A. Notations

We begin the appendix with a restatement of the notations. Denote \( c, c', c_i \) as some universal positive constants. Notice that their values may not necessarily the same even for those with same notations. We denote \( a \lesssim b \) if there exists some positive constant \( c_0 > 0 \) such that \( a \leq c_0 b \). Similarly we define \( a \gtrsim b \) provided \( a \geq c_0 b \) for some positive constant \( c_0 \). We write \( a \asymp b \) when \( a \lesssim b \) and \( a \gtrsim b \) hold simultaneously.

For an arbitrary matrix \( X \), we denote \( X_{i,:} \) as the \( i \)-th row, \( X_{:,j} \) as its \( i \)-th column, and \( X_{ij} \) as the \((i,j)\)-th element. The Frobenius norm of \( X \) is defined as \( \|X\|_F \) while the operator norm is denoted as \( \|X\|_{op} \), whose definition can be found in Section 2.3 of Golub and Loan (2013) (P71). Its stable rank \( \rho(X) \) is defined as the ratio \( \|X\|_F^2/\|X\|_{op}^2 \) (Section 2.1.15 in Tropp (2015)). The inner product \( \langle A, C \rangle \) is defined as \( \sum_{ij} A_{ij} C_{ij} \).

Associate with each permutation matrix \( \Pi \), we define the operator \( \pi(\cdot) \) that transforms index \( i \) to \( \pi(i) \). The Hamming distance \( d_H(\Pi_1, \Pi_2) \) between permutation matrix \( \Pi_1 \) and \( \Pi_2 \) is defined as \( \sum_{i=1}^n \mathbb{1}(\pi_1(i) \neq \pi_2(i)) \). Additionally, we denote \( \mathcal{E} \) as the complement of the event \( \mathcal{E} \) and the signal-to-noise-ratio (SNR) as \( \text{SNR} = \|B\|_F^2/(m\sigma^2) \).

B. Problem Restatement

To begin with, we recall the problem formulation, which reads as

\[
Y = \Pi^t X B + W,
\]

where \( Y \in \mathbb{R}^{n \times m} \) represents the observation, \( \Pi \in \mathbb{R}^{n \times n} \) denotes the unknown permutation matrix, \( X \in \mathbb{R}^{n \times p} \) is the sensing matrix (design matrix) with \( X_{ij} \sim \mathcal{N}(0, 1) \) being a standard normal random variable (RV), \( B \in \mathbb{R}^{p \times m} \) is the matrix of regression coefficients, and \( W \in \mathbb{R}^{n \times m} \) is the additive Gaussian noise matrix such that \( W_{ij} \sim \mathcal{N}(0, \sigma^2) \).

Our goal is to reconstruct the pair \( (\hat{\Pi}, \hat{B}) \) from the observation \( Y \) and sensing matrix (design matrix) \( X \). The proposed one-step estimator can be written as

\[
\hat{\Pi} = \arg\max_{\Pi \in \mathcal{P}_n} \left\{ \Pi, YY^\top XX^\top \right\},
\]

\[
\hat{B} = (X^\dagger)^\top \hat{\Pi}^\top Y,
\]

where \( X^\dagger = (X^\top X)^{-1}X^\top \) denotes the pseudo-inverse of \( X \). In the following, we will separately investigate its properties under the single observation model \((m = 1)\) and multiple observations model \((m > 1)\). The formal statement is packaged in Theorem 1 and Theorem 2.

C. Appendix for Section 3

This section focuses on the special case where \( p = 1, m = 1 \). Consider \( X \in \mathbb{R}^n \) to be a Gaussian distributed RV such that \( X \sim \mathcal{N}(0, I_{n \times n}) \), and permutation matrix \( \Pi^t \) which satisfies \( d_H(I, \Pi^t) = h \leq n/4 \).

C.1. Notations

First we define the following events \( \mathcal{E}_i \), \((1 \leq i \leq 5)\), which reads

\[
\mathcal{E}_1 \triangleq \left\{ \langle X, \Pi^t X \rangle \geq c_0 n \right\},
\]

\[
\mathcal{E}_2 \triangleq \left\{ \|X\|_2 \leq 2\sqrt{n} \right\},
\]

\[
\mathcal{E}_3(\Pi) \triangleq \left\{ W^\top X X^\top \left( \Pi^t - \Pi \right) W \lesssim \sigma^2 n^2 \log n \right\},
\]

\[
\mathcal{E}_4(\Pi) \triangleq \left\{ \langle W, X \rangle \langle \Pi^t X, (\Pi^t - \Pi)^\top X \rangle + \langle W, (\Pi^t - \Pi)^\top X \rangle \langle \Pi^t X, X \rangle \lesssim \sigma n^2 \sqrt{\log n} \right\}
\]

\[
\mathcal{E}_5(\Pi; \ell) \triangleq \left\{ \|X - \Pi X\|_2 \geq \frac{12\ell}{5cn^2 \sigma}, \ d_H(I, \Pi) = \ell \right\},
\]

where \( \Pi \) is an arbitrary permutation matrix, and \( c_0 > 0 \) is some positive constant.
C.2. Outline of proof

We will prove that ground truth permutation matrix $\Pi^\sharp$ will be returned with high probability under the assumptions in Theorem 1. The formal statement is shown in Theorem 1. Before we delve into the proof details, we give a roadmap of the proof, which is

- **Step I:** Under the events $\mathcal{E}_i \cap \Pi (\mathcal{E}_3(\Pi) \cap \mathcal{E}_4(\Pi) \cap \mathcal{E}_5(\Pi; \ell))$, we have

  $$\langle \Pi^\sharp, yy^T XX^T \rangle - \langle \Pi, yy^T XX^T \rangle \geq \frac{c_0 \beta^2}{n^{19}} - c_1 \beta \sigma n^2 \sqrt{\log n} - c_2 \sigma^2 n^2 \log n.$$ 

  Notice that under assumptions in Theorem 1, we conclude that $\langle \Pi^\sharp, yy^T XX^T \rangle > \langle \Pi, yy^T XX^T \rangle, \forall \Pi$, which suggests that $\Pi^\sharp$ will always be returned by our estimator in Eq. (3).

- **Step II:** We upper-bound the probability $\mathbb{P}(\Pi \neq \Pi^\sharp)$ by $\mathbb{P}(\mathcal{E}_1 \cup \Pi (\mathcal{E}_3(\Pi) \cup \mathcal{E}_4(\Pi) \cup \mathcal{E}_5(\Pi; \ell)))$ and complete the proof by showing it is at most $cn^{-1}$.

Having illustrated the proof strategy, we turn to the proof details. The main proof is attached in Section C.3 while the supporting lemmas bounding $\mathbb{P}(\mathcal{E}_i), (1 \leq i \leq 5)$, are put in Section C.4.

C.3. Proof of Theorem 1

**Proof 1** For an arbitrary permutation matrix $\Pi$, we can expand the term $\langle \Pi, yy^T XX^T \rangle$ as

$$\langle \Pi, yy^T XX^T \rangle = \mathcal{T}_1(\Pi) + \beta \mathcal{T}_2(\Pi) + \beta^2 \mathcal{T}_3(\Pi),$$

where $\mathcal{T}_i(\Pi), (1 \leq i \leq 3)$, are defined as

$$\mathcal{T}_1(\Pi) = \langle W, \Pi^T X \rangle \langle X, W \rangle;$$
$$\mathcal{T}_2(\Pi) = \langle W, X \rangle \langle \Pi^T X, \Pi^T X \rangle + \langle W, \Pi^T X \rangle \langle \Pi^T X, X \rangle;$$
$$\mathcal{T}_3(\Pi) = \langle \Pi^T X, \Pi X \rangle \langle \Pi^T X, X \rangle.$$

**Step I:** We rewrite the difference $\langle \Pi^\sharp, yy^T XX^T \rangle - \langle \Pi, yy^T XX^T \rangle$ as

$$\langle \Pi^\sharp, yy^T XX^T \rangle - \langle \Pi, yy^T XX^T \rangle = \mathcal{T}_1(\Pi^\sharp) - \mathcal{T}_1(\Pi) + \beta \left( \mathcal{T}_2(\Pi^\sharp) - \mathcal{T}_2(\Pi) \right) + \beta^2 \left( \mathcal{T}_3(\Pi^\sharp) - \mathcal{T}_3(\Pi) \right);$$

$$\geq \frac{\beta^2}{2} \langle \Pi^T X, X \rangle \|X - \Pi^T \Pi X\|_2^2 + \beta \left( \mathcal{T}_2(\Pi^\sharp) - \mathcal{T}_2(\Pi) \right) + \mathcal{T}_3(\Pi^\sharp) - \mathcal{T}_3(\Pi);$$

$$\geq \frac{\beta^2}{2} c_0 n \frac{24}{\beta^2 n} - \beta \left| \mathcal{T}_2(\Pi^\sharp) - \mathcal{T}_2(\Pi) \right| - \left| \mathcal{T}_3(\Pi^\sharp) - \mathcal{T}_3(\Pi) \right| \geq 0,$$

where in ① we rewrite $\|X\|_2^2 - \langle \Pi^T X, \Pi X \rangle$ as

$$\|X\|_2^2 - \langle \Pi^T X, \Pi X \rangle = \frac{1}{2} \left( \|X\|_2^2 + \|\Pi^T \Pi X\|_2^2 - 2 \langle \Pi^T X, \Pi X \rangle \right) = \frac{1}{2} \left( \|X - \Pi^T \Pi X\|_2^2 \right);$$

in ② we condition on event $\mathcal{E}_1, \mathcal{E}_5(\Pi; \ell)$ and have $\|X - \Pi X\|_2^2 \geq \frac{12\ell}{\beta^2 n^3} \geq \frac{24}{\beta^2 n^3}$, in ③ we condition on $\mathcal{E}_3(\Pi), \mathcal{E}_4(\Pi)$, and in ④ we use the assumption $\log(\text{SNR}) \geq \log n$ in Theorem 1.
Step II: The error probability $P\left(\hat{\Pi} \neq \Pi^\natural\right)$ is hence be upper-bounded as

\[
P\left(\hat{\Pi} \neq \Pi^\natural\right) \leq P\left( \bigcup_{\Pi} \left( \bigcup_{\Pi} \left( E_3(\Pi) \cup E_4(\Pi) \cup E_5(\Pi; \ell) \right) \right) \cup E_1 \right) + P\left( E_2 \right) + P\left( E_1 \right)
\]

\[
\leq \sum_{\Pi \neq \Pi} P\left( E_3(\Pi) \cup E_4(\Pi) \cup E_5(\Pi; \ell) \right) + \sum_{\Pi \neq \Pi} P\left( E_4(\Pi) \cap E_3(\Pi) \right) + \sum_{\ell \geq 2} P\left( E_5(\Pi; \ell) \cap E_3(\Pi) \right) + 8n^{-1} + 2e^{-c_0n}
\]

\[
\leq 2n^{-n} + 3 \sum_{\ell \geq 2} \left( \frac{n}{\ell} \right) \ell! n^{-2\ell} + 8n^{-1} + 2e^{-c_0n}
\]

\[
\leq c_0n^{-n} + n^{-1} + 3 \sum_{\ell \geq 2} n^\ell n^{-2\ell} \leq c_0n^{-1} + \frac{3}{n(n-1)} \leq n^{-1},
\]

where in \(\circ\) we use the union bound, in \(\circ\) we complete the proof with Lemma I and the fact $P\left( E_2 \right) \leq e^{-0.8n}$, in \(\circ\) we invoke Lemma 2, Lemma 3, Lemma 4, and in \(\circ\) we use $n!/(n-\ell)! \leq n^\ell$ and complete the proof.

C.4. Supporting Lemmas for Theorem 1

This subsection collects the supporting lemmas for the proof of Theorem 1.

Lemma 1 We have $P\left( E_1 \right) \leq 8n^{-1} + e^{-0.238n}$ when $n$ is sufficiently large.

Proof 2 Different from the proof in Lemma 9, we consider the case where $X \in \mathbb{R}^n$ is a vector and would lower-bound $\left\langle X, \Pi^\natural X \right\rangle$. W.l.o.g. we assume the first $h$ entries are permuted and expand the inner product $\left\langle X, \Pi^\natural X \right\rangle$ as

\[
\left\langle X, \Pi^\natural X \right\rangle = \sum_{i=1}^{h} X_i X_{\pi(i)} + \sum_{i=h+1}^{n} X_i^2.
\]

With union bound, we can upper bound $P\left( \left\langle X, \Pi^\natural X \right\rangle \leq c_0n \right)$ as

\[
P\left( \left\langle X, \Pi^\natural X \right\rangle \leq c_0n \right) \overset{(\dagger)}{\leq} P\left( \sum_{i=h+1}^{n} X_i^2 \leq \frac{1}{4} (n-h) \right) + P\left( \sum_{i=1}^{h} X_i X_{\pi(i)} \leq -\frac{4\sqrt{2} + \sqrt{35}}{\sqrt{2}} \sqrt{n \log n} \right),
\]

where $c_0 > 0$ is some positive constant, in \(\dagger\) we use the fact

\[
\frac{n-h}{4} - \frac{4\sqrt{2} + \sqrt{35}}{\sqrt{2}} \sqrt{n \log n} \geq \frac{3n}{16} - \frac{4\sqrt{2} + \sqrt{35}}{\sqrt{2}} \sqrt{n \log n} \geq c_0n,
\]

when $n$ is large. We finish the proof by separately upper-bounding $\zeta_1 \leq e^{-0.2386n}$ and $\zeta_2 \leq 8n^{-1}$. The detailed computation comes as follows.

Phase I: For $\zeta_1$, we can view $\sum_{i=h+1}^{n} X_i^2$ as a $\chi^2$-$RV$ with $(n-h)$ freedom and have

\[
\zeta_1 \overset{(2)}{=} \exp\left( \frac{n-h}{2} \left( \log \left( \frac{1}{4} - \frac{1}{4} + 1 \right) \right) \right) \overset{(3)}{\leq} e^{-0.2386n},
\]
According to the Bernstein inequality, we have

$$P \left( \sum_{i=1}^{h} X_i X_{\pi(i)} \leq - \frac{4\sqrt{2} + \sqrt{35}}{3\sqrt{2}} \sqrt{n \log n}, \ |X_i| \leq 2\sqrt{\log n}, \ \forall i \right) + P \left( |X_i| \geq 2\sqrt{\log n}, \ \exists i \right) \leq 3 \cdot \frac{4\sqrt{2} + \sqrt{35}}{3\sqrt{2}} \sqrt{n \log n}, \ |X_i| \leq 2\sqrt{\log n}, \ \forall i,$$

where in (4) we use the union bound for $$\sum_{i=1}^{h} X_i X_{\pi(i)}$$ and the tail bounds for Gaussian distributed $$X_i$$. Then we define $$Z_i = \sum_{j \in I_i} X_j X_{\pi(j)}$$ and bound $$\zeta_{2,i}$$ via the Bernstein inequality (Theorem 2.8.4 in Vershynin (2018)).

Then, we verify that $$E \left( X_j X_{\pi(j)} \right) = (E X_j) (E X_{\pi(j)}) = 0$$. Meanwhile, we compute $$\sigma^2 = \sum_{j \in I_i} E \left( X_j X_{\pi(j)} \right)^2 = h_i$$. According to the Bernstein inequality, we have

$$\left| \sum_{j \in I_i} X_j X_{\pi(j)} \right| \geq \frac{4}{3} (\log n)^2 + \sqrt{\frac{16}{9} (\log n)^4 + 2(\log n) h_i},$$

holds with probability $$2n^{-1}$$. Meanwhile, we can upper bound as

$$\frac{4}{3} (\log n)^2 + \sqrt{\frac{16}{9} (\log n)^4 + 2(\log n) h_i} \leq \frac{4}{3} (\log n)^2 + \sqrt{\frac{16}{9} (\log n)^4 + \frac{n \log n}{6}} \leq \frac{4\sqrt{2} + \sqrt{35}}{3\sqrt{2}} \sqrt{n \log n},$$

where (5) is because $$n \geq \log^3(n)$$ for $$n \geq 95$$. Hence, we conclude that $$\zeta_{2,i} \leq 2n^{-1}$$ and complete the proof by combining the bound for $$\zeta_1$$ and $$\zeta_2$$.

**Lemma 2** We have $$P \left( \mathcal{E}_3(\Omega) \cap \mathcal{E}_2 \right) \leq n^{-2n}$$.

**Proof 3** For the conciseness of notation, we define $$\Xi$$ as $$\Xi = XX^T (\Pi^2 - \Pi)$$. Due to the independence of the $$X$$ and $$W$$, we can condition on $$X$$ and bound $$P(\mathcal{E}_3(\Omega) \cap \mathcal{E}_2)$$ as

$$P \left( \mathcal{E}_3(\Omega) \cap \mathcal{E}_2 \right) \leq P \left( \mathcal{E}_2 \cap \mathcal{E}_3 \right) \leq \exp \left( - \left( c_0 n^4 \log^2 n \wedge c_1 n^2 \log n \right) \right) \leq n^{-2n},$$

where in (1) we condition on $$\mathcal{E}_2$$ and use the fact

$$\Xi W^T \Xi W + c \sigma^2 n^2 \log n \leq \sigma^2 \|X\|_2^2 + c \sigma^2 n^2 \log n \leq \sigma^2 n^2 \log n,$$

in (2) we use Hanson-Wright inequality (Theorem 6.2.1 in Vershynin (2018)), and in (3) we condition on $$\mathcal{E}_2$$ and use $$\|\Xi\|_2 \leq \|X\|_2 \leq n$$.

**Lemma 3** We have $$P \left( \mathcal{E}_4(\Omega) \cap \mathcal{E}_2 \right) \leq n^{-2n}$$. 

**Proof 4** Due to the independence between $$W$$ and $$X$$, we would like to condition on $$X$$ and bound $$P(\mathcal{E}_4(\Omega) \cap \mathcal{E}_2)$$ as

$$P \left( \mathcal{E}_4(\Omega) \cap \mathcal{E}_2 \right) \leq \exp \left( - \frac{4c \sigma^2 n^4 \log n}{2\sigma^2 \Pi} \right),$$
where $\sigma_\Pi^2$ is defined as

$$\sigma_\Pi^2 = \sigma^2 \left\langle \Pi^T X, (\Pi^T - \Pi)^T X \right\rangle + \left\langle \Pi^T X, (\Pi^T - \Pi) X \right\rangle_2^2,$$

Notice under $\mathcal{E}_2$, we have $\sigma_\Pi^2 \lesssim \sigma^2 \left( \frac{4 \left\| X \right\|_2^2}{2\sigma_\Pi^2} \right)^2 = c \sigma^2 n^3$, and complete the proof by showing

$$\exp \left( -\frac{4c\sigma^2 n^3 \log n}{2\sigma_\Pi^2} \right) \leq \exp \left( -\frac{4c\sigma^2 n^3 \log n}{2c\sigma^2 n^3} \right) = n^{-2n}.$$

**Lemma 4** We have $\mathbb{P} \left( \mathcal{E}_5(\Pi; \ell) \right) \leq 3n^{-2\ell}$.

**Proof 5** Adopting a similar approach as in proving Lemma 1, we can decompose the index sets $\{ j : j \neq \pi(j) \}$ into 3 disjoint sets $\mathcal{I}_i$ ($1 \leq i \leq 3$) such that: (1) $j$ and $\pi(j)$ do not lie within the same index set $\mathcal{I}_i$; and (2) the cardinality $\ell_i$ of $\mathcal{I}_i$ satisfies $|\ell/5| \leq \ell_i \leq \ell/3$. Then we can bound $\mathbb{P}(\mathcal{E}_5(\Pi; \ell))$ as

$$\mathbb{P} \left( \left\| X - \Pi^T X \right\|_2^2 \leq \frac{12\ell}{5\sigma_\Pi^2} \right) \overset{\mathbb{E}}{=} \sum_{i=1}^{3} \mathbb{P} \left( \sum_{j \in \mathcal{I}_i} (X_j - X_{\pi(j)})^2 \leq \frac{4\ell}{5\sigma_\Pi^2} \right) \overset{\mathbb{E}}{=} \sum_{i=1}^{3} \exp \left( \frac{\ell_i}{2} \left( \log \frac{2\ell}{5\sigma_\Pi^2} + 1 \right) \right) \overset{\mathbb{E}}{=} \frac{3n^{-2\ell}}{\frac{\ell_i}{2} \left( \log \frac{2\ell}{5\sigma_\Pi^2} + 1 \right)} \leq \frac{\ell_i}{2} \left( \log \frac{2\ell}{5\sigma_\Pi^2} - 20 \log n \right) \leq -10\ell_i \log n \leq -2\ell \log n.$$

where $\mathbb{E}$ is due to the decomposition $\mathcal{I}_i$, $1 \leq i \leq 3$, $\mathbb{E}$ is because $\sum (X_j - X_{\pi(j)})^2 / 2$ is a $\chi^2$ RV with freedom $\ell_i$ and $\mathbb{E}$, $\mathbb{E}$ is due to $|\ell/5| \leq \ell_i \leq \ell/3$ and hence

$$\mathbb{E} \overset{\mathbb{E}}{=} \frac{\ell_i}{2} \left( \log \frac{2\ell}{5\sigma_\Pi^2} - 20 \log n \right) \leq -10\ell_i \log n \leq -2\ell \log n.$$

**D. Appendix for Section 4**

This section provides theoretical analysis for the multiple observations model, i.e., $m > 1$. We will show that our estimator in Eq. (3) gives correct permutation matrix $\Pi^t$ once

$$\log(SNR) \geq \frac{\log n}{\rho(B^*)} + \log \log n.$$

The formal statement is packaged in Theorem 2.

**D.1. Notations**

Before our discussion, first we define $\overline{B}$ and $B^*$ respectively as

$$\overline{B} = (n - h)^{-1} X^T \Pi^T X B^2,$$

$$B^* = (n - h)^{-1} X^T Y = \overline{B} + (n - h)^{-1} X^T W,$$

where $h$ is denoted as the Hamming distance between identity matrix $I$ and the ground truth permutation matrix $\Pi^t$, i.e., $h = d_H(I, \Pi^t)$. Similar as in Section C, we define events $\mathcal{E}_i$, ($6 \leq i \leq 9$) as

$$\mathcal{E}_6 \overset{\mathbb{E}}{=} \left\{ \left\| X_{i,:} \right\|_2 \leq 2 \sqrt{p \log n}, \forall i \right\};$$

$$\mathcal{E}_7 \overset{\mathbb{E}}{=} \left\{ \left\| X_{i,:} \right\|_2 \leq 6 c_0 \frac{(\log n)^{3/2}(\log p) \sqrt{n}}{\sqrt{m}} \right\} + c_1 \sqrt{m}(\log n)\sigma \left( 1 + \frac{p}{n} \right), \forall i \};$$

$$\mathcal{E}_8 \overset{\mathbb{E}}{=} \left\{ \left\| X_{i,:} \right\|_2 \leq 6 c_0 \frac{(\log n)^{3/2}(\log p) \sqrt{n}}{\sqrt{m}} \right\} + \Delta, \exists i, j \};$$

$$\mathcal{E}_9 \overset{\mathbb{E}}{=} \left\{ \left\| X_{i,:} \right\|_2 \leq 6 c_0 \frac{(\log n)^{3/2}(\log p) \sqrt{n}}{\sqrt{m}} \right\} + \Delta, \exists i, j \};$$

where $\Delta$ is defined as

$$\Delta = 16\sqrt{c_0} \sigma \frac{(\log n)^{3/2}(\log p) \sqrt{n}}{\sqrt{m}} \left\| B^2 \right\|_F + 16c_1 \sqrt{2m}(\log n)\sigma^2 \left( 1 + \frac{p}{n} \right) + 4\sqrt{c_2}(\log n)\sigma \left\| B^2 \right\|_F.$$
D.2. Outline of proof

In front of the rigorous proof in Section D.3, we first illustrate our proof strategy as

- **Step I:** We relax the wrong recovery \( \{ \hat{\Pi} \neq \Pi^k \} \) to event \( \mathcal{E} \), i.e. \( \{ \hat{\Pi} \neq \Pi^k \} \subseteq \mathcal{E} \), which reads as

\[
\mathcal{E} \triangleq \left\{ \| Y_{i,:} - X_{n\pi(i),:} B^* \|_2^2 \geq \| Y_{i,:} - X_{j,:} B^* \|_2^2, \quad \exists \ i, j \right\}.
\]

(14)

The physical meaning of \( \mathcal{E} \) is that we may reduce the residual \( \| Y - \hat{\Pi}^k X B^* \|_F \) by changing \( \pi^k(i) \) to \( j \). Same relaxation has been previously used in Collier and Dalalyan (2016); Slawski et al. (2019a); Zhang et al. (2019a,b).

- **Step II:** The core in this step lies in how to lower bound \( \mathbb{P}(\mathcal{E}) \). First we decompose \( \mathcal{E} \) into \( \mathcal{E}_8 \cup \mathcal{E}_9 \) with some simple algebraic manipulations. Under the SNR assumption in Eq. (7), we show both \( \mathbb{P}(\mathcal{E}_8) \) and \( \mathbb{P}(\mathcal{E}_9) \) are approximately \( \mathbb{P}(\mathcal{E}) \), as in Lemma 5 and Lemma 6, respectively.

To show \( \mathbb{P}(\mathcal{E}) \) is with high probability, in another words, \( \mathbb{P}(\mathcal{E}_7) \) is highly likely, we prove the following relations hold with high probability under \( \mathcal{E}_6 \),

\[
\| X_{i,:} (\hat{B} - B^*) \|_2 \lesssim \frac{p(\log n)^{3/2} (\log p)}{\sqrt{n}} \| B^* \|_F;
\]

\[
\| X_{i,:} X^\top W \|_2 \lesssim \sqrt{m}(\log n)\sigma(n + p),
\]

whose proof are in Lemma 9 and Lemma 10, respectively, and hence finish the proof by

\[
\| X_{i,:} (B^* - B^*) \|_2 \leq \| X_{i,:} (\hat{B} - B^*) \|_2 + \frac{1}{n-h} \| X_{i,:} X^\top W \|_2.
\]

In particular, we would like to mention the technique used in bounding \( \| X_{i,:} X^\top W \|_2 \). First we review the widely-used bounding procedure, which proceeds as

\[
\| X_{i,:} X^\top W \|_2 \leq \| X_{i,:} \|_2 \| X \|_2 \| W \|_2 \overset{\text{1}}{\lesssim} \sqrt{p \log n} (\sqrt{n} + \sqrt{p}) \| X \|_2 \lesssim \sqrt{p \log n}, \quad \| X \|_2 \lesssim \sqrt{n} + \sqrt{p}, \quad \| W \|_2 \lesssim (\sqrt{n} + \sqrt{m}) \| W \|_2 \overset{\text{2}}{\lesssim} \sigma(\sqrt{n} + \sqrt{m})
\]

where in 1 we use the fact \( \| X_{i,:} \|_2 \lesssim \sqrt{p \log n} \), \( \| X \|_2 \lesssim \sqrt{n} + \sqrt{p} \), \( \| W \|_2 \lesssim \sigma(\sqrt{n} + \sqrt{m}) \) hold with high probability, and in 2 we use \( p \gg n \). Comparing with our results in Lemma 10, this bound experience inflations when \( m \ll n \) and will lift the SNR requirement to \( \log(\text{SNR}) \gtrsim \log n \), which hides the role of \( \rho(B^*) \) compared with our current result in Theorem 2. To handle such problem, we adopt the leave-one-out trick as in El Karoui (2013; 2018); Chen et al. (2019); Sur et al. (2019) and refer to Lemma 10 for the technical details.

Having illustrated our proof strategies, we leave the detailed calculation to Section D.3.

D.3. Proof of Theorem 2

**Proof** We restate the definition of event \( \mathcal{E} \) as

\[
\mathcal{E} \triangleq \left\{ \| Y_{i,:} - X_{n\pi(i),:} B^* \|_2^2 \geq \| Y_{i,:} - X_{j,:} B^* \|_2^2, \quad \exists \ i, j \right\}.
\]

**Step I:** First we verify that

\[
\hat{\Pi} = \arg\min_{\Pi} \| Y - \Pi X B^* \|_F
\]

returns the same permutation matrix \( \hat{\Pi} \) as that by Eq. (3). Hence, correct recovery of the ground truth permutation matrix \( \Pi^k \) suggests that

\[
\| Y - \Pi^k X B^* \|_F < \| Y - \Pi X B^* \|_F, \quad \forall \Pi \neq \Pi^k.
\]
Then we finish the proof by showing that $\mathcal{E} \subseteq \{\tilde{\Pi} = \Pi^2\}$. Assuming the claim is not true, which means we have matrix $\Pi$ such that

$$\|Y - \Pi^2XB^*\|_F^2 \geq \|Y - \Pi XB^*\|_F^2,$$

conditional on event $\mathcal{E}$. Meanwhile we have

$$\|Y - \Pi^2XB^*\|_F^2 = \sum_{i=1}^{n} \|Y_{i,:} - X_{\pi^*(i,:),B^*}\|_2^2 \leq \sum_{i=1}^{n} \|Y_{i,:} - X_{\pi(i,:),B^*}\|_2^2 = \|Y - \Pi XB^*\|_F^2,$$

which leads to contradiction, where in $1$ we use the definition of $\mathcal{E}$.

**Step II:** We verify that $\|Y_{i,:} - X_{\pi^*(i,:),B^*}\|_2^2 \geq \|Y_{i,:} - X_{j,:}B^*\|_2^2$ is equivalent to

$$2\langle W_{i,:} (X_{j,:} - X_{\pi^*(i,:),B^*}) B^* \rangle \geq \|X_{\pi^*(i,:),B^*}\|_2^2 + \|X_{j,:} (B^* - B^*)\|_2^2 + 2\langle (X_{\pi^*(i,:),B^*} - X_{j,:}B^*) (B^* - B^*) \rangle - \|X_{\pi^*(i,:),B^*}\|_2^2,$$

which suggests that $\mathbb{P}(\mathcal{E}) \leq \mathbb{P}(\mathcal{E}_8) + \mathbb{P}(\mathcal{E}_9)$ and completes the proof with Lemma 5 and Lemma 6.

**Lemma 5** We have $\mathbb{P}(\mathcal{E}_8) \leq c_0 \exp(\langle -((\log n)^4 \wedge (\log n)^2 \rho(B^*)\rangle) + c_1 n^{-1} + c_2 n e^{-c_3 n} + c_4 n e^{-c_0 m} + 2e^{-p} + 6p^{-2}.$

**Proof 7** For the conciseness of notation, we define $\Delta_1$ and $\Delta_2$ as

$$\Delta_1 = 4c_0 \frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}} B^* F + 4c_1 \sqrt{m}(\log n) \sigma \left(1 + \frac{p}{n}\right);$$

$$\Delta_2 = c_2 (\log n) \|B^*\|_F.$$

Then we can bound $\mathbb{P}(\mathcal{E}_8)$ as

$$\begin{align*}
\mathbb{P}(\mathcal{E}_8) & \leq \mathbb{P}\left(\|X_{j,:} - X_{\pi^*(i,:),B^*}\|_2^2 \geq \Delta_1 + \Delta_2, \exists i,j\right) + \exp\left(-\frac{\Delta^2}{2\sigma^2 (\Delta_1 + \Delta_2)^2}\right) \\
& \leq \mathbb{P}\left(\|X_{j,:} - X_{\pi^*(i,:),B^* - B^*}\|_2^2 \geq \Delta_1, \exists i,j\right) + \mathbb{P}\left(\|X_{j,:} - X_{\pi^*(i,:),B^*}\|_2^2 \geq \Delta_2, \exists i,j\right) + n^{-8}, \\
& \leq \mathbb{P}\left(\|X_{j,:} - X_{\pi^*(i,:),B^* - B^*}\|_2^2 \geq \Delta_1, \exists i,j\right) (15)
\end{align*}$$

where in $1$ we use the independence between $W$ and $X$ and condition on $X$, in $2$ we use the relation $\Delta = 4\sqrt{2}\sigma (\Delta_1 + \Delta_2)$. Then we will prove that $\zeta_1 \leq \mathbb{P}(\mathcal{E}_7)$ and $\zeta_2 \leq e^{-((\log n)^4 \wedge (\log n)^2 \rho(B^*))}$.

**Phase I: bounding $\zeta_1$** Conditional on $\mathcal{E}_7$, we have

$$\|X_{j,:} - X_{\pi^*(i,:),B^* - B^*}\|_2 \leq \|X_{j,:} (B^* - B^*)\|_2 + \|X_{\pi^*(i,:),B^* - B^*}\|_2 \leq 2c_0 \frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}} B^* F + 2c_1 \sqrt{m}(\log n) \sigma \left(1 + \frac{p}{n}\right) < \frac{\Delta_1}{2},$$

and obtain $\zeta_1 = 0$, where $3$ is due to the definition of $\mathcal{E}_7$. Then we conclude that $\zeta_1 \leq \mathbb{P}(\mathcal{E}_7)$.

**Phase II: bounding $\zeta_2$** For $\zeta_2$, we upper-bound it as

$$\begin{align*}
\zeta_2 & \leq \sum_{\pi^*(i),j} \mathbb{P}\left(Z \geq c_2 (\log n)^2 \|B^*\|_F^2\right) \leq n^2 \mathbb{P}\left(Z - EZ \geq c_3 (\log n)^2 \|B^*\|_F^2\right) \\
& \leq n^2 \exp\left(-\frac{(\log n)^4 \|B^*\|_F^4}{\|B^* B^*\|_F^2} \wedge \frac{(\log n)^2 \|B^*\|_F^2}{\|B^* B^*\|_F \|B^*\|_F}\right) = n^2 e^{-((\log n)^4 \wedge (\log n)^2 \rho(B^*))} \leq e^{-((\log n)^4 \wedge (\log n)^2 \rho(B^*))}, (16)
\end{align*}$$
where in 4 we define $Z \triangleq \| (X_{j,} - X_{t-1,}) B^* \|_2^2$, in 5 we have $\mathbb{E} Z = 4 \| B^* \|^2$ and use $c_2 (\log n)^2 \| B^* \|^2_F \geq (4 + c_3 (\log n)^2) \| B^* \|^2$ when $n$ is sufficiently large, and in 6 we use the Hanson-Wright inequality (Theorem 6.2.1 in Vershynin (2018)). Combining Eq. (15), Eq. (16) and Lemma 8 together, we complete the proof.

Lemma 6 Consider the same setting of Theorem 2. Provided the SNR satisfies

$$\log(\text{SNR}) \geq \frac{6 \log n}{\rho(B^*)} + \log \log n,$$

we have $P(\mathcal{E}_0) \leq 2e^{-p} + ne^{-c_1 m} + c_2 p^{-2} + c_3 ne^{-c_2 n}$, when $n$ is sufficiently large, where $c_i > 0$, $0 \leq i \leq 4$ are some positive constants.

Proof

8 We upper bound $P(\mathcal{E}_0)$ as

$$P(\mathcal{E}_0) \leq P \left( \left\| (X_{s-1,} - X_{t,}) B^* \right\|_2^2 - 2 \left\| (X_{s-1,} - X_{t,}) B^* \right\|_2 \left\| X_{s-1,} (B^* - B) \right\|_2 - \left\| X_{t,} (B^* - B) \right\|_2 \leq \Delta, \ \exists i, j \right) \ \leq \ P \left( \left\| (X_{s-1,} - X_{t,}) B^* \right\|_2 \leq \delta, \ \exists i, j \right) + \ P \left( \left\| (X_{s-1,} (B^* - B)) \right\|_2 \leq \frac{2}{\delta}, \ \exists i, j \right) \ .$$

Setting $\delta$ as $\| B^* \|_F n^{-\frac{3}{6 (\log p)^2}}$, we would like to show $\zeta_1 \lesssim n^{-1}$ and $\zeta_2 \leq P(\mathcal{E}_7)$ under the assumptions in Lemma 6.

Phase I: bounding $\zeta_1$

We set $\delta$ as $\| B^* \|_F n^{-\frac{3}{6 (\log p)^2}}$, and can upper bound $\zeta_1$ as

$$\zeta_1 \leq \sum_{i=1}^{n} \sum_{j \neq \pi(i)}^{n} P \left( \left\| (X_{s-1,} - X_{t,}) B^* \right\|_2 \leq \delta \right) \ \leq \ \sum_{i=1}^{n} \sum_{j \neq \pi(i)}^{n} n^{-3} \lesssim n^{-1}, \tag{17}$$

where 1 comes from the small ball probability as in Lemma 2.6 in Latala et al. (2007), which is also stated as Lemma 12.

Phase II: bounding $\zeta_2$

Then we prove that $\zeta_2$ can be arbitrarily small under the SNR requirement in Eq. (7). Conditional on event $\mathcal{E}_7$, we have

$$\left\| (X_{s-1,} (B^* - B)) \right\|_2 \leq \frac{2c_0 p^{-2} (\log n)^3 (\log p)^2}{\delta^2} \| B^* \|_F^2 + \frac{2c_1 m (\log n)^2 \sigma^2 (1 + p/n)}{\frac{n}{\text{SNR}}} \| B^* \|_F^2 n^{-\frac{3}{6 (\log p)^2}} \right\|_F^2 \ .$$

in 2 we use the fact $p \leq n$. Since we have $n \geq \frac{p^4 (\log n)^6 (\log p)^4}{\rho(B^*^2)}$ and $\rho(B^2) \geq 18/c$, we conclude $\eta_1 \rightarrow 0$ as $n$ goes to infinity. Meanwhile, because of the assumptions in Eq. (7), we have $\eta_2$ to be a small positive constant.

Additionally, we can expand $\Delta/\delta^2$ as

$$\Delta \lesssim \frac{n^{\frac{6}{(\log p)^2}}}{\| B^* \|_F^2} \left( \frac{p (\log n)^3/2 (\log p)}{\sqrt{m}} \right) \frac{c_0 \sqrt{m} (\log n) \sigma (1 + \frac{p}{n}) + c_2 (\log n) \| B^* \|_F^2}{\text{SNR}} \ .$$

Following similar procedures as above, we can prove $\Delta/\delta^2$ to be a small positive constant given Eq. (7). Combining Eq. (18) and Eq. (19) together, we conclude

$$\eta_1 + \eta_2 + 2 \sqrt{\eta_1 + \eta_2} + \frac{\Delta}{\delta^2} < 1,$$

which suggests that $\zeta_2$ equals zero conditional on events $\mathcal{E}_7$. Therefore, we obtain

$$\zeta_2 \leq P(\mathcal{E}_7) \lesssim 2e^{-p} + 6p^{-2} + ne^{-c_0 m} + c_0 n^{-1} + c_1 ne^{-c_2 n} \lesssim 2e^{-p} + ne^{-c_0 m} + c_0 p^{-2} + c_1 ne^{-c_2 n}$$

and completes the proof together with Eq. (17), where 3 is due to Lemma 8, and 4 is because of $n \gtrsim p^2$.  

Optimal Estimator for Unlabeled Linear Regression
D.4. Supporting Lemmas for Theorem 2

Lemma 7 For arbitrary row $X_{i,:}$, we have

$$\|X_{i,:}\|_2 \leq 2\sqrt{p \log n},$$

with probability exceeding $1 - n^{-p}$.

Proof 9 Notice that $\|X_{i,:}\|^2$ is a $\chi^2$-RV with freedom $p$, we have

$$\mathbb{P} \left( \|X_{i,:}\|^2 \geq 4p \log n \right) \leq \exp \left( \frac{p}{2} (\log(4p \log n) - 4 \log n + 1) \right) \leq \exp (-p \log n) = n^{-p},$$

where in (1) we use $2 \log n \geq \log (4 \log n) + 1$, when $n \geq 4$.

Lemma 8 We have $\mathbb{P}(\mathcal{E}_T) \geq 1 - 2e^{-p} - 6p^{-2} - ne^{-c_0 \log n} - c_0 n^{-1} - c_1 n e^{-c_2 n}$.

Proof 10 Invoking Lemma 10, we have

$$\mathbb{P} \left( \|X_{i,:}^\top W\|_2 \leq c_0 \sqrt{m} \log n \sigma (n + p), \forall i \right) = 1 - \mathbb{P} \left( \|X_{i,:}^\top W\|_2 > c_0 \sqrt{m} \log n \sigma (n + p), \exists i \right) \geq 1 - \sum_i \mathbb{P} \left( \|X_{i,:}^\top W\|_2 > c_0 \sqrt{m} \log n \sigma (n + p) \right) \geq 1 - n^{1-p} - ne^{-c_0 \log n} - n^{-1} - c_1 n e^{-c_2 n}.$$ 

Then we conclude

$$\|X_{i,:} (B^* - B^\dagger)\|_2 \leq \left\|X_{i,:} \left( \widetilde{B} - B^\dagger \right) \right\|_2 + \frac{1}{n - h} \|X_{i,:}^\top W\|_2$$

$$\leq \|X_{i,:}\|_2 \left\|\widetilde{B} - B^\dagger \right\|_F + \frac{1}{n - h} \|X_{i,:}^\top W\|_2$$

$$\overset{1}{\leq} c_0 \frac{p \log n \sigma^2 (\log p)}{\sqrt{n}} \left\|B^\dagger \right\|_F + \frac{c_1 \sqrt{m} \log n \sigma (n + p)}{n - h}$$

$$\overset{2}{\leq} c_0 \frac{p \log n \sigma^2 (\log p)}{\sqrt{n}} \left\|B^\dagger \right\|_F + \frac{4}{5} c_1 \sqrt{m} \log n \sigma \left( 1 + \frac{p}{n} \right),$$

where in (1) we condition on Lemma 9 and Eq. (20), and in (2) we use the fact $h \leq n/4$.

Lemma 9 Provided that $n \geq p^2$, $h \leq n/4$, we have

$$\left\|\widetilde{B} - B^\dagger \right\|_F \leq \sqrt{\frac{p}{n \sigma}} \left\|B^\dagger \right\|_F \left( 4 \sqrt{6} + (\log n)(\log p) \right),$$

with probability at least $1 - 2e^{-p} - 6p^{-2}$ when $n, p$ are sufficiently large.

Proof 11 We assume that the first $h$ rows of $X$ are permuted w.l.o.g. First, we expand $X^\top \Pi^2 X$ as

$$X^\top \Pi^2 X = \sum_{i=1}^{h} X_{\pi(i),:}^\top X_{i,:} + \sum_{i=h+1}^{n} X_{i,:}^\top X_{i,:},$$
Optimal Estimator for Unlabeled Linear Regression

and obtain

$$\mathbb{P}\left( \left\| B^2 - \bar{B} \right\|_2 \geq \sqrt{\frac{p}{n}} \left\| B^2 \right\|_F \left( 4\sqrt{6} + (\log n)(\log p) \right) \right)$$

$$\leq \mathbb{P}\left( \frac{1}{n-h} \sum_{i=1}^{h} X_{\pi(i):j}. : B^2 \right\|_F + \frac{1}{n-h} \sum_{i=h+1}^{n} \left( X_{\pi(i):j}. : - I \right) B^2 \right\|_F \geq \sqrt{\frac{p}{n}} \left\| B^2 \right\|_F \left( 4\sqrt{6} + (\log n)(\log p) \right)$$

$$\leq \mathbb{P}\left( \frac{1}{n-h} \sum_{i=1}^{h} X_{\pi(i):j}. : B^2 \right\|_F \geq \frac{(\log n)(\log p)\sqrt{p}}{\sqrt{n}} \left\| B^2 \right\|_F \right)$$

$$\leq \mathbb{P}\left( \frac{1}{n-h} \sum_{i=h+1}^{n} \left( X_{\pi(i):j}. : - I \right) B^2 \right\|_F \geq 4\sqrt{\frac{6p}{n}} \left\| B^2 \right\|_F \right)$$

where (1) is because of the union bound. Then we separately bound $\zeta_1$ and $\zeta_2$.

**Phase I: Bounding $\zeta_1$** According to Lemma 8 in Pananjady et al. (2017a) (restated as Lemma 13), we can decompose the set $\{j : \pi(j) \neq j\}$ into three disjoint sets $\mathcal{I}_i$, $1 \leq i \leq 3$, such that $j$ and $\pi(j)$ does not lie in the same set. And the cardinality of set $\mathcal{I}_i$ is $h_i$ satisfies $|h_i| \leq h_i \leq h/3$. Adopting the union bound, we can upper-bound $\zeta_1$ as

$$\zeta_1 \leq \sum_{i=1}^{3} \mathbb{P}\left( \frac{1}{n-h} \left\| \sum_{j \in \mathcal{I}_i} X_{\pi(j):j}. : B^2 \right\|_F \geq \frac{(\log n)(\log p)\sqrt{p}}{3\sqrt{n}} \left\| B^2 \right\|_F \right)$$

$$\leq \sum_{i=1}^{3} \mathbb{P}\left( \frac{1}{n-h} \left\| \sum_{j \in \mathcal{I}_i} X_{\pi(j):j}. : \right\|_F \geq \frac{(\log n)(\log p)\sqrt{p}}{3\sqrt{n}} \left\| B^2 \right\|_F \right)$$

Defining $Z_i$ as $Z_i = \sum_{j \in \mathcal{I}_i} X_{\pi(j):j}.$, we would bound the above probability by invoking the matrix Bernstein inequality (cf. Thm 7.3.1 in Tropp (2015)). First, we have

$$\mathbb{E}\left( X_{\pi(j):j}. : \right) = (EX_{\pi(j):j}.)^T (EX_{j,:}) = 0,$$

due to the independence between $X_{\pi(j):j}. :$ and $X_{j,:}$. Then we upper bound $\left\| X_{\pi(j):j}. : \right\|_2$ as

$$\left\| X_{\pi(j):j}. : \right\|_2 \leq \left\| X_{\pi(j):j}. : \right\|_F \equiv \left\| X_{\pi(j):j}. : \right\|_2 \leq 4p \log n,$$

where (3) is because $X_{\pi(j):j}. :$ is rank-1, (4) is due to the fact $\left\| uv^T \right\|_F^2 = \text{Tr}(uv^Tvu^T) = \left\| u \right\|_2^2 \left\| v \right\|_2^2$ for arbitrary vector $u, v \in \mathbb{R}^p$, and (4) is because of Lemma 7.

In the end, we compute $\mathbb{E}\left( Z_i Z_i^T \right)$ and $\mathbb{E}\left( Z_i^T Z_i \right)$ as

$$\mathbb{E}\left( Z_i Z_i^T \right) = \mathbb{E}\left( \sum_{j_1, j_2 \in \mathcal{I}_i} X_{\pi(j_1):j_1}. : X_{j_2,:} X_{\pi(j_2):j_2}. : X_{j_2,:} X_{\pi(j_1):j_1}. : \right) \equiv \mathbb{E}\left( \sum_{j \in \mathcal{I}_i} X_{\pi(j):j}. : X_{j,:} X_{\pi(j):j}. : X_{j,:} \right)$$

$$\mathbb{E}\left( \sum_{j \in \mathcal{I}_i} X_{\pi(j):j}. : X_{j,:} X_{\pi(j):j}. : X_{j,:} \right) = p \mathbb{E}\left( \sum_{j \in \mathcal{I}_i} EX_{\pi(j):j}. : X_{\pi(j):j}. : \right) = p h_i I_{p \times p} = \mathbb{E}\left( ZZ^T \right),$$

where (5) and (6) is because of the fact such that $j$ and $\pi(j)$ are not within the set $\mathcal{I}_i$ simultaneously. To sum up, we invoke the matrix Bernstein inequality (cf. Thm 7.3.1 in Tropp (2015)) and have

$$\frac{1}{n-h} \left\| \sum_{j \in \mathcal{I}_i} X_{\pi(j):j}. : \right\|_2 \leq \frac{1}{3} \left( 4p \log n)(\log p) \right) + \frac{p \sqrt{16(\log n)^2(\log p)^2 + 6h_i \log p/p}}{n-h}$$
The proof is completed via combing the results in Phase I and Phase II: bounding

\[
\frac{p \sqrt{16 \log n}^2 (\log p)^2 + 6h \log p}{n - h} \leq \frac{4p}{3n} \sqrt{16 \log n}^2 (\log p)^2 + \frac{n}{2p} (\log n)(\log p) \leq \frac{4 \sqrt{p}}{3 \sqrt{n}} \times (\log n)(\log p),
\]

in (7) we \( n \geq p^2 \geq 32p \) and hence

\[
\frac{1}{n - h} \left\| \sum_{j \in \mathcal{I}} X_{\pi(j),:} X_{j,:) \right\|_2 \leq (\log n)(\log p) \left( \frac{16p}{9n} + \frac{4 \sqrt{p}}{9 \sqrt{n}} \right) \leq \sqrt{\frac{p}{n}} (\log n)(\log p),
\]

holds with probability exceeding \( 1 - 6p^{-2} \), where in (8) we use \( n \geq 256p/25 \).

**Phase II: Bounding** \( \zeta_2 \) We upper bound \( \zeta_2 \) as

\[
\zeta_2 \leq \mathbb{P} \left( \frac{1}{n - h} \left\| \sum_{i=h+1}^{n} (X_{i,:}^T X_{i,:} - I) B^2 \right\|_F \geq 4 \sqrt{\frac{6p}{n}} \| B^2 \|_F \right) \leq \mathbb{P} \left( \frac{1}{n - h} \left\| \sum_{i=h+1}^{n} (X_{i,:}^T X_{i,:} - I) \right\|_F \| B^2 \|_F \geq 4 \sqrt{\frac{6p}{n}} \| B^2 \|_F \right) \leq 2e^{-p}.
\]

where (9) is because of \( (n - h)^{-1} \left\| \sum_{i=h+1}^{n} (X_{i,:}^T X_{i,:} - I) \right\|_2 \leq 6 \sqrt{2p/(n - h)} \) with probability \( 2e^{-p} \) in Example 6.1 in Wainwright (2019) (also listed as Lemma 14) and \( h \leq n/4 \).

The proof is completed via combing the results in Phase I and Phase II.

**Lemma 10** For an arbitrary index \( i \), we have

\[
\mathbb{P} \left( \left\| X_{i,:} X_{i,:}^T W \right\|_2 \geq c_0 \sqrt{m} (\log n) \sigma (n + p) \right) \leq e^{-c_0 m} + n^{-2} + c_1 e^{-c_2 n}.
\]

**Proof 12** For the conciseness of notation, we define \( \delta = c_0 \sqrt{m} (\log n) \sigma (n + p) \). In addition, we assume that \( i = 1 \) w.l.o.g and prove this lemma with the leave-one-out trick, which is previously used in El Karoui (2013); El Karoui et al. (2013); El Karoui (2018); Chen et al. (2019); Sur et al. (2019). First we define a perturbed matrix \( \tilde{X} \) such that \( \tilde{X}_{j,:} = X_{j,:} \), \( 2 \leq j \leq n \), while \( X_{1,:} \in \mathbb{R}^{1 \times p} \) is a independent identically distributed Gaussian vector as \( X_{1,:} \), namely, \( N(0, I) \).

Then we can upper-bound the probability as

\[
\mathbb{P} \left( \left\| X_{1,:} X_{1,:}^T W \right\|_2 \geq \delta \right) \leq \mathbb{P} \left( \left\| X_{1,:} \tilde{X}^T W \right\|_2 + \left\| X_{1,:} \left( X - \tilde{X} \right)^T W \right\|_2 \right) \geq \delta \right)
\leq \mathbb{P} \left( \left\| X_{1,:} \left( X - \tilde{X} \right)^T W \right\|_2 \right) \geq 4p (\log n) \sqrt{\sigma m} \mathbb{P} \left( \left\| X_{1,:} \tilde{X}^T W \right\|_2 \geq \delta - 4p (\log n) \sqrt{\sigma m} \right).
\]

**Phase I: Bounding** \( \zeta_1 \) To bound \( \zeta_1 \), easily we can verify the following relation

\[
\left\| X_{1,:} \left( X - \tilde{X} \right)^T W \right\|_2 \leq \left\| X_{1,:} \left( X - \tilde{X} \right)^T W \right\|_F \leq \left\| X_{1,:} \right\|_F \left\| \left( X - \tilde{X} \right)^T W \right\|_2 \leq 4p (\log n) \sqrt{\sigma m},
\]

with probability exceeding \( 1 - n^{-p} - e^{-c_0 m} \), where (1) is because only the first row of \( X - \tilde{X} \) is nonzero, and (2) conditions on \( \mathcal{E}_0 \) and \( \left\| W \right\|_2 \leq 2 \sqrt{\sigma m} \) holds with probability at least \( 1 - e^{-c_0 m} \).

**Phase II: Bounding** \( \zeta_2 \) Since \( \delta - 4p (\log n) \sqrt{\sigma m} \geq n (\log n) \sqrt{\sigma m} \), we can upper-bound \( \zeta_2 \) as

\[
\zeta_2 \leq \mathbb{P} \left( \left\| X_{1,:} \tilde{X}^T W \right\|_2 \geq c_1 n (\log n) \sqrt{\sigma m} \right).
\]
Due to the construction of $\bar{X}$, we have $X_{1,:}$ to be independent of $\bar{X}$. Hence, we condition on $\bar{X}^\top W$ and obtain

$$
\zeta_2 \leq P \left( \left\| X_{1,:} \bar{X}^\top W \right\|_2 \geq c_1 n (\log n) \sqrt{m \sigma}, \; \left\| \bar{X}^\top W \right\|_F < 8 n \sqrt{m \sigma} \right) + P \left( \left\| \bar{X}^\top W \right\|_F \geq 8 n \sqrt{m \sigma} \right)
$$

$$
\leq E_{\bar{X}^\top W} \mathbb{1} \left( \left\| X_{1,:} \bar{X}^\top W \right\|_2 \geq c_2 (\log n) \left\| \bar{X}^\top W \right\|_F \right) + P \left( \left\| \bar{X}^\top W \right\|_F \geq 8 n \sqrt{m \sigma} \right).
$$

For $\zeta_{2,1}$, we define $Z = \left\| X_{1,:} \bar{X}^\top W \right\|_2^2$ and have

$$
\zeta_{2,1} \leq E_{\bar{X}^\top W} \mathbb{1} \left( \left| Z - E Z \right| \geq c_3 (\log n)^2 \left\| \bar{X}^\top W \right\|_F^2 \right)
$$

$$
\leq E_{\bar{X}^\top W} \exp \left( - \frac{(\log n)^4 \left\| \bar{X}^\top W \right\|_F^4}{\left\| \bar{X}^\top W \bar{W}^\top \bar{X} \right\|_F^2} \wedge \frac{(\log n)^2 \left\| \bar{X}^\top W \right\|_F^2}{\left\| \bar{X}^\top W \bar{W}^\top \bar{X} \right\|_F} \right) \leq n^{-2},
$$

where $\mathcal{R}$ is because of the Hanson-Wright inequality (Theorem 6.2.1 in Vershynin (2018)), and $\mathcal{R}$ is due to the stable rank $\rho(\bar{X}^\top W) \geq 1$. Meanwhile we upper-bound $\zeta_{2,2}$ as

$$
P \left( \left\| \bar{X}^\top W \right\|_2 \geq 8 n \sqrt{m \sigma} \right) \leq P \left( \left\| \bar{X} \right\|_{op} \left\| W \right\|_F \geq 8 n \sqrt{m \sigma} \right)
$$

$$
\leq P \left( \left\| \bar{X} \right\|_{op} \geq 2 \left( \sqrt{n} + \sqrt{p} \right) \right) + P \left( \left\| W \right\|_F \geq \frac{8 n \sqrt{m \sigma}}{2 \left( \sqrt{n} + \sqrt{p} \right)}, \; \left\| \bar{X} \right\|_{op} \leq 2 \left( \sqrt{n} + \sqrt{p} \right) \right)
$$

$$
\leq P \left( \left\| \bar{X} \right\|_{op} \geq 2 \left( \sqrt{n} + \sqrt{p} \right) \right) + P \left( \left\| W \right\|_F \geq \sqrt{2 n m \sigma} \right) \leq e^{-\kappa_0 n} + e^{-0.8 n m},
$$

where $\mathcal{R}$ is because of the union bound, in $\mathcal{R}$ we use $p \leq n$, and in $\mathcal{R}$ we use $\left\| \bar{X} \right\|_{op} \geq 2 \left( \sqrt{n} + \sqrt{p} \right)$ with probability less than $e^{-\kappa_0 n}$ (Chandrasekaran et al., 2012) and the fact $\left\| W \right\|_F \geq \sigma^2$ is a $\chi^2$-RV with $n m$ freedom, and Lemma 11.

### E. Useful Facts

This section lists some useful facts for the sake of self-containing.

**Lemma 11.** For a $\chi^2$-RV $Z$ with $\ell$ freedom, we have

$$
P \left( Z \leq t \right) \leq \exp \left( \frac{\ell}{2} \left( \log \frac{t}{\ell} - t + 1 \right) \right), \; t < \ell;
$$

$$
P \left( Z \geq t \right) \leq \exp \left( \frac{\ell}{2} \left( \log \frac{t}{\ell} - t + 1 \right) \right), \; t > \ell.
$$

**Lemma 12 (Small ball probability, Lemma 2.6 in Latała et al. (2007)).** Given an arbitrary fixed vector $y \in \mathbb{R}^n$, we have

$$
P \left( \left\| y - A g \right\|_2 \leq \alpha \left\| A \right\|_F \right) \leq \exp \left( \kappa \log(\alpha) \rho(A) \right), \; \forall \alpha \in (0, \alpha_0),
$$

where $g$ is a Gaussian RV following $\mathcal{N}(0, I_{n \times n})$, $A \in \mathbb{R}^{n \times n}$ is a non-zero matrix, and $\alpha_0 \in (0, 1)$ and $\kappa > 0$ are some universal constants.

**Lemma 13 (Lemma 8 in Pananjady et al. (2017a)).** Consider an arbitrary permutation map $\pi$ with Hamming distance $k$ from the identity map, i.e., $d_H(\pi, I) = k$. We define the index set $\{ i : i \neq \pi(i) \}$ and can decompose it into 3 independent sets $\mathcal{I}_j$ ($1 \leq j \leq 3$), i.e., $i$ and $\pi(i)$ are in different sets $\mathcal{I}_j$ for arbitrary $i \in \{ i : i \neq \pi(i) \}$, such that the cardinality of each set satisfies $|\mathcal{I}_j| \geq |k/3| \geq k/5$.

**Lemma 14 (Example 6.1 in Wainwright (2019)).** Let $G \in \mathbb{R}^{n_1 \times n_2}$ be generated with iid standard normal random variables, we have $\left\| G \right\|_{op} \leq 4 \sqrt{n_2/n_1}$, hold with probability exceeding $1 - 2e^{-n_2/2}$.