A. Additional Experiments

In this section, we provide additional experimental results, including simulations under the rectangular setting, the relative error versus CPU time plots and multi-label learning experiments on Yahoo datasets.

**Simulations under Rectangular setting.** The data generation approach for rectangular setting follows the same procedure in Section 6.1. Specifically, the unknown low-rank matrix \( \mathbf{L}^* \in \mathbb{R}^{d_1 \times d_2} \) is modeled as \( \mathbf{L}^* = \mathbf{X} \mathbf{M} \mathbf{X}^\top \) with dimensions \( d_1 = 5000, d_2 = 2000 \) and rank \( r = 5 \). The feature dimensions \( n_1, n_2 \) are specified as \( n_1 = 100 \) and \( n_2 = 50 \). We compare the performance of our proposed algorithm with existing (inductive) matrix completion methods, as explained in Section 6.1, in terms of relative error and effective number of data passes, which is illustrated in Figure 3. Note that the sampling rate \( p \) is chosen from the range \{0.25\%, 0.5\%, 1\%, 2\%\}. The results show that our proposed algorithm has better performance under the rectangular setting compared with existing methods.

![Figure 3](image)

*Figure 3.* Plots of logarithm relative error vs. number of effective data passes for different (inductive) matrix completion algorithms under the setting \( d_1 = 5000, d_2 = 2000, n_1 = 100, n_2 = 50 \) and \( r = 5 \) with sampling rate \( p \) varied in the range \{0.25\%, 0.5\%, 1\%, 2\%\}.

**Relative Error versus CPU Time.** To further demonstrate the computational advantage of the proposed method, we compare our algorithm with existing (inductive) matrix completion methods in terms of relative error and CPU time for all of the aforementioned simulation settings. It can be seen from Figure 4 that our proposed method achieves the lowest relative error with respect to the same CPU time under all settings, which again confirms the superiority of our proposed algorithm.

![Figure 4](image)

*Figure 4.* Plots of relative error vs. CPU time for different (inductive) matrix completion algorithms under the settings: (i) \( d = 1000, n = 100 \) and \( r = 10 \) with sampling rate \( p \) selected from \{2\%, 5\%, 10\%\} in the top panel. (ii) \( d_1 = 5000, d_2 = 2000, n_1 = 100, n_2 = 50 \) and \( r = 5 \) with \( p \) varied in the range \{0.25\%, 0.5\%, 1\%, 2\%\} in the bottom panel.
Multi-Label Learning on Yahoo Datasets. We provide the experimental results for multi-label learning on the Web page classification Yahoo datasets obtained from Ueda & Saito (2003), including Arts, Education, Health and Science. In particular, each dataset has 5000 different Web pages, which are associated with 462-743 different features and labeled with 26-40 different predefined categories. We extract the top-100 principal components to reconstruct the feature matrix so that our estimation model will not overfit. The comparison results are demonstrated in Table 3, which shows that our method achieves higher prediction accuracy and shorter computational time against existing inductive matrix completion approaches under all of the experimental settings (cf. Section 6.2 for detailed information).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>( p% = 10% )</th>
<th>( p% = 25% )</th>
<th>( p% = 50% )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>averaged AP (std)</td>
<td>time (s)</td>
<td>averaged AP (std)</td>
<td>time (s)</td>
</tr>
<tr>
<td>Arts</td>
<td>BR-linear</td>
<td>0.4333 (0.0055)</td>
<td>1.63 \times 10^3</td>
<td>0.4608 (0.0059)</td>
</tr>
<tr>
<td></td>
<td>Maxide</td>
<td>0.5131 (0.0062)</td>
<td>3.53 \times 10^9</td>
<td>0.5698 (0.0046)</td>
</tr>
<tr>
<td></td>
<td>AltMin</td>
<td>0.5232 (0.0060)</td>
<td>3.04 \times 10^9</td>
<td>0.5647 (0.0056)</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>0.5412 (0.0056)</td>
<td>1.16 \times 10^9</td>
<td>0.5768 (0.0050)</td>
</tr>
<tr>
<td>Education</td>
<td>BR-linear</td>
<td>0.4515 (0.0057)</td>
<td>1.52 \times 10^3</td>
<td>0.4637 (0.0073)</td>
</tr>
<tr>
<td></td>
<td>Maxide</td>
<td>0.5451 (0.0054)</td>
<td>4.95 \times 10^9</td>
<td>0.5851 (0.0052)</td>
</tr>
<tr>
<td></td>
<td>AltMin</td>
<td>0.5338 (0.0057)</td>
<td>2.03 \times 10^9</td>
<td>0.5731 (0.0075)</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>0.5555 (0.0042)</td>
<td>1.05 \times 10^9</td>
<td>0.5906 (0.0058)</td>
</tr>
<tr>
<td>Health</td>
<td>BR-linear</td>
<td>0.6402 (0.0064)</td>
<td>1.39 \times 10^3</td>
<td>0.6859 (0.0028)</td>
</tr>
<tr>
<td></td>
<td>Maxide</td>
<td>0.7210 (0.0073)</td>
<td>6.10 \times 10^9</td>
<td>0.7454 (0.0059)</td>
</tr>
<tr>
<td></td>
<td>AltMin</td>
<td>0.6961 (0.0091)</td>
<td>1.73 \times 10^9</td>
<td>0.7370 (0.0042)</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>\textbf{0.7305} (0.0053)</td>
<td>0.76 \times 10^9</td>
<td>\textbf{0.7496} (0.0036)</td>
</tr>
<tr>
<td>Science</td>
<td>BR-linear</td>
<td>0.4517 (0.0113)</td>
<td>1.80 \times 10^3</td>
<td>0.4671 (0.0089)</td>
</tr>
<tr>
<td></td>
<td>Maxide</td>
<td>0.4960 (0.0075)</td>
<td>7.19 \times 10^9</td>
<td>0.5297 (0.0042)</td>
</tr>
<tr>
<td></td>
<td>AltMin</td>
<td>0.4888 (0.0081)</td>
<td>4.38 \times 10^9</td>
<td>0.5341 (0.0052)</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>\textbf{0.5095} (0.0069)</td>
<td>1.11 \times 10^9</td>
<td>\textbf{0.5391} (0.0039)</td>
</tr>
</tbody>
</table>

B. Proofs of the Main Results

In this section, we provide the proofs for our main theoretical results. To begin with, we introduce some notations to simplify our proof. Let \( \mathbb{I} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^{d_1 \times d_2} \) be the identity map, i.e., \( \mathbb{I}(A) = A \). Denote the elementwise infinity norm of matrix \( A \) by \( \| A \|_{\infty, \infty} \). For any \( Z \in \mathbb{R}^{(n_1 + n_2) \times r} \), we denote \( Z = [U | V] \), where \( U \in \mathbb{R}^{n_1 \times r} \) and \( V \in \mathbb{R}^{n_2 \times r} \). According to (4.1), our objective is equivalent to minimizing the following regularized loss function in terms of \( Z \)

\[
\tilde{f}_\Omega(Z) := f_\Omega(U, V) = \frac{1}{2p} \| P_\Omega(X_L U V^\top X_R^\top - L) \|_F^2 + \frac{1}{8} \| U^\top U - V^\top V \|_F^2. \tag{B.1}
\]

Let \( \text{Sym} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^{(d_1 + d_2) \times (d_1 + d_2)} \) be the lifting operator, such that for any matrix \( A \in \mathbb{R}^{d_1 \times d_2} \)

\[
\text{Sym}(A) = \begin{bmatrix} 0 & A^\top \\ A & 0 \end{bmatrix}.
\]

For any block matrices \( A \in \mathbb{R}^{(d_1 + d_2) \times (d_1 + d_2)} \) with partitions

\[
A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \text{ where } A_{ij} \in \mathbb{R}^{d_i \times d_j}, i, j \in \{1, 2\},
\]

define linear operators \( P_{\text{diag}} \) and \( P_{\text{off}} : \mathbb{R}^{(d_1 + d_2) \times (d_1 + d_2)} \rightarrow \mathbb{R}^{(d_1 + d_2) \times (d_1 + d_2)} \) as

\[
P_{\text{diag}}(A) = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \quad \text{and} \quad P_{\text{off}}(A) = \begin{bmatrix} 0 & A_{12} \\ A_{21} & 0 \end{bmatrix}.
\]

Similarly, for any block matrices \( B \in \mathbb{R}^{(n_1 + n_2) \times (n_1 + n_2)} \) with partitions

\[
B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \text{ where } B_{ij} \in \mathbb{R}^{n_i \times n_j}, i, j \in \{1, 2\},
\]
define operators $\mathcal{P}_{\text{diag}}$ and $\mathcal{P}_{\text{off}} : \mathbb{R}^{(n_1 + n_2) \times (n_1 + n_2)} \rightarrow \mathbb{R}^{(n_1 + n_2) \times (n_1 + n_2)}$ as

$$
\mathcal{P}_{\text{diag}}(B) = \begin{bmatrix} B_{11} & 0 \\ 0 & B_{22} \end{bmatrix} \quad \text{and} \quad \mathcal{P}_{\text{off}}(B) = \begin{bmatrix} 0 & B_{12} \\ B_{21} & 0 \end{bmatrix}.
$$

In addition, let $\Omega \in [d_1 + d_2] \times [d_1 + d_2]$ be the corresponding index set of observed entries in the lifted space, then the observed matrix in the lifted space can be expressed as

$$
\mathcal{P}_\Omega (\text{Sym}(L)) = \begin{bmatrix} 0 \\ (\mathcal{P}_\Omega (L))^\top \\ 0 \end{bmatrix}.
$$

And we let $X = X_L \begin{bmatrix} I_{n_1} & 0 \\ 0 & -I_{n_2} \end{bmatrix}.$

Thus with the notations above, the regularized loss function $\tilde{f}_\Omega$ in (B.1) can be rewritten as

$$
\tilde{f}_\Omega(Z) = \frac{1}{4p} \| \mathcal{P}_\Omega (XZZ^\top X^\top - \text{Sym}(L)) \|_F^2 + \frac{1}{8} \| Z^\top DZ \|_F^2, \quad (B.2)
$$

where $D$ is defined as

$$
D = \begin{bmatrix} I_{n_1} & 0 \\ 0 & -I_{n_2} \end{bmatrix}.
$$

Recall that $Z^* = [U^*; V^*]$ and $\mathcal{P}_\Omega (L) = \mathcal{P}_\Omega (L^*).$, then the gradient of $\tilde{f}_\Omega$ can be formulated as

$$
\nabla \tilde{f}_\Omega(Z) = \frac{1}{p} X^\top \mathcal{P}_\Omega (X[ZZ^\top - Z^*Z^*\top]X^\top)XZ
$$

$$
+ \frac{1}{2} (\mathcal{P}_{\text{diag}} - \mathcal{P}_{\text{off}})(ZZ^\top)Z. \quad (B.3)
$$

B.1. Proof of Theorem 5.3

Proof. According to the initialization phase of Algorithm 1, we have

$$
\| U_{\text{init}} V_{\text{init}}^\top - M^\star \|_2 = \| P_{0}^{-1} X_L^\top P_{\Omega_0} (L^*) X_R - M^\star \|_2
$$

$$
= \| X_L^\top (P_{0}^{-1} P_{\Omega_0} (L^*) - L^*) X_R \|_2
$$

$$
= \| \sum_{(i,j)\in(1,1)} (P_{0}^{-1} \xi_{ij} - 1) L_{ij}^* \cdot X_L^\top e_i e_j^\top X_R \|_2
$$

$$
: = \| \sum_{(i,j)\in(1,1)} S_{ij} \|_2, \quad (B.4)
$$

where $\xi_{ij} = 1, \text{ if } (i,j) \in \Omega_0; \xi_{ij} = 0, \text{ otherwise.}$ Next we are going to apply Matrix Bernstein to the right hand side of (B.4). Note that $\mathbb{E}[S_{ij}] = 0$, and we have

$$
\| S_{ij} \|_2 \leq \sum_{(i,j)\in(1,1)} \| L_{ij}^* \|_2 \cdot \| X_L^\top e_i \|_2 \cdot \| X_R^\top e_j \|_2 \leq \frac{1}{p_0} \| L^* \|_\infty,\infty \cdot \| X_L \|_2,\infty \cdot \| X_R \|_2,\infty.
$$

By Assumptions 3.1 and 3.2, we further obtain

$$
\| S_{ij} \|_2 \leq \frac{1}{p_0} \| X_L U^* \|_2,\infty \cdot \| \Sigma^\star \|_2 \cdot \| X_R V^* \|_2,\infty \cdot \sqrt{\frac{\mu_0^2 n_1 n_2}{d_1 d_2}} \leq \frac{\mu_0 \mu_1 r \sqrt{n_1 n_2} \sigma_1}{p_0 d_1 d_2}. \quad (B.5)
$$
Denote $S = \sum_{i,j} S_{ij}$. To apply Matrix Bernstein, it remains to bound the terms $\|E(S^T S)\|_2$ and $\|E(S^T)\|_2$ respectively. Since $S_{ij}$’s are independent, we have

$$
\|E(S^T S)\|_2 = \left\| \sum_{(i,j)=(1,1)}^{(d_1,d_2)} E(S_{ij} S_{ij}^\top) \right\|_2
= \frac{1-p_0}{p_0} \left\| \sum_{(i,j)=(1,1)}^{(d_1,d_2)} L_{ij}^2 (X_L^\top e_i) \cdot \|X_R^\top e_j\|^2 \cdot (e_i^\top X_L) \right\|_2
\leq \frac{1}{p_0} \left\| \sum_{i=1}^{d_1} e_i e_i^\top \cdot \left( \sum_{j=1}^{d_2} L_{ij}^2 \cdot \|X_R^\top e_j\|^2 \right) \right\|_2
\leq \frac{1}{p_0} \cdot \max_{i \in [d_1]} \left( \sum_{j=1}^{d_2} L_{ij}^2 \cdot \|X_R^\top e_j\|^2 \right) \leq \frac{\mu_1 n_1}{p_0 d_2} \cdot \|L^*\|_{2,\infty}^2, \tag{B.5}
$$

where the first inequality is due to the fact that $\|A\|_2 \leq \|A\| \cdot \|B\|_2$ and $X_L$ is orthonormal, the last inequality follows from Assumption 3.2. According to the SVD of $L^*$ and Assumption 3.1, we have

$$
\|L^*\|_{2,\infty} = \|(X_L U^\top) \cdot \Sigma^*\|_{2,\infty} \leq \|X_L U^\top\|_{2,\infty} \cdot \|\Sigma^*\|_2 \leq \sqrt{\frac{\mu_0 \sigma_1}{d_1}} \sigma_1^*.
$$

(B.6)

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

$$
\|E(S^T S)\|_2 \leq \frac{\mu_1 n_1 \sigma_1^*}{p_0 d_1 d_2}. \tag{B.5}
$$

Similarly, we can obtain the upper bound of $\|E(S^T S)\|_2$, which implies

$$
\max \left\{ \left\| E(S^T S) \right\|_2, \left\| E(S^T) \right\|_2 \right\} \leq \frac{\mu_1 n_1 (n_1 + n_2)}{p_0 d_1 d_2} \sigma_1^*.
$$

Applying Matrix Bernstein Lemma E.1, under condition that $p_0 \geq c \mu_1 \sigma_1^* d/\sqrt{d}$, we have

$$
P\left\{ \left\| \sum_{(i,j)=(1,1)}^{(d_1,d_2)} S_{ij} \right\|_2 \geq \gamma \cdot \sigma_r^* / \sqrt{r} \right\} \leq (d_1 + d_2) \cdot \exp(-c \gamma \log d) \leq \frac{1}{d},
$$

where $c, c', \gamma > 0$ are some constants. This further implies that with probability at least $1 - 1/d$, we have $\|U_{init} V_{init}^\top - M^*\|_2 \leq \gamma \sqrt{\sigma_r^*}$. Finally, according to Lemma 5.14 in Tu et al. (2015), we obtain

$$
D^2(Z_{init}, Z^*) \leq \frac{2}{\sqrt{2} - 1} \cdot \frac{\|U_{init} V_{init}^\top - M^*\|^2_E}{\sigma_r^* (M^*)} \leq \frac{10r}{\sigma_r^*} \frac{\|U_{init} V_{init}^\top - M^*\|^2_F}{\sigma_r^*} \leq 10 \gamma^2 \sigma_r^*,
$$

where the second inequality holds because $\text{rank}(U_{init} V_{init}^\top - M^*)$ is at most $2r$. □

### B.2. Proof of Theorem 5.4

Before proceeding to the main proof, we introduce the following notations and facts. Recall that $M^* = U^\top \Sigma^* V^\top$ and $Z^* = [U^\top, V^\top] \Sigma^{*1/2}$, we denote $\tilde{Z}^* = [U^\top, -V^\top] \Sigma^{*1/2}$. Note that $M^*$ and $L^*$ have the same set of singular values, thus for any $\ell \in [r]$, $\sigma_\ell^2(Z^*) = \sigma_\ell^2(\tilde{Z}^*) = 2 \sigma_\ell^2$. We further note that $Z^*^\top \tilde{Z}^* = \tilde{Z}^*^\top Z^* = 0$, and $\text{Sym}(M^*) = (Z^* Z^*^\top - \tilde{Z}^* \tilde{Z}^*^\top) / 2$. Define reference function $G(Z)$ as $G(Z) = \|Z Z^* - Z^* Z^*^\top\|_2^2 / 4$, then the gradient of $G$ is given by

$$
\nabla G(Z) = (Z Z^* - Z^* Z^*^\top) Z.
$$

(B.7)

Thus according to (B.3), we have

$$
\nabla f_{\Omega}(Z) = \frac{1}{2} \nabla G(Z) + \frac{1}{2} (\tilde{Z}^* \tilde{Z}^*^\top) Z + X^\top \left( \frac{1}{p} \mathcal{P}_{\Omega} - \mathcal{P}_{\text{off}} \right) (X^\top ZZ^*^\top - Z^* Z^*^\top X^\top) X Z.
$$

(B.8)
The following lemmas demonstrate the local curvature and local smoothness properties of $\hat{f}_{1,\Omega}$, which are proved in Sections C.1 and C.2, respectively. In both lemmas, for any $Z \in \mathbb{R}^{(n_1+n_2) \times r}$, we let $R = \text{argmin}_{R \in Q_r} \|Z - Z^* R\|_F$, and denote $H = Z - Z^* R$.

**Lemma B.1** (Local curvature). Under the previously stated assumptions in Theorem 5.4, for any fixed $Z = [U; V] \in \mathbb{R}^{(n_1+n_2) \times r}$ such that $D(Z, Z^*) \leq \sqrt{r \sigma_r^2}/4$ and $\|X_L U\|_{2,\infty} \leq 2 \sqrt{\mu_r \sigma_1^r}/d_1$, $\|X_R V\|_{2,\infty} \leq 2 \sqrt{\mu_r \sigma_1^r}/d_2$, there exist constants $c_1, c_2$ such that if $|\Omega| \geq c_1 \max\{\mu_0^2 \sigma_r^2, \mu_0 \mu_1 \tau kn\} \log d$, then with probability at least $1 - c_2/d$, we have

$$\langle \nabla \hat{f}_{1,\Omega}(Z), H \rangle \geq \frac{1}{20} \|ZZ^* - Z^* Z^* H\|_F^2 + \frac{1}{4\sigma_1^r} \|Z^* Z^* H\|_F^2 + \frac{\sigma_1^r}{8} \|H\|_F^2 - 40 \|H\|_F^4,$$

where $Z^* = [U^*; -V^*]$.

**Lemma B.2** (Local smoothness). Under the previous stated assumptions as in Theorem 5.4, for any fixed $Z = [U; V] \in \mathbb{R}^{(n_1+n_2) \times r}$ such that $D(Z, Z^*) \leq \sqrt{r \sigma_r^2}/4$ and $\|X_L U\|_{2,\infty} \leq 2 \sqrt{\mu_r \sigma_1^r}/d_1$, $\|X_R V\|_{2,\infty} \leq 2 \sqrt{\mu_r \sigma_1^r}/d_2$, there exist constants $c_1, c_2$ such that if $|\Omega| \geq c_1 \mu_0 \mu_1 \tau kn \log d$, then with probability at least $1 - c_2/d$, we have

$$\|\nabla \hat{f}_{1,\Omega}(Z)\|_F \leq (16r + 4\sigma_1^r) \|ZZ^* - Z^* Z^* H\|_F^2 + \|Z^* Z^* H\|_F^2 + 4r \sigma_1^r \|H\|_F^2,$$

where $Z^* = [U^*; -V^*]$.

Now we are ready to prove Theorem 5.4.

**Proof.** Theorem 5.4 will be proved by induction. Consider the $s$-th iteration in Phase 2 of Algorithm 1, for any $s \geq 1$. Suppose the previous iterate $Z_{s-1}$ is sufficiently close to $Z^*$, i.e., $D(Z_{s-1}, Z^*) \leq \alpha \sqrt{r}$, where $\alpha$ is defined in Theorem 5.4. In the following discussions, we are going to show that the following contraction result with respect to the $s$-th iterate:

$$D^2(Z_s, Z^*) \leq \left(1 - \frac{\eta \sigma_1^r}{16}\right) \cdot D^2(Z_{s-1}, Z^*) + 3 \delta \cdot \alpha \sqrt{r} + 2 \delta^2$$

holds with probability at least $1 - c_1/d$, where $Z_s = [U_s; V_s]$ and $Z_{s-1} = [U_{s-1}; V_{s-1}]$.

Denote the optimal rotation with respect to $Z_{init}$ by $R_{init}$ such that $R_{init} = \text{argmin}_{R \in Q_r} \|Z_{init} - Z^* R\|_F$. Since the initial iterate $Z_{init}$ satisfies $D(Z_{init}, Z^*) \leq \sqrt{r \sigma_r^2}/40$, we have $\|Z_{init} - Z^* R_{init}\|_2 \leq \sqrt{r \sigma_r^2}/40$, which implies

$$\sqrt{r} \leq \|Z^* R_{init}\|_2 - \|Z_{init}\|_2 \leq \|Z^* R_{init}\|_2 \leq \|Z_{init}\|_2 \leq 2 \sqrt{r \sigma_r^2} / 40.$$

Thus, according to the definition of $C_1, C_2$ in (4.2) and Assumption 3.1, we have

$$\|X_L U^*\|_{2,\infty} \leq \|X_L U^*\|_{2,\infty} \cdot \|\Sigma^*\|_{2,\infty} \leq \sqrt{\frac{\mu^r \sigma_1^r}{d_3}} \leq \sqrt{\frac{\mu^r \sigma_1^r}{d_3}} \|Z_{init}\|_2,$$

which implies that $U^* \in C_1$. Similarly, we can derive that $V^* \in C_2$. In addition, based on the definition of the constraint sets $C_1$ and $C_2$ in (4.2), we further have $\|U\|_{2,\infty} \leq 2 \sqrt{\mu_r \sigma_1^r}/d_1$ and $\|V\|_{2,\infty} \leq 2 \sqrt{\mu_r \sigma_1^r}/d_2$. For any $s \in \{1, 2, \ldots, S\}$, we denote $R_s = \text{argmin}_{R \in Q_r} \|Z_s - Z^* R\|_F$ as the optimal rotation with respect to $Z_s$, and we let $H_s = Z_s - Z^* R_s$. Consider the $s$-iteration of Phase 2 in Algorithm 1, we let $\tilde{U}_s = U_{s-1} - \eta \nabla_U f_{\Omega_s}(U_{s-1}, V_{s-1})$ and $\tilde{V}_s = V_{s-1} - \eta \nabla V f_{\Omega_s}(U_{s-1}, V_{s-1})$. Thus based on the update rule, we have

$$D^2(Z_s, Z^*) \leq \|U_{s-1} - U^* R_{s-1}\|_F^2 + \|V_{s-1} - V^* R_{s-1}\|_F^2 = \|P_{C_1}(\tilde{U}_s, \delta) - U^* R_{s-1}\|_F^2 + \|P_{C_2}(\tilde{V}_s, \delta) - V^* R_{s-1}\|_F^2,$$

where the first inequality follows from Definition 5.1, and the second inequality follows from the update rule. As for the first term $I_1$, we have

$$I_1 = \|P_{C_1}(\tilde{U}_s) - U^* R_{s-1}\|_F^2 + 2 \langle P_{C_1}(\tilde{U}_s, \delta) - P_{C_1}(\tilde{U}_s), P_{C_1}(\tilde{U}_s) - U^* R_{s-1}\rangle + \|P_{C_1}(\tilde{U}_s, \delta) - P_{C_1}(\tilde{U}_s)\|_F^2 \leq \|P_{C_1}(\tilde{U}_s) - U^* R_{s-1}\|_F^2 + 2 \delta \cdot \|P_{C_1}(\tilde{U}_s) - U^* R_{s-1}\|_F + \delta^2 \leq \|U_{s-1} - \eta \nabla_U f_{\Omega_s}(U_{s-1}, V_{s-1}) - U^* R_{s-1}\|_F^2 + 2 \delta \cdot \|U_{s-1} - \eta \nabla_U f_{\Omega_s}(U_{s-1}, V_{s-1}) - U^* R_{s-1}\|_F + \delta^2.$$
where the first inequality holds because $P_{C_1}(\bar{U}_s, \delta)$ is the $\delta$-approximate solution and $P_{C_1}(\bar{U}_s)$ is the exact solution to the same optimization problem, and the second inequality is due to the non-expansive property of projection $P_{C_1}$ and the fact that $U^* \in C_1$. Based on the similar technique, we can upper bound $I_2$ as follows

$$I_2 \leq \| V_{s-1} - \eta \nabla f_{\Omega_s} (U_{s-1}, V_{s-1}) - V^* R_{s-1} \|_F^2 + 2\delta \cdot \| V_{s-1} - \eta \nabla f_{\Omega_s} (U_{s-1}, V_{s-1}) - V^* R_{s-1} \|_F + \delta^2.$$ 

Therefore, combining the upper bounds of $I_1$ and $I_2$, we have

$$D^2(Z_s, Z^*) \leq D^2(Z_{s-1}, Z^*) - 2\eta \langle \nabla \tilde{f}_{\Omega_s}(Z_{s-1}), H_{s-1} \rangle + \eta^2 \| \nabla \tilde{f}_{\Omega_s}(Z_{s-1}) \|_F^2 + 2\sqrt{2\delta} \cdot \left( D^2(Z_{s-1}, Z^*) - 2\eta \langle \nabla \tilde{f}_{\Omega_s}(Z_{s-1}), H_{s-1} \rangle + \eta^2 \| \nabla \tilde{f}_{\Omega_s}(Z_{s-1}) \|_F^2 \right)^{1/2} + \sqrt{8\delta},$$

where $\tilde{f}_{\Omega_s}$ is defined in (B.3), and the inequality follows from the triangle inequality.

It is worth noting that the subsampling technique ensures the previous iterate $Z_{s-1}$ is independent of the samples $\Omega_s$ used in the $s$-th iteration. According to the assumptions of Theorem 5.4, $|\Omega_s| = |\Omega|/(2S) \leq c_1 \max\{n, \mu_0 r^2, \mu_1 \alpha \}$, we can directly apply Lemmas A.1 and A.2. More specifically, with probability at least $1 - c_1/d$, we have

$$\langle \nabla \tilde{f}_{\Omega_s}(Z_{s-1}), H_{s-1} \rangle \geq \frac{1}{20} \| Z_{s-1} Z_{s-1}^T - Z^* Z^* \|_F^2 + \frac{1}{4}\sigma^2 \| \bar{Z}^* \bar{Z}^* - Z_{s-1} Z_{s-1} \|_F^2 + \frac{\sigma^2}{8} \| H_{s-1} \|_F^2 - 40 \| H_{s-1} \|_F^2,$$

where $\bar{Z}^* = [U^*; -V^*]$. In addition, we have

$$\| \nabla \tilde{f}_{\Omega_s}(Z_{s-1}) \|_F \leq (16r + 4)\sigma^2 \| Z_{s-1} Z_{s-1}^T - Z^* Z^* \|_F^2 + \| \bar{Z}^* \bar{Z}^* - Z_{s-1} Z_{s-1} \|_F^2 + 4r\sigma^2 \| H_{s-1} \|_F^2.$$

Thus, by setting step size $\eta \leq 1/(200r\sigma^2)$, we obtain

$$-2\eta \langle \nabla \tilde{f}_{\Omega_s}(Z_{s-1}), H_{s-1} \rangle + \eta^2 \| \nabla \tilde{f}_{\Omega_s}(Z_{s-1}) \|_F^2 \leq -\frac{\eta\sigma^2}{8} \| H_{s-1} \|_F^2 + 80\eta \| H_{s-1} \|_F^2,$$

which implies that the induction hypothesis $D(Z_{s-1}, Z^*) = \| H_{s-1} \|_F \leq \alpha\sqrt{\sigma^2}$, we have

$$D^2(Z_s, Z^*) \leq \left( 1 - \frac{\eta\sigma^2}{16} \right) \cdot D^2(Z_{s-1}, Z^*) + 2\delta \cdot \sqrt{2} \left( 1 - \frac{\eta\sigma^2}{16} \right) - D^2(Z_{s-1}, Z^*) + 2\delta^2$$

$$\leq \left( 1 - \frac{\eta\sigma^2}{16} \right) \cdot D^2(Z_{s-1}, Z^*) + 2\delta \cdot \alpha \sqrt{\sigma^2} + 2\delta^2,$$

which completes the proof of (B.9).

Moreover, according to the assumption that $D(Z_{init}, Z^*) \leq \alpha\sqrt{\sigma^2}/2$, if we choose $\delta \leq \alpha\sqrt{\sigma^2}/(2\sqrt{2})$, then we have

$$D^2(Z_0, Z^*) \leq \| U_0 - U^* R_{init} \|_F^2 + \| V_0 - V^* R_{init} \|_F^2$$

$$\leq \| P_{C_1}(U_{init}) - U^* R_{init} \|_F^2 + \| P_{C_2}(V_{init}) - V^* R_{init} \|_F^2$$

$$+ 2\delta \cdot \left( \| P_{C_1}(U_{init}) - U^* R_{init} \|_F + \| P_{C_2}(V_{init}) - V^* R_{init} \|_F \right) + 2\delta^2$$

$$\leq D^2(Z_{init}, Z^*) + 2\sqrt{2\delta} \cdot D(Z_{init}, Z^*) + 2\delta^2 \leq \alpha^2 \sigma^2,$$

where we let $R_{init}$ be the optimal rotation with respect to $Z_{init}$, and the last inequality follows from the non-expansiveness property of projection onto convex set. Thus, we have shown that the induction hypothesis $D(Z_{s-1}, Z^*) \leq \alpha\sqrt{\sigma^2}$ holds for the first iterate.

To this end, it remains to verify the induction step, or more specifically, $D(Z_{s-1}, Z^*) \leq \alpha\sqrt{\sigma^2}$ implies $D(Z_s, Z^*) \leq \alpha\sqrt{\sigma^2}$, for any $s \geq 1$. This step can be proved based on (B.9): with high probability, we have

$$D^2(Z_s, Z^*) \leq \left( 1 - \frac{\eta\sigma^2}{16} \right) \cdot D^2(Z_{s-1}, Z^*) + 3\delta \cdot \alpha \sqrt{\sigma^2} + 2\delta^2$$

$$\leq \left( 1 - \frac{\eta\sigma^2}{16} \right) \cdot \alpha^2 \sigma^2 + 3\delta \cdot \alpha \sqrt{\sigma^2} + 2\delta^2 \leq \alpha^2 \sigma^2,$$
provided that $\delta \leq c_2 \sqrt{\sigma^2/(r\kappa)}$ with constant $c_2$ sufficiently small. Finally, by induction and union bound, we obtain

$$D^2(Z_{S}, Z^*) \leq \left(1 - \frac{\eta \sigma^*}{16}\right)^{S} \cdot D^2(Z_{0}, Z^*) + \frac{16}{\eta \sigma^*} \cdot (3\delta \cdot \alpha \sqrt{\sigma^2_r} + 2\delta^2)$$

holds with probability at least $1 - c_1 S/d$, which completes the proof.

\[\square\]

B.3. Proof of Theorem 5.5

The proof of Theorem 5.5 will be similar to the proof of Theorem 5.4. The only difference is that we do not require sample splitting in Phase 3, thus the iterates are no longer independent from the subset of samples. The following curvature and smoothness of $f_{\Omega}$ are proved for all $Z \in \mathbb{R}^{(n_1+n_2) \times r}$ satisfying $D(Z, Z^*) \leq c_0 \sqrt{\sigma^2_r}/(\mu_1 n)$. The proofs are presented in Sections C.3 and C.4 respectively.

**Lemma B.3 (Local curvature).** Under the previously stated assumptions in Theorem 5.5, for all $Z = [U; V] \in \mathbb{R}^{(n_1+n_2) \times r}$ such that $D(Z, Z^*) \leq c_0 \sqrt{\sigma^2_r}/(\mu_1 n)$ with constant $c_0$ small enough, there exist constants $c_1, c_2$ such that if $|\Omega| \geq c_1 \mu_0 \mu_1 r n \log d$, then with probability at least $1 - c_2/d$, we have

$$\langle \nabla f_\Omega(Z), H \rangle \geq \frac{1}{20} \|Z Z^T - Z^* Z^*\|_F^2 + \frac{1}{4\sigma^*_1} \|Z^* Z^{*\top} Z\|_F^2 + \frac{\sigma^*_r}{8} \|H\|_F^2 + 20\|H\|_F^2,$$

where $\tilde{Z}^* = [U^*; -V^*]$.

**Lemma B.4 (Local smoothness).** Under the previous stated assumptions as in Theorem 5.5, for all $Z = [U; V] \in \mathbb{R}^{(n_1+n_2) \times r}$ such that $D(Z, Z^*) \leq c_0 \sqrt{\sigma^2_r}/(\mu_1 n)$ with constant $c_0$ small enough, there exist constants $c_1, c_2$ such that if $|\Omega| \geq c_1 \mu_0 \mu_1 r n \log d$, then with probability at least $1 - c_2/d$, we have

$$\|\nabla f_\Omega(Z)\|_F^2 \leq 84\sigma^*_1 \|Z Z^T - Z^* Z^*\|_F^2 + \|Z^* Z^{*\top} Z\|_F^2 + 4\sigma^*_1 \|H\|_F^2 + 40\sigma^*_1 \|H\|_F^2,$$

where $\tilde{Z}^* = [U^*; -V^*]$.

Now we are ready to prove Theorem 5.5.

**Proof.** For any $t \in \{0, 1, \ldots, T\}$, we denote $R^t = \text{argmin}_{R \in \mathbb{R}^{n_1 \times r}} \|Z^t - Z^* R\|_F$ as the optimal rotation with respect to $Z^t$, and we let $H^t = Z^t - Z^* R^t$. Note that the initial iterate of Phase 3 in Algorithm 1 satisfying $D(Z^0, Z^*) \leq c_0 \sqrt{\sigma^2_r}/(\mu n)$. Assume the induction hypothesis $D(Z^t, Z^*) \leq c_0 \sqrt{\sigma^2_r}/(\mu n)$ holds for $s = 1, 2, \ldots, t$. Consider the $t$-th iteration, based on the update rule, we have

$$D^2(Z^{t+1}, Z^*) \leq \|U^{t+1} - U^* R^t\|_F^2 + \|V^{t+1} - V^* R^t\|_F^2$$

$$= \|U^t - \tau \nabla_U f_{\Omega}(U^t, V^t) - U^* R^t\|_F^2 + \|V^t - \tau \nabla_V f_{\Omega}(U^t, V^t) - V^* R^t\|_F^2$$

$$= D^2(Z^t, Z^*) - \tau \langle \nabla f_{\Omega}(Z^t), H^t \rangle + \tau^2 \|\nabla f_{\Omega}(Z^t)\|_F^2,$$

where the inequality follows from Definition 5.1. According to the assumptions of Theorem 5.5, we can directly apply Lemmas B.3 and B.4. More specifically, with probability at least $1 - c/d$, we have

$$\langle \nabla f_{\Omega}(Z^t), H^t \rangle \geq \frac{1}{20} \|Z^t Z^{t\top} - Z^* Z^{*\top}\|_F^2 + \frac{1}{4\sigma^*_1} \|Z^* Z^{*\top} Z^t\|_F^2 + \frac{\sigma^*_r}{8} \|H^t\|_F^2,$$

where $\tilde{Z}^* = [U^*; -V^*]$. In addition, we have

$$\|\nabla f_{\Omega}(Z^t)\|_F^2 \leq 84\sigma^*_1 \|Z^t Z^{t\top} - Z^* Z^{*\top}\|_F^2 + \|Z^* Z^{*\top} Z^t\|_F^2 + 4\sigma^*_1 \|H^t\|_F^2 + 40\sigma^*_1 \|H^t\|_F^2,$$

Thus, by setting step size $\tau \leq c_1/\sigma^*_1$ with $c_1 \leq 1/840$, we obtain

$$-2\tau \langle \nabla f_{\Omega}(Z^t), H^t \rangle + \tau^2 \|\nabla f_{\Omega}(Z^t)\|_F^2 \leq -\frac{\eta \sigma^*_r}{8} \|H^t\|_F^2 + 50\eta \|H^t\|_F^2,$$

which implies that under condition that $D(Z^t, Z^*) = \|H^t\|_F \leq c_2 \sqrt{\sigma^2_r}$ with $c_2 \leq 1/30$, we have

$$D^2(Z^{t+1}, Z^*) \leq \left(1 - \frac{\tau \sigma^*_r}{16}\right) \cdot D^2(Z^t, Z^*),$$

which implies the $(t + 1)$-th iterate $Z^{t+1}$ still satisfies $D^2(Z^{t+1}, Z^*) \leq c_0 \sqrt{\sigma^2_r}/(\mu n)$. Thus by induction, we complete the proof. \[\square\]
B.4. Proof of Theorem 5.2

Proof. To prove the overall theoretical guarantee of Algorithm 1, we need to examine the conditions required by previous main theorems regarding the three phases. More specifically, to ensure the \(O(\sqrt{\sigma_p^2})\) initial ball assumption \(D(Z_{\text{init}}, Z^*) \leq \alpha\sqrt{\sigma_p^2}/2\) in Theorem 5.4, it suffices to set \(\delta \leq \alpha/8\) in Theorem 5.3, which implies the sample complexity required by Phase 1 is \(|\Omega_0| = O(r^2\kappa^2n \log d)\). In addition, according to Theorem 5.4, we have

\[
D^2(Z_S, Z^*) \leq \left(1 - \frac{c_2}{16c^2\kappa}\right)^S \cdot \alpha^2\sigma_p^2 + c_4\delta \cdot r\kappa\sqrt{\sigma_p^2}.
\]

Thus, in order to guarantee the \(O(\sqrt{\sigma_p^2}/n)\) initial ball assumption \(D(Z_S, Z^*) \leq c_0\sqrt{\sigma_p^2}/(\mu_1n)\) holds in Theorem 5.5, it is sufficient to perform \(S = O(r\kappa\log n)\) iterations in Phase 2 of Algorithm 1 and let the approximation error \(\delta = O(1/(r\kappa n^2))\). Based on Theorem 5.4, we derive the sample complexity required by Phase 2 is \(c \cdot \max\{\mu_1n, \mu_0r\kappa\} \mu_0r^2\kappa^2 \log n \log d\). Together with Theorem 5.5, we conclude that the overall sample complexity of Algorithm 1.

Finally, as for the reconstruction error \(\|M^T - M^*\|_F\), let \(R^T\) be the optimal rotation between \(Z^T\) and \(Z^*\), then we have

\[
\|U^T(V^T)^\top - U^*V^*\|_F \leq \|U^T(V^T - V^*R^T)^\top\|_F + \|(U^T - U^*R^T)(V^*R^T)^\top\|_F
\]

\[
\leq \|U^T\|_2 \cdot \|V^T - V^*R^T\|_F + \|V^*R^T\|_2 \cdot \|U^T - U^*R^T\|_F
\]

\[
\leq 3\sqrt{\sigma_p^2} \cdot D(Z^T, Z^*), \quad (B.10)
\]

where the second inequality is due to \(\|AB\|_F \leq \|A\|_2 \cdot \|B\|_F\), and the last inequality follows from the fact that \(\|Z^T - Z^*R^T\|_F \leq D(Z_S, Z^*) \leq \sqrt{\sigma_p^2}(\mu_1n) \leq \sqrt{\sigma_p^2}\). Noticing that according to Theorem 5.5, Phase 3 achieves linear rate of convergence, which implies that with \(T = O(\kappa \log(1/\epsilon))\) iterations, we have \(D(Z^T, Z^*) \leq \epsilon\). Combining with (B.10), we complete the proof.

\[\square\]

C. Proofs of the Technical Lemmas in Section B

In this section, we provide the theoretical proofs of the technical lemmas used in Section B.

C.1. Proof of Lemma B.1

To prove Lemma B.1, we need to make use of the following auxiliary lemmas. Inspired by Tu et al. (2015), we show that the reference function \(G(Z)\) has a similar local curvature property in Lemma C.1. Based on Matrix Bernstein Inequality, Lemma C.2 generalizes the results of Theorem 4.1 in Candès & Recht (2009) to inductive setting, while Lemma C.3 provides the high probability bound on the remaining term.

Lemma C.1. Let \(Z, Z^* \in \mathbb{R}^{(n_1 + n_2)\times r}\), and \(G(Z) = \|Z\|_{F}^2 - Z^*\|Z^*\|_{F}^2/4\). For any \(Z\) satisfying

\[
\|Z - Z^*R\|_F^2 \leq \sigma_p^2(Z^*)/5, \text{ where } R = \text{argmin}_{R \in O_r} \|Z - Z^*R\|_F,
\]

we have

\[
\langle \nabla G(Z), Z - Z^*R \rangle \geq \frac{\sigma_p^2(Z^*)}{4} \|Z - Z^*R\|_F^2 + \frac{1}{4} \|ZZ^\top - ZZ^*\|_F^2.
\]

Lemma C.2. Assume the previously stated assumptions in Lemma B.1 hold. Define subspace

\[
\mathcal{T} = \{ U^T A^\top + B\tilde{V}^\top, \text{ for some } A \in \mathbb{R}^{n_2 \times r} \text{ and } B \in \mathbb{R}^{n_1 \times r} \}.
\]

Let \(\mathcal{P}_T : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^{n_1 \times n_2}\) be the Euclidean projection onto \(\mathcal{T}\). Specifically, for any \(Z \in \mathbb{R}^{n_1 \times n_2}\), we have

\[
\mathcal{P}_T(Z) = U^T U^*Z + \tilde{Z}\tilde{V}^\top - U^T U^*Z\tilde{V}^\top.
\]

(C.1)

For any \(\gamma \in (0, 1)\), there exist constants \(c_1, c_2\) such that if \(|\Omega| \geq c_1\mu_0 \mu_1 r n \log d / \gamma^2\), then for all \(Z \in \mathbb{R}^{n_1 \times n_2}\), with probability at least \(1 - c_2/d\), we have

\[
\|\mathcal{P}_T(Z) - p^{-1}\mathcal{P}_T(X_L^p P_{\Omega}(X_L^p \mathcal{P}_T(Z)X_L^p) X_L^p)\|_F \leq \gamma \|Z\|_F.
\]

(C.2)
Moreover for all $Z_1, Z_2 \in \mathcal{T}$, we have

$$\left\langle \mathbf{X}_L \mathbf{Z}_1 \mathbf{X}_R^T - p^{-1} \mathcal{P}_\Omega(\mathbf{X}_L \mathbf{Z}_1 \mathbf{X}_R^T), \mathbf{X}_L \mathbf{Z}_2 \mathbf{X}_R^T \right\rangle \leq \gamma \|Z_1\|_F \cdot \|Z_2\|_F,$$

and for all $Z \in \mathcal{T}$, we have

$$p^{-1} \left\| \mathcal{P}_\Omega(\mathbf{X}_L \mathbf{Z} \mathbf{X}_R^T) \right\|_F^2 \leq (1 + \gamma) \|Z\|_F^2.$$

**Lemma C.3.** Assume the previously stated assumptions in Lemma B.1 hold. For any fixed $U \in \mathbb{R}^{n_1 \times r}$, $V \in \mathbb{R}^{n_2 \times r}$ satisfying $\|X_L U\|_{2, \infty} \leq 3 \sqrt{\mu_0 \sigma^*_r / d_1}$ and $\|X_R V\|_{2, \infty} \leq 3 \sqrt{\mu_0 \sigma^*_r / d_2}$, there exist constants $c_1, c_2$ such that with probability $1 - c_1 / d$, we have

$$\frac{1}{p} \left\| \mathcal{P}_\Omega(\mathbf{X}_L \mathbf{Z} \mathbf{U}^T \mathbf{X}_R^T) \right\|_F^2 \leq \|\mathbf{U}^T\|_F^2 + \gamma \sigma^*_r \cdot (\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2),$$

provided that $|\Omega| \geq c_2 \max\{\mu_0^2 \kappa^2, \mu_0 \mu_1 \kappa n\} \log d / \gamma^2$.

Now we are ready to prove Lemma B.1.

**Proof of Lemma B.1.** Recall that $R = \arg\min_{\tilde{R} \in \mathbb{R}^r} \|Z - Z^* \tilde{R}\|_F$ and $H = Z - Z^* R$. According to the gradient of $\tilde{f}_\Omega$ in (B.8), we have

$$\langle \nabla \tilde{f}_\Omega(Z) , H \rangle = \frac{1}{2} \langle \nabla G(Z), H \rangle + \frac{1}{2} \langle \tilde{Z}^* \tilde{Z}^T Z, H \rangle + \frac{p^{-1} \left( \mathcal{P}_\Omega - \mathcal{P}_{\text{stat}} \right)(\mathbf{X} [\mathbf{Z} \mathbf{Z}^T - \mathbf{Z}^* \mathbf{Z}^*^T] \mathbf{X}^T), \mathbf{X} \mathbf{H} (\mathbf{X} \mathbf{Z})^T \rangle \tag{I_1}$$

In the sequel, we are going to bound the terms $I_1$ and $I_2$, respectively.

**Lower bound of $I_1$.** According to the assumption $\|H\|_F \leq \sqrt{2} \sigma^*_r / \sqrt{p}$, we have $\|Z - Z^* R\|_2^2 \leq \sigma^2_r (Z^*) / 5$. Thus, we can apply Lemma C.1 directly

$$I_1 = \frac{1}{2} \langle \nabla G(Z), H \rangle + \frac{1}{2} \text{tr}(Z^T \tilde{Z}^* \tilde{Z}^T Z) \geq \frac{\sigma_r^2 \langle Z^* \rangle}{8} \|H\|_F^2 \geq \frac{1}{8} \|Z^* Z^T - \tilde{Z}^* \tilde{Z}^T \|_F^2 \geq \frac{1}{2} \|\tilde{Z}^* \tilde{Z}^T Z\|_F^2$$

where the first inequality holds because $Z^* Z^T \tilde{Z}^* = 0$, the second inequality follows from Lemma C.1 and the fact that $\|Z^* Z^T\|_F^2 \leq \|Z^*\|_2^2 \cdot \|Z^T\|_2^2$, and the third equality holds because $\sigma^2_r (Z^*) = 2 \sigma^*_r$ and $\|\tilde{Z}^*\|_2^2 = 2 \sigma^*_r$.

**Upper bound of $|I_2|$.** Note that $Z = [U; V]$ and $M^* = U^* V^*^T$, we denote $M = U V^T$, $\tilde{U} = U^* R$ and $\tilde{V} = V^* R$. Besides, let $H_U \in \mathbb{R}^{n_1 \times \kappa}$, $H_V \in \mathbb{R}^{n_2 \times \kappa}$ be the top $n_1$ and bottom $n_2$ rows of $H$, respectively, then we have $U = \tilde{U} + H_U$ and $V = \tilde{V} + H_V$. Note that $U V^T - \tilde{U} \tilde{V}^T = \tilde{U} H_V^T + H_U \tilde{V}^T + H_U H_V^T$, and $H_U V^T + \tilde{U} H_V^T = \tilde{U} H_V^T + H_U \tilde{V}^T + 2 H_U H_V^T$. Based on the above notations, we can reformulate $I_2$ as follows

$$I_2 = \langle (p^{-1} \mathcal{P}_\Omega - \mathcal{I})(\mathbf{X}_L [\mathbf{U} \mathbf{V}^T - \tilde{U} \tilde{V}^T] \mathbf{X}_R^T), \mathbf{X}_L (H_U \mathbf{V}^T + \tilde{U} H_V^T X_R) \rangle$$

and

$$= \langle \mathbf{X}_L [\mathbf{U} \mathbf{V}^T - \tilde{U} \tilde{V}^T] \mathbf{X}_R^T, \mathbf{X}_L \mathbf{H}_U \mathbf{V}^T \mathbf{X}_R^T \rangle \tag{I_{21}}$$

$$+ 3 \langle (p^{-1} \mathcal{P}_\Omega - \mathcal{I})(\mathbf{X}_L [\tilde{U} \mathbf{H}_V^T + H_U \tilde{V}^T] \mathbf{X}_R^T), \mathbf{X}_L \mathbf{H}_U \mathbf{H}_V^T \mathbf{X}_R^T \rangle \tag{I_{22}}$$

and

$$+ 2 \langle (p^{-1} \mathcal{P}_\Omega - \mathcal{I})(\mathbf{X}_L \mathbf{H}_U \mathbf{V}^T \mathbf{X}_R^T), \mathbf{X}_L \mathbf{H}_U \mathbf{H}_V^T \mathbf{X}_R^T \rangle \tag{I_{23}}.$$
Note that $\bar{U}_V^T + H_U \bar{V}^T$ falls into the subspace $\mathcal{T}$ defined in Lemma C.2. Thus according to Lemma C.2, we obtain
\begin{equation}
|I_{21}| \leq \gamma \|\bar{U}_V^T + H_U \bar{V}^T\|_F^2 \leq 2\gamma \|M - M^*\|_F^2 + \frac{\gamma}{2} \|H\|_F^2 \tag{C.4}
\end{equation}
holds with probability at least $1 - c_2/d$, provided that $|\Omega| \geq c_1 n \log d / \gamma^2$, where $\gamma$ is a constant such that $\gamma \in (0, 1)$. Here, the second inequality holds because $\|A + B\|_F^2 \leq 2\|A\|_F^2 + 2\|B\|_F^2$ and $\|H_U H_U^T\|_F \leq \|H\|_F^2 / 2$. As for the term $I_{22}$, we have
\begin{align}
|I_{22}| &\leq \frac{1}{p} \|\langle \Omega, (\bar{U}_V^T + H_U \bar{V}^T)X^\dagger \rangle, X_L H_U H_U^T X_{R}^{-}\rangle + \langle X_L (\bar{U}_V^T + H_U \bar{V}^T)X^\dagger, X_L H_U H_U^T X_{R}^{-}\rangle \|
\leq \frac{1}{2\beta} \|P_{\Omega}(X_L \bar{U}_V^T + H_U \bar{V}^T)X^\dagger\|_F^2 + \frac{\beta}{4} \|P_{\Omega}(X_L H_U H_U^T X_{R}^{-}\rangle\|_F^2
\end{align}
where constant $\beta > 0$ will be specified later. Here, the second inequality follows from the Young’s Inequality, and the last inequality is due to Lemma C.2 and the fact that $\|H_U H_U^T\|_F \leq \|H_U\|_F \leq \|H\|_F^2 / 2$. Since we have $\|X_L H_U\|_{2,\infty} \leq \|X_L U\|_{2,\infty} + \|X_L \bar{U}\|_{2,\infty} \leq 3\sqrt{\mu_0 r \sigma_1^4 / d_1}$, and similarly we have $\|X_R H_U\|_{2,\infty} \leq 3\sqrt{\mu_0 r \sigma_1^4 / d_2}$, according to Lemma C.3, we further obtain
\begin{equation}
\frac{1}{p} \|P_{\Omega}(X_L H_U H_U^T X_{R}^{-}\rangle\|_F^2 \leq \|H_U H_V\|_F^2 + \gamma \sigma_r^* \|H\|_F^2 \leq \frac{1}{4} \|H\|_F^2 + \gamma \sigma_r^* \|H\|_F^2 \tag{C.6}
\end{equation}
holds with probability at least $1 - c_1 / d$, provided that $|\Omega| \geq c_2 \max\{\mu_0 r^2 \kappa^2, \mu_0 \mu_1 \rho \kappa n\} \log d / \gamma^2$. Thus, plugging (C.6) into (C.5), we have
\begin{align}
|I_{22}| &\leq \frac{\beta}{2} \|M - M^*\|_F^2 + \frac{2\gamma}{2\beta} \|\bar{U}_V^T + H_U \bar{V}^T\|_F^2 + \frac{\beta}{4} \|H\|_F^2
\leq \frac{\beta}{2} \|M - M^*\|_F^2 + \frac{\beta}{2} \|\bar{U}_V^T + H_U \bar{V}^T\|_F^2 + \frac{\beta}{4} \|H\|_F^2, \tag{C.7}
\end{align}
where $M = UV^T$. Similarly, according to Lemma C.3, we can upper bound $|I_{23}|$
\begin{align}
|I_{23}| &\leq \frac{1}{p} \|P_{\Omega}(X_L H_U H_U^T X_{R}^{-}\rangle, X_L H_U H_U^T X_{R}^{-}\rangle + \langle X_L H_U H_U^T X_{R}^{-}\rangle, X_L H_U H_U^T X_{R}^{-}\rangle\|
\leq \frac{1}{p} \|P_{\Omega}(X_L H_U H_U^T X_{R}^{-}\rangle\|_F^2 + \|H_U H_U^T\|_F^2
\leq \frac{1}{2} \|H\|_F^2 + \gamma \sigma_r^* \|H\|_F^2 \tag{C.8}
\end{align}
Therefore, combining (C.4), (C.7) and (C.8), we obtain the upper bound of $|I_2|
\begin{equation}
|I_2| \leq |I_{21}| + 3|I_{22}| + 2|I_{23}|
\leq \left(2\gamma + \frac{3(2 + \gamma)}{\beta}\right) \|M - M^*\|_F^2 + \frac{\gamma}{2} \|\bar{U}_V^T + H_U \bar{V}^T\|_F^2 + \frac{3(2 + \gamma)}{4} \|H\|_F^2 + \frac{3\beta}{2} \|H\|_F^2, \tag{C.9}
\end{equation}
Finally, set $\beta = 48$, then combining (C.3) and (C.9), we obtain
\begin{equation}
\langle \nabla f_0(Z), H \rangle \geq \left(1 - \frac{3\gamma}{2}\right) \|ZZ^T - Z^* Z^*\|_F^2 + \frac{1}{4\sigma_r^*} \|\tilde{Z}^T \tilde{Z}^T Z\|_F^2 + \left(\frac{\sigma_r^*}{4} - 75 \gamma \sigma_r^*\right) \|H\|_F^2 - 40 \|H\|_F^2,
\end{equation}
where the inequality holds because $\|M - M^*\|_F^2 \leq 40 \|H\|_F^2$. Thus, by choosing constant $\gamma$ to be sufficiently small, we complete the proof.
C.2. Proof of Lemma B.2

Proof. Recall the gradient of \( \tilde{f}_\Omega \) in (B.3), we have

\[
\| \nabla \tilde{f}_\Omega(Z) \|_F^2 \leq 2 \left( \sum_{i=1}^{p_1} (p_1)^{1/2} \| X_i \|_F \right)^2 + \frac{1}{2} \left( \| P - P_{\text{off}} \|_F \right)^2 \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \| X_i X_j^T - X_i X_j^T \|_F^2,
\]

where the inequality holds because \( \| A + B \|_F^2 \leq 2 (\| A \|_F^2 + \| B \|_F^2) \). In the sequel, we will upper bound \( I_1 \) and \( I_2 \), respectively.

Upper bound of \( I_1 \). Recall that \( Z = [U; V] \), \( R = \arg \min_{R \in \mathbb{Q}_r} \| Z - Z^* R \|_F \), and \( H = Z - Z \cdot R \). Denote \( M = U V^T \), then based on the above notations, we have

\[
I_1 = \| p^{-1} X_L^T P_{\Omega}(X_L \| M - M^* \| X_R^H) X_R V \|_F^2 + \| p^{-1} X_L^T (P_{\Omega}(X_L \| M - M^* \| X_R^T)^T X_L U \|_F^2
\]

\[
\leq r \cdot \| p^{-1} X_L^T P_{\Omega}(X_L \| M - M^* \| X_R^H) X_R V \|_F^2 + r \cdot \| p^{-1} X_L^T (P_{\Omega}(X_L \| M - M^* \| X_R^T)^T X_L U \|_F^2,
\]

where the inequality holds because both \( U \) and \( V \) have rank at most \( r \). Consider the term \( I_{11} \) first, we observe

\[
\frac{1}{p} X_L^T P_{\Omega}(X_L \| M - M^* \| X_R^H) X_R V \|_F^2 = \sum_{(i,j) : (i,j) = (1,1)} (p^{-1} \xi_{ij} - 1) \cdot \| X_L \|_{M - M^* \| X_R^H} \|_{ij} \cdot \| X_L \|_{M - M^* \| X_R^T} \|_{ij} \cdot \| X_R V \|_{ij} \cdot := \sum_{(i,j) = (1,1)} A_{ij},
\]

where \( \xi_{ij} = 1 \) if \( (i,j) \in \Omega \); \( \xi_{ij} = 0 \), otherwise. We are going to apply matrix bernstein inequality to the above summation. Due to sample splitting, the randomness only comes from \( \Omega \), thus \( A_{ij} \)'s are independent and \( \mathbb{E}[A_{ij}] = 0 \). Denote \( U = U^* R \) and \( V = V^* R \), and let \( H_U \in \mathbb{R}^{n_1 \times r} \), \( H_V \in \mathbb{R}^{n_2 \times r} \) be the top \( n_1 \) and bottom \( n_2 \) rows of \( H \). Then we have \( M - M^* = U H_V^T + H_U V^T + H_U H_V \). For any \( (i,j) \), we have the following upper bound of \( \| A_{ij} \|_2 \)

\[
\| A_{ij} \|_2 \leq \frac{1}{p} \left( \| X_L \|_{M - M^* \| X_R^H} \|_{ij} \cdot \| X_L \|_{M - M^* \| X_R^T} \|_{ij} \cdot \| X_R V \|_{ij} \right),
\]

\[
\leq \frac{1}{p} \left( \| X_L \|_{M - M^* \| X_R^H} \|_{ij} \cdot \| X_L \|_{M - M^* \| X_R^T} \|_{ij} \cdot \| X_R V \|_{ij} \right) \leq \frac{10 \mu_0 \mu_1 r \sigma_1^*}{pd_1 d_2} \cdot \| H \|_F^2,
\]

where the last inequality is due to Assumptions 3.1, 3.2 and the fact that \( \| X_L H_U \|_{2,\infty} \leq \| X_L U \|_{2,\infty} + \| X_L \|_{2,\infty} \leq 3 \sqrt{\mu_0 \sigma_1^* / d_1} \) and \( \| X_R V \|_{2,\infty} \leq 2 \sqrt{\mu_0 \sigma_1^* / d_2} \). To apply matrix bernstein, it remains to bound \( \| \sum_{i,j} \mathbb{E}[A_{ij}] A_{ij} \|_2 \) and \( \| \sum_{i,j} \mathbb{E}[A_{ij}] A_{ij} \|_2 \). In particular, we have

\[
\left\| \sum_{(i,j) = (1,1)} \mathbb{E}[A_{ij}] A_{ij} \right\|_2 \leq \frac{1}{p} \left( \sum_{(i,j) = (1,1)} \| X_L \|_{M - M^* \| X_R^H} \|_{ij} \cdot \| X_L \|_{M - M^* \| X_R^T} \|_{ij} \cdot \| X_R V \|_{ij} \right),
\]

\[
\leq \frac{1}{p} \left( \sum_{i=1}^{d_1} \| X_L \|_{M - M^* \| X_R^H} \|_{ij} \cdot \| X_L \|_{M - M^* \| X_R^T} \|_{ij} \cdot \| X_R V \|_{ij} \right),
\]

\[
\leq \frac{1}{p} \max_{i \in [d_1]} \left\| \sum_{j=1}^{d_2} \| X_L \|_{M - M^* \| X_R^H} \|_{ij} \cdot \| X_R V \|_{ij} \right\|_2,\infty \]

\[
\leq \frac{4 \mu_0 \mu_1 r \sigma_1^*}{pd_1 d_2} \cdot \| X_L \|_{M - M^* \| X_R^H} \|_{ij} \cdot \| X_R V \|_{ij} \|_{2,\infty} \leq \frac{c \mu_0 \mu_1 r \sigma_1^* n}{pd_1 d_2} \cdot \| M - M^* \|_F^2,
\]
where the last inequality holds because \( \|AB\|_{2,\infty} \leq \|A\|_{2,\infty} \cdot \|B\|_F \). Similarly, we have

\[
\left\| \sum_{(i,j) = (1,1)}^{(d_1,d_2)} \mathbf{E}[A_{ij}^T A_{ij}] \right\|_2 \leq \frac{1}{p} \left\| \sum_{(i,j) = (1,1)}^{(d_1,d_2)} \|X_L(M - M^*)X_{ij}^\top\|_2 \cdot \|X_R V\|_{j^*} \cdot \|X_L\|_{i^*} \cdot \|X_R V\|_{j^*} \right\|_2 \\
\leq \frac{1}{p} \|X_L\|_{2,\infty} \cdot \|X_R V\|_{2,\infty} \cdot \sum_{(i,j) = (1,1)}^{(d_1,d_2)} \|X_L(M - M^*)X_{ij}^\top\|_2 \\
\leq \frac{4 \mu_0 \mu_1 \sigma^*_n}{pd_1 d_2} \cdot \|M - M^*\|_F^2,
\]

where the second inequality follows from the definition of spectral norm. Therefore, according to Lemma E.1, with probability at least 1 - c/d, we have

\[
\left\| \sum_{(i,j) = (1,1)}^{(d_1,d_2)} A_{ij} \right\|_2 \leq \gamma \sigma^*_r^{1/2} \|M - M^*\|_F + \gamma \sigma^*_r \|H\|_F,
\]

under condition that \( |\Omega| \geq c' \mu_0 \mu_1 \alpha n c/\gamma^2 \), where \( c, c' \) are both constants. Hence, by triangle’s inequality, we obtain the upper bound of \( I_{11} \)

\[
I_{11} \leq (\|M - M^*\|V\|_2 + \gamma \sigma^*_r^{1/2} \|M - M^*\|_F + \gamma \sigma^*_r \|H\|_F)^2 \leq 8 \sigma^*_r M - M^*\|_F^2 + \sigma^*_r^2 \|H\|_F^2,
\]

where the last inequality holds because \( \|V\|_2 \leq \|V^*\|_2 + \|V - V^* R\|_2 \leq 2 \sqrt{\sigma^*_r} \) and \( \gamma \in (0,1/2) \). Similarly, we obtain the upper bound of \( I_{12} \)

\[
I_{12} \leq (\|U^T (M - M^*)\|_2 + \gamma \sigma^*_r^{1/2} \|M - M^*\|_F + \gamma \sigma^*_r^2 \|H\|_F)^2 \leq 8 \sigma^*_r M - M^*\|_F^2 + \sigma^*_r^2 \|H\|_F^2.
\]

Plugging (C.11) and (C.12) into (C.10), we have

\[
I_1 \leq 2 \cdot (8 \sigma^*_r M - M^*\|_F^2 + \sigma^*_r^2 \|H\|_F^2).
\]

Upper bound of \( I_2 \). As for \( I_2 \), we obtain

\[
I_2 = \left\| \tilde{Z}^* \tilde{Z}^\top Z - (\overline{P}_{\text{diag}} - \overline{P}_{\text{off}})(\tilde{Z}^* \tilde{Z}^\top Z - ZZ^\top)Z \right\|_F^2 \\
\leq 2 \left\| \tilde{Z}^* \tilde{Z}^\top Z \right\|_F^2 + 2 \left( \|\overline{P}_{\text{diag}} - \overline{P}_{\text{off}}\|_{\text{F}} \right) \|\tilde{Z}^* \tilde{Z}^\top Z - ZZ^\top\|_F \cdot \|Z\|_2^2 \\
\leq 2 \left\| \tilde{Z}^* \tilde{Z}^\top Z \right\|_F^2 + 8 \sigma^*_r \|ZZ^\top - Z^* Z^\top\|_F^2,
\]

where the first inequality holds because \( \|A - B\|_F^2 \leq 2 (\|A\|_F^2 + \|B\|_F^2) \) and \( \|AB\|_F \leq \|A\|_F \cdot \|B\|_2 \), and the second inequality holds because \( \|Z\|_2 \leq 2 \sqrt{\sigma^*_r} \).

Finally, combining (C.13) and (C.14), we obtain

\[
\| \nabla f_{\Omega}(Z) \|_F^2 \leq 32 \sigma^*_r \|M - M^*\|_F^2 + 4 \sigma^*_r^2 \|H\|_F^2 + \|\tilde{Z}^* \tilde{Z}^\top Z\|_F^2 + 4 \sigma^*_r \|ZZ^\top - Z^* Z^\top\|_F^2 \\
\leq (16 \sigma^*_r^2 \|ZZ^\top - Z^* Z^\top\|_F^2 + \|\tilde{Z}^* \tilde{Z}^\top Z\|_F^2 + 4 \sigma^*_r^2 \|H\|_F^2,
\]

where the second inequality holds because \( \|M - M^*\|_F^2 \leq \|ZZ^\top - Z^* Z^\top\|_F^2/2 \), which completes the proof.

C.3. Proof of Lemma B.3

**Proof of Lemma B.3.** Similar to the proof of Lemma B.1, we have

\[
\langle \nabla f_{\Omega}(Z), H \rangle = \frac{1}{2} \langle \nabla G(Z), H \rangle + \frac{1}{2} \langle \tilde{Z}^* \tilde{Z}^\top Z, H \rangle + \langle (p^{-1} \overline{P}_{\text{diag}} - \overline{P}_{\text{off}})(X [ZZ^\top - Z^* Z^\top] X^\top), XH(XZ)^\top \rangle.
\]
According to Lemma C.1, we have
\[
I_1 \geq \frac{\sigma_r^2}{4} \|H\|_F^2 + \frac{1}{8} \|ZZ^T - Z^*Z^{*T}\|_F^2 + \frac{1}{4\sigma_1^2} \|\tilde{Z}^*\|_F^2. \tag{C.15}
\]

Recall the notations in the proof of Lemma B.1, we let \(Z = [U; V]\) and \(M^* = UV^*\), we denote \(M = UV^T\), \(\tilde{U} = U^r\) and \(\tilde{V} = V^r\). Besides, let \(H_U \in \mathbb{R}^{n_1 \times r}\), \(H_V \in \mathbb{R}^{n_2 \times r}\) be the top \(n_1\) and bottom \(n_2\) rows of \(H\). Then we can reformulate \(I_2\) as follows
\[
I_2 = \langle (p^{-1}P_\Omega(I)X_L(\tilde{U}H_V^T + H_U\tilde{V}^T)X_R), X_L(\tilde{U}H_V^T + H_U\tilde{V}^T)X_R^T) \rangle_{I_21}
+ 3 \langle (p^{-1}P_\Omega(I)(\tilde{U}H_V^T + H_U\tilde{V}^T)X_R), X_L(\tilde{U}H_V^T + H_U\tilde{V}^T)X_R^T) \rangle_{I_22}
+ 2 \langle (p^{-1}P_\Omega(I)X_LH_UH_V^T X_R^T), X_LH_UH_V^T X_R^T \rangle_{I_23}.
\]

Note that \(\tilde{U}H_V^T + H_U\tilde{V}^T\) falls into the subspace \(\mathcal{T}\) defined in Lemma C.2. Thus according to Lemma C.2, we can still obtain the same upper bound of \(I_{21}\)
\[
|I_{21}| \leq \gamma \|\tilde{U}H_V^T + H_U\tilde{V}^T\|_F^2 \leq 2\gamma \|M - M^*\|_F^2 + \frac{\gamma}{2} \|H\|_F^4
\]
holds with probability at least \(1 - c_2/d\), provided that \(|\Omega| \geq c_1 \mu_0 \mu_1 r n \log d/\gamma^2\). As for the term \(I_{22}\), similarly we have
\[
|I_{22}| \leq \frac{1}{p} \langle P_\Omega(I)(X_L\tilde{U}H_V^T + H_U\tilde{V}^T)X_R), X_LH_UH_V^T X_R^T \rangle + |\langle X_L(\tilde{U}H_V^T + H_U\tilde{V}^T)X_R^T, X_LH_UH_V^T X_R^T \rangle|
\leq \frac{\beta}{2} \cdot \frac{1}{p} \|P_\Omega(I)X_LH_UH_V^T X_R\|_F^2 + \frac{2 + \gamma}{2\beta} \|\tilde{U}H_V^T + H_U\tilde{V}^T\|_F^2 + \frac{\beta}{8} \|H\|_F^4,
\]
where constant \(\beta > 0\) will be specified later. Here, the second inequality follows from the Young’s Inequality and Lemma C.2. According to bernstein-type inequality for Bernoulli random variables, we further obtain
\[
\frac{1}{p} \|P_\Omega(I)(X_LH_UH_V^T X_R^T)\|_F^2 \leq \frac{1}{p} \|\Omega\| \cdot \|X_LH_UH_V^T X_R^T\|_{\infty, \infty}^2 \leq \frac{3}{2} \mu_1^2 n_1 n_2 \cdot \|H_UH_V^T\|_F^2 \leq \beta_0 \sigma_r^* \|H\|_F^2
\]
holds with probability at least \(1 - c_2/d\), where the second inequality follows from Lemma E.2 and the incoherence Assumptions 3.2, and the last inequality holds because \(\|H\|_F \leq D^2(\tilde{Z}, Z^*) \leq \beta_0 \sigma_r^*/(\mu_1 n^2)\). Therefore, we obtain the upper bound of \(I_{22}\)
\[
|I_{22}| \leq \frac{2 + \gamma}{2\beta} \|\tilde{U}H_V^T + H_U\tilde{V}^T\|_F^2 + \frac{\beta}{2} \cdot \beta_0 \sigma_r^* \|H\|_F^2 + \frac{\beta}{8} \|H\|_F^4
\leq \frac{2 + \gamma}{\beta} \|M - M^*\|_F^2 + \frac{\beta}{2} \cdot \beta_0 \sigma_r^* \|H\|_F^2 + \left(\frac{2 + \gamma}{4\beta} + \frac{\beta}{8}\right) \|H\|_F^4.
\]

Similarly, we can upper bound \(I_{23}\)
\[
|I_{23}| \leq \frac{1}{p} \|P_\Omega(I)(X_LH_UH_V^T X_R^T)\|_F^2 + \|H_UH_V^T\|_F^2 \leq \beta_0 \sigma_r^* \|H\|_F^2 + \frac{1}{4} \|H\|_F^4.
\]

Hence, we obtain
\[
|I_2| \leq |I_{21}| + 3|I_{22}| + 2|I_{23}|
\leq \left(2\gamma + \frac{3(2 + \gamma)}{\beta}\right) \cdot \|M - M^*\|_F^2 + \left(2 + \frac{3\beta}{2}\right) \cdot \beta_0 \sigma_r^* \|H\|_F^2 + \left(\frac{3(2 + \gamma)}{4\beta} + \frac{3\beta}{8} + \frac{1}{2}\right) \cdot \|H\|_F^4. \tag{C.16}
\]
Set \(\beta = 48\), and let constants \(c_0, \gamma\) be sufficiently small. Combining (C.15) and (C.16), we complete the proof. \(\square\)
C.4. Proof of Lemma B.4

In order to prove Lemma B.4, we need to make use of the following lemma.

**Lemma C.4.** Let $X \in \mathbb{R}^{d_1 \times n_1}$, $Y \in \mathbb{R}^{d_2 \times n_2}$ be the feature matrices, which are orthonormal and self-incoherent with parameter $\mu_1$, and $\Omega \subseteq [d_1] \times [d_2]$ be an index set followed Bernoulli Model (3.1) with $p = |\Omega|/(d_1 d_2)$. For any $\gamma \in (0, 1)$, there exist constants $c_1$ and $c_2$ such that, under condition $|\Omega| \geq c_1 \mu_1 n \log d/\gamma^2$, for all $Z \in \mathbb{R}^{n_1 \times n_2}$

$$
\|p^{-1}X^T \mathcal{P}_\Omega (XZ\xi^T)Y\|_F \leq (1 + \mu_1 n \gamma) \|Z\|_F,
$$

holds with probability at least $1 - c_2/d$.

Now we are ready to prove Lemma B.4.

**Proof of Lemma B.4.** According to the gradient of $\tilde{f}_\Omega (Z)$ in (B.3), we have

$$
\left\| \nabla \tilde{f}_\Omega (Z) \right\|_F^2 \leq 2 \left\| p^{-1}X^T \mathcal{P}_\Omega (X[ZZ^T - Z^* Z^{* \top}]X^T) XX^T \right\|_F^2 + \frac{1}{2} \left( \| \mathcal{P}_{\text{diag}} - \mathcal{P}_{\text{off}} \| \|Z^T Z\|_F^2 \right).
$$

In the following discussions, we will upper bound $I_1$ and $I_2$ respectively. As for $I_1$, similar to the proof of Lemma B.2, we have

$$
I_1 = \left| p^{-1}X^T \mathcal{P}_\Omega (X[L - M^\top]X^T)X^T \right\|_F^2 + \left| p^{-1}X^T \mathcal{P}_\Omega (X[L - M^\top]X^T)X^T \right\|_F^2,
$$

Consider the first term $I_{11}$. Note that $M - M^\top = H_U \tilde{V}^\top + \tilde{U} H_V^\top + H_V H_V^\top$. Recall the definition of $\mathcal{P}_T$ in Lemma C.2, we have $\mathcal{P}_{\text{T}} (M) + \mathcal{P}_{\text{T}} (M) = M$, for any $M \in \mathbb{R}^{n_1 \times n_2}$, where $\mathcal{P}_{\text{T}} (M) = (I - UU^\top)M(I - VV^\top)$. Thus we obtain

$$
I_{11} \leq 2 \left| p^{-1}X^T \mathcal{P}_\Omega (X[L - M^\top]X^T)X^T \right\|_F^2 + 2 \left| p^{-1}X^T \mathcal{P}_\Omega (X[L - M^\top]X^T)X^T \right\|_F^2
\leq 4 \left| p^{-1}X^T \mathcal{P}_\Omega (X[L - M^\top]X^T)X^T \right\|_F^2
\leq 4 \left| p^{-1}X^T \mathcal{P}_\Omega (X[L - M^\top]X^T)X^T \right\|_F^2
+ 4 \left| p^{-1}X^T \mathcal{P}_\Omega (X[L - M^\top]X^T)X^T \right\|_F^2,
$$

where the inequality follows from $\|A + B\|_F^2 \leq 2 \|A\|_F^2 + 2 \|B\|_F^2$, and the equality is due to the definition of $\mathcal{P}_{\text{T}}$ and the fact that $(I - \tilde{V}^\top \tilde{V}^\top) \tilde{V} = 0$. In the sequel, we will upper bound the terms $J_1$, $J_2$ and $J_3$ respectively. Note that $\tilde{U} H_V^\top + H_U \tilde{V}^\top \mathcal{T} \leq 0$, according to Lemma C.2, with probability at least $1 - c'/d$, we have

$$
J_1 \leq (1 + \gamma)^2 \| \tilde{U} H_V^\top + H_U \tilde{V}^\top \|_F^2 \leq 4 (1 + \gamma)^2 \sigma_2^\top \cdot \| \tilde{U} H_V^\top + H_U \tilde{V}^\top \|_F^2,
$$

provided that $|\Omega| \geq c_1 \mu_1 n \log d/\gamma^2$, where the first inequality holds because $\|AB\|_F \leq \|A\|_F \cdot \|B\|_2$, and the second inequality follows from the fact that $\|H\|_F \leq \|H^\top\|_2 + \|H_V \|_2 \leq 2 \sqrt{\sigma_2^\top}$. Consider the second term $J_2$, we have

$$
J_2 \leq \left| p^{-1}X^T \mathcal{P}_\Omega (X[L - M^\top]X^T)X^T \right\|_F \| H_V \|_F^2
\leq \left| p^{-1}X^T \mathcal{P}_\Omega (X[\tilde{U} H_V^\top + H_U \tilde{V}^\top]X^T)X^T \right\|_F \| H_V \|_F^2
\leq (1 + \mu_1 n \gamma)^2 \cdot \| \tilde{U} H_V^\top + H_U \tilde{V}^\top \|_F^2 \| H_V \|_F^2 \leq 2 \gamma^2 \sigma_2^\top \cdot \| H_V \|_F^2 \leq 2 \gamma^2 \sigma_2^\top \cdot \| H_V \|_F^2 \leq 2 \gamma^2 \sigma_2^\top \| H_V \|_F^2,
$$

where the first inequality holds because $\|AB\|_F \leq \|A\|_F \cdot \|B\|_2$, the second inequality is due to the non-expansiveness of $\mathcal{P}_{\text{T}}$, the third inequality follows from Lemma C.4, and the last inequality is due to $\|H\|_F \leq c_0 \sqrt{\sigma_2^\top}/(\mu_1 n)$. According to Lemma C.4, we can upper bound the last term $J_3$ as follows

$$
J_3 \leq \left| p^{-1}X^T \mathcal{P}_\Omega (X[L - M^\top]X^T)X^T \right\|_F \| V \|_2
\leq (1 + \mu_1 n \gamma)^2 \cdot \| H_U \|_F^2 \| V \|_2 \leq 2 \gamma^2 \sigma_2^\top \| H \|_F^2,
$$
where the second inequality follows from Lemma C.4, and the last inequality is due to \(\|H'h'V\|_F \leq \|H\|^2_2/2\) and \(\|H\|_F \leq \sqrt{\sigma_1^2}/(\mu_{111})\). Therefore, we obtain the upper bound of \(I_{11}\)

\[
I_{11} \leq 4J_1 + 4J_2 + 2J_3 \leq 20\sigma_1^2\|UH'_V + H'h'V\|^2_2 + \sigma_1^4\|H\|^2_2
\]

\[
\leq 40\sigma_1^4\|M - M'\|^2_2 + 10\sigma_1^4\|H\|^2_2 + \sigma_1^4\|H\|^2_2,
\]

where we set \(\gamma\) to be small enough in the first inequality, and the last inequality follows from \(\|A + B\|^2_2 \leq 2\|A\|^2_2 + 2\|B\|^2_2\).

By symmetry, we can use the same techniques to bound the term \(I_{12}\), which will yields the same upper bound and implies

\[
I_1 \leq 80\sigma_1^4\|M - M'\|^2_2 + 20\sigma_1^4\|H\|^2_2 + 2\sigma_1^4\|H\|^2_2.
\]

The upper bound of the remaining term \(I_2\) is as follows

\[
I_2 \leq 2\|Z^*Z^*^T Z\|^2_2 + 8\sigma_1^4\|ZZ^T - Z^*Z^*^T\|^2_2,
\]

where the detailed proof can be found in the proof of Lemma B.2. Hence, we obtain

\[
\|\nabla f_1(Z)\|_2^2 \leq 2I_1 + \frac{1}{2}I_2 \leq 84\sigma_1^4\|ZZ^T - Z^*Z^*^T\|^2_2 + \|Z^*Z^*^T Z\|^2_2 + 4\sigma_1^4\|H\|^2_2 + 40\sigma_1^4\|H\|^2_2,
\]

which completes the proof.

\[\square\]

**D. Proof of Technical Lemmas**

In this section, we prove the technical lemmas used in Section C.

**D.1. Proof of Lemma C.1**

*Proof.* We begin our proof with some properties regarding the optimal solution. Let \(H = Z - Z^* R\) and \(A\Sigma B^T\) be the SVD of \(Z^T Z\), then we have \(R = AB^T\). Thus,

\[
Z^T Z^* R = B\Sigma B^T = (Z^* R)^T Z,
\]

which implies that \(Z^T Z^* R\) is symmetric and positive definite. Moreover, we have

\[
\]

which implies \(H^T Z^* R\) is also symmetric. Without loss of generality, we assume \(R = I\), then \(Z^T Z^*\) is positive definite, and \(H^T Z^*\) is symmetric. Thus, to prove Lemma C.1, it is sufficient to prove

\[
\langle (Z^* H^T + HZ^*^T + HH^T)(Z^* + H), H \rangle \geq \frac{\sigma_1^2(Z^*)}{4} \|H\|^2_2 + \frac{1}{4} \|Z^* H^T + HZ^*^T + HH^*\|^2_2,
\]

which is equivalent to

\[
0 \leq \text{tr} \left( (H^T H)^2 + 3H^T HH^T Z^* + (H^T Z^*)^2 + H^T HZ^*^T Z^* - \frac{1}{4} (H^T H)^2 + 4H^T HH^T Z^* + 2(H^T Z^*)^2 + 2H^T HZ^*^T Z^* - \frac{\sigma_1^2(Z^*)}{4} H^T H \right).
\]

Combining terms, we have

\[
0 \leq \text{tr} \left( \frac{3}{4} (H^T H)^2 + 2H^T HH^T Z^* + \frac{1}{2} (H^T Z^*)^2 + \frac{1}{2} H^T HZ^*^T Z^* - \frac{\sigma_1^2(Z^*)}{4} H^T H \right),
\]

which is further equivalent to

\[
0 \leq \text{tr} \left( \frac{1}{2} (H^T Z^* + 2H^T H)^2 - \frac{5}{4} (H^T H)^2 + \frac{1}{2} H^T HZ^*^T Z^* - \frac{\sigma_1^2(Z^*)}{4} H^T H \right). \tag{D.1}
\]

Note that \(\text{tr}(H^T H)^2 = \|H^T H\|^2_2 \leq \|H\|^2_2 \cdot \|H\|^2_2\), and \(\text{tr}(H^T HZ^*^T Z^*) \geq \sigma_1^2(Z^*) \cdot \|H\|^2_2\). Therefore, in order to prove (D.1), it is sufficient to require that

\[
\|H\|^2_2 \leq \frac{\sigma_1^2(Z^*)}{5},
\]

which completes the proof.

\[\square\]
D.2. Proof of Lemma C.2

Proof. For any \(i \in [d_1]\), denote \(X_{i,*} \in \mathbb{R}^{1 \times n_1}\) as the \(i\)-th row of \(X\). Similarly, for any \(j \in [d_2]\), denote \(Y_{j,*} \in \mathbb{R}^{1 \times n_2}\) as the \(j\)-th row of \(Y\). Thus, for any \(Z \in \mathbb{R}^{n_1 \times n_2}\), we have

\[
X^T P_\Omega (XP_T(Z)Y^T)Y = \sum_{(i,j) \in \Omega} \langle XP_T(Z)Y^T, e_i e_j^T \rangle X^T e_i e_j^T Y
\]

\[
= \sum_{(i,j) \in \Omega} \langle P_T(Z), X_{i,*} Y_{j,*} \rangle X_{i,*} Y_{j,*}
\]

\[
= \sum_{(i,j) \in \Omega} \langle Z, P_T(X_{i,*}^T Y_{j,*}) \rangle X_{i,*} Y_{j,*},
\]

where the last equality holds because \(\langle P_T(A), B \rangle = \langle A, P_T(B) \rangle\). Besides, for any \((i, j) \in [d_1] \times [d_2], \) let \(\xi_{ij} = 1\), if \((i, j) \in \Omega\), and zero otherwise. Note that both \(X\) and \(Y\) are orthonormal, thus according to (D.2), we have

\[
P_T(Z) - p^{-1} P_T \left(X^T P_\Omega (XP_T(Z)Y^T)Y \right) = P_T \left(X^T (I - p^{-1} P_\Omega) (XP_T(Z)Y^T)Y \right)
\]

\[
= \sum_{(i,j) \in [d_1] \times [d_2]} (1 - p^{-1} \xi_{ij}) \langle Z, P_T(X_{i,*}^T Y_{j,*}) \rangle P_T(X_{i,*}^T Y_{j,*}),
\]

where \(I : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}\) is the identity mapping. For any \((i, j) \in [d_1] \times [d_2], \) define linear operator \(S_{ij} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}, \) such that

\[
S_{ij}(Z) = (1 - p^{-1} \xi_{ij}) \langle Z, P_T(X_{i,*}^T Y_{j,*}) \rangle P_T(X_{i,*}^T Y_{j,*}).
\]

Define \(S_{ij} \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}\) as the corresponding matrix to the linear operator \(S_{ij}\) such that

\[
S_{ij} = (1 - p^{-1} \xi_{ij}) \text{vec}(P_T(X_{i,*}^T Y_{j,*})) \text{vec}(P_T(X_{i,*}^T Y_{j,*}))^T,
\]

then we can easily show that \(\|\sum_{i,j} S_{ij}(Z)\|_F^2 = \|\sum_{i,j} S_{ij} \cdot \text{vec}(Z)\|_F^2, \) for any \(Z \in \mathbb{R}^{d_1 \times d_2}\). Next, we are going to apply matrix Bernstein to the summation \(\sum_{i,j} S_{ij}\). We note that \(E[S_{ij}] = 0\), and \(S_{ij}\) is symmetric, for any \((i, j) \in [d_1] \times [d_2]\).

Since \(XM^*Y^T\) is \(\mu_0\)-incoherent and \(M^* = \mathbf{U} \Sigma \mathbf{V}^T\), we have \(\|XU^*\|_{2,\infty} \leq \sqrt{\mu_0 r/d_1}, \|YV^*\|_{2,\infty} \leq \sqrt{\mu_0 r/d_2}\).

According to the definition of \(P_T\) in (C.1), for any \((i, j) \in [d_1] \times [d_2], \) we obtain

\[
\|P_T(X_{i,*}^T Y_{j,*})\|_F^2 = \langle U^T U^T X_{i,*} Y_{j,*} + V^T V^T Y_{j,*} - U^T U^T X_{i,*} Y_{j,*}, V^T V^T Y_{j,*}, X_{i,*} Y_{j,*} \rangle
\]

\[
= \|U^T X_{i,*} Y_{j,*}\|_F^2 + \|V^T Y_{j,*}\|_F^2 - \|U^T X_{i,*} Y_{j,*} V^T \|_F^2
\]

\[
\leq \|X_{i,*} U^T\|_2 \cdot \|Y_{j,*}\|_2 + \|Y_{j,*} V^T\|_2 \cdot \|X_{i,*}\|_2
\]

\[
\leq \|X\|_2^{2,\infty} \|Y\|_2^{2,\infty} + \|Y\|_2^{2,\infty} \cdot \|X\|_2^{2,\infty} \leq \frac{\mu_0 \mu_1 r(n_1 + n_2)}{d_1 d_2},
\]

where the first equality holds because \(\|P_T(A)\|_F^2 = \langle P_T(A), A \rangle\), the first inequality holds because for any vectors \(x, y \in \mathbb{R}^n, \|x y^T\|_F \leq \|x\|_2 \cdot \|y\|_2\), and the last inequality holds because both \(X, Y\) are \(\mu_1\) self-incoherent. To apply matrix Bernstein inequality, we need to bound \(\|S_{ij}\|_2\) and \(\|\sum_{i,j} E(S_{ij})\|_2\), respectively. To begin with, for any \((i, j) \in [d_1] \times [d_2], \) according to definition of \(S_{ij}\), we have

\[
\|S_{ij}\|_2 \leq \frac{1}{p} \|\text{vec}(P_T(X_{i,*}^T Y_{j,*}))\|_2 = \frac{1}{p} \|P_T(X_{i,*}^T Y_{j,*})\|_F^2 \leq \frac{\mu_0 \mu_1 r(n_1 + n_2)}{pd_1 d_2},
\]
where the first inequality holds because $|\langle A, B \rangle| \leq \|A\|_F \cdot \|B\|_F$, and the second inequality follows from (D.3). Similarly, for any $Z \in \mathbb{R}^{n_1 \times n_2}$, we have

$$\left\| \sum_{i,j} \mathbb{E}[S_{ij}^2] \right\|_2 = \frac{1-p}{p} \left\| \sum_{i,j} \text{vec}((P_T(X_i^*,Y_{j,*})) \cdot (P_T(X_i^*,Y_{j,*}))^T \right\|_2^2$$

$$\leq \frac{\mu_0 \mu_1 r(n_1 + n_2)}{pd_1 d_2} \cdot \sup_{\|Z\|_F = 1} \left\| \sum_{i,j} \text{vec}(P_T(X_i^*,Y_{j,*})) \cdot \text{vec}(P_T(X_i^*,Y_{j,*}))^T \right\|_2$$

$$= \frac{\mu_0 \mu_1 r(n_1 + n_2)}{pd_1 d_2} \cdot \sup_{\|Z\|_F = 1} \left\| P_T \left( \sum_{i,j} \langle P_T(Z), X_i^*, Y_{j,*} \rangle X_i^*, Y_{j,*} \rangle \right) \right\|_F$$

$$\leq \frac{\mu_0 \mu_1 r(n_1 + n_2)}{pd_1 d_2} \cdot \sup_{\|Z\|_F = 1} \left\| P_T(X^T BP_T(Z) Y^T Y) \right\|_F \leq \frac{\mu_0 \mu_1 r(n_1 + n_2)}{pd_1 d_2},$$

where the first equality follows from the definition of $S_{ij}$, the first inequality follows from (D.3) and the definition of spectral norm, the second equality holds because $\langle P_T(A), B \rangle = \langle A, P_T(B) \rangle$, and the second inequality holds because $X$, $Y$ are orthonormal and the projection operator $P_T$ is non-expansive. Thus, we obtain

$$\left\| \mathbb{E}[S_{ij}^2] \right\|_2 \leq \frac{\mu_0 \mu_1 r(n_1 + n_2)}{pd_1 d_2}. \tag{D.5}$$

Therefore, combining (D.4) and (D.5), according to Lemma E.1, for any $\gamma \in (0, 1)$, we have

$$P \left\{ \left\| \sum_{i,j} S_{ij} \right\|_2 \geq \gamma \right\} \leq (n_1 + n_2) \cdot \exp \left( \frac{-\gamma^2 / 2}{(1 + \gamma/3)\mu_0 \mu_1 r(n_1 + n_2)/(pd_1 d_2)} \right) \leq c' / d,$$

under condition $p \geq c\mu_0 \mu_1 r n \log(d) / (\gamma^2 d_1 d_2)$, where $c$ is a constant. Note that for all $Z \in \mathbb{R}^{d_1 \times d_2}$, we have

$$\left\| \sum_{i,j} S_{ij}(Z) \right\|_F = \left\| \sum_{i,j} S_{ij} \cdot \text{vec}(Z) \right\|_F \leq \left\| \sum_{i,j} S_{ij} \right\|_2 \cdot \|Z\|_F,$$

we complete the proof of (C.2). Furthermore, for all $Z_1, Z_2 \in T$, we have

$$\left| \langle XZ_1 Y^T - p^{-1}P_{\Omega}(XZ_1 Y^T), XZ_2 Y^T \rangle \right| = \left| \langle P_T(Z_1) - p^{-1}P_T \left( X^T P_T(XP_T(Z_1) Y^T) Y^T \right), Z_2 \rangle \right|$$

$$\leq \left\| P_T(Z_1) - p^{-1}P_T \left( X^T P_T(XP_T(Z_1) Y^T) Y^T \right) \right\|_F \cdot \|Z_2\|_F$$

$$\leq \gamma \|Z_1\|_F \cdot \|Z_2\|_F,$$

where the first equality holds because $\langle A, P_T(B) \rangle = \langle P_T(A), B \rangle$, the second inequality holds because $\langle A, B \rangle \leq \|A\|_F \cdot \|B\|_F$, and the last inequality follows from (C.2). Finally, for all $Z \in T$, we have

$$p^{-1}\|P_{\Omega}(XZY^T)\|_F^2 \leq \|Z - p^{-1}X^T P_{\Omega}(XZY^T) Y, Z\|_F^2 \leq (1 + \gamma)\|Z\|_F^2,$$

which complete the proof.

D.3. Proof of Lemma C.3

Proof. Note that we have

$$\frac{1}{p} \left\| P_{\Omega}(X_L U V^T X_R^T) \right\|_F^2 = \left\| X_L U V^T X_R^T \right\|_F^2 + (p^{-1}P_{\Omega}(X_L U V^T X_R^T) - X_L U V X_R^T) \cdot (X_L U V X_R^T)$$

$$= \left\| U V^T \right\|_F^2 + \left( (p^{-1}P_{\Omega} - I)(X_L U V^T X_R^T), X_L U V X_R^T \right)$$

$$= \left\| U V^T \right\|_F^2 + \sum_{(i,j) = (1,1)} \left( \xi_{ij} / p - 1 \right) \cdot (X_L U V^T X_R^T)_{ij}^2.$$
where \( \xi_{ij} = 1 \), if \((i, j) \in \Omega; \xi_{ij} = 0\), otherwise. Thus, it is sufficient to bound the second term on the right hand side. For simplicity, we let \( \alpha_{ij} = (p^{-1}\xi_{ij} - 1) \cdot \|X_L UV^T X_R^T\|^2_{2,\infty} \). Note that \( U, V \) are fixed, we have \( E[\alpha_{ij}] = 0 \). In addition, we can upper bound \( |\alpha_{ij}| \) by

\[
|\alpha_{ij}| \leq \frac{1}{p} \|X_L UV^T X_R^T\|_{2,\infty} \leq \frac{1}{p} \max \{ \|X_L\|^2_{2,\infty} \cdot \|U\|^2_{p} \cdot \|X_R V\|^2_{2,\infty} \cdot \|X_L U\|^2_{2,\infty} \cdot \|V\|^2_{p} \cdot \|X_R\|^2_{2,\infty} \} \\
\leq \frac{9\mu_0 r_2 \sigma_1^2 (n_1 + n_2)}{pd_1 d_2} \cdot (\|U\|_F^2 + \|V\|_F^2),
\]

where the second inequality follows from Assumption 3.2. Next, we are going to bound the variance

\[
\text{Var} \left( \sum_{(i,j)=(1,1)}^{(d_1,d_2)} \alpha_{ij} \right) = \sum_{(i,j)=(1,1)}^{(d_1,d_2)} \text{Var}(\alpha_{ij}) \\
\leq \frac{1}{p} \sum_{(i,j)=(1,1)}^{(d_1,d_2)} \|X_L UV^T X_R^T\|_{2,\infty}^4 \\
\leq \frac{1}{p} \|X_L UV^T X_R^T\|_{2,\infty}^2 \cdot \|X_L UV^T X_R^T\|_F^2 \\
\leq \frac{9\mu_0^2 r_2^2 \sigma_1^2}{pd_1 d_2} \cdot (\|U\|_F^2 + \|V\|_F^2)^2,
\]

which completes the proof. \( \square \)

**D.4. Proof of Lemma C.4**

**Proof.** For any \( i \in [d_1] \), we denote \( X_{i,*} \in \mathbb{R}^{1 \times n_1} \) as the \( i \)-th row vector of \( X \). Similarly, for any \( j \in [d_2] \), we denote \( Y_{j,*} \in \mathbb{R}^{1 \times n_2} \) as the \( j \)-th row vector of \( Y \). Besides, for any \((i, j) \in [d_1] \times [d_2]\), let \( \xi_{ij} = 1 \), if \((i, j) \in \Omega\), and zero otherwise. Note that \( X, Y \) are orthonormal, thus for any \( Z \in \mathbb{R}^{n_1 \times n_2} \), we have

\[
Z - p^{-1} X^T P_{ij} (XZX^T) Y = X^T (I - p^{-1} P_{ij}) (XZX^T) Y \\
= \sum_{(i,j) \in [d_1] \times [d_2]} (1 - p^{-1} \xi_{ij}) (XZX^T, e_i e_j^T) X^T e_i e_j^T Y \\
= \sum_{(i,j) \in [d_1] \times [d_2]} (1 - p^{-1} \xi_{ij}) (Z, X_{i,*} Y_{j,*}) X_{i,*} Y_{j,*},
\]

where \( I : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^{d_1 \times d_2} \) is the identity mapping, and \( e_i \) denotes the \( i \)-th standard basis. For any \((i, j) \in [d_1] \times [d_2]\), define linear operator \( S_{ij} : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^{n_1 \times n_2} \), such that

\[
S_{ij}(Z) = (1 - p^{-1} \xi_{ij}) (Z, X_{i,*} Y_{j,*}) X_{i,*} Y_{j,*},
\]

Define \( S_{ij} \in \mathbb{R}^{d_1 d_2 \times d_1 d_2} \) as the corresponding matrix to the linear operator \( S_{ij} \) such that

\[
S_{ij} = (1 - p^{-1} \xi_{ij}) \text{vec}(X_{i,*} Y_{j,*}) \text{vec}(X_{i,*} Y_{j,*})^T,
\]

then we can easily show that \( \| \sum_{(i,j) \in [d_1] \times [d_2]} S_{ij}(Z) \|_F^2 = \| \sum_{(i,j) \in [d_1] \times [d_2]} \text{vec}(Z) \|_2^2 \), for any \( Z \in \mathbb{R}^{d_1 \times d_2} \). Obviously, we have \( \mathbb{E}[S_{ij}] = 0 \) and \( S_{ij} \) is symmetric. For any \((i, j) \in [d_1] \times [d_2]\), according to the definition of \( S_{ij} \) in (D.6), we have

\[
\|S_{ij}\|_2 \leq \frac{1}{p} \|\text{vec}(X_{i,*} Y_{j,*})\|_2^2 = \frac{1}{p} \|X_{i,*} Y_{j,*}\|_F^2 \leq \frac{1}{p} \|X_{i,*}\|_{2,\infty}^2 \cdot \|Y_{j,*}\|_{2,\infty}^2 \leq \frac{\mu_0^2 m_1 n_2}{pd_1 d_2}
\]

(D.7)
Assume that each random matrix satisfies with probability at least \( 1 - \Omega \), where the second inequality holds because \( \| X_{i,*}^T Y_{j,*} \|_F^2 \leq \mu_1 n_2^2 / (d_1 d_2) \), and the last inequality follows from the fact that \( X, Y \) are \( \mu_1 \) self-incoherent. Moreover, for any \( Z \in \mathbb{R}^{n_1 \times n_2} \), we have

\[
\left\| \sum_{i,j} \mathbb{E}[S_{ij}^2] \right\|_2 = \frac{1}{p} \left\| \sum_{i,j} \text{vec}(X_{i,*}^T Y_{j,*}) \cdot \|X_{i,*}^T Y_{j,*}\|_2 \cdot \text{vec}(X_{i,*}^T Y_{j,*})^\top \right\|_2 \\
\leq \frac{\mu_1^2 n_1 n_2}{pd_1 d_2} \sup_{\|Z\|_F=1} \left\| \sum_{i,j} \text{vec}(X_{i,*}^T Y_{j,*}) \text{vec}(X_{i,*}^T Y_{j,*})^\top \text{vec}(Z) \right\|_2 \\
= \frac{\mu_1^2 n_1 n_2}{pd_1 d_2} \sup_{\|Z\|_F=1} \left\| \sum_{i,j} (Z, X_{i,*}^T Y_{j,*}) X_{i,*}^T Y_{j,*} \right\|_F \\
= \frac{\mu_1^2 n_1 n_2}{pd_1 d_2} \sup_{\|Z\|_F=1} \left\| X^T Z Y \right\|_F \leq \frac{\mu_1^2 n_1 n_2}{pd_1 d_2},
\]

where the first equality follows from the definition of \( S_{ij} \) in (D.6), the first inequality follows from the fact that \( \| X_{i,*}^T Y_{j,*} \|_F^2 \leq \mu_1^2 n_1 n_2 / (d_1 d_2) \), and the last equality holds because \( X, Y \) are orthonormal. Therefore, according to matrix Bernstein inequality as in Lemma E.1, for any \( \gamma \in (0, 1) \), we have

\[
P \left\{ \left\| \sum_{i,j} S_{ij} \right\|_2 \geq \mu_1 n \gamma \right\} \leq 2 n_1 n_2 \cdot \exp \left( \frac{-\gamma^2 \mu_1^2 n_2^2 / 2}{(1 + \gamma \mu_1 n / 3) \mu_1^2 n_1 n_2 / (pd_1 d_2)} \right) \leq c' / d,
\]

under condition \( |\Omega| \geq c \mu_1 n \log d / \gamma^2 \), where \( c \) is a constant. Thus according to the definition of \( S_{ij} \), for all \( Z \in \mathbb{R}^{n_1 \times n_2} \), with probability at least \( 1 - c' / d \), we have

\[
\| Z - p^{-1} X^T P_{\Omega}(XZY^T) Y \|_F = \left\| \sum_{i,j} S_{ij}(Z) \right\|_F = \left\| \sum_{i,j} S_{ij} \cdot \text{vec}(Z) \right\|_2 \leq \left\| \sum_{i,j} S_{ij} \right\|_2 \cdot \|Z\|_F \leq \mu_1 n \gamma \|Z\|_F.
\]

By triangle’s inequality, we complete the proof. \( \square \)

### E. Additional Auxiliary Lemmas

In this section, we provide the Bernstein inequalities used in the proofs for our main results.

**Lemma E.1.** (Tropp, 2012) Consider a finite sequence \( \{Z_k\} \) of independent random matrices with dimension \( d_1 \times d_2 \).

Assume that each random matrix satisfies

\[
\mathbb{E}(Z_k) = 0 \quad \text{and} \quad \|Z_k\|_2 \leq R \quad \text{almost surely}.
\]

Define

\[
\sigma^2 = \max \left\{ \left\| \sum_k \mathbb{E}(Z_k Z_k^\top) \right\|_2, \left\| \sum_k \mathbb{E}(Z_k^\top Z_k) \right\|_2 \right\}.
\]

Then, for all \( t \geq 0 \), we have

\[
P \left\{ \left\| \sum_k Z_k \right\|_2 \geq t \right\} \leq \left( d_1 + d_2 \right) \cdot \exp \left( \frac{-t^2 / 2}{\sigma^2 + \|Rt/3\|} \right).
\]

**Lemma E.2.** Assume the index set \( \Omega \) follows Bernoulli model (3.1). There exists constants \( c, c' \) such that under condition that \( |\Omega| \geq c \log d \), with probability at least \( 1 - c' / d \), we have

\[
|\Omega| - pd_1 d_2 \leq \frac{1}{2} pd_1 d_2,
\]

where \( p = |\Omega| / (d_1 d_2) \).

Lemma E.2 can be directly derived from the Bernstein-type inequality for independent random variables.