S d^2(\widetilde{\mathbf{Z}}^0, \mathbf{Z}^*) + c_7 \max\{\gamma_\alpha^2, r\beta^2\sigma_1^2\}\frac{rd\log d}{N}, \tag{A.6}$$

where the contraction parameter $\rho < 1$.

**Remark A.2.** For one-bit matrix completion, our algorithm achieves $O\big(r\sqrt{d\log d/N}\big)$ statistical error after $O\big(\log(N/(r^2 d\log d))\big)$ number of outer loop iterations. We note that this statistical error is near optimal, compared with the minimax lower bound of one-bit matrix completion $O\big(\sqrt{rd\log d/N}\big)$ established in (Davenport et al., 2014; Cai & Zhou, 2013). Moreover, Remark 3.10 tells us that for our estimator $\widetilde{\mathbf{Z}}^S$ to achieve $\epsilon$ accuracy, the overall computational complexity required by our algorithm is $O\big((N+\kappa^2 b)r^3 d\log(1/\epsilon)\big)$. Nevertheless, the overall computational complexity for the state-of-the-art gradient descent based algorithm (Wang et al., 2016) to obtain $\epsilon$ accuracy is $O\big(N\kappa r^3 d\log(1/\epsilon)\big)$. Therefore, as long as we have $\kappa \leq n$, our approach is more efficient than the state-of-the-art gradient descent method. Furthermore, the overall computational complexities for the state-of-the-art projected gradient descent algorithm (Chen & Wainwright, 2015) and the conditional gradient descent (a.k.a., Frank-Wolfe) algorithm (Ni & Gu, 2016) to obtain $\epsilon$ accuracy are both $O\big(Nr^2\log(1/\epsilon)\big)$[2]. If we have $\kappa^2 \leq nr$, our method clearly has a lower computational complexity than theirs.

### A.3. Experimental Results for Matrix Sensing and One-bit Matrix Completion

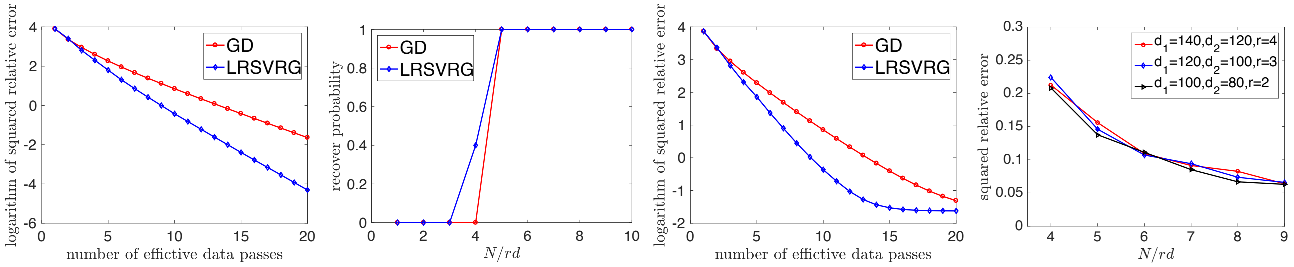
In this section, we present our experimental results for matrix sensing and one-bit matrix completion respectively.

#### A.3.1. Matrix Sensing

For matrix sensing, we use the same procedure as in matrix completion to generate the unknown low-rank matrix $\mathbf{X}^*$. Then, we obtain linear measurements from the following observation model $y_i = \langle \mathbf{A}_i, \mathbf{X}^*\rangle + \epsilon_i$, where each element of the sensing matrix $\mathbf{A}_i$ follows i.i.d. standard normal distribution. We also consider the same noisy and noiseless settings as in matrix completion.

For the results of the convergence rate, Figure 2(a) and 2(c) illustrate the squared relative error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/\|\mathbf{X}^*\|_F^2$ in log scale versus number of effective data passes for both methods under setting (i). These results show the linear convergence rate of our method. Most importantly, it clearly demonstrates the superiority of our approach, since our algorithm shows better performance after the same number of effective data passes compared with the state-of-the-art gradient descent algorithm (Zheng & Lafferty, 2015; Wang et al., 2016). Since we get results with similar patterns for other settings, we leave them out for simplicity. Figure 2(b) shows the empirical recovery probability of different methods under setting (i). The result implies a phase transition around $N = 3rd$, which is consistent with the optimal sample complexity that $N$ is linear with $rd$. Besides, since we get results with similar patterns for other settings, we leave them out to save space. For the results of statistical error, Figure 2(d) shows, in the noisy case, how the estimation errors scale with the rescaled sample size $N/(rd)$, which confirms our theoretical results.

[2] Note that the overall computational complexities for the projected gradient descent (Chen & Wainwright, 2015) and conditional gradient descent (Ni & Gu, 2016) algorithms also depend on some problem dependent parameters, which we omit here but actually can make their computational complexities worse. Please refer to their papers for more accurate complexity results.



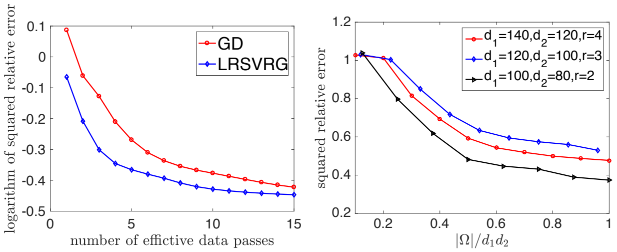
(a) Noiseless Case (b) Sample Complexity (c) Noisy Case (d) Statistical Error

*Figure 2.* Numerical results for matrix sensing. (a) and (c) Convergence rates for matrix sensing in the noiseless and noisy case, respectively: logarithm of $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/\|\mathbf{X}^*\|_F^2$ versus number of effective data passes. They demonstrate the linear convergence rate and the superiority of our method; (b) Empirical probability of exact recovery versus $N/(rd)$, which confirms the optimal sample complexity in the noiseless case that $N = O(rd)$; (d) Statistical error: $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/\|\mathbf{X}^*\|_F^2$ versus $N/(rd)$, which matches the statistical error of our theory.

### A.3.2. One-bit Matrix Completion

We use the same settings of $\mathbf{X}^*$ for one-bit matrix completion as before. In order to obtain $\mathbf{X}^*$, we adopt the similar procedure as in (Davenport et al., 2014; Bhaskar & Javanmard, 2015; Ni & Gu, 2016). In detail, we first randomly generate $\mathbf{U}^* \in \mathbb{R}^{d_1 \times r}, \mathbf{V}^* \in \mathbb{R}^{d_2 \times r}$ from a uniform distribution on $[-1/2, 1/2]$. Then we get $\mathbf{X}^*$ by $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^{*\top}$. Finally, we scale $\mathbf{X}^*$ to make it satisfies $\|\mathbf{X}^*\|_\infty = \alpha = 1$. Here we consider the uniform observation model with function $f(X_{ij}) = \Phi(\mathbf{X}_{ij}/\sigma)$ in (A.1), where $\Phi$ is the cumulative distribution function of the standard normal distribution, and $\sigma$ is the noise level, which we set it to be $\sigma = 0.5$.

For the results of convergence rate, we compute the logarithm of the squared relative error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/\|\mathbf{X}^*\|_F^2$, which are displayed in Figure 3(a). Note that, for the ease of illustration, we show the results of convergence rate after the first data pass. The results not only confirm the linear rate of convergence of our algorithm, but also demonstrate the effectiveness of our method after the same number of effective data passes. Besides, since we get results with similar patterns for other settings, we leave them out for simplicity. For the results of statistical error, Figure 3(b) illustrates that with the same percentage of observations, the squared relative error decreases as the ratio $r/d$ decreases. Although our theoretical results give $O(r^2 d \log d/|\Omega|)$ statistical error, the simulation results suggest that our method can achieve the minimax statistical error.



(a) Convergence Rate (b) Statistical Error

*Figure 3.* Numerical results for one-bit matrix completion. (a) Convergence rates for one-bit matrix completion: logarithm of squared relative error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/\|\mathbf{X}^*\|_F^2$ versus number of effective data passes. It illustrates the superiority of our method after the same number of effective data passes; (b) Statistical error for one-bit matrix completion: squared relative error $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F^2/\|\mathbf{X}^*\|_F^2$ versus $|\Omega|/(d_1 d_2)$, which verifies the statistical rate.

## B. Proof of the Main Theory

We provide the proof of our main theoretical results in this section. Since we aim to minimize the following objective function in terms of $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$

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

Therefore, we obtain the corresponding gradient

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

### B.1. Proof of Theorem 3.8

In order to prove Theorem 3.8, we need following lemmas, and we present the corresponding proofs in Sections D.2 and D.3, respectively.

**Lemma B.1** (Local Curvature Condition)**.** Suppose the sample loss function $\mathcal{L}_N$ satisfies Conditions 3.3 and 3.4. For any matrix $\mathbf{Z} = [\mathbf{U}; \mathbf{V}] \in \mathbb{R}^{(d_1+d_2)\times r}$, where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2\times r}$, denote $\widetilde{\mathbf{Z}} = [\mathbf{U}; -\mathbf{V}]$. In addition, we use $\mathbf{R} = \mathrm{argmin}_{\widetilde{\mathbf{R}}\in\mathbb{Q}_r} \|\mathbf{Z} - \mathbf{Z}^*\widetilde{\mathbf{R}}\|_F$ to denote the optimal rotation with respect to $\mathbf{Z}$, and $\mathbf{H} = \mathbf{Z} - \mathbf{Z}^*\mathbf{R}$, then the following inequality holds

$$\begin{aligned} \langle \nabla \widetilde{F}_N(\mathbf{Z}), \mathbf{H}\rangle \geq \frac{\mu}{8}\|\mathbf{X} - \mathbf{X}^*\|_F^2 + \frac{\mu'\sigma_r}{10}\|\mathbf{H}\|_F^2 + \frac{1}{16}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2 \\ - \frac{3L+1}{8}\|\mathbf{H}\|_F^4 - \left(\frac{4r}{\mu} + \frac{r}{2L}\right)\cdot\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2, \end{aligned}$$

where $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, and $\mu' = \min\{\mu, 1\}$.

**Lemma B.2** (Local Smoothness Condition)**.** Assume the component loss function $\mathcal{L}_i$ satisfies Condition 3.7. Suppose we randomly pick $i \in [n]$. For any $\mathbf{U} \in \mathbb{R}^{d_1\times r}$, $\mathbf{V} \in \mathbb{R}^{d_2\times r}$ and rank-$r$ matrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{d_1\times d_2}$, we denote $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$ and $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. Let $\mathbf{G}_U = \nabla_{\mathbf{U}} F_i(\mathbf{U}, \mathbf{V}) - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\mathbf{V} + \nabla\mathcal{L}_N(\widetilde{\mathbf{X}})\mathbf{V}$, $\mathbf{G}_V^t = \nabla_{\mathbf{V}} F_i\big(\mathbf{U}, \mathbf{V}) - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})^\top\mathbf{U} + \nabla\mathcal{L}_N(\mathbf{X})^\top\mathbf{U}$, and $\mathbf{G} = [\mathbf{G}_U; \mathbf{G}_V]$. Then we have

$$\begin{aligned} \mathbb{E}\|\mathbf{G}\|_F^2 \leq 24\big(2L'^2\|\widetilde{\mathbf{X}} - \mathbf{X}^*\|_F^2 + (2L'^2 + L^2)\cdot\|\mathbf{X} - \mathbf{X}^*\|_F^2\big)\cdot\|\mathbf{Z}\|_2^2 \\ + \|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2\cdot\|\mathbf{Z}\|_2^2 + 12r\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2\cdot\|\mathbf{Z}\|_2^2. \end{aligned}$$

*Proof of Theorem 3.8.* According to stochastic variance reduced gradient descent Algorithm 1, consider iteration $t$ in the inner loop, we have the following update

$$\mathbf{U}^{t+1} = \mathcal{P}_{\mathcal{C}_1}(\mathbf{U}^t - \eta\mathbf{G}_U^t), \text{ and } \mathbf{V}^{t+1} = \mathcal{P}_{\mathcal{C}_2}(\mathbf{V}^t - \eta\mathbf{G}_V^t),$$

where we denote

$$\begin{aligned} \mathbf{G}_U^t &= \nabla_{\mathbf{U}} F_{i_t}(\mathbf{U}^t, \mathbf{V}^t) - \nabla\mathcal{L}_{i_t}(\widetilde{\mathbf{X}})\mathbf{V}^t + \nabla\mathcal{L}_N(\widetilde{\mathbf{X}})\mathbf{V}^t, \\ \mathbf{G}_V^t &= \nabla_{\mathbf{V}} F_{i_t}\big(\mathbf{U}^t, \mathbf{V}^t) - \nabla\mathcal{L}_{i_t}(\widetilde{\mathbf{X}})^\top\mathbf{U}^t + \nabla\mathcal{L}_N(\mathbf{X}^t)^\top\mathbf{U}^t. \end{aligned}$$

Since $i_t$ is uniformly picked from $[n]$, we have $\mathbb{E}[\mathbf{G}_U^t] = \nabla_{\mathbf{U}} F_N(\mathbf{U}^t, \mathbf{V}^t)$ and $\mathbb{E}[\mathbf{G}_V^t] = \nabla_{\mathbf{V}} F_N(\mathbf{U}^t, \mathbf{V}^t)$, where the expectation is taken with respect to $i_t$. Recall $\mathbf{Z}^t = [\mathbf{U}^t; \mathbf{V}^t]$, and $\mathbf{R}^t = \mathrm{argmin}_{\mathbf{R}\in\mathbb{Q}_r}\|\mathbf{Z}^t - \mathbf{Z}^*\mathbf{R}\|_F$ as the optimal rotation with respect to $\mathbf{Z}^t$. Denote $\mathbf{H}^t = \mathbf{Z}^t - \mathbf{Z}^*\mathbf{R}^t$ and $\mathbf{G}^t = [\mathbf{G}_U^t; \mathbf{G}_V^t]$. By induction, for any $t \geq 0$, we assume $\mathbf{Z}^t \in \mathbb{B}(c_2\sqrt{\sigma_r})$. Thus, by taking the expectation of $\mathbf{H}^{t+1}$ over $i_t$ conditioned on $\mathbf{Z}^t$, we have

$$\begin{aligned} \mathbb{E}\|\mathbf{H}^{t+1}\|_F^2 &\leq \mathbb{E}\|\mathcal{P}_{\mathcal{C}_1}(\mathbf{U}^t - \eta\mathbf{G}_U^t) - \mathbf{U}^*\mathbf{R}^t\|_F^2 + \mathbb{E}\|\mathcal{P}_{\mathcal{C}_2}(\mathbf{V}^t - \eta\mathbf{G}_V^t) - \mathbf{V}^*\mathbf{R}^t\|_F^2 \\ &\leq \mathbb{E}\|\mathbf{U}^t - \eta\mathbf{G}_U^t - \mathbf{U}^*\mathbf{R}^t\|_F^2 + \mathbb{E}\|\mathbf{V}^t - \eta\mathbf{G}_V^t - \mathbf{V}^*\mathbf{R}^t\|_F^2 \\ &= \|\mathbf{H}^t\|_F^2 - 2\eta\mathbb{E}\langle\mathbf{G}^t, \mathbf{H}^t\rangle + \eta^2\mathbb{E}\|\mathbf{G}^t\|_F^2 \\ &= \|\mathbf{H}^t\|_F^2 - 2\eta\langle\nabla\widetilde{F}_N(\mathbf{Z}^t), \mathbf{H}^t\rangle + \eta^2\mathbb{E}\|\mathbf{G}^t\|_F^2, \end{aligned} \tag{B.3}$$

where the first inequality follows from the definition of $\mathbf{H}^t$, the second inequality follows from the non-expansive property of the projection $\mathcal{P}_{\mathcal{C}_i}$ onto $\mathcal{C}_i$ and the fact that $\mathbf{U}^* \in \mathcal{C}_1, \mathbf{V}^* \in \mathcal{C}_2$, and the last equality holds because conditioned on $\mathbf{Z}^t$, $\mathbb{E}\langle\mathbf{H}^t, \mathbf{G}^t\rangle = \langle\mathbf{H}^t, \mathbb{E}\mathbf{G}^t\rangle = \langle\mathbf{H}^t, \nabla\widetilde{F}_N(\mathbf{Z}^t)\rangle$, where $\widetilde{F}_N$ is defined in (B.1) . According to Lemma B.1, we can obtain the lower bound of $\langle\nabla\widetilde{F}_N(\mathbf{Z}^t), \mathbf{H}^t\rangle$.

$$\begin{aligned} \langle\nabla\widetilde{F}_N(\mathbf{Z}^t), \mathbf{H}^t\rangle \geq \frac{\mu}{8}\|\mathbf{X}^t - \mathbf{X}^*\|_F^2 + \frac{\mu'\sigma_r}{10}\|\mathbf{H}^t\|_F^2 + \frac{1}{16}\|\mathbf{U}^{t\top}\mathbf{U}^t - \mathbf{V}^{t\top}\mathbf{V}^t\|_F^2 - \frac{3L+1}{8}\|\mathbf{H}^t\|_F^4 \\ - \left(\frac{4r}{\mu} + \frac{r}{2L}\right)\cdot\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2, \end{aligned} \tag{B.4}$$

where $\mu' = \min\{\mu, 1\}$. According to Lemma B.2, we have

$$\begin{aligned}\mathbb{E}\|\mathbf{G}^t\|_F^2 \leq 24\big(2L'^2\|\widetilde{\mathbf{X}}^t - \mathbf{X}^*\|_F^2 + (2L'^2 + L^2)\cdot\|\mathbf{X}^t - \mathbf{X}^*\|_F^2\big)\cdot\|\mathbf{Z}^t\|_2^2 \\ + \|\mathbf{U}^{t\top}\mathbf{U}^t - \mathbf{V}^{t\top}\mathbf{V}^t\|_F^2\cdot\|\mathbf{Z}^t\|_2^2 + 12r\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2\cdot\|\mathbf{Z}^t\|_2^2. \end{aligned} \tag{B.5}$$

Note that for any $\mathbf{Z} \in \mathbb{B}(\sqrt{\sigma_r}/4)$, denote $\mathbf{R}$ as the optimal rotation with respect to $\mathbf{Z}$, we have $\|\mathbf{Z}\|_2 \leq \|\mathbf{Z}^*\|_2 + \|\mathbf{Z} - \mathbf{Z}^*\mathbf{R}\|_2 \leq 2\sqrt{\sigma_1}$. Thus, we have $\|\mathbf{Z}^t\|_2^2 \leq 4\sigma_1$. Denote $L_m = \max\{L, L'\}$, and we let $\eta = c_1/\sigma_1$, where $c_1 \leq \min\{1/32, \mu/(1152L_m^2)\}$. Therefore, combining (B.4) and (B.5), we have

$$\begin{aligned}-2\eta\langle\nabla\widetilde{F}_N(\mathbf{Z}), \mathbf{H}\rangle + \eta^2\mathbb{E}\|\mathbf{G}^t\|_F^2 \leq -\frac{\eta\mu'\sigma_r}{5}\|\mathbf{H}^t\|_F^2 + \frac{\eta(3L+1)}{4}\|\mathbf{H}^t\|_F^4 + 192\eta^2\sigma_1 L'^2\|\widetilde{\mathbf{X}}^t - \mathbf{X}^*\|_F^2 \\ + \eta\bigg(\frac{8r}{\mu} + \frac{r}{L}\bigg)\cdot\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2 + 48\eta^2\sigma_1 r\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2.\end{aligned}$$

Note that according to our assumption, $\|\mathbf{H}^t\|_F^2 \leq c_2^2\sigma_r$ with $c_2^2 \leq 2\mu'/(5(3L+1))$. Thus, according to Condition 3.6, we further have

$$-2\eta\langle\nabla\widetilde{F}_N(\mathbf{Z}), \mathbf{H}\rangle + \eta^2\mathbb{E}\|\mathbf{G}^t\|_F^2 \leq -\frac{\eta\mu'\sigma_r}{10}\|\mathbf{H}^t\|_F^2 + 192\eta^2\sigma_1 L'^2\|\widetilde{\mathbf{X}}^t - \mathbf{X}^*\|_F^2 + c_3\eta r\epsilon^2(N,\delta), \tag{B.6}$$

holds with probability at least $1-\delta$, where $c_3 \geq 48c_1 + 8/\mu + 1/L$. Therefore, plugging (B.6) into (B.3), with probability at least $1-\delta$, we have

$$\mathbb{E}\|\mathbf{H}^{t+1}\|_F^2 \leq \bigg(1 - \frac{\eta\mu'\sigma_r}{10}\bigg)\cdot\|\mathbf{H}^t\|_F^2 + 192\eta^2\sigma_1 L'^2\|\widetilde{\mathbf{X}}^t - \mathbf{X}^*\|_F^2 + c_3\eta r\epsilon^2(N,\delta). \tag{B.7}$$

Finally, for a fixed stage of $s$, we have $\widetilde{\mathbf{X}} = \widetilde{\mathbf{X}}^{s-1}$ accordingly. Denote $\widetilde{\mathbf{Z}}^s = [\widetilde{\mathbf{U}}^s; \widetilde{\mathbf{V}}^s]$, for any $s$. According to Algorithm 1, we randomly choose $\widetilde{\mathbf{Z}}^s$ after all of the updates are completed. Therefore, we first take summation of the previous inequality (B.7) over $t \in \{0, 1, \cdots, m-1\}$, and then take expectation with regard to all the history, we can get

$$\mathbb{E}\|\mathbf{H}^m\|_F^2 - \mathbb{E}\|\mathbf{H}^0\|_F^2 \leq -\frac{\eta\mu'\sigma_r}{10}\sum_{t=0}^{m-1}\mathbb{E}\|\mathbf{H}^t\|_F^2 + 192\eta^2\sigma_1 L'^2 m\mathbb{E}\|\widetilde{\mathbf{X}}^{s-1} - \mathbf{X}^*\|_F^2 + c_3\eta m r\epsilon^2(N,\delta).$$

For any $s$, we denote $\widetilde{\mathbf{R}}^s = \operatorname{argmin}_{\mathbf{R}\in\mathbb{Q}_r}\|\widetilde{\mathbf{Z}}^s - \mathbf{Z}^*\mathbf{R}\|_F$ and $\widetilde{\mathbf{H}}^s = \widetilde{\mathbf{Z}}^s - \mathbf{Z}^*\widetilde{\mathbf{R}}^s$. According to the choice of $\widetilde{\mathbf{Z}}^s$ in Algorithm 1, we have

$$\mathbb{E}\|\widetilde{\mathbf{H}}^s\|_F^2 = \frac{1}{m}\sum_{t=0}^{m-1}\mathbb{E}\|\mathbf{H}^t\|_F^2.$$

Note that according to Algorithm 1, we have $\mathbf{H}^0 = \widetilde{\mathbf{H}}^{s-1}$, thus we further obtain

$$\mathbb{E}\|\mathbf{H}^m\|_F^2 - \mathbb{E}\|\widetilde{\mathbf{H}}^{s-1}\|_F^2 \leq -\frac{\eta m\mu'\sigma_r}{10}\mathbb{E}\|\widetilde{\mathbf{H}}^s\|_F^2 + 192\eta^2\sigma_1 L'^2 m\mathbb{E}\|\widetilde{\mathbf{X}}^{s-1} - \mathbf{X}^*\|_F^2 + c_3\eta m r\epsilon^2(N,\delta).$$

Note that $\widetilde{\mathbf{Z}}^{s-1} \in \mathbb{B}(\sqrt{\sigma_r}/4)$, thus according to Lemma F.3, we have

$$\|\widetilde{\mathbf{X}}^{s-1} - \mathbf{X}^*\|_F^2 \leq 3\|\mathbf{Z}^*\|_2^2\cdot d^2(\widetilde{\mathbf{Z}}^{s-1}, \mathbf{Z}^*) = 6\sigma_1\|\widetilde{\mathbf{H}}^{s-1}\|_F^2.$$

where the first inequality follows from Lemma F.3 and the second inequality holds because $\|\mathbf{Z}^*\|_2 = \sqrt{2\sigma_r}$. Therefore, we obtain

$$\frac{\eta m\mu'\sigma_r}{10}\mathbb{E}\|\widetilde{\mathbf{H}}^s\|_F^2 \leq (1152\eta^2\sigma_1^2 L'^2 m + 1)\cdot\mathbb{E}\|\widetilde{\mathbf{H}}^{s-1}\|_F^2 + c_3\eta m r\epsilon^2(N,\delta),$$

holds with probability at least $1-\delta$, which gives us following contraction parameter

$$\rho = \frac{10\kappa}{\mu'}\bigg(\frac{1}{\eta m\sigma_1} + 1152\eta L'^2\bigg).$$

Note that $\eta = c_1/\sigma_1$, hence we can let $\rho \in (0,1)$ by choosing sufficiently small constant $c_1$ and sufficiently large number of iterations $m$. Therefore, with probability at least $1-\delta$, we can get

$$\mathbb{E}\|\widetilde{\mathbf{H}}^s\|_F^2 \leq \rho\mathbb{E}\|\widetilde{\mathbf{H}}^{s-1}\|_F^2 + \frac{10c_3}{\mu'\sigma_r}\cdot r\epsilon^2(N,\delta).$$

□

## C. Proofs of Specific Models

### C.1. Proof of Corollary 3.12

In order to prove the theoretical guarantees for matrix sensing, we only need to verify the restricted strong convexity and smoothness conditions for sample loss function $\mathcal{L}_N$, the restricted strong smoothness condition for each component function $\mathcal{L}_{\mathcal{S}_i}$ and the upper bound of $\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2$. In the following discussions, we use $\|\mathbf{A}\|_* = \sum_{i=1}^r \sigma_i(\mathbf{A})$ to denote the nuclear norm of matrix $\mathbf{A}$, where $r$ is the rank of $\mathbf{A}$.

First, we briefly introduce the definition of $\boldsymbol{\Sigma}$-ensemble which has been used in (Negahban & Wainwright, 2011) to verify the similar property of random sensing matrix $\mathbf{A}_i$ with dependent elements. Let $\mathrm{vec}(\mathbf{A}_i) \in \mathbb{R}^{d_1 d_2}$ be the vectorization of sensing matrix $\mathbf{A}_i$. If $\mathrm{vec}(\mathbf{A}_i) \sim N(0, \boldsymbol{\Sigma})$, we say that the sensing matrix $\mathbf{A}_i$ is sampled from $\boldsymbol{\Sigma}$-ensemble. In addition, we define $\pi^2(\boldsymbol{\Sigma}) = \sup_{\|\mathbf{u}\|_2=1, \|\mathbf{v}\|_2=1} \mathrm{Var}(\mathbf{u}^\top \mathbf{A}\mathbf{v})$. Specifically, in classical matrix sensing, we have $\boldsymbol{\Sigma} = \mathbf{I}$ and $\pi(\mathbf{I}) = 1$.

Recall that we have linear measurement operators $\mathcal{A}_N(\mathbf{X}) = (\langle\mathbf{A}_1, \mathbf{X}\rangle, \langle\mathbf{A}_2, \mathbf{X}\rangle, \ldots, \langle\mathbf{A}_N, \mathbf{X}\rangle)^\top$, and $\mathcal{A}_{\mathcal{S}_i}(\mathbf{X}) = (\langle\mathbf{A}_{i_1}, \mathbf{X}\rangle, \langle\mathbf{A}_{i_2}, \mathbf{X}\rangle, \ldots, \langle\mathbf{A}_{i_b}, \mathbf{X}\rangle)^\top$ for $i = 1, \ldots, n$. In order to prove the restricted strongly convex and smooth conditions of our objective function, we need to ultilize the following lemma, which has been used in (Agarwal et al., 2010; Negahban & Wainwright, 2011).

**Lemma C.1.** Suppose each sensing matrix $\mathbf{A}_i$ of the linear measurement operator $\mathcal{A}_M$ is sampled from $\boldsymbol{\Sigma}$-ensemble, where $M$ is the number of sensing matrices. Then there exist constants $C_0$ and $C_1$, such that the following inequalities hold for all $\boldsymbol{\Delta} \in \mathbb{R}^{d_1\times d_2}$ with probability at least $1 - \exp(-C_0 M)$

$$\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2^2}{M} \geq \frac{1}{2}\big\|\sqrt{\boldsymbol{\Sigma}}\mathrm{vec}(\boldsymbol{\Delta})\big\|_2^2 - C_1\pi^2(\boldsymbol{\Sigma})\frac{d}{M}\|\boldsymbol{\Delta}\|_*^2, \tag{C.1}$$

$$\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2^2}{M} \leq \frac{1}{2}\big\|\sqrt{\boldsymbol{\Sigma}}\mathrm{vec}(\boldsymbol{\Delta})\big\|_2^2 + C_1\pi^2(\boldsymbol{\Sigma})\frac{d}{M}\|\boldsymbol{\Delta}\|_*^2, \tag{C.2}$$

where $d = \max\{d_1, d_2\}$.

In order to bound the gradient of sample loss function $\nabla\mathcal{L}_M(\mathbf{X}^*)$ with respect to $M$ observations, we need to ultilize the following lemma, which has been used in (Negahban & Wainwright, 2011).

**Lemma C.2.** Suppose each sensing matrix $\mathbf{A}_i$ of the linear measurement operator $\mathcal{A}_M$ is sampled from $\boldsymbol{\Sigma}$-ensemble, where $M$ is the number of sensing matrices. Furthermore, suppose the noise vector $\boldsymbol{\epsilon}$ satisfies that $\|\boldsymbol{\epsilon}\|_2 \leq 2\nu\sqrt{M}$. Then we have the following inequality

$$\bigg\|\frac{1}{M}\sum_{i=1}^M \epsilon_i\mathbf{A}_i\bigg\|_2 \leq C\nu\sqrt{\frac{d}{M}},$$

holds with probability at least $1 - C_1\exp(-C_2 d)$, where $C, C_1$ and $C_2$ are universal constants.

Note that Lemma C.2 requires the noise vector $\boldsymbol{\epsilon}$ satisfies $\|\boldsymbol{\epsilon}\|_2 \leq 2\nu\sqrt{M}$ for some constant $\nu$. For any bounded noise vector, this condition obviously holds. And if the noise vector follows sub-Gaussian distribution with parameter $\nu$, it has been proved in (Vershynin, 2010) that this condition holds with high probability.

*Proof of Corollary 3.12.* First, we prove the restricted strong convexity condition for sample loss function $\mathcal{L}_N$. First, we have

$$\mathcal{L}_N(\mathbf{X}) = \frac{1}{2N}\sum_{i=1}^N \big(\langle\mathbf{A}_i, \mathbf{X}\rangle - y_i\big)^2 = \frac{1}{2N}\sum_{i=1}^N \big(\langle\mathbf{A}_i, \mathbf{X} - \mathbf{X}^*\rangle - \epsilon_i\big)^2.$$

Consider two rank-$r$ matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d_1\times d_2}$. Let $\boldsymbol{\Delta} = \mathbf{Y} - \mathbf{X}$, then we have the following equality

$$\begin{aligned}
&\mathcal{L}_N(\mathbf{Y}) - \mathcal{L}_N(\mathbf{X}) - \langle\nabla\mathcal{L}_N(\mathbf{X}), \boldsymbol{\Delta}\rangle \\
&= \frac{1}{2N}\sum_{i=1}^N \Big(\langle\mathbf{A}_i, \mathbf{Y} - \mathbf{X}^*\rangle^2 - \langle\mathbf{A}_i, \mathbf{X} - \mathbf{X}^*\rangle^2 - 2\langle\mathbf{A}_i, \mathbf{X} - \mathbf{X}^*\rangle\langle\mathbf{A}_i, \boldsymbol{\Delta}\rangle\Big) \\
&= \frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2^2}{2N}.
\end{aligned} \tag{C.3}$$

Therefore, according to (C.3), in order to establish the restricted strongly convex and smooth conditions for $\mathcal{L}_N$, we need to bound the term $\|\mathcal{A}(\mathbf{\Delta})\|_2^2/N$. According to (C.1) in Lemma C.1, we get

$$\frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{N} \geq \frac{1}{2}\big\|\sqrt{\mathbf{\Sigma}}\text{vec}(\mathbf{\Delta})\big\|_2^2 - C_1\pi^2(\mathbf{\Sigma})\frac{d}{N}\big\|\mathbf{\Delta}\big\|_*^2.$$

Furthermore, note that $\mathbf{\Delta} = \mathbf{Y} - \mathbf{X}$ has rank at most $2r$. Thus, we conclude that $\|\mathbf{\Delta}\|_* \leq \sqrt{2r}\|\mathbf{\Delta}\|_F$, which further implies

$$\frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{N} \geq \bigg\{\frac{\lambda_{\min}(\mathbf{\Sigma})}{2} - 2C_1 r\pi^2(\mathbf{\Sigma})\frac{d}{N}\bigg\}\big\|\mathbf{\Delta}\big\|_F^2.$$

Therefore, as long as $N \geq C_3\pi^2(\mathbf{\Sigma})rd/\lambda_{\min}(\mathbf{\Sigma})$ for some sufficiently large constant $C_3$, we get

$$\frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{N} \geq \frac{4\lambda_{\min}(\mathbf{\Sigma})}{9}\|\mathbf{\Delta}\|_F^2.$$

Since $\mathbf{\Sigma} = \mathbf{I}$ in matrix sensing, we obtain the restricted strongly convex parameter $\mu = 4/9$.

Second, we prove the restricted strong smoothness condition for $\mathcal{L}_N$ using (C.2) in Lemma C.1. Similar to the proof of the restricted strong convexity condition, we get

$$\frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{N} \leq \frac{5\lambda_{\max}(\mathbf{\Sigma})}{9}\|\mathbf{\Delta}\|_F^2,$$

as long as $N \geq C_3\pi^2(\mathbf{\Sigma})rd/\lambda_{\min}(\mathbf{\Sigma})$ for some sufficiently large constant $C_3$. Therefore, because $\mathbf{\Sigma} = \mathbf{I}$ in matrix sensing, we accordingly obtain $L = 5/9$.

Next, we prove each component loss function $\mathcal{L}_{\mathcal{S}_i}$ is restricted strongly smooth, for $i = 1, \ldots, n$. Recall that we have

$$\mathcal{L}_{\mathcal{S}_i}(\mathbf{X}) = \frac{1}{2b}\|\mathbf{y}_{\mathcal{S}_i} - \mathcal{A}_{\mathcal{S}_i}(\mathbf{U}\mathbf{V}^\top)\|_2^2 = \frac{1}{2b}\sum_{j\in\Omega_{\mathcal{S}_i}}\big(\langle\mathbf{A}_j, \mathbf{X} - \mathbf{X}^*\rangle - \epsilon_j\big)^2.$$

Thus, for each component loss function $\mathcal{L}_{\mathcal{S}_i}$, where $i = 1, \ldots, n$, we have

$$\mathcal{L}_{\mathcal{S}_i}(\mathbf{Y}) - \mathcal{L}_{\mathcal{S}_i}(\mathbf{X}) - \langle\nabla\mathcal{L}_{\mathcal{S}_i}(\mathbf{X}), \mathbf{\Delta}\rangle = \frac{\|\mathcal{A}_{\mathcal{S}_i}(\mathbf{\Delta})\|_2^2}{2b}. \tag{C.4}$$

Following the same steps as in the proof of restricted strong smoothness condition for $\mathcal{L}_N$, for each component function $\mathcal{L}_{\mathcal{S}_i}$, we get

$$\frac{\|\mathcal{A}_{\mathcal{S}_i}(\mathbf{\Delta})\|_2^2}{b} \leq C_5\lambda_{\max}(\mathbf{\Sigma})\|\mathbf{\Delta}\|_F^2,$$

if $b \geq C_4\pi^2(\mathbf{\Sigma})rd/\lambda_{\min}(\mathbf{\Sigma})$ for some sufficiently large constant $C_4$. Therefore we have $L' = C_5$ since $\mathbf{\Sigma} = \mathbf{I}$.

Finally, we bound the statistical error term $\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2$. According to the definition of $\mathcal{L}_N$, we have

$$\nabla\mathcal{L}_N(\mathbf{X}^*) = \frac{1}{N}\sum_{i=1}^N \epsilon_i\mathbf{A}_i.$$

Based on Lemma C.2, we have the following inequality holds with probability at least $1 - C_1'\exp(-C_2'd)$

$$\bigg\|\frac{1}{N}\sum_{i=1}^N \epsilon_i\mathbf{A}_i\bigg\|_2 \leq C\nu\sqrt{\frac{d}{N}},$$

which implies that

$$\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2 \leq C^2\nu^2\frac{d}{N}.$$

□

### C.2. Proof of Corollary 3.14

In order to prove the theoretical guarantees for matrix completion we only need to verify the restricted strong convexity and smoothness conditions for sample loss function $\mathcal{L}_\Omega$, the restricted strong smoothness condition for each component function $\mathcal{L}_{\Omega_{\mathcal{S}_i}}$ and the upper bound of $\|\nabla\mathcal{L}_\Omega(\mathbf{X}^*)\|_2$.

To establish the restricted strong convexity and smoothness conditions for $\mathcal{L}_\Omega$, and the restricted strong smoothness condition for $\mathcal{L}_{\Omega_{\mathcal{S}_i}}$ we need to ultilize the following lemma, which used in (Negahban & Wainwright, 2012).

**Lemma C.3.** Suppose the number of observations $M$ satisfying $M > c_1 rd\log d$. Furthermore, if for all $\boldsymbol{\Delta}\in\mathbb{R}^{d_1\times d_2}$, we have

$$\sqrt{\frac{d_1d_2}{r}}\frac{\|\boldsymbol{\Delta}\|_{\infty,\infty}}{\|\boldsymbol{\Delta}\|_F}\cdot\frac{\|\boldsymbol{\Delta}\|_*}{\|\boldsymbol{\Delta}\|_F}\le\frac{1}{c_2}\sqrt{n/(d\log d)}, \tag{C.5}$$

then the following inequality holds with probability at least $1-c_3/d$

$$\left|\frac{\|\mathcal{A}(\boldsymbol{\Delta})\|_2}{\sqrt{n}}-\frac{\|\boldsymbol{\Delta}\|_F}{\sqrt{d_1d_2}}\right|\le c_4\frac{\|\boldsymbol{\Delta}\|_F}{\sqrt{d_1d_2}}\left(1+\frac{c_5\sqrt{d_1d_2}\|\boldsymbol{\Delta}\|_{\infty,\infty}}{\sqrt{n}\|\boldsymbol{\Delta}\|_F}\right).$$

where $c_1,c_2,c_3,c_4,c_5$ are universal constants .

Moreover, in order to upper bound of the gradient $\nabla\mathcal{L}_\Omega$ at $\mathbf{X}^*$, we need to use the following lemma.

**Lemma C.4.** (Negahban & Wainwright, 2012) Suppose $\mathbf{A}_i$ is uniformly distributed over $\mathcal{X}$. In addition, each noise $\epsilon_i$ follows i.i.d. zero mean distribution with variance $\nu^2$. Then the following inequality holds with probability at least $1-c_1/d$

$$\left\|\frac{1}{M}\sum_{i=1}^{M}\epsilon_i\mathbf{A}_i\right\|_2\le c_2\nu\sqrt{\frac{d\log d}{d_1d_2M}},$$

where $M$ is the number of observations, and $c_1,c_2$ are universal constants.

*Proof of Corollary 3.14.* Let $|\Omega|=N$, $|\Omega_{\mathcal{S}_i}|=b$. For any $(j,k)\in\Omega$, we denote $\mathbf{A}_{jk}=\mathbf{e}_j\mathbf{e}_k^\top$, where $\mathbf{e}_j,\mathbf{e}_k$ are unit vectors with dimensionality $d_1$ and $d_2$ respectively. Similarly, for any $(j,k)\in\Omega_{\mathcal{S}_i}$, we let $\mathbf{A}_{jk}^i=\mathbf{e}_j^i\mathbf{e}_k^{i\top}$. Thus, we can rewrite the sample loss function as follows (here for simplicity, we use $\mathcal{L}_N$ and $\mathcal{L}_{\mathcal{S}_i}$ to denote $\mathcal{L}_\Omega$ and $\mathcal{L}_{\Omega_{\mathcal{S}_i}}$ respectively)

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

where $p=N/(d_1d_2)$. In addition, we can rewrite each component loss function as follows

$$\mathcal{L}_{\mathcal{S}_i}(\mathbf{X}):=\frac{1}{2p'}\sum_{(j,k)\in\Omega_{\mathcal{S}_i}}\big(\langle\mathbf{A}_{jk}^i,\mathbf{X}\rangle-Y_{jk}\big)^2,$$

where $p'=b/(d_1d_2)$. For simplicity we let $\mathcal{A}$ and $\mathcal{A}_{\mathcal{S}_i}$ be the corresponding transformation operator with respect to $\mathcal{L}_N$ and $\mathcal{L}_{\mathcal{S}_i}$, respectively. First, we prove the restricted strong convexity and smoothness conditions for $\mathcal{L}_N$. Consider any two rank-$r$ matrices $\mathbf{X},\mathbf{Y}$, which satisfy the incoherence condition. In the following discussion, denote $\boldsymbol{\Delta}=\mathbf{Y}-\mathbf{X}$.

**Case** 1**:** If condition (C.5) is violated. Then we obtain

$$\|\boldsymbol{\Delta}\|_F^2\le C_0\big(\sqrt{d_1d_2}\|\boldsymbol{\Delta}\|_\infty\big)\cdot\|\boldsymbol{\Delta}\|_*\sqrt{\frac{d\log d}{rN}}\le 2C_0\alpha\sqrt{d_1d_2}\|\boldsymbol{\Delta}\|_*\sqrt{\frac{d\log d}{rN}},$$

where $\alpha=\beta r\sigma_1/\sqrt{d_1d_2}$. Furthermore, we get

$$\|\boldsymbol{\Delta}\|_F^2\le 2C_0\sqrt{2}\alpha\sqrt{d_1d_2}\|\boldsymbol{\Delta}\|_F\sqrt{\frac{d\log d}{N}},$$

where the inequality holds because $\mathrm{rank}(\mathbf{\Delta}) \leq 2r$, which implies the following bound

$$\frac{1}{d_1 d_2}\|\mathbf{\Delta}\|_F^2 \leq C\alpha^2 \frac{d \log d}{N}. \tag{C.6}$$

**Case** 2**:** If condition (C.5) is satisfied. We first establish the restricted strongly convex condition for $\mathcal{L}_N$. In particular, we have

$$\begin{aligned}
&\mathcal{L}_N(\mathbf{Y}) - \mathcal{L}_N(\mathbf{X}) - \langle \nabla \mathcal{L}_N(\mathbf{X}), \mathbf{\Delta} \rangle \\
&= \frac{1}{2p} \sum_{(j,k)\in\Omega} \big( \langle \mathbf{A}_{jk}, \mathbf{Y} - \mathbf{X}^* \rangle^2 + \langle \mathbf{A}_{jk}, \mathbf{X} - \mathbf{X}^* \rangle^2 - 2\langle \mathbf{A}_{jk}, \mathbf{X} - \mathbf{X}^* \rangle \langle \mathbf{A}_{jk}, \mathbf{\Delta} \rangle \big) \\
&= \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{2p},
\end{aligned} \tag{C.7}$$

Thus, as long as $c_5\sqrt{d_1 d_2}\|\mathbf{\Delta}\|_{\infty,\infty}/\|\mathbf{\Delta}\|_F \geq \sqrt{N}$, by the definition of spikiness ration $\alpha_{\mathrm{sp}}(\mathbf{\Delta})$, we get

$$\frac{1}{d_1 d_2}\|\mathbf{\Delta}\|_F^2 \leq c'\alpha^2 \frac{1}{N}. \tag{C.8}$$

If $c_5\sqrt{d_1 d_2}\|\mathbf{\Delta}\|_{\infty,\infty}/\|\mathbf{\Delta}\|_F \leq \sqrt{N}$, according to Lemma C.3, we obtain

$$\frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{p} \geq \frac{8}{9}\|\mathbf{\Delta}\|_F^2,$$

which implies the restricted strong convexity parameter $\mu = 8/9$.

Next, for sample loss function $\mathcal{L}_N$, we establish the restricted strong smoothness condition by similar proof. According to (C.7) and Lemma C.3, as long as $c_5\sqrt{d_1 d_2}\|\mathbf{\Delta}\|_{\infty,\infty}/\|\mathbf{\Delta}\|_F \leq \sqrt{N}$, we have

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

which gives us the restricted strong smoothness parameter $L = 10/9$.

Similarly, we show the restricted strong smoothness condition for each component loss function $\mathcal{L}_{\mathcal{S}_i}$, where $i = 1, \ldots, n$. Since we have

$$\mathcal{L}_{\mathcal{S}_i}(\mathbf{Y}) - \mathcal{L}_{\mathcal{S}_i}(\mathbf{X}) - \langle \nabla \mathcal{L}_{\mathcal{S}_i}(\mathbf{X}), \mathbf{\Delta} \rangle = \frac{\|\mathcal{A}_{\mathcal{S}_i}(\mathbf{\Delta})\|_2^2}{2p'},$$

thus according to Lemma C.3, as long as $c_5\sqrt{d_1 d_2}\|\mathbf{\Delta}\|_{\infty,\infty}/\|\mathbf{\Delta}\|_F/\sqrt{b} \leq c_6$, we have

$$\frac{\|\mathcal{A}_{\mathcal{S}_i}(\mathbf{\Delta})\|_2^2}{p'} \leq c_7\|\mathbf{\Delta}\|_F^2,$$

which implies that $L' = c_7$. Otherwise, it is sufficient to ensure $\alpha = O(1/\sqrt{n})$.

Finally, for the statistical error term $\|\nabla \mathcal{L}_N(\mathbf{X}^*)\|_2^2$, according to the definition of $\mathcal{L}_N$, we have

$$\nabla \mathcal{L}_N(\mathbf{X}^*) = \frac{1}{p} \sum_{(j,k)\in\Omega} \epsilon_{jk}\mathbf{A}_{jk}.$$

Remember that each elements of the noise matrix follows i.i.d. Gaussian distribution with variance $\nu^2/d_1 d_2$. Therefore, according to Lemma C.4, we obtain

$$\bigg\| \frac{1}{p} \sum_{(j,k)\in\Omega} \epsilon_{jk}\mathbf{A}_{jk} \bigg\|_2 \leq C\nu\sqrt{\frac{d \log d}{N}},$$

holds with probability at least $1 - C'/d$, which implies that

$$\|\nabla \mathcal{L}_N(\mathbf{X}^*)\|_2^2 \leq C^2\nu^2 \frac{d \log d}{N}, \tag{C.9}$$

holds with probability at least $1 - C'/d$. Combining error bounds (C.6), (C.8) and (C.9), we conclude the following upper bound in Condition 3.6 as $C_1 \max\{r\beta^2\sigma_1, \nu^2\} d \log d/|\Omega|$. □

### C.3. Proof of Corollary A.1

In order to prove the theoretical guarantees for one-bit matrix completion, we only need to verify the restricted strong convexity and smoothness conditions for the sample loss function $\mathcal{L}_\Omega$, the restricted strong smoothness condition for each component function $\mathcal{L}_{\Omega_{\mathcal{S}_i}}$, and the upper bound of $\|\nabla \mathcal{L}_\Omega(\mathbf{X}^*)\|_2$. Note that we have the estimator $\mathbf{X}$ satisfies incoherence condition such that $\|\mathbf{X}\|_{\infty,\infty} \leq r\beta\sigma_1/\sqrt{d_1 d_2}$. Thus we should consider the twice differentiable function $f(x) = g(x/\tau)$, where $\tau$ is a scale parameter with order $O(\nu/\sqrt{d_1 d_2})$.

*Proof of Corollary A.1.* Let $|\Omega| = N$, $|\Omega_{\mathcal{S}_i}| = b$. For any $(j,k) \in \Omega$, we denote $\mathbf{A}_{jk} = \mathbf{e}_j \mathbf{e}_k^\top$, where $\mathbf{e}_i, \mathbf{e}_j$ are unit vectors with $d_1$ and $d_2$ dimensions. Similarly, for any $(j,k) \in \Omega_{\mathcal{S}_i}$, we let $\mathbf{A}_{jk}^i = \mathbf{e}_j^i \mathbf{e}_k^{i\top}$. Note that for simplicity we use $\mathcal{A}$ and $\mathcal{A}_{\mathcal{S}_i}$ to denote the corresponding transformation operator with respect to $\mathcal{L}_N$ and $\mathcal{L}_{\mathcal{S}_i}$, respectively. We can rewrite the sample loss function as follows (here for simplicity, we use $\mathcal{L}_N$ and $\mathcal{L}_{\mathcal{S}_i}$ to denote $\mathcal{L}_\Omega$ and $\mathcal{L}_{\Omega_{\mathcal{S}_i}}$ respectively)

$$\mathcal{L}_N(\mathbf{X}) := -\frac{1}{N} \sum_{(j,k)\in\Omega} \Big\{ \mathbb{1}\{Y_{jk} = 1\} \log\big(g(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)\big) + \mathbb{1}\{Y_{jk} = -1\} \log\big(1 - g(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)\big) \Big\},$$

Therefore, we have each component loss function $\mathcal{L}_{\mathcal{S}_i}$ as

$$\mathcal{L}_{\mathcal{S}_i}(\mathbf{X}) := -\frac{1}{b} \sum_{(j,k)\in\Omega_{\mathcal{S}_i}} \Big\{ \mathbb{1}\{Y_{jk} = 1\} \log\big(g(\langle \mathbf{A}_{jk}^i, \mathbf{X}\rangle/\tau)\big) + \mathbb{1}\{Y_{jk} = -1\} \log\big(1 - g(\langle \mathbf{A}_{jk}^i, \mathbf{X}\rangle/\tau)\big) \Big\},$$

Therefore, we get

$$\nabla \mathcal{L}_N(\mathbf{X}) = \frac{\sqrt{d_1 d_2}}{N\nu} \sum_{(j,k)\in\Omega} \bigg( -\frac{g'(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)}{g(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)} \mathbb{1}\{Y_{jk} = 1\} + \frac{g'(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)}{1 - g(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)} \mathbb{1}\{Y_{jk} = -1\} \bigg) \mathbf{A}_{jk}. \tag{C.10}$$

Furthermore, we obtain

$$\nabla^2 \mathcal{L}_N(\mathbf{X}) = \frac{1}{p\nu^2} \sum_{(j,k)\in\Omega} B_{jk}(\mathbf{X}) \text{vec}(\mathbf{A}_{jk}) \text{vec}(\mathbf{A}_{jk})^\top, \tag{C.11}$$

where we have

$$\begin{aligned} B_{jk}(\mathbf{X}) = \bigg[ &\bigg( \frac{g'^2(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)}{g^2(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)} - \frac{f''(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle)}{g(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)} \bigg) \mathbb{1}\{Y_{jk} = 1\} \\ &+ \bigg( \frac{f''(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle)}{1 - g(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)} - \frac{g'^2(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)}{(1 - g(\langle \mathbf{A}_{jk}, \mathbf{X}\rangle/\tau)^2} \bigg) \mathbb{1}\{Y_{jk} = -1\} \bigg]. \end{aligned}$$

First, we establish the strong convexity and smoothness conditions for $\mathcal{L}_N$. For any $\mathbf{X}, \mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$, let $\mathbf{W} = \mathbf{M} + a(\mathbf{X} - \mathbf{M})$ for $a \in [0,1]$, $\mathbf{x} = \text{vec}(\mathbf{X})$ and $\mathbf{m} = \text{vec}(\mathbf{M})$. According to the mean value theorem, we get

$$\mathcal{L}_N(\mathbf{X}) = \mathcal{L}_N(\mathbf{M}) + \langle \nabla \mathcal{L}_N(\mathbf{M}), \mathbf{X} - \mathbf{M} \rangle + \frac{1}{2} (\mathbf{x} - \mathbf{m})^\top \nabla^2 \mathcal{L}_N(\mathbf{W}) (\mathbf{x} - \mathbf{m}),$$

Moreover, according to (C.11), we further obtain

$$\begin{aligned} (\mathbf{x} - \mathbf{m})^\top \nabla^2 \mathcal{L}_N(\mathbf{W}) (\mathbf{x} - \mathbf{m}) &= \frac{1}{p\nu^2} \sum_{(j,k)\in\Omega} B_{jk}(\mathbf{W}) \langle \text{vec}(\mathbf{A}_{jk})^\top (\mathbf{x} - \mathbf{m}), \text{vec}(\mathbf{A}_{jk})^\top (\mathbf{x} - \mathbf{m}) \rangle \\ &= \frac{1}{p\nu^2} \sum_{(j,k)\in\Omega} B_{jk}(\mathbf{W}) \langle \mathbf{A}_{jk}, \mathbf{\Delta} \rangle^2, \end{aligned}$$

where $\mathbf{\Delta} = \mathbf{X} - \mathbf{M}$. Thus, according to the definition of $\mu_\alpha$ in (A.3), we obtain

$$\frac{1}{p\nu^2} \sum_{(j,k)\in\Omega} B_{jk}(\mathbf{W}) \langle \mathbf{A}_{jk}, \mathbf{\Delta} \rangle^2 \geq \mu_\alpha \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{p\nu^2},$$

Therefore, following the same steps in the proof of matrix completion, we have

$$\mu_\alpha \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{p\nu^2} \geq C_1\mu_\alpha\|\mathbf{\Delta}\|_F^2,$$

which implies

$$\mathcal{L}_N(\mathbf{X}) \geq \mathcal{L}_N(\mathbf{M}) + \langle \nabla\mathcal{L}_N(\mathbf{M}), \mathbf{X}-\mathbf{M}\rangle + \frac{1}{2}C_1\mu_\alpha\|\mathbf{\Delta}\|_F^2.$$

Therefore it gives us the restricted strong convexity parameter $\mu = C_1\mu_\alpha$. On the other hand, according to the definition of $L_\alpha$ in (A.4), we obtain

$$\frac{1}{p\nu^2}\sum_{(j,k)\in\Omega} B_{jk}(\mathbf{W})\langle \mathbf{A}_{jk}, \mathbf{\Delta}\rangle^2 \leq L_\alpha \frac{\|\mathcal{A}(\mathbf{\Delta})\|_2^2}{p\nu^2},$$

Therefore, following the same steps in the proof of the restricted strong smoothness condition in matrix completion, we get

$$\mathcal{L}_N(\mathbf{X}) \leq \mathcal{L}_N(\mathbf{M}) + \langle \nabla\mathcal{L}_N(\mathbf{M}), \mathbf{X}-\mathbf{M}\rangle + \frac{1}{2}C_2L_\alpha\|\mathbf{\Delta}\|_F^2,$$

which implies the restricted strong smoothness parameter $L = C_2L_\alpha$. By the similar procedure, for each component function, we can derive that

$$\mathcal{L}_{\mathcal{S}_i}(\mathbf{X}) \leq \mathcal{L}_{\mathcal{S}_i}(\mathbf{M}) + \langle \nabla\mathcal{L}_{\mathcal{S}_i}(\mathbf{M}), \mathbf{X}-\mathbf{M}\rangle + \frac{1}{2}C_3L_\alpha\|\mathbf{\Delta}\|_F^2.$$

Finally, for the term $\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2$, according to Lemma C.4, we have

$$\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2 \leq C\gamma_\alpha\sqrt{\frac{d\log d}{|\Omega|}}, \tag{C.12}$$

where $\gamma_\alpha$ is defined in (A.5). In addition, we also have the following bounds, which have been shown in proofs of matrix completion when condition (C.5) is not satisfied

$$\frac{1}{d_1d_2}\|\mathbf{\Delta}\|_F^2 \leq \max\left\{C\alpha^2\frac{1}{|\Omega|}, C'\alpha^2\frac{rd\log d}{|\Omega|}\right\}. \tag{C.13}$$

Therefore, combining error bounds (C.13) and (C.12), we have the following bound in Condition 3.6 $C\max\{r\beta^2\sigma_1, \gamma_\alpha^2\}d\log d/|\Omega|$. □

## D. Proofs of Technical Lemmas

In this section, we present the proofs of several technical lemmas. Before proceeding to the theoretical proof, we first introduce the following notations and definitions, which are essential for proving the following lemmas. For any $\mathbf{Z}\in\mathbb{R}^{(d_1+d_2)\times r}$, we denote $\mathbf{Z} = [\mathbf{U};\mathbf{V}]$, where $\mathbf{U}\in\mathbb{R}^{d_1\times r}$ and $\mathbf{V}\in\mathbb{R}^{d_2\times r}$. Denote $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. Let $\mathbf{R} = \text{argmin}_{\widetilde{\mathbf{R}}\in\mathbb{Q}_r}\|\mathbf{Z}-\mathbf{Z}^*\widetilde{\mathbf{R}}\|_F$ be the optimal rotation with respect to $\mathbf{Z}$, and $\mathbf{H} = \mathbf{Z}-\mathbf{Z}^*\mathbf{R} = [\mathbf{H}_U;\mathbf{H}_V]$, where $\mathbf{H}_U\in\mathbb{R}^{d_1\times r}$, $\mathbf{H}_V\in\mathbb{R}^{d_2\times r}$.

Moreover, let $\overline{\mathbf{U}}_1, \overline{\mathbf{U}}_2, \overline{\mathbf{U}}_3$ be the left singular matrices of $\mathbf{X}, \mathbf{U}, \mathbf{H}_U$, respectively. Define $\widetilde{\mathbf{U}}$ as the matrix spanned by the column of $\overline{\mathbf{U}}_1, \overline{\mathbf{U}}_2$ and $\overline{\mathbf{U}}_3$ such that

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

Note that for the above subspace, each column vector of $\widetilde{\mathbf{U}}$ is a basis vector. In addition, we define the sum of two subspaces $\mathbf{U}_1, \mathbf{U}_2$ as $\mathbf{U}_1+\mathbf{U}_2 = \{\mathbf{u}_1+\mathbf{u}_2 \mid \mathbf{u}_1\in\mathbf{U}_1, \mathbf{u}_2\in\mathbf{U}_2\}$. Obviously, $\widetilde{\mathbf{U}}$ is an orthonormal matrix with at most $3r$ columns.

Similarly, let $\overline{\mathbf{V}}_1, \overline{\mathbf{V}}_2, \overline{\mathbf{V}}_3$ be the right singular matrices of $\mathbf{X}, \mathbf{V}, \mathbf{H}_V$, respectively. Define $\widetilde{\mathbf{V}}$ as the matrix spanned by the column of $\overline{\mathbf{V}}_1, \overline{\mathbf{V}}_2$ and $\overline{\mathbf{V}}_3$ such that

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

where $\widetilde{\mathbf{V}}$ has at most $3r$ columns.

### D.1. Proof of Lemma 3.5

*Proof.* Define reference function $f_{\mathbf{Y}} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ such that

$$f_{\mathbf{Y}}(\widetilde{\mathbf{X}}) = \mathcal{L}_N(\widetilde{\mathbf{X}}) - \mathcal{L}_N(\mathbf{Y}) - \langle \nabla \mathcal{L}_N(\mathbf{Y}), \widetilde{\mathbf{X}} - \mathbf{Y} \rangle. \tag{D.3}$$

Since $\mathcal{L}_N$ satisfies restricted strong convexity Condition 3.3, we have $f_{\mathbf{Y}}(\mathbf{X}) \geq 0$, for any matrix $\widetilde{\mathbf{X}}$ with rank at most $3r$. Obviously, $f_{\mathbf{Y}}(\mathbf{Y}) = 0$. Recall the SVD of $\mathbf{X}$ is $\mathbf{X} = \overline{\mathbf{U}}_1 \boldsymbol{\Sigma}_1 \overline{\mathbf{V}}_1^\top$. Since $\operatorname{col}(\overline{\mathbf{U}}_1) \subseteq \operatorname{col}(\widetilde{\mathbf{U}})$ and $\overline{\mathbf{U}}_1^\top \overline{\mathbf{U}}_1 = \mathbf{I}_{r_1}$, we have $\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top \mathbf{X} = \mathbf{X}$. Thus we have

$$\begin{aligned}
0 = f_{\mathbf{Y}}(\mathbf{Y}) &\leq \min_{\eta} f_{\mathbf{Y}}\big(\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top [\mathbf{X} - \eta \nabla f_{\mathbf{Y}}(\mathbf{X})]\big) \\
&= \min_{\eta} f_{\mathbf{Y}}\big(\mathbf{X} - \eta \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X})\big) \\
&\leq \min_{\eta} \Big\{ \mathcal{L}_N\big(\mathbf{X} - \eta \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X})\big) - \mathcal{L}_N(\mathbf{Y}) - \langle \nabla \mathcal{L}_N(\mathbf{Y}), \mathbf{X} - \eta \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X}) - \mathbf{Y} \rangle \Big\},
\end{aligned} \tag{D.4}$$

where the first inequality holds because $\operatorname{rank}(\mathbf{AB}) \leq \min\{\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\}$ and $\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top$ has rank at most $3r$, and the last inequality follows from (D.3). Since $\mathcal{L}_N$ satisfies restricted strong smoothness Condition 3.4, we have

$$\mathcal{L}_N\big(\mathbf{X} - \eta \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X})\big) \leq \mathcal{L}_N(\mathbf{X}) + \langle \nabla \mathcal{L}_N(\mathbf{X}), -\eta \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X}) \rangle + \frac{L}{2} \|\eta \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X})\|_F^2. \tag{D.5}$$

Thus, plugging (D.5) into (D.4), we have

$$\begin{aligned}
f_{\mathbf{Y}}(\mathbf{Y}) &\leq \min_{\eta} \Big\{ f_{\mathbf{Y}}(\mathbf{X}) - \eta \langle \nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{Y}), \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X}) \rangle + \frac{\eta^2 L}{2} \|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X})\|_F^2 \Big\} \\
&= f_{\mathbf{Y}}(\mathbf{X}) + \min_{\eta} \Big\{ -\eta \|\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X})\|_F^2 + \frac{\eta^2 L}{2} \|\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X})\|_F^2 \Big\} \\
&= f_{\mathbf{Y}}(\mathbf{X}) - \frac{1}{2L} \|\widetilde{\mathbf{U}}^\top \nabla f_{\mathbf{Y}}(\mathbf{X})\|_F^2,
\end{aligned} \tag{D.6}$$

where the first equality follows from (D.3), the second inequality holds because $\widetilde{\mathbf{U}}^\top \widetilde{\mathbf{U}} = \mathbf{I}_{r_1}$ and $\nabla f_{\mathbf{Y}}(\mathbf{X}) = \nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{Y})$, and the last equality holds because the minimizer is $\eta = 1/L$. Thus, plugging the definition of $f_{\mathbf{Y}}$ into (D.6), we obtain

$$\mathcal{L}_N(\mathbf{X}) - \mathcal{L}_N(\mathbf{Y}) - \langle \nabla \mathcal{L}_N(\mathbf{Y}), \mathbf{X} - \mathbf{Y} \rangle - \frac{1}{2L} \|\widetilde{\mathbf{U}}^\top (\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{Y}))\|_F^2 \geq 0. \tag{D.7}$$

Since $\widetilde{\mathbf{V}}$ is orthonormal matrix with at most $3r$ columns and $\mathbf{X}\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top = \mathbf{X}$, following the same techniques, we obtain

$$\mathcal{L}_N(\mathbf{X}) - \mathcal{L}_N(\mathbf{Y}) - \langle \nabla \mathcal{L}_N(\mathbf{Y}), \mathbf{X} - \mathbf{Y} \rangle - \frac{1}{2L} \|(\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{Y}))\widetilde{\mathbf{V}}\|_F^2 \geq 0. \tag{D.8}$$

Therefore, combining (D.7) and (D.8), we complete the proof. □

### D.2. Proof of Lemma B.1

In order to prove the local curvature condition, we need to make use of the following lemmas. In Lemma D.1, we denote $\widetilde{\mathbf{Z}} \in \mathbb{R}^{(d_1+d_2)\times r}$ as $\widetilde{\mathbf{Z}} = [\mathbf{U}; -\mathbf{V}]$. Recall $\mathbf{Z} = [\mathbf{U}; \mathbf{V}]$, then we have $\|\mathbf{U}^\top \mathbf{U} - \mathbf{V}^\top \mathbf{V}\|_F^2 = \|\widetilde{\mathbf{Z}}^\top \mathbf{Z}\|_F^2$, and $\nabla_{\mathbf{Z}}(\|\mathbf{U}^\top \mathbf{U} - \mathbf{V}^\top \mathbf{V}\|_F^2) = 4\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top \mathbf{Z}$. We refer Wang et al. (2016) to readers for a detailed proof of Lemma D.1. Lemma D.2, proved in Section E.1, is a variation of the regularity condition of the sample loss function $\mathcal{L}_N$ (Tu et al., 2015), which is essential to derive the linear convergence rate in our main theorem.

**Lemma D.1.** (Wang et al., 2016) Let $\mathbf{Z}, \mathbf{Z}^* \in \mathbb{R}^{(d_1+d_2)\times r}$. Denote the optimal rotation with respect to $\mathbf{Z}$ as $\mathbf{R} = \operatorname{argmin}_{\widetilde{\mathbf{R}} \in \mathbb{Q}_r} \|\mathbf{Z} - \mathbf{Z}^* \widetilde{\mathbf{R}}\|_F$, and $\mathbf{H} = \mathbf{Z} - \mathbf{Z}^* \mathbf{R}$. Consider the gradient of the regularization term $\|\widetilde{\mathbf{Z}}^\top \mathbf{Z}\|_F^2$, we have

$$\langle \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top \mathbf{Z}, \mathbf{H} \rangle \geq \frac{1}{2} \|\widetilde{\mathbf{Z}}^\top \mathbf{Z}\|_F^2 - \frac{1}{2} \|\widetilde{\mathbf{Z}}^\top \mathbf{Z}\|_F \cdot \|\mathbf{H}\|_F^2,$$

where $\widetilde{\mathbf{Z}} = [\mathbf{U}; -\mathbf{V}]$.

**Lemma D.2.** Suppose the sample loss function $\mathcal{L}_N$ satisfies Conditions 3.3 and 3.4. For any rank-$r$ matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d_1 \times d_2}$, let the singular value decomposition of $\mathbf{X}$ be $\overline{\mathbf{U}}_1 \boldsymbol{\Sigma}_1 \overline{\mathbf{V}}_1^\top$, then we have

$$\begin{aligned}\langle \nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{Y}), \mathbf{X} - \mathbf{Y} \rangle \geq \frac{1}{4L} \|\widetilde{\mathbf{U}}^\top (\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{Y}))\|_F^2 \\ + \frac{1}{4L} \|(\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{Y}))\widetilde{\mathbf{V}}\|_F^2 + \frac{\mu}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2,\end{aligned}$$

where $\widetilde{\mathbf{U}} \in \mathbb{R}^{d_1 \times r_1}$ is an orthonormal matrix with $r_1 \leq 3r$ satisfying $\text{col}(\overline{\mathbf{U}}_1) \subseteq \text{col}(\widetilde{\mathbf{U}})$, and $\widetilde{\mathbf{V}} \in \mathbb{R}^{d_2 \times r_2}$ is an orthonormal matrix with $r_2 \leq 3r$ satisfying $\text{col}(\overline{\mathbf{V}}_1) \subseteq \text{col}(\widetilde{\mathbf{V}})$.

Now, we are ready to prove Lemma B.1.

*Proof of Lemma B.1.* According to (B.2), we have

$$\langle \nabla \widetilde{F}_N(\mathbf{Z}), \mathbf{H} \rangle = \underbrace{\langle \nabla_{\mathbf{U}} \mathcal{L}_N(\mathbf{U}\mathbf{V}^\top), \mathbf{H}_U \rangle + \langle \nabla_{\mathbf{V}} \mathcal{L}_N(\mathbf{U}\mathbf{V}^\top), \mathbf{H}_V \rangle}_{I_1} + \frac{1}{2} \underbrace{\langle \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top \mathbf{Z}, \mathbf{H} \rangle}_{I_2}, \tag{D.9}$$

where $\widetilde{\mathbf{Z}} = [\mathbf{U}; -\mathbf{V}]$. Recall that $\mathbf{X}^* = \mathbf{U}^* \mathbf{V}^{*\top}$, and $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. Note that $\nabla_{\mathbf{U}} \mathcal{L}_N(\mathbf{U}\mathbf{V}^\top) = \nabla \mathcal{L}_N(\mathbf{X})\mathbf{V}$, and $\nabla_{\mathbf{V}} \mathcal{L}_N(\mathbf{U}\mathbf{V}^\top) = \nabla \mathcal{L}_N(\mathbf{X})^\top \mathbf{U}$. Thus, for the term $I_1$ in (D.9), we have

$$\begin{aligned} I_1 &= \langle \nabla \mathcal{L}_N(\mathbf{X}), \mathbf{U}\mathbf{V}^\top - \mathbf{U}^*\mathbf{V}^{*\top} + \mathbf{H}_U \mathbf{H}_V^\top \rangle \\ &= \underbrace{\langle \nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*), \mathbf{X} - \mathbf{X}^* + \mathbf{H}_U \mathbf{H}_V^\top \rangle}_{I_{11}} + \underbrace{\langle \nabla \mathcal{L}_N(\mathbf{X}^*), \mathbf{X} - \mathbf{X}^* + \mathbf{H}_U \mathbf{H}_V^\top \rangle}_{I_{12}}. \end{aligned} \tag{D.10}$$

First, we consider the term $I_{11}$ in (D.10). Recall the definition of $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ in (D.1) and (D.2), respectively. According to Lemma D.2, we have

$$\begin{aligned}\langle \nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*), \mathbf{X} - \mathbf{X}^* \rangle \geq \frac{1}{4L} \|\widetilde{\mathbf{U}}^\top (\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*))\|_F^2 \\ + \frac{1}{4L} \|(\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*))\widetilde{\mathbf{V}}\|_F^2 + \frac{\mu}{2}\|\mathbf{X} - \mathbf{X}^*\|_F^2.\end{aligned} \tag{D.11}$$

Second, for the remaining term in $I_{11}$, we have

$$\begin{aligned} \big|\langle \nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*), \mathbf{H}_U \mathbf{H}_V^\top \rangle\big| &= \big|\langle \widetilde{\mathbf{U}}^\top (\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*)), \widetilde{\mathbf{U}}^\top \mathbf{H}_U \mathbf{H}_V^\top \rangle\big| \\ &\leq \|\widetilde{\mathbf{U}}^\top (\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*))\|_F \cdot \|\widetilde{\mathbf{U}}^\top\|_2 \cdot \|\mathbf{H}_U \mathbf{H}_V^\top\|_F \\ &\leq \frac{1}{2}\|\widetilde{\mathbf{U}}^\top (\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*))\|_F \cdot \|\mathbf{H}\|_F^2, \end{aligned} \tag{D.12}$$

where the equality holds because $\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top \mathbf{H}_U = \mathbf{H}_U$, the first inequality holds because $|\langle \mathbf{A}, \mathbf{B} \rangle| \leq \|\mathbf{A}\|_F \cdot \|\mathbf{B}\|_F$ and $\|\mathbf{A}\mathbf{B}\|_F \leq \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_F$, and the second inequality holds because $2\|\mathbf{A}\mathbf{B}\|_F \leq \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2$ and $\widetilde{\mathbf{U}}$ is orthonormal. Similarly, we have

$$\big|\langle \nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*), \mathbf{H}_U \mathbf{H}_V^\top \rangle\big| \leq \frac{1}{2}\|(\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*))\widetilde{\mathbf{V}}\|_F \cdot \|\mathbf{H}\|_F^2. \tag{D.13}$$

Thus combining (D.12) and (D.13), we have

$$\begin{aligned}\big|\langle \nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*), \mathbf{H}_U \mathbf{H}_V^\top \rangle\big| \leq \frac{1}{4}\|\widetilde{\mathbf{U}}^\top (\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*))\|_F \cdot \|\mathbf{H}\|_F^2 \\ + \frac{1}{4}\|(\nabla \mathcal{L}_N(\mathbf{X}) - \nabla \mathcal{L}_N(\mathbf{X}^*))\widetilde{\mathbf{V}}\|_F \cdot \|\mathbf{H}\|_F^2.\end{aligned} \tag{D.14}$$

Therefore, combining (D.11) and (D.14), the term $I_{11}$ can be lower bounded by

$$\begin{aligned} I_{11} &\geq \frac{1}{4L}\big(\|\widetilde{\mathbf{U}}^\top(\nabla\mathcal{L}_N(\mathbf{X})-\nabla\mathcal{L}_N(\mathbf{X}^*))\|_F^2+\|(\nabla\mathcal{L}_N(\mathbf{X})-\nabla\mathcal{L}_N(\mathbf{X}^*))\widetilde{\mathbf{V}}\|_F^2\big)+\frac{\mu}{2}\|\mathbf{X}-\mathbf{X}^*\|_F^2 \\ &\quad -\frac{1}{4}\big(\|\widetilde{\mathbf{U}}^\top(\nabla\mathcal{L}_N(\mathbf{X})-\nabla\mathcal{L}_N(\mathbf{X}^*))\|_F+\|(\nabla\mathcal{L}_N(\mathbf{X})-\nabla\mathcal{L}_N(\mathbf{X}^*))\widetilde{\mathbf{V}}\|_F\big)\cdot\|\mathbf{H}\|_F^2 \\ &\geq \frac{\mu}{2}\|\mathbf{X}-\mathbf{X}^*\|_F^2-\frac{L}{8}\|\mathbf{H}\|_F^4, \end{aligned} \tag{D.15}$$

where the last inequality holds because $2ab \leq ca^2+b^2/c$, for any $c>0$. Next, for the term $I_{12}$ in (D.10), we have

$$\big|\langle\nabla\mathcal{L}_N(\mathbf{X}^*),\mathbf{X}-\mathbf{X}^*\rangle\big| \leq \|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2\cdot\|\mathbf{X}-\mathbf{X}^*\|_* \leq \sqrt{2r}\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2\cdot\|\mathbf{X}-\mathbf{X}^*\|_F, \tag{D.16}$$

where the first inequality is due to the Von Neumann trace inequality, and the second inequality is due to the fact that $\mathrm{rank}(\mathbf{X}-\mathbf{X}^*)\leq 2r$. Similar for the remaining term in $I_{12}$, we have

$$\big|\langle\nabla\mathcal{L}_N(\mathbf{X}^*),\mathbf{H}_U\mathbf{H}_V^\top\rangle\big| \leq \sqrt{2r}\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2\cdot\|\mathbf{H}_U\mathbf{H}_V^\top\|_F. \tag{D.17}$$

Thus, combining (D.16) and (D.17), the term $I_{12}$ can be lower bounded by

$$\begin{aligned} I_{12} &\geq -\sqrt{2r}\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2\cdot\left(\|\mathbf{X}-\mathbf{X}^*\|_F+\frac{1}{2}\|\mathbf{H}\|_F^2\right) \\ &\geq -\frac{\mu}{8}\|\mathbf{X}-\mathbf{X}^*\|_F^2-\frac{L}{4}\|\mathbf{H}\|_F^4-\left(\frac{4r}{\mu}+\frac{r}{2L}\right)\cdot\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2, \end{aligned} \tag{D.18}$$

where the first inequality follows from the fact that $2\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2$, and the last inequality is due to $2ab \leq ca^2+b^2/c$, for any $c>0$. Therefore, plugging (D.15) and (D.18) into (D.10), we obtain the lower bound of $I_1$

$$I_1 \geq \frac{3\mu}{8}\|\mathbf{X}-\mathbf{X}^*\|_F^2-\frac{3L}{8}\|\mathbf{H}\|_F^4-\left(\frac{4r}{\mu}+\frac{r}{2L}\right)\cdot\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2. \tag{D.19}$$

On the other hand, for the term $I_2$ in (D.9), according to lemma D.1, we have

$$I_2 \geq \frac{1}{2}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2-\frac{1}{2}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F\cdot\|\mathbf{H}\|_F^2 \geq \frac{1}{4}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2-\frac{1}{4}\|\mathbf{H}\|_F^4, \tag{D.20}$$

where the last inequality holds because $2ab\leq a^2+b^2$. By plugging (D.19) and (D.20) into (D.9), we have

$$\langle\nabla\widetilde{F}_N(\mathbf{Z}),\mathbf{H}\rangle \geq \frac{3\mu}{8}\|\mathbf{X}-\mathbf{X}^*\|_F^2+\frac{1}{8}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2-\frac{3L+1}{8}\|\mathbf{H}\|_F^4-\left(\frac{4r}{\mu}+\frac{r}{2L}\right)\cdot\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2. \tag{D.21}$$

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

$$\begin{aligned} \|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2 &= \langle\mathbf{Z}\mathbf{Z}^\top-\mathbf{Z}^*\mathbf{Z}^{*\top},\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top-\widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top}\rangle+\langle\mathbf{Z}^*\mathbf{Z}^{*\top},\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top\rangle+\langle\mathbf{Z}\mathbf{Z}^\top,\widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top}\rangle \\ &\geq \langle\mathbf{Z}\mathbf{Z}^\top-\mathbf{Z}^*\mathbf{Z}^{*\top},\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top-\widetilde{\mathbf{Z}}^*\widetilde{\mathbf{Z}}^{*\top}\rangle \\ &= \|\mathbf{U}\mathbf{U}^\top-\mathbf{U}^*\mathbf{U}^{*\top}\|_F^2+\|\mathbf{V}\mathbf{V}^\top-\mathbf{V}^*\mathbf{V}^{*\top}\|_F^2-2\|\mathbf{U}\mathbf{V}^\top-\mathbf{U}^*\mathbf{V}^{*\top}\|_F^2, \end{aligned} \tag{D.22}$$

where the first equality is due to $\widetilde{\mathbf{Z}}^{*\top}\mathbf{Z}^*=0$, and the inequality is due to $\langle\mathbf{A}\mathbf{A}^\top,\mathbf{B}\mathbf{B}^\top\rangle=\|\mathbf{A}^\top\mathbf{B}\|_F^2\geq 0$. Thus, according to Lemma F.2, we have

$$4\|\mathbf{X}-\mathbf{X}^*\|_F^2+\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2=\|\mathbf{Z}\mathbf{Z}^\top-\mathbf{Z}^*\mathbf{Z}^{*\top}\|_F^2\geq 4(\sqrt{2}-1)\sigma_r\|\mathbf{H}\|_F^2, \tag{D.23}$$

where the first inequality holds because of (D.22), and the second inequality is due to Lemma F.2 and the fact that $\sigma_r^2(\mathbf{Z}^*)=2\sigma_r$. Denote $\mu'=\min\{\mu,1\}$. Therefore, plugging (D.23) into (D.21), we have

$$\begin{aligned} \langle\nabla\widetilde{F}_N(\mathbf{Z}),\mathbf{H}\rangle &\geq \frac{\mu}{8}\|\mathbf{X}-\mathbf{X}^*\|_F^2+\frac{\mu'\sigma_r}{10}\|\mathbf{H}\|_F^2+\frac{1}{16}\|\widetilde{\mathbf{Z}}^\top\mathbf{Z}\|_F^2 \\ &\quad -\frac{3L+1}{8}\|\mathbf{H}\|_F^4-\left(\frac{4r}{\mu}+\frac{r}{2L}\right)\cdot\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2, \end{aligned}$$

which completes the proof. □

### D.3. Proof of Lemma B.2

*Proof.* Consider the term $\mathbf{G}_U$ first. Denote $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. According to the definition of $\mathbf{G}_U$, we have

$$\begin{aligned}
\|\mathbf{G}_U\|_F^2 &= \|\nabla_{\mathbf{U}} F_i(\mathbf{U},\mathbf{V}) - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\mathbf{V} + \nabla\mathcal{L}_N(\widetilde{\mathbf{X}})\mathbf{V}\|_F^2 \\
&= \bigg\|\nabla\mathcal{L}_i(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\mathbf{V} + \nabla\mathcal{L}_N(\widetilde{\mathbf{X}})\mathbf{V} + \frac{1}{2}\mathbf{U}(\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V})\bigg\|_F^2 \\
&\le 2\underbrace{\|\nabla\mathcal{L}_i(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\mathbf{V} + \nabla\mathcal{L}_N(\widetilde{\mathbf{X}})\mathbf{V}\|_F^2}_{I_1} + \frac{1}{2}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 \cdot \|\mathbf{U}\|_2^2, \qquad \text{(D.24)}
\end{aligned}$$

where the second equality follows from definition of $\mathcal{F}_i$ in (2.4), and the inequality holds because $\|\mathbf{A}+\mathbf{B}\|_F^2 \le 2\|\mathbf{A}\|_F^2 + 2\|\mathbf{B}\|_F^2$ and $\|\mathbf{A}\mathbf{B}\|_F \le \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_F$. As for the term $I_1$ in (D.24), we further have

$$\begin{aligned}
I_1 &= \|\nabla\mathcal{L}_i(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\mathbf{V} + \nabla\mathcal{L}_N(\widetilde{\mathbf{X}})\mathbf{V} - \nabla\mathcal{L}_N(\mathbf{X})\mathbf{V} + \nabla\mathcal{L}_N(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_N(\mathbf{X}^*)\mathbf{V} + \nabla\mathcal{L}_N(\mathbf{X}^*)\mathbf{V}\|_F^2 \\
&\le 3\|\widetilde{\mathbf{G}}_U\|_F^2 + 3\|\nabla\mathcal{L}_N(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_N(\mathbf{X}^*)\mathbf{V}\|_F^2 + 3\|\nabla\mathcal{L}_N(\mathbf{X}^*)\mathbf{V}\|_F^2 \\
&\le 3\|\widetilde{\mathbf{G}}_U\|_F^2 + 3\|\nabla\mathcal{L}_N(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_N(\mathbf{X}^*)\mathbf{V}\|_F^2 + 3r\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2 \cdot \|\mathbf{V}\|_2^2, \qquad \text{(D.25)}
\end{aligned}$$

where we define $\widetilde{\mathbf{G}}_U = \nabla\mathcal{L}_i(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_N(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\mathbf{V} + \nabla\mathcal{L}_N(\widetilde{\mathbf{X}})\mathbf{V}$, the second inequality holds because $\|\mathbf{A}+\mathbf{B}+\mathbf{C}\|_F^2 \le 3\|\mathbf{A}\|_F^2 + 3\|\mathbf{B}\|_F^2 + 3\|\mathbf{C}\|_F^2$, and the last inequality holds because $\|\mathbf{A}\mathbf{B}\|_F \le \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_F$ and $\mathbf{V}$ has rank $r$. Thus combining (D.24) and (D.25), we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{G}_U\|_F^2 \le 6\underbrace{\mathbb{E}\|\widetilde{\mathbf{G}}_U\|_F^2}_{I_2} + 6\underbrace{\|\nabla\mathcal{L}_N(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_N(\mathbf{X}^*)\mathbf{V}\|_F^2}_{I_3} \\
+ \frac{1}{2}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 \cdot \|\mathbf{U}\|_2^2 + 6r\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2 \cdot \|\mathbf{V}\|_2^2, \qquad \text{(D.26)}
\end{aligned}$$

where the expectation is taken with respect to $i$. Next, we are going to upper bound $I_2$ and $I_3$, respectively. First, let us consider $I_2$ in (D.26). Since $i$ is uniformly picked from $[n]$, we have $\mathbb{E}[\nabla\mathcal{L}_i(\mathbf{X})\mathbf{V}] = \nabla\mathcal{L}_N(\mathbf{X})\mathbf{V}$ and $\mathbb{E}[\nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\mathbf{V}] = \nabla\mathcal{L}_N(\widetilde{\mathbf{X}})\mathbf{V}$. Recall the definition of $\widetilde{\mathbf{V}}$ in (D.2), we have

$$\begin{aligned}
\mathbb{E}\|\widetilde{\mathbf{G}}_U\|_F^2 &= \mathbb{E}\big\|[\nabla\mathcal{L}_i(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\mathbf{V}] - \mathbb{E}[\nabla\mathcal{L}_i(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\mathbf{V}]\big\|_F^2 \\
&\le \mathbb{E}\big\|\nabla\mathcal{L}_i(\mathbf{X})\mathbf{V} - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\mathbf{V}\big\|_F^2 \\
&\le \mathbb{E}\|(\nabla\mathcal{L}_i(\mathbf{X}) - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}}))\widetilde{\mathbf{V}}\|_F^2 \cdot \|\widetilde{\mathbf{V}}^\top\mathbf{V}\|_2^2 \\
&\le \frac{1}{n}\sum_{i=1}^n \big\|\big(\nabla\mathcal{L}_i(\mathbf{X}) - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\big)\widetilde{\mathbf{V}}\big\|_F^2 \cdot \|\mathbf{V}\|_2^2, \qquad \text{(D.27)}
\end{aligned}$$

where the first inequality holds because $\mathbb{E}\|\boldsymbol{\xi} - \mathbb{E}\boldsymbol{\xi}\|_2^2 \le \mathbb{E}\|\boldsymbol{\xi}\|_2^2$ for any random vector $\boldsymbol{\xi}$, the second inequality holds because $\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\mathbf{V} = \mathbf{V}$ and $\|\mathbf{A}\mathbf{B}\|_F \le \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_F$, and the last inequality holds because $\|\mathbf{A}\mathbf{B}\|_2 \le \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2$ and $\|\widetilde{\mathbf{V}}\|_2 = 1$. Similarly, as for the term $I_3$ in (D.26), we have

$$I_3 = \big\|\big(\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{X}^*)\big)\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\mathbf{V}\big\|_F^2 \le \big\|\big(\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{X}^*)\big)\widetilde{\mathbf{V}}\|_F^2 \cdot \|\mathbf{V}\|_2^2, \qquad \text{(D.28)}$$

where the equality holds because $\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top\mathbf{V} = \mathbf{V}$, and the inequality holds because $\widetilde{\mathbf{V}}$ is orthonormal. Plugging (D.27) and (D.28) into (D.26), we obtain

$$\begin{aligned}
\mathbb{E}\|\mathbf{G}_U\|_F^2 \le \frac{6}{n}\sum_{i=1}^n \big\|\big(\nabla\mathcal{L}_i(\mathbf{X}) - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\big)\widetilde{\mathbf{V}}\big\|_F^2 \cdot \|\mathbf{V}\|_2^2 + 6\big\|\big(\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{X}^*)\big)\widetilde{\mathbf{V}}\|_F^2 \cdot \|\mathbf{V}\|_2^2 \\
+ \frac{1}{2}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 \cdot \|\mathbf{U}\|_2^2 + 6r\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2 \cdot \|\mathbf{V}\|_2^2. \qquad \text{(D.29)}
\end{aligned}$$

As for $\mathbb{E}\|\mathbf{G}_V\|_F^2$, by the same techniques, we have

$$\begin{aligned}\mathbb{E}\|\mathbf{G}_V\|_F^2 \leq \frac{6}{n}\sum_{i=1}^n \big\|\widetilde{\mathbf{U}}^\top\big(\nabla\mathcal{L}_i(\mathbf{X}) - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\big)\big\|_F^2 \cdot \|\mathbf{U}\|_2^2 + 6\big\|\widetilde{\mathbf{U}}^\top\big(\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{X}^*)\big)\big\|_F^2 \cdot \|\mathbf{U}\|_2^2 \\ + \frac{1}{2}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 \cdot \|\mathbf{V}\|_2^2 + 6r\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2 \cdot \|\mathbf{U}\|_2^2, \qquad \text{(D.30)}\end{aligned}$$

where $\widetilde{\mathbf{U}}$ is defined in (D.1). Recall $\mathbf{Z} = [\mathbf{U};\mathbf{V}]$ and note that $\max\{\|\mathbf{U}\|_2, \|\mathbf{V}\|_2\} \leq \|\mathbf{Z}\|_2$. Thus combining (D.29) and (D.30), we obtain the upper bound of $\mathbb{E}\|\mathbf{G}\|_F^2$

$$\begin{aligned}\mathbb{E}\|\mathbf{G}\|_F^2 \leq \frac{12}{n}\sum_{i=1}^n \big(\underbrace{\big\|\big(\nabla\mathcal{L}_i(\mathbf{X}) - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\big)\widetilde{\mathbf{V}}\big\|_F^2 + \big\|\widetilde{\mathbf{U}}^\top\big(\nabla\mathcal{L}_i(\mathbf{X}) - \nabla\mathcal{L}_i(\widetilde{\mathbf{X}})\big)\big\|_F^2}_{I_4}\big) \cdot \|\mathbf{Z}\|_2^2 \\ + 12\big(\underbrace{\big\|\big(\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{X}^*)\big)\widetilde{\mathbf{V}}\|_F^2 + \big\|\widetilde{\mathbf{U}}^\top\big(\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{X}^*)\big)\|_F^2}_{I_5}\big) \cdot \|\mathbf{Z}\|_2^2 \\ + \|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 \cdot \|\mathbf{Z}\|_2^2 + 12r\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2 \cdot \|\mathbf{Z}\|_2^2. \qquad \text{(D.31)}\end{aligned}$$

Finally, according to the Lemma 3.5 and the restricted strong smoothness Conditions 3.4 and 3.7, we obtain the upper bound of $I_4$ and $I_5$

$$I_4 \leq 4L'\big(\mathcal{L}_i(\mathbf{X}) - \mathcal{L}_i(\widetilde{\mathbf{X}}) - \langle\nabla\mathcal{L}_i(\widetilde{\mathbf{X}}), \mathbf{X} - \widetilde{\mathbf{X}}\rangle\big) \leq 2L'^2\|\mathbf{X} - \widetilde{\mathbf{X}}\|_F^2 \leq 4L'^2(\|\widetilde{\mathbf{X}} - \mathbf{X}^*\|_F^2 + \|\mathbf{X} - \mathbf{X}^*\|_F^2), \qquad \text{(D.32)}$$

where the last inequality holds because $\|\mathbf{A} + \mathbf{B}\|_F^2 \leq 2\|\mathbf{A}\|_F^2 + 2\|\mathbf{B}\|_F^2$. Similarly we have

$$I_5 \leq 4L\big(\mathcal{L}_N(\mathbf{X}) - \mathcal{L}_N(\mathbf{X}^*) - \langle\nabla\mathcal{L}_N(\mathbf{X}^*), \mathbf{X} - \mathbf{X}^*\rangle\big) \leq 2L^2\|\mathbf{X} - \mathbf{X}^*\|_F^2. \qquad \text{(D.33)}$$

Hence, plugging (D.32) and (D.33) into (D.31), we obtain

$$\begin{aligned}\mathbb{E}\|\mathbf{G}\|_F^2 \leq \big(48L'^2\|\widetilde{\mathbf{X}} - \mathbf{X}^*\|_F^2 + 24(2L'^2 + L^2)\|\mathbf{X} - \mathbf{X}^*\|_F^2\big) \cdot \|\mathbf{Z}\|_2^2 \\ + \big(\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 + 12r\|\nabla\mathcal{L}_N(\mathbf{X}^*)\|_2^2\big) \cdot \|\mathbf{Z}\|_2^2,\end{aligned}$$

which completes the proof. □

# E. Proof of Technical Lemma in Appendix D

## E.1. Proof of Lemma D.2

*Proof.* By the restricted strong convexity of $\mathcal{L}_N$ in Condition 3.3, we have

$$\mathcal{L}_N(\mathbf{Y}) \geq \mathcal{L}_N(\mathbf{X}) + \langle\nabla\mathcal{L}_N(\mathbf{X}), \mathbf{Y} - \mathbf{X}\rangle + \frac{\mu}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2. \qquad \text{(E.1)}$$

Besides, according to lemma 3.5, we have

$$\begin{aligned}\mathcal{L}_N(\mathbf{X}) - \mathcal{L}_N(\mathbf{Y}) \geq \langle\nabla\mathcal{L}_N(\mathbf{Y}), \mathbf{X} - \mathbf{Y}\rangle + \frac{1}{4L}\|\widetilde{\mathbf{U}}^\top(\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{Y}))\|_F^2 \\ + \frac{1}{4L}\|(\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{Y}))\widetilde{\mathbf{V}}\|_F^2. \qquad \text{(E.2)}\end{aligned}$$

Therefore, combining (E.1) and (E.2), we have

$$\begin{aligned}\langle\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{Y}), \mathbf{X} - \mathbf{Y}\rangle \geq \frac{1}{4L}\|\widetilde{\mathbf{U}}^\top(\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{Y}))\|_F^2 \\ + \frac{1}{4L}\|(\nabla\mathcal{L}_N(\mathbf{X}) - \nabla\mathcal{L}_N(\mathbf{Y}))\widetilde{\mathbf{V}}\|_F^2 + \frac{\mu}{2}\|\mathbf{X} - \mathbf{Y}\|_F^2,\end{aligned}$$

which completes the proof. □

## F. Auxiliary lemmas

For the completeness of our proofs, we provide several auxiliary lemmas in this section, which are originally proved in Tu et al. (2015).

**Lemma F.1.** (Tu et al., 2015) Assume $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d_1 \times d_2}$ are two rank-$r$ matrices. Suppose they have singular value decomposition $\mathbf{X} = \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^\top$ and $\mathbf{Y} = \mathbf{U}_2 \mathbf{\Sigma}_2 \mathbf{V}_2^\top$. Suppose $\|\mathbf{X} - \mathbf{Y}\|_2 \leq \sigma_r(\mathbf{X})/2$, then we have

$$d^2\left([\mathbf{U}_2; \mathbf{V}_2]\mathbf{\Sigma}_1^{1/2}, [\mathbf{U}_1; \mathbf{V}_1]\mathbf{\Sigma}_2^{1/2}\right) \leq \frac{2}{\sqrt{2}-1} \frac{\|\mathbf{Y} - \mathbf{X}\|_F^2}{\sigma_r(\mathbf{X})}.$$

**Lemma F.2.** (Tu et al., 2015) For any matrices $\mathbf{Z}, \mathbf{Z}' \in \mathbb{R}^{(d_1+d_2)\times r}$, we have the following inequality

$$d^2(\mathbf{Z}, \mathbf{Z}') \leq \frac{1}{2(\sqrt{2}-1)\sigma_r^2(\mathbf{Z}')} \|\mathbf{Z}\mathbf{Z}^\top - \mathbf{Z}'\mathbf{Z}'^\top\|_F^2.$$

**Lemma F.3.** (Tu et al., 2015) For any matrices $\mathbf{Z}, \mathbf{Z}' \in \mathbb{R}^{(d_1+d_2)\times r}$, which satisfy $d(\mathbf{Z}, \mathbf{Z}') \leq \|\mathbf{Z}'\|_2/4$, we have the following inequality

$$\|\mathbf{Z}\mathbf{Z}^\top - \mathbf{Z}'\mathbf{Z}'^\top\|_F \leq \frac{9}{4}\|\mathbf{Z}'\|_2 \cdot d(\mathbf{Z}, \mathbf{Z}').$$