A Convex Optimization Framework for Bi-Clustering

A. Proofs

A.1. Proof of Theorem 1

Let \( R = \{(i, j) : Y_{ij}^* = b_i\} \). Let \( \mathcal{P}_R \) be the projection operator on matrices such that \((\mathcal{P}_R Z)_{ij} = Z_{ij}\) if \((i, j) \in R\) and \((\mathcal{P}_R Z)_{ij} = 0\) otherwise.

Recall that \( Y^* = USV^T \) where \( U = UCUB \) and \( V = VCVB \). We use the index \( p_i \) to refer to the cluster to which node \( i \) belong. Note that

\[
(UU^T)_{ij} = \frac{(U^p_i, U^p_j)}{\sqrt{K_{p_i}K_{p_j}}} \leq \frac{u_1}{K}, \tag{2}
\]

\[
(VV^T)_{ij} = \frac{(V^q_i, V^q_j)}{\sqrt{L_{q_i}L_{q_j}}} \leq \frac{u_2}{L} \tag{3}
\]

and

\[
(UV^T)_{ij} = \frac{(U^p_i, V^q_j)}{\sqrt{K_{p_i}L_{q_j}}} \leq \max\left\{\frac{u_1}{K}, \frac{u_2}{L}\right\}. \tag{4}
\]

Furthermore, if \( U_B \) is orthogonal, we have \((UU^T)_{ij} = \frac{1}{K_{p_i}}\) if \( p_i = p_j \) and \((UU^T)_{ij} = 0\) otherwise. Similarly, if \( V_B \) is orthogonal, we have \((VV^T)_{ij} = \frac{1}{L_{q_i}}\) if \( q_i = q_j \) and \((VV^T)_{ij} = 0\) otherwise.

Define the projection operators

\[
\mathcal{P}_TZ := UU^TZ + ZVV^T - UU^TZVV^T \tag{5}
\]

and

\[
\mathcal{P}_{T\perp}Z := Z - \mathcal{P}_TZ. \tag{6}
\]

For any matrix \( X \) such that \( \|X\| \leq \lambda, \quad UU^T + \frac{1}{\lambda}\mathcal{P}_T\perp X \)

is a subgradient of \( \| \cdot \|_* \) at \( Y^* \). For any feasible \( Y \), we therefore have

\[
\|Y\|_* \geq \|Y^*\|_* + \langle UV^T + \frac{1}{\lambda}\mathcal{P}_T\perp X, Y - Y^*\rangle
\]

which gives

\[
\langle X, Y^* - Y\rangle + \lambda\|\|Y\|_* - \|Y^*\|_*\| \geq \langle \mathcal{P}_T X - \lambda UV^T, Y^* - Y\rangle. \tag{7}
\]

**Lemma 1.** Suppose that \( B \) is full-rank. For any \( n_1 \times n_2 \) matrix \( Z \), we have

\[
\langle \mathcal{P}_T Z, Z\rangle = \frac{1}{K_{p_i}} \sum_{i' \in C_{r_i}} Z_{i'j} + \sum_{t=1}^{n_2} \left( \frac{V^q_{i'} V^q_j}{\sqrt{L_{q_i}L_{q_j}}} Z_{it} \right) - \frac{1}{K_{p_i}} \sum_{i' \in C_{r_i}} \left( \frac{V^q_{i'} V^q_j}{\sqrt{L_{q_i}L_{q_j}}} Z_{i't} \right) \tag{8}
\]

if \( r_1 \leq r_2 \) and

\[
\langle \mathcal{P}_T Z, Z\rangle = \frac{1}{L_{q_j}} \sum_{j' \in D_{c_{q_j}}} \langle Z_{ij}, Z_{ij'} \rangle - \frac{1}{L_{q_j}} \sum_{j' \in D_{c_{q_j}}} \left( \frac{1}{K_{p_i}} \sum_{i=1}^{n_1} \frac{\langle U^p_i, U^p_j \rangle}{\sqrt{K_{p_i}K_{p_j}}} Z_{ij'} \right) \tag{9}
\]

if \( r_2 \leq r_1 \).

**Proof.** By the assumption that \( B \) is full rank, it has to be that either \( U_B \) is \( r_1 \times r_1 \) orthogonal (when \( r_1 \leq r_2 \)) or \( V_B \)

is \( r_2 \times r_2 \) orthogonal (when \( r_2 \leq r_1 \)). Suppose \( r_1 \leq r_2 \), then we obtain (7) from (2), (3) and (5). Similarly, for the case when \( r_2 \leq r_1 \), we obtain (8).

**Lemma 2.** With probability at least \( 1 - n^{-\beta} \) the followings hold:

\[
\|W - EW\| \leq \lambda \tag{10}
\]

and for all \( i, j \)

\[
\|\langle \mathcal{P}_T (W - EW) \rangle_{ij}\| \leq \lambda \max\left\{\frac{u_1}{K}, \frac{u_2}{L}\right\} \tag{11}
\]

where

\[
\lambda = c_1 \left( b\beta \log n + \sqrt{\beta VN \log n} \right). \tag{12}
\]

**Proof.** For (9), consider \( W - EW \) as the sum of independent, zero-mean random matrices:

\[
W - EW = \sum_{i,j} X_{i,j} \tag{13}
\]

where

\[
X_{i,j} = W_{ij} e_i^T - EW_{ij} e_i^T \tag{14}
\]

and \( e_i \) denotes the \( i \)-th vector of the standard basis. Note that

\[
\|X_{i,j}\| \leq 2b \quad \forall i, j \tag{15}
\]

and

\[
\max\left\{\sum_{i,j} \|EX_{i,j}^T X_{i,j}\|, \sum_{i,j} \|EX_{i,j}^T X_{i,j}\|\right\} \leq VN \tag{16}
\]

By applying matrix Bernstein inequality, we obtain (9).

For (10), let \( Z = W - EW \) be a zero-mean random matrix where we have \( |Z_{ij}| \leq 2b, \forall (i,j) \). Suppose \( r_1 \leq r_2 \), then \( \langle \mathcal{P}_T Z \rangle_{ij} \) is given by (7). Note that \( u_1 = 1 \) in this case. The first summation term in (7) is the average of \( K_{p_i} \)

independent random variables \( Z_{i't} \). By applying standard Bernstein inequality, we have that

\[
\left| \frac{1}{K_{p_i}} \sum_{i' \in C_{r_i}} Z_{i't} \right| \leq c \left( b\beta \log n + \sqrt{R\beta VN \log n} \right) \leq \frac{u_1}{K} \tag{17}
\]

where

\[
R = \sum_{i,j} \|EX_{i,j}^T X_{i,j}\| \leq VN \tag{18}
\]

and

\[
\sum_{i,j} \|EX_{i,j}^T X_{i,j}\| \leq VN \tag{19}
\]

By applying matrix Bernstein inequality, we obtain (10).
with probability at least $1 - n^{-\beta}$. The second summation term in (7) is the sum of $n_2$ zero-mean random variables $Z_i$ with

$$Z_i = \frac{(V^*_B Y)_i}{\sqrt{L_n u_2}} Z_i.$$  

We have that $|\hat{Z}_i| \leq \frac{2 u_b}{L}$ and Var $Z_i \leq \frac{u_2^2}{L} V$. Again, by standard Bernstein inequality, we have that

$$\sum_{i=1}^{n_2} |\hat{Z}_i| \leq c u_2 b \log n + u_2 \sqrt{\beta V} n_2 \log n \leq \lambda \frac{u_2}{L}$$

with probability at least $1 - n^{-\beta}$. The last summation term of (7) can be bounded similarly by noting that the magnitude of the average is no larger than the magnitude of the individual terms.

The case of $r_2 \leq r_1$ can be bounded similarly for each corresponding term of (8). The proof is completed by applying a union bound over all $i, j$. □

We show that with probability at least $1 - n^{-\beta}$ the following holds for all feasible $Y \neq Y^*$:

$$(W, Y^* - Y + \lambda(\|Y\|_* - \|Y^*\|_*)) > 0$$

which implies that $Y^*$ is the unique solution of program (1):

$$(W, Y^* - Y) + \lambda(\|Y\|_* - \|Y^*\|_*) = (EW, Y^* - Y) + (W - EW, Y^* - Y) + \lambda(\|Y\|_* - \|Y^*\|_*)$$

$$\geq \min\{E_1, E_0\} \|Y^* - Y\|_1 + \lambda(\|Y\|_* - \|Y^*\|_*)$$

$$\geq \min\{E_1, E_0\} \|Y^* - Y\|_1 + \lambda \left(\frac{u_2}{L}\right) \|Y^* - Y\|_1$$

$$\geq \min\{E_1, E_0\} \|Y^* - Y\|_1 - \lambda \max\left\{\frac{u_1}{K}, \frac{u_2}{L}\right\} \|Y^* - Y\|_1$$

$$= \left(\min\{E_1, E_0\} - \lambda \max\left\{\frac{u_1}{K}, \frac{u_2}{L}\right\}\right) \|Y^* - Y\|_1$$

where $\lambda \geq 0$.

A.2. Proof of Theorem 2

We shall apply Theorem 1 and establish an upper-bound for $\max\{u_1/K, u_2/L\}$.

Since $r_1 \leq r_2$, $B$ is at most rank-$r_1$. By Theorem 1.1 of (Rudelson & Vershynin, 2009), the smallest singular value of $B$ is at least $c_1 (\sqrt{r_2} - \sqrt{r_1})$ for some universal constant $c_1$, $0 < c_1 < 1$ with probability at least $1 - \frac{1}{2} n^{-\beta}$ provided $r_1 \geq \frac{\beta \log r_1}{c_1}$. This implies that $B$ is full rank with $\text{rank}(B) = r_1$.

Recall that we defined $U_B S_B V_B^T$ as an SVD of $KBL$. Since $K$ and $L$ do not change the rank of $B$, we have that $U_B$ is orthogonal and therefore $u_1 = 1$. Furthermore, the smallest singular value of $S_B$ is at least $c_1 \sqrt{KL}(\sqrt{r_2} - \sqrt{r_1})$.

We upper-bound $u_2$ by using the lower-bound on the singular value of $KBL$. First, note that the $(p, q)$-entry of $KBL$ is either $+\sqrt{KL}r_q$ or $-\sqrt{KL}r_q$, we therefore have

$$n_1 L_q = \sum_{p=1}^{r_1} K_q L_q$$

$$= \|\langle KBL \rangle e_q\|^2$$

$$= \|U_B S_B V_B^T e_q\|^2$$

$$= \|S_B V_B^T e_q\|^2$$

$$\geq c_1^2 K L (\sqrt{r_2} - \sqrt{r_1} - 1)^2 \|V_B^T\|^2.$$  

Define $\gamma$ such that $L = \gamma K$. Continuing from the above, we have

$$\|V_B^T\|^2 \leq \frac{n_1 L_q}{c_1^2 K L (\sqrt{r_2} - \sqrt{r_1} - 1)^2}$$

$$\leq \frac{\psi_1}{\gamma n_1} \frac{r_1}{r_1 - 1}$$

where in (a) we use the relationship $r_2 \geq \frac{\psi_1}{\gamma n_1} r_1 \geq \frac{\psi n_2}{\gamma n_1} (r_1 - 1)$ and that $\frac{r_1}{r_1 - 1} \leq 2$. Since the above bound for $\|V_B^T\|^2$ holds for all $q$, we now have

$$\frac{u_2}{L} = \frac{u_2}{\gamma K} \leq \frac{\lambda}{K} \min \left\{ \frac{1}{\gamma}, \frac{2}{c_1^2 \psi \left(\frac{\sqrt{\psi n_2}}{n_1} - \sqrt{\gamma}\right)^2} \right\}$$

$$\leq \frac{8 n_1}{c_1^2 \psi^2 n_2} \frac{1}{K}$$

where the last inequality can be obtained by considering the case of $\gamma < \frac{c_1^2 \psi n_2}{8 n_1}$ and $\gamma \geq \frac{c_1^2 \psi n_2}{8 n_1}$ respectively.

Applying Theorem 1 with failure probability $\frac{1}{2} n^{-\beta}$ and a union bound completes the proof.

A.3. Proofs of Theorem 3 and Theorem 4

We refer the reader to (Lim et al., 2014) for the proofs of the analogous results in clustering. Theorem 3 corresponds to Theorem 2 and Corollary 1 in (Lim et al., 2014), while Theorem 4 corresponds to Theorem 4 in (Lim et al., 2014). An additional element of Theorem 3 in the present paper is the introduction of $u_1$ and $u_2$. These present no difficulty by simply observing their respective range $\frac{1}{r_1} \leq u_1 \leq 1$ and $\frac{1}{r_2} \leq u_2 \leq 1$.  
