Supplementary Material of " Crowdsourcing via Annotator Co-occurrence Imputation and Provable Symmetric Nonnegative Matrix Factorization"

A. Notation

Notation	Definition
x	scalar in $\mathbb{R}$
$\boldsymbol{x}$	vector in $\mathbb{R}^n$, i.e., $\boldsymbol{x} = [x_1, \dots, x_n]^\top$
$\boldsymbol{X}$	matrix in $\mathbb{R}^{m \times n}$ with $\boldsymbol{X}(i,j) = x_{i,j}$
$[\boldsymbol{X}]_{i,j}$ or $\boldsymbol{X}(i,j)$	(i,j)th entry of $\boldsymbol{X}$
$\boldsymbol{X} \geq \boldsymbol{0}$	$\boldsymbol{X}(i,j) \geq 0 \,\forall\, (i,j)$
$\kappa(\boldsymbol{X})$	condition number of $\boldsymbol{X}$
$\sigma_{\max}(\boldsymbol{X})$	maximum singular value of $\boldsymbol{X}$
$\sigma_{\min}(\boldsymbol{X})$	minimum singular value of $\boldsymbol{X}$
$\|\boldsymbol{X}\|_2$	2-norm of $\boldsymbol{X}$ (same as $\sigma_{\max}(\boldsymbol{X})$)
$\|\boldsymbol{X}\|_{\rm F}$	Frobenius norm of $\boldsymbol{X}$
$\mathcal{R}(\boldsymbol{X})$	range space of $\boldsymbol{X}$
$\mathsf{cone}(\boldsymbol{X})$	conic hull of $\boldsymbol{X}$: $\{\boldsymbol{y} \mid \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta},\ \forall \boldsymbol{\theta} \geq \boldsymbol{0}\}$
$\|\boldsymbol{x}\|_2$	ℓ_2-norm of $\boldsymbol{x}$
$\|\boldsymbol{x}\|_1$	ℓ_1-norm of $\boldsymbol{x}$
$\mathsf{Diag}(\boldsymbol{x})$	diagonal matrix with $x_1, \dots, x_n$ in the diagonal
$\dagger$	pseudo-inverse
$\top$	transpose
$\lvert\mathcal{C}\rvert$	the cardinality of the set $\mathcal{C}$
$[T]$	$\{1, \dots, T\}$ for an integer T
$\boldsymbol{I}$	identity matrix with proper size
$\boldsymbol{1}$	all-one vector with proper size
$\boldsymbol{0}$	all-zero vector or matrix with proper size
$\boldsymbol{e}_i$	unit vector with the ith element being 1
$\mathbb{R}^n_+$	nonnegative orthant of $\mathbb{R}^n$

B. More Details of The Robust Co-occurrence Imputation Algorithm

B.1. Iteratively Reweighted Algorithm for Robust Co-occurrence Imputation

In order to design an algorithm for solving Problem (9), we approximate (9) using a smooth version of the objective function. Specifically, we propose to use

$$\underset{\boldsymbol{U}_m, \boldsymbol{U}_j,\, \forall (m,j) \in \boldsymbol{\Omega}}{\text{minimize}} \sum_{(m,j) \in \boldsymbol{\Omega}} \left(\|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{U}_m \boldsymbol{U}_j^\top\|_{\rm F}^2 + \xi \right)^{\frac{1}{2}} \tag{17a}$$

$$\text{subject to } \|\boldsymbol{U}_m\|_{\rm F} \leq D,\ \|\boldsymbol{U}_j\|_{\rm F} \leq D,\ \forall m, \tag{17b}$$

where $\xi > 0$ is a small number.

We update $\boldsymbol{U}_m$ by fixing $\{w_{m,j}\}_{(m,j) \in \boldsymbol{\Omega}}$ and $\boldsymbol{U}_j$'s where $j \neq m$. Then, we can update $\{w_{m,j}\}_{(m,j) \in \boldsymbol{\Omega}}$, by fixing $\boldsymbol{U}_m$ and $\boldsymbol{U}_j$, for all $(m,j) \in \boldsymbol{\Omega}$. In each iteration t, the sub-problem to solve $\boldsymbol{U}_m$ can be written as

$$\underset{\boldsymbol{U}_m}{\text{minimize}} \sum_{j \in \mathcal{S}_m} w_{m,j}^{(t)} \|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{U}_m (\boldsymbol{U}_j^{(t)})^\top\|_{\rm F}^2 \tag{18a}$$

$$\text{subject to } \|\boldsymbol{U}_m\|_{\rm F} \leq D, \tag{18b}$$

where $\mathcal{S}_m = \{j \mid \widehat{\boldsymbol{R}}_{m,j} \text{ or } \widehat{\boldsymbol{R}}_{j,m} \text{ is observed}\}$. The problem in (18) is a second-order cone-constrained quadratic program, and can be solved using any off-the-shelf convex optimization algorithm. We propose to use the *projected gradient descent*

(PGD) algorithm due to its simplicity. Specifically, in iteration r of the PGD conducted during the tth outer iteration, $\boldsymbol{U}_m$ is updated via

$$\boldsymbol{U}_m^{(t,r+1)} \leftarrow \mathsf{Proj}_{\mathcal{D}}\left(\boldsymbol{U}_m^{(t,r)} - \beta \boldsymbol{G}_m^{(t,r)}\right),$$

where $\beta > 0$ is the step size, $\mathsf{Proj}_{\mathcal{D}}(\cdot) : \mathbb{R}^{K\times K} \rightarrow \mathbb{R}^{K\times K}$ denotes the orthogonal projection onto the set $\mathcal{D} = \{\boldsymbol{X} \in \mathbb{R}^{K\times K} \mid \|\boldsymbol{X}\|_{\rm F} \leq D\}$, and $\boldsymbol{G}_m^{(t,r)}$ is the gradient of the objective function (18a) w.r.t to $\boldsymbol{U}_m$. Specifically, we have

$$\boldsymbol{G}_m^{(t,r)} = \sum_{j\in\mathcal{S}_m} w_{m,j}^{(t)} \left(\boldsymbol{U}_m^{(t,r)}(\boldsymbol{U}_j^{(t)})^\top \boldsymbol{U}_j^{(t)} - \widehat{\boldsymbol{R}}_{m,j}\boldsymbol{U}_j^{(t)}\right).$$

The step size is selected as the inverse of the Lipschitz constant of the gradient. In addition, the projection is simply re-scaling; i.e., for any $\boldsymbol{Z} \in \mathbb{R}^{K\times K}$,

$$\mathsf{Proj}_{\mathcal{D}}(\boldsymbol{Z}) = \begin{cases} \boldsymbol{Z}, & \boldsymbol{Z} \in \mathcal{D} \\ \frac{\boldsymbol{Z}}{\|\boldsymbol{Z}\|_F}, & \boldsymbol{Z} \notin \mathcal{D}. \end{cases}$$

Note that we let

$$\boldsymbol{U}_m^{(t+1)} \leftarrow \boldsymbol{U}_m^{(t,r_t^\star)}, \quad \boldsymbol{U}_m^{(t+1,0)} \leftarrow \boldsymbol{U}_m^{(t+1)},$$

where $r_t^\star$ is the number of iterations where the PGD stops for updating $\boldsymbol{U}_m^{(t)}$. After $\boldsymbol{U}_m$ for all m are updated using PGD, we update $w_{m,j}$, for all $(m,j) \in \boldsymbol{\Omega}$, by the following:

$$w_{m,j}^{(t+1)} \leftarrow \left(\|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{U}_m^{(t)}(\boldsymbol{U}_j^{(t)})^\top\|_{\rm F}^2 + \xi\right)^{-\frac{1}{2}}, \ \forall (m,j) \in \boldsymbol{\Omega}.$$

B.2. Complexity and Convergence

The per-iteration complexity of the algorithm is often not large, due to its first-order optimization nature. The complexity-dominating step are the computation of the step size and constructing the gradient, which both cost $O(MK^3)$ flops. This is acceptable since K is normally small.

Iteratively reweighted algorithms' stationary-point convergence properties have been well understood. By a connection between the algorithm and the *block successive upper bound minimization* (BSUM) (Razaviyayn et al., 2013), it is readily seen that the solution sequence converges to a stationary point of (17). Although global optimality of the algorithm may be much harder to establish, such a procedure often works well in practice—which presents a valuable heuristic for tackling the stability-guaranteed co-occurrence imputation criterion in Theorem 2, i.e., Problem (9).

C. More Details of Experiments

Parameters. The stopping criterion for all the iterative algorithms in the experiments is set such that the algorithms are terminated when the relative change of their respective cost functions is less than 10^{-6}. For the proposed SymNMF algorithm, we set $\alpha_{(0)} = 10^{-6}$, and we use two $\alpha_{(t)}$ scheduling rules in our simulations and real data experiments, respectively. Specifically, for simulations that demonstrate the convergence properties of the proposed algorithm, we use $\alpha_{(t)} = \psi^{t+1}$ where $0 < \psi < 1$. For the rest of the simulations and real data experiments, we let $\alpha_{(t)} = \alpha_{(0)}$ for simplicity. We run all the experiments in Matlab 2018b on Windows 10 on an Intel I7 CPU running at 3.40 GHZ.

C.1. Synthetic Data Simulations

C.1.1. Identifiability

In this section, we analyze the D&S model identifiability of the proposed framework using synthetic data experiments.

First, we consider the noiseless case where we directly generate $\boldsymbol{R}_{m,j} = \boldsymbol{A}_m\boldsymbol{D}\boldsymbol{A}_j^\top$ for $(m,j) \in \boldsymbol{\Omega}$ and observe if the confusion matrices and the prior can be identified by the algorithms up to a common column permutation. We fix $M = 25$ annotators and the number of classes $K = 3$. An annotator is chosen randomly from M annotators and is made as a "class specialist" of all the classes $1,\ldots,K$. This is achieved by setting its confusion matrix $\boldsymbol{A}_m$ to be close to an identity matrix. Specifically, for the chosen "class specialist", we set $\|\boldsymbol{A}_m(k,:) - \boldsymbol{e}_k^\top\|_2 \leq \varepsilon.\forall k$, with $\varepsilon = 0.10$. In this way, the $\boldsymbol{H}$ matrix

Table 6. Average MSE of the proposed methods for $M = 25$, $K = 3$ with different block missing proportions (noiseless case).

Algorithms	Miss=70%	Miss=50%	Miss=30%
`RobSymNMF`	4.10×10^{-3}	1.70×10^{-3}	3.44×10^{-4}
`DesSymNMF`	2.84×10^{-4}	4.59×10^{-4}	3.05×10^{-4}

Table 7. Average MSE and the runtime of the proposed methods and baselines for $M = 25, K = 3, p = 0.3$ for different values of N

Algorithms	$N = 1000$	$N = 5000$	$N = 10000$	Time (s)
`RobSymNMF`	**0.0099**	**0.0019**	**0.0012**	0.342
`DesSymNMF`	**0.0127**	0.0038	0.0029	0.072
`MultiSPA`	0.2248	0.1645	0.1575	0.0148
`CNMF`	0.0314	**0.0036**	**0.0009**	22.475
`TensorADMM`	0.0218	0.0041	0.0011	27.263
`Spectral-D&S`	0.0465	0.0259	0.0050	17.492
`MV-EM`	0.0495	0.0866	0.1051	0.055

as defined in (6) approximately satisfy the SSC (see Definition 1). The columns of the confusion matrices for the rest of the annotators and the prior probability vector $\boldsymbol{\lambda} \in \mathbb{R}^K$ are generated using Dirichlet distribution with parameter $\boldsymbol{\mu} = \mathbf{1} \in \mathbb{R}^K$. We generate different missing proportions by observing each pairwise blocks with a probability smaller than one. Using these observed pairwise blocks, the proposed algorithms are run and the mean squared error (MSE) of the confusion matrices and the prior vector are estimated. The MSE is computed as follows:

$$\text{MSE} = \min_{\boldsymbol{\Pi}} \frac{1}{MK+1} \left(\|\boldsymbol{\Pi}^\top \boldsymbol{\lambda} - \widehat{\boldsymbol{\lambda}}\|_2^2 + \sum_{m=1}^{M} \|\boldsymbol{A}_m \boldsymbol{\Pi} - \widehat{\boldsymbol{A}}_m\|_{\text{F}}^2 \right), \tag{19}$$

where $\boldsymbol{\Pi}$ is a permutation matrix and $\widehat{\boldsymbol{A}}_m, m = [M]$ and $\widehat{\boldsymbol{\lambda}}$ are the outputs by the algorithms.

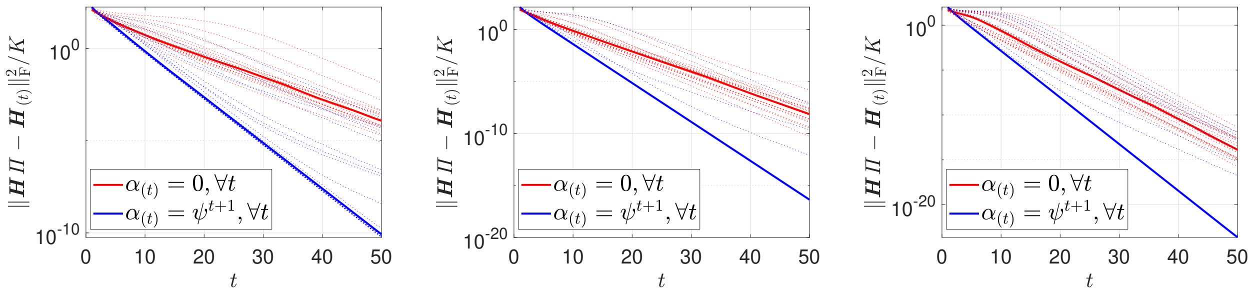
Table 6 presents the MSE of the proposed methods for different proportions of the missing co-occurrences, averaged over 20 different trials. Both the proposed methods output low MSE values in all the cases. One can see that the MSE of the `RobSymNMF` decreases when more blocks are observed, which is consistent with Theorem 2. Since we consider the noiseless case by observing $\boldsymbol{R}_{m,j} = \boldsymbol{A}_m \boldsymbol{D} \boldsymbol{A}_j^\top$ for all $(m, j) \in \boldsymbol{\Omega}$, the algorithm `DesSymNMF` is able to impute all the missing pairwise co-occurrences accurately via (7)-(8). Therefore, the MSE of the `DesSymNMF` is more or less unaffected with changing co-occurrence missing proportions.

Table 7 presents the average MSE and the runtime of the methods under test using various numbers of data items. We fix $M = 25$, $K = 3$ and vary the number of data items N. The generating process for the confusion matrices and the prior vector is the same as that used in Table 6. Once the confusion matrices $\boldsymbol{A}_m, m = [M]$ are generated, the labels from each annotator m for a data item with true label $c \in [K]$ is randomly chosen from $[K]$ using the probability distribution $\boldsymbol{A}_m(:, c)$. An annotator label for each data item is retained with probability $p < 1$ which is fixed at 0.3. Using such labels, the co-occurrences are estimated via (3). In all the cases in Table 7, there are 4% of the pairwise co-occurrences missing. One can see that the proposed methods, especially `RobSymNMF`, outperform the other methods in most of the cases and also enjoy promising runtime performance. The `DesSymNMF` imputes all the missing blocks, even though there are no designated annotators and still provides good performance. This is because most co-occurrences are available and the conditions for using (7)-(8) are almost always satisfied. Particularly, the MSEs of the proposed methods are at least 40% lower than the best-performing baseline, when the number of data items are small (see $N = 1000$). This shows the advantages of the pairwise co-occurrence based methods in the sample-starved regime. As N increases, the MSEs of all the methods become better and closer.

C.1.2. Convergence

In this section, we compare the convergence behaviors of the proposed SymNMF algorithm [cf. Eq. (13)] and the SymNMF algorithm proposed in (Huang et al., 2014). The proposed algorithm uses a shifted ReLU function for the $\boldsymbol{H}$ update with $\alpha_{(t)} > 0$. The algorithm in (Huang et al., 2014) has nonnegative thresholding, i.e., a ReLU function with $\alpha_{(t)} = 0$ for all t.

In our proof of Theorem 3, we assume that $\alpha_{(t)}$ is chosen such that a key condition is always satisfied; see Eqs. (49) and (50). In practice, these conditions may not be checkable. A heuristic way of selecting $\{\alpha_{(t)}\}$ is to use a diminishing sequence

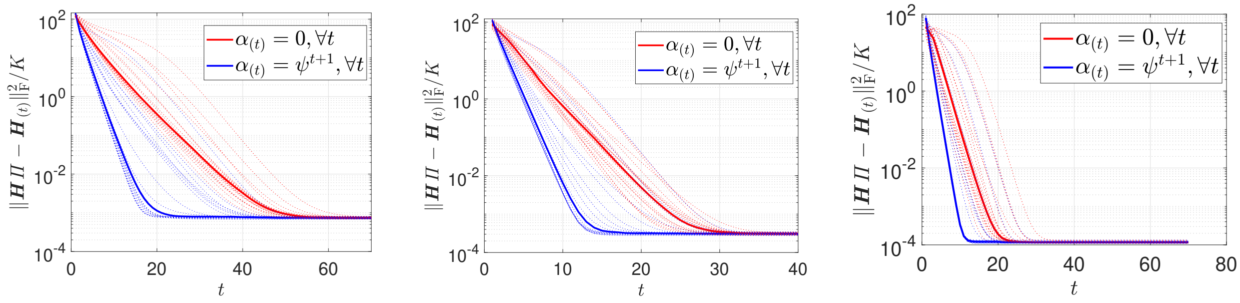


(a) $\varphi = 0.7, \psi = 0.75$

(b) $\varphi = 0.5, \psi = 0.75$

(c) $\varphi = 0.3, \psi = 0.75$

Figure 2. Convergence of the SymNMF algorithm with $\alpha_{(t)} = \psi^{t+1}$ (proposed) and $\alpha_t = 0$ for different levels of sparsity of $\boldsymbol{H} \in \mathbb{R}^{1000\times 3}$ (noiseless case). Dashed line represents each trial and the bold line denotes the median of the 20 independent trials.



(a) $\varphi = 0.7, \psi = 0.60$

(b) $\varphi = 0.5, \psi = 0.60$

(c) $\varphi = 0.3, \psi = 0.60$

Figure 3. Convergence of the SymNMF algorithm with $\alpha_{(t)} = \psi^{t+1}$ (proposed) and $\alpha_t = 0$ for different levels of sparsity of $\boldsymbol{H} \in \mathbb{R}^{1000\times 3}$ and SNR=30dB. Dashed line represents each trial and the bold line denotes the median of the 20 independent trials.

$\{\alpha_{(t)}\}$. In simulations, we found that using such sequences $\{\alpha_{(t)}\}$ can often accelerate convergence.

We consider a nonnegative matrix $\boldsymbol{H} \in \mathbb{R}_+^{J\times K}$ and control its sparsity (i.e., the number of zero entries in $\boldsymbol{H}$) using a parameter φ such that $1 - \varphi = \mathsf{Pr}([\boldsymbol{H}]_{j,k} = 0)$. The nonzero entries are randomly sampled from a uniform distribution between 0 and 1. Using the matrix $\boldsymbol{H}$, the symmetric nonnegative matrix $\boldsymbol{X} \in \mathbb{R}_+^{J\times J}$ is formed by $\boldsymbol{X} = \boldsymbol{HH}^\top$ and its rank-K square root decomposition is performed, i.e., $\boldsymbol{X} = \boldsymbol{UU}^\top$. The matrix $\boldsymbol{U} \in \mathbb{R}^{J\times K}$ resulted from the rank-K square root decomposition is input to the algorithms. Both the algorithms are initialized by $\boldsymbol{Q}_{(0)} = \boldsymbol{I}$.

Fig. 2 shows $\|\boldsymbol{H\Pi} - \boldsymbol{H}_{(t)}\|_{\mathrm{F}}^2/K$, where $\boldsymbol{\Pi}$ is a permutation matrix, against the iteration index t. One can see that for different sparsity levels, the proposed SymNMF algorithm converges faster. It can also be observed that as the sparsity level increases (i.e., φ decreases), both SymNMF algorithms converge quickly to low MSE levels.

Fig. 3 shows the convergence behaviour of the algorithms when zero-mean i.i.d. Gaussian noise with variance σ^2 is added to the matrix $\boldsymbol{X}$. The signal-to-noise ratio (SNR) in dB is defined as $\mathsf{SNR} = 10\log_{10}\left(\frac{\|\boldsymbol{X}\|_{\mathrm{F}}^2/J^2}{\sigma^2}\right)$. The rank-$K$ square root decomposition is performed on the resulted noisy matrix $\widehat{\boldsymbol{X}}$, i.e., $\widehat{\boldsymbol{X}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top$ and the matrix $\widehat{\boldsymbol{U}}$ is input to the algorithms. In this case as well, one can observe faster convergence for the proposed SymNMF for different sparsity levels.

C.2. Details of The UCI Data Experiments

MATLAB Classifiers for UCI Data Experiments. For UCI data (`https://archive.ics.uci.edu/ml/datasets.php`) experiments, we choose 10 different classifiers from the MATLAB statistics and machine learning toolbox (`https://www.mathworks.com/products/statistics.html`); see Table 8.

Table 8. Ten Classifiers used As Machine Annotators.

Coarse k-nearest neighbor classifier
Medium k-nearest neighbor classifier
Fine k-nearest neighbor classifier
Cosine k-nearest neighbor classifier
Coarse decision tree classifier
Medium decision tree classifier
Fine decision tree classifier
Linear support vector machine (SVM) classifier
Quadratic support vector machine (SVM) classifier
Coarse Gaussian support vector machine (SVM) classifier

Table 9. Classification error ($\%$) and runtime (sec.) on the LabelMe dataset ($N = 1000$, $M = 59$, $K = 8$). The "`SymNMF`" family are the proposed methods.

Algorithms	**Error** ($\%$)	**Time** (s)
`RobSymNMF`	32.10	1.25
`RobSymNMF-EM`	22.10	1.29
`DesSymNMF`	29.10	0.11
`DesSymNMF-EM`	22.20	0.20
`CrowdLayer`	20.90	15.80
`DL-MV`	23.10	14.31

Simulation Setup of Table 3. For the experiment in Table 3, we employ the following strategy in order to generate different proportions of the missing blocks:

1. Consider N items to be labeled by the annotators (machine classifiers). We split the test data into three disjoint parts having sizes of $0.1N, 0.3N$ and $0.6N$, respectively.
2. Each disjoint part of the test data is co-labeled by only P annotators, which are chosen randomly from M available annotators and $P \ll M$. We also make sure that every annotator labels at least one part out of the three test data parts.

By varying P for the three test data parts, we are able to control the proportions of missing co-occurrences. For each column of the table, we adjust P and generate the cases such that the corresponding missing proportion (Miss) is achieved.

In addition, since we have chosen different sizes for the three sets, different annotator pairs co-label varying number of data items. This makes the estimation accuracy for the pairwise statistics $\widehat{\boldsymbol{R}}_{m,j}$'s unbalanced—and we use this setting to test the robustness of our co-occurrence imputation algorithm.

C.3. Additional Real-Data Experiment

In this section, we present an additional real-data experiment. Specifically, we compare the proposed algorithms with a number of deep learning (DL)-based crowdsourcing methods, namely, `CrowdLayer` and `DL-MV` from the work in (Rodrigues & Pereira, 2018).

Note that the DL-based methods are implemented under fairly different settings relative to classic D&S learning methods. For example, both DL baselines train a deep neural networks using data items (e.g., images) as (part of the) input, whereas the classic D&S methods do not need to know or see the data items.

The dataset used in this experiment is the LabelMe data that is posted by the authors of (Rodrigues & Pereira, 2018). We use 1,000 data items that belong to 8 classes and are labeled by 59 annotators. The methods `CrowdLayer` and `DL-MV` are trained with 50 epochs. Table 9 presents the results of the algorithms under test. In the table, the results of `CrowdLayer` and our method are averaged from 100 trials (to observe performance under random initialization and stochastic algorithms). We observed that `CrowdLayer`'s and our method's average error rates are close, but `CrowdLayer` has an almost 10 times larger standard deviation (`RobSymNMF-EM` $22.1\% \pm 0.5\%$ v.s. `CrowdLayer` $20.9\% \pm 4.7\%$). The proposed method is also around 12 times faster (1.3 sec. vs 15.8 sec.).

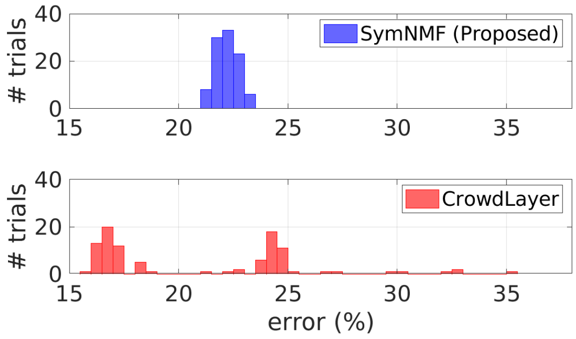


Figure 4. Histograms of error rates from 100 trials. The `CrowdLayer` method could work very well to attain low error rate in some trials, but multiple failed trials with error rate$\geq 30\%$ are also observed.

Fig. 4 presents the histogram of the error rates for our method and `CrowdLayer`. From Fig. 4, one can see that there are trials where `CrowdLayer` offers impressively low error rate, but there are also multiple trials where `CrowdLayer` gives high error rates ($\sim 30\% - 37\%$). The large variance is perhaps because DL methods' computational problem is more challenging, since DL algorithms such as SGD/Adam may not always converge well. However, the proposed method with convergence guarantees offers stable results.

D. Proof of Theorem 1

Theorem 1 *Assume that $\widehat{\boldsymbol{R}}_{m,n}$ is estimated by (7)-(8) using the sample-estimated $\widehat{\boldsymbol{R}}_{m,r}$, $\widehat{\boldsymbol{R}}_{n,\ell}$ and $\widehat{\boldsymbol{R}}_{\ell,r}$ [using (3) with at least S items]. Also assume that $\kappa(\boldsymbol{A}_m) \leq \gamma$ and $\operatorname{rank}(\boldsymbol{A}_m) = \operatorname{rank}(\boldsymbol{D}) = K$ for all $m \in [M]$. Let $\varrho = \min_{(m,j)\in\boldsymbol{\Omega}} \sigma_{\min}(\boldsymbol{R}_{m,j})$. Suppose that $S = \Omega\left(\frac{K^2\gamma^2\log(1/\delta)}{\varrho^4}\right)$ for $\delta > 0$. Then, for any $(m,n) \notin \boldsymbol{\Omega}$, with probability of at least $1-\delta$, we have:*

$$\|\widehat{\boldsymbol{R}}_{m,n} - \boldsymbol{R}_{m,n}\|_{\mathrm{F}} = O\left(\frac{K^2\gamma^3\sqrt{\log(1/\delta)}}{\varrho^2\sqrt{S}}\right),$$

where $\boldsymbol{R}_{m,n} = \boldsymbol{A}_m\boldsymbol{D}\boldsymbol{A}_n^\top$ is the missing ground-truth.

The missing pairwise co-occurrence $\boldsymbol{R}_{m,n}$ is imputed by (7)-(8) using available co-occurrences $\boldsymbol{R}_{m,r}$, $\boldsymbol{R}_{n,\ell}$ and $\boldsymbol{R}_{\ell,r}$. In practice, we do not observe the true pairwise co-occurrences $\boldsymbol{R}_{m,r}$, $\boldsymbol{R}_{n,\ell}$ and $\boldsymbol{R}_{\ell,r}$. Therefore, we first form the matrix $\widehat{\boldsymbol{C}}$ by using the corresponding sample estimated co-occurrences as below:

$$\widehat{\boldsymbol{C}} = [\widehat{\boldsymbol{R}}_{m,r}^\top, \widehat{\boldsymbol{R}}_{\ell,r}^\top]^\top.$$

To characterize $\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_{\mathrm{F}}$, we use Lemma 13 from (Zhang et al., 2016) which gives the result that, with probability at least $1-\delta$,

$$\|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{R}_{m,j}\|_{\mathrm{F}} \leq \frac{1+\sqrt{\log(1/\delta)}}{\sqrt{S}} := \phi,\ \forall m \neq j, \tag{20}$$

where $S > 0$ is the number of samples that the annotators m and j have co-labeled. Then we have

$$\begin{aligned} &\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_{\mathrm{F}}^2 = \|\widehat{\boldsymbol{R}}_{m,r} - \boldsymbol{R}_{m,r}\|_{\mathrm{F}}^2 + \|\widehat{\boldsymbol{R}}_{\ell,r} - \boldsymbol{R}_{\ell,r}\|_{\mathrm{F}}^2 \leq 2\phi^2 \\ \implies &\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_{\mathrm{F}} \leq \sqrt{2}\phi. \end{aligned} \tag{21}$$

Let us denote the thin SVD operation on $\widehat{\boldsymbol{C}}$ as follows:

$$\widehat{\boldsymbol{C}} = [\widehat{\boldsymbol{U}}_m^\top, \widehat{\boldsymbol{U}}_\ell^\top]^\top \widehat{\boldsymbol{\Sigma}}_{m,\ell,r}\widehat{\boldsymbol{V}}_r^\top. \tag{22}$$

We consider the below lemma to characterize this SVD operation:

Lemma 1 *(Yu et al., 2014) Let $\boldsymbol{C} \in \mathbb{R}^{m\times n}$ and $\widehat{\boldsymbol{C}} \in \mathbb{R}^{m\times n}$ have singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)}$ and $\widehat{\sigma}_1 \geq \widehat{\sigma}_2 \geq \ldots \widehat{\sigma}_{\min(m,n)}$, respectively. Fix $1 \leq t \leq s \leq \mathrm{rank}(\boldsymbol{C})$ and assume that $\min(\sigma_{t-1}^2 - \sigma_t^2, \sigma_s^2 - \sigma_{s+1}^2) > 0$, where $\sigma_0^2 := \infty$ and $\sigma_{\mathrm{rank}(\boldsymbol{C})+1} := 0$. Let $q := s - t + 1$ and let $\boldsymbol{U} = \left[\boldsymbol{u}_t, \boldsymbol{u}_{t+1}, \ldots, \boldsymbol{u}_s\right] \in \mathbb{R}^{m\times q}$ and $\widehat{\boldsymbol{U}} = \left[\widehat{\boldsymbol{u}}_t, \widehat{\boldsymbol{u}}_{t+1}, \ldots, \widehat{\boldsymbol{u}}_s\right] \in \mathbb{R}^{m\times q}$ have orthonormal columns satisfying $\boldsymbol{C}^\top \boldsymbol{u}_j = \sigma_j \boldsymbol{v}_j$ and $\widehat{\boldsymbol{C}}^\top \widehat{\boldsymbol{u}}_j = \widehat{\sigma}_j \widehat{\boldsymbol{v}}_j$ for $j = t, t+1, \ldots, s$ and let $\boldsymbol{V} = \left[\boldsymbol{v}_t, \boldsymbol{v}_{t+1}, \ldots, \boldsymbol{v}_s\right] \in \mathbb{R}^{n\times q}$ and $\widehat{\boldsymbol{V}} = \left[\widehat{\boldsymbol{v}}_t, \widehat{\boldsymbol{v}}_{t+1}, \ldots, \widehat{\boldsymbol{v}}_s\right] \in \mathbb{R}^{n\times q}$ have orthonormal columns satisfying $\boldsymbol{C}\boldsymbol{v}_j = \sigma_j \boldsymbol{u}_j$ and $\widehat{\boldsymbol{C}}\widehat{\boldsymbol{v}}_j = \widehat{\sigma}_j \widehat{\boldsymbol{u}}_j$ for $j = t, t+1, \ldots, s$. Then there exists an orthogonal matrix $\boldsymbol{O} \in \mathbb{R}^{q\times q}$ such that*

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{O}\|_{\mathrm{F}} \leq \frac{2^{3/2}(2\sigma_1 + \|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_2)\min(q^{1/2}\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_2, \|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_{\mathrm{F}})}{\min(\sigma_{t-1}^2 - \sigma_t^2, \sigma_s^2 - \sigma_{s+1}^2)}$$

and the same upper bound holds for $\|\widehat{\boldsymbol{V}} - \boldsymbol{V}\boldsymbol{O}\|_{\mathrm{F}}$.

For now, let us assume

$$\mathrm{rank}(\boldsymbol{C}) = K, \quad \|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_2 \leq \|\boldsymbol{C}\|_2 = \sigma_{\max}(\boldsymbol{C}). \tag{23}$$

By applying Lemma 1 in (22), we get

$$\|\widehat{\boldsymbol{U}}_m - \boldsymbol{U}_m\boldsymbol{O}\|_{\mathrm{F}} \leq \frac{2^{3/2}\sqrt{K}3\sigma_{\max}(\boldsymbol{C})\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_2}{\sigma_{\min}^2(\boldsymbol{C})}.$$

where $\boldsymbol{O} \in \mathbb{R}^{K\times K}$ is orthogonal.

By substituting the bound (21) in the above, we get that with probability of at least $1 - \delta$,

$$\|\widehat{\boldsymbol{U}}_m - \boldsymbol{U}_m\boldsymbol{O}\|_{\mathrm{F}} \leq \frac{12\sqrt{K}\sigma_{\max}(\boldsymbol{C})\phi}{\sigma_{\min}^2(\boldsymbol{C})}, \tag{24}$$

$$\|\widehat{\boldsymbol{U}}_\ell - \boldsymbol{U}_\ell\boldsymbol{O}\|_{\mathrm{F}} \leq \frac{12\sqrt{K}\sigma_{\max}(\boldsymbol{C})\phi}{\sigma_{\min}^2(\boldsymbol{C})}. \tag{25}$$

The missing co-occurrence $\boldsymbol{R}_{m,n}$ is imputed by using $\widehat{\boldsymbol{U}}_m$, $\widehat{\boldsymbol{U}}_\ell$ and $\widehat{\boldsymbol{R}}_{n,\ell}$ via the following operation:

$$\widehat{\boldsymbol{R}}_{m,n} = \widehat{\boldsymbol{U}}_m\widehat{\boldsymbol{U}}_\ell^{-1}\widehat{\boldsymbol{R}}_{n,\ell}^\top.$$

The first term $\widehat{\boldsymbol{U}}_m$ is characterized by (24). To characterize the term $\widehat{\boldsymbol{U}}_\ell^{-1}$, we use the following lemma:

Lemma 2 *Consider any matrices $\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{E} \in \mathbb{R}^{K\times K}$ such that $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{E}$ and $\boldsymbol{Y}$ is invertible. Suppose that $\mathrm{rank}(\boldsymbol{Z}) = \mathrm{rank}(\boldsymbol{Y})$ and that $\|\boldsymbol{E}\|_2 \leq \sigma_{\min}(\boldsymbol{Y})/2$. Then, we have*

$$\|\boldsymbol{Z}^{-1} - \boldsymbol{Y}^{-1}\|_2 \leq \frac{2\|\boldsymbol{E}\|_2}{\sigma_{\min}^2(\boldsymbol{Y})}.$$

The proof of the lemma can be found in Section G.

Applying Lemma 2 by letting $\boldsymbol{Y} := \boldsymbol{U}_\ell\boldsymbol{O}$ and $\boldsymbol{Z} := \widehat{\boldsymbol{U}}_\ell$, we get

$$\|\widehat{\boldsymbol{U}}_\ell^{-1} - (\boldsymbol{U}_\ell\boldsymbol{O})^{-1}\|_2 \leq \frac{2}{\sigma_{\min}^2(\boldsymbol{U}_\ell)}\|\widehat{\boldsymbol{U}}_\ell - (\boldsymbol{U}_\ell\boldsymbol{O})\|_2. \tag{26}$$

We proceed to characterize $\sigma_{\min}(\boldsymbol{U}_\ell)$ in the above relation by utilizing the following result:

Lemma 3 *Suppose that $\kappa(\boldsymbol{A}_m) \leq \gamma$, for all m. Then, we have*

$$\sigma_{\min}(\boldsymbol{U}_\ell) \geq \frac{1}{\sqrt{2K}\gamma}, \quad \sigma_{\max}(\boldsymbol{U}_\ell) \leq \gamma$$

and the above bounds are applicable for $\boldsymbol{U}_m$ as well.

The proof of the lemma can be found in Section H.

Applying Lemma (3) in (26), we get

$$\begin{aligned}\|\widehat{\boldsymbol{U}}_{\ell_1}^{-1} - (\boldsymbol{U}_{\ell_1}\boldsymbol{O})^{-1}\|_2 &\leq 4K\gamma^2\|\widehat{\boldsymbol{U}}_{\ell_1} - (\boldsymbol{U}_{\ell_1}\boldsymbol{O})\|_2 \\ &\leq \frac{48K\sqrt{K}\gamma^2\sigma_{\max}(\boldsymbol{C})\phi}{\sigma^2_{\min}(\boldsymbol{C})}, \end{aligned} \tag{27}$$

where we applied (25) in the last inequality.

Using the above derived upper bounds, we proceed to bound the following term:

$$\|\widehat{\boldsymbol{R}}_{m,n} - \boldsymbol{R}_{m,n}\|_2 = \|\widehat{\boldsymbol{U}}_m\widehat{\boldsymbol{U}}_\ell^{-1}\widehat{\boldsymbol{R}}^\top_{n,\ell} - \boldsymbol{U}_m\boldsymbol{U}_\ell^{-1}\boldsymbol{R}^\top_{n,\ell}\|_2,$$

where we can see that $\boldsymbol{U}_m\boldsymbol{U}_\ell^{-1}\boldsymbol{R}^\top_{n,\ell} = \boldsymbol{U}_m\boldsymbol{O}(\boldsymbol{U}_\ell\boldsymbol{O})^{-1}\boldsymbol{R}^\top_{n,\ell}$ for the orthogonal matrix $\boldsymbol{O}$. To simplify the notations, let us define $\boldsymbol{Z}_1 := \boldsymbol{U}_m\boldsymbol{O}$, $\boldsymbol{Z}_2 := (\boldsymbol{U}_\ell\boldsymbol{O})^{-1}$ and $\boldsymbol{Z}_3 := \boldsymbol{R}^\top_{n,\ell}$. We also define $\widehat{\boldsymbol{Z}}_1 := \widehat{\boldsymbol{U}}_m$, $\widehat{\boldsymbol{Z}}_2 := \widehat{\boldsymbol{U}}_\ell^{-1}$ and $\boldsymbol{Z}_3 := \widehat{\boldsymbol{R}}^\top_{n,\ell}$. Using these notations, we have the following set of relations:

$$\begin{aligned}\left\|\widehat{\boldsymbol{Z}}_1\widehat{\boldsymbol{Z}}_2\widehat{\boldsymbol{Z}}_3 - \boldsymbol{Z}_1\boldsymbol{Z}_2\boldsymbol{Z}_3\right\|_2 &= \left\|\widehat{\boldsymbol{Z}}_1\widehat{\boldsymbol{Z}}_2\widehat{\boldsymbol{Z}}_3 - \boldsymbol{Z}_1\boldsymbol{Z}_2\boldsymbol{Z}_3 - \widehat{\boldsymbol{Z}}_1\boldsymbol{Z}_2\boldsymbol{Z}_3 + \widehat{\boldsymbol{Z}}_1\boldsymbol{Z}_2\boldsymbol{Z}_3\right\|_2 \\ &= \left\|\left(\widehat{\boldsymbol{Z}}_1 - \boldsymbol{Z}_1\right)\boldsymbol{Z}_2\boldsymbol{Z}_3 + \widehat{\boldsymbol{Z}}_1\left(\widehat{\boldsymbol{Z}}_2\widehat{\boldsymbol{Z}}_3 - \boldsymbol{Z}_2\boldsymbol{Z}_3\right)\right\|_2 \\ &\leq \left\|\left(\widehat{\boldsymbol{Z}}_1 - \boldsymbol{Z}_1\right)\boldsymbol{Z}_2\boldsymbol{Z}_3\right\|_2 + \left\|\widehat{\boldsymbol{Z}}_1\left(\widehat{\boldsymbol{Z}}_2\widehat{\boldsymbol{Z}}_3 - \boldsymbol{Z}_2\boldsymbol{Z}_3\right)\right\|_2 \\ &= \left\|\left(\widehat{\boldsymbol{Z}}_1 - \boldsymbol{Z}_1\right)\boldsymbol{Z}_2\boldsymbol{Z}_3\right\|_2 + \left\|\widehat{\boldsymbol{Z}}_1(\widehat{\boldsymbol{Z}}_2 - \boldsymbol{Z}_2)\boldsymbol{Z}_3 + \widehat{\boldsymbol{Z}}_1\widehat{\boldsymbol{Z}}_2\left(\widehat{\boldsymbol{Z}}_3 - \boldsymbol{Z}_3\right)\right\|_2 \\ &\leq \|\boldsymbol{Z}_2\|_2\|\boldsymbol{Z}_3\|_2\left\|\widehat{\boldsymbol{Z}}_1 - \boldsymbol{Z}_1\right\|_2 + \|\widehat{\boldsymbol{Z}}_1\|_2\|\boldsymbol{Z}_3\|_2\left\|\widehat{\boldsymbol{Z}}_2 - \boldsymbol{Z}_2\right\|_2 + \|\widehat{\boldsymbol{Z}}_1\|_2\|\widehat{\boldsymbol{Z}}_2\|_2\left\|\widehat{\boldsymbol{Z}}_3 - \boldsymbol{Z}_3\right\|_2,\end{aligned}$$

where we have used triangle inequality to obtain the first inequality and used the fact that $\|\boldsymbol{X}\boldsymbol{Y}\|_2 \leq \|\boldsymbol{X}\|_2\|\boldsymbol{Y}\|_2$ in the last inequality. Applying this result, we get

$$\begin{aligned}\|\widehat{\boldsymbol{U}}_m\widehat{\boldsymbol{U}}_\ell^{-1}\widehat{\boldsymbol{R}}^\top_{n,\ell} - \boldsymbol{U}_m\boldsymbol{O}(\boldsymbol{U}_\ell\boldsymbol{O})^{-1}\boldsymbol{R}^\top_{n,\ell}\|_2 \leq\ & \|(\boldsymbol{U}_\ell\boldsymbol{O})^{-1}\|_2\|\boldsymbol{R}_{n,\ell}\|_2\|\widehat{\boldsymbol{U}}_m - \boldsymbol{U}_m\boldsymbol{O}\|_2 \\ &+ \|\widehat{\boldsymbol{U}}_m\|_2\|\boldsymbol{R}_{n,\ell}\|_2\|\widehat{\boldsymbol{U}}_\ell^{-1} - (\boldsymbol{U}_\ell\boldsymbol{O})^{-1}\|_2 \\ &+ \|\widehat{\boldsymbol{U}}_m\|_2\|\widehat{\boldsymbol{U}}_\ell^{-1}\|_2\|\widehat{\boldsymbol{R}}_{n,\ell} - \boldsymbol{R}_{n,\ell}\|_2.\end{aligned} \tag{28}$$

In (28), we need to apply the below characterizations to derive the final bound:

1. **Upper bound for $\|\widehat{\boldsymbol{U}}_m\|_2$**

$$\begin{aligned}\|\widehat{\boldsymbol{U}}_m\|_2 &= \|\widehat{\boldsymbol{U}}_m - \boldsymbol{U}_m\boldsymbol{O} + \boldsymbol{U}_m\boldsymbol{O}\|_2 \leq \|\widehat{\boldsymbol{U}}_m - \boldsymbol{U}_m\boldsymbol{O}\|_2 + \|\boldsymbol{U}_m\boldsymbol{O}\|_2 \\ &\leq \sigma_{\max}(\boldsymbol{U}_m) + \sigma_{\max}(\boldsymbol{U}_m) = 2\sigma_{\max}(\boldsymbol{U}_m) \leq 2\gamma.\end{aligned}$$

   where we have used triangle inequality for the first inequality, used the assumption that $\|\widehat{\boldsymbol{U}}_m - \boldsymbol{U}_m\boldsymbol{O}\|_2 \leq \sigma_{\min}(\boldsymbol{U}_m)/2$ for the second inequality and invoked Lemma 3 for the last inequality.

2. **Upper bound for $\|(\boldsymbol{U}_\ell\boldsymbol{O})^{-1}\|_2$**

$$\|(\boldsymbol{U}_\ell\boldsymbol{O})^{-1}\|_2 = 1/\sigma_{\min}(\boldsymbol{U}_\ell) \leq \sqrt{2K}\gamma,$$

   where we have applied Lemma 3 for the last inequality.

3. **Upper bound for $\|\widehat{\boldsymbol{U}}_\ell^{-1}\|_2$**

$$\|\widehat{\boldsymbol{U}}_\ell^{-1}\|_2 = 1/\sigma_{\min}(\widehat{\boldsymbol{U}}_\ell) \leq 2/\sigma_{\min}(\boldsymbol{U}_\ell) \leq 2\sqrt{2K}\gamma,$$

   where we have used the assumption that $\|\widehat{\boldsymbol{U}}_\ell - \boldsymbol{U}_\ell\boldsymbol{O}\|_2 \leq \sigma_{\min}(\boldsymbol{U}_\ell)/2$ for the first inequality and invoked Lemma 3 for the last inequality.

4. **Upper bound for $\|\boldsymbol{R}_{n,\ell}\|_2$**

$$\|\boldsymbol{R}_{n,\ell}\|_2 \leq \|\boldsymbol{R}_{n,\ell}\|_{\mathrm{F}} \leq 1,$$

where we used the fact that the entries of the matrix $\boldsymbol{R}_{n,\ell}$ are nonnegative and sum to one and therefore $\|\boldsymbol{R}_{n,\ell}\|_{\mathrm{F}}^2 \leq 1$.

Applying these upper bounds to (28), we attain the following:

$$\begin{aligned} &\|\widehat{\boldsymbol{R}}_{m,n} - \boldsymbol{R}_{m,n}\|_2 \leq \sqrt{2K}\gamma\|\widehat{\boldsymbol{U}}_m - \boldsymbol{U}_m\boldsymbol{O}\|_2 + 2\gamma\|\widehat{\boldsymbol{U}}_\ell^{-1} - (\boldsymbol{U}_\ell\boldsymbol{O})^{-1}\|_2 + 4\sqrt{2K}\gamma^2\|\widehat{\boldsymbol{R}}_{n,\ell} - \boldsymbol{R}_{n,\ell}\|_2 \\ \implies &\|\widehat{\boldsymbol{R}}_{m,n} - \boldsymbol{R}_{m,n}\|_{\mathrm{F}} \leq \sqrt{2}K\gamma\|\widehat{\boldsymbol{U}}_m - \boldsymbol{U}_m\boldsymbol{O}\|_{\mathrm{F}} + 2\sqrt{K}\gamma\|\widehat{\boldsymbol{U}}_\ell^{-1} - (\boldsymbol{U}_\ell\boldsymbol{O})^{-1}\|_2 + 4\sqrt{2}K\gamma^2\|\widehat{\boldsymbol{R}}_{n,\ell} - \boldsymbol{R}_{n,\ell}\|_{\mathrm{F}}, \end{aligned}$$

where we used the matrix norm equivalence $\|\boldsymbol{X}\|_2 \leq \|\boldsymbol{X}\|_{\mathrm{F}} \leq \sqrt{K}\|\boldsymbol{X}\|_2$, for a matrix $\boldsymbol{X}$ of rank K, in the last inequality. By substituting the bounds (20), (24) and (27) in the above, we get

$$\|\widehat{\boldsymbol{R}}_{m,n} - \boldsymbol{R}_{m,n}\|_{\mathrm{F}} \leq \frac{12\sqrt{2K}K\gamma\sigma_{\max}(\boldsymbol{C})\phi}{\sigma_{\min}(\boldsymbol{C})^2} + \frac{96K^2\gamma^3\sigma_{\max}(\boldsymbol{C})\phi}{\sigma_{\min}^2(\boldsymbol{C})} + 4\sqrt{2}K\gamma^2\phi.$$

where we have $\boldsymbol{C} = [\boldsymbol{R}_{m,r}^\top, \boldsymbol{R}_{\ell,r}^\top]^\top$ and can immediately see that $\|\boldsymbol{C}\|_{\mathrm{F}}^2 \leq 2$, which implies that $\sigma_{\max}(\boldsymbol{C}) \leq \|\boldsymbol{C}\|_{\mathrm{F}} \leq \sqrt{2}$ and $\phi = \frac{1+\sqrt{\log(1/\delta)}}{\sqrt{S}}$. Combining this, we get that with probability at least $1-\delta$, for a certain constant $C_1 > 0$,

$$\|\widehat{\boldsymbol{R}}_{m,n} - \boldsymbol{R}_{m,n}\|_{\mathrm{F}} \leq \frac{C_1K^2\gamma^3\sqrt{\log(1/\delta)}}{\sigma_{\min}^2(\boldsymbol{C})\sqrt{S}}. \tag{29}$$

Finally, we will summarize the conditions to be satisfied to obtain (29). From (23), we can see that the below condition needs to be satisfied:

$$\begin{aligned} \|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_2 \leq \|\boldsymbol{C}\|_2 = \sigma_{\max}(\boldsymbol{C}) \implies \sqrt{2}\phi \leq \sigma_{\max}(\boldsymbol{C}) \\ \implies S \leq \frac{2(1+\sqrt{\log(1/\delta)})^2}{\sigma_{\max}^2(\boldsymbol{C})}. \end{aligned} \tag{30}$$

From Lemma 2, the condition to be satisfied is:

$$\|\widehat{\boldsymbol{U}}_\ell - \boldsymbol{U}_\ell\boldsymbol{O}\|_2 \leq \sigma_{\min}(\boldsymbol{U}_\ell)/2. \tag{31}$$

By applying (25) and Lemma 3 in the left and right hand sides of (31), respectively, the condition to be satisfied can be re-written as:

$$\begin{aligned} \frac{12\sqrt{K}\sigma_{\max}(\boldsymbol{C})\phi}{\sigma_{\min}(\boldsymbol{C})^2} \leq \frac{1}{\sqrt{2K}\gamma} \implies \frac{12\sqrt{2}\sqrt{K}\sigma_{\max}(\boldsymbol{C})(1+\sqrt{\log(1/\delta)})}{\sigma_{\min}(\boldsymbol{C})^2\sqrt{S}} \leq \frac{1}{\sqrt{2K}\gamma}, \\ \implies S \geq \frac{C_2K^2\gamma^2\log(1/\delta)}{\sigma_{\min}(\boldsymbol{C})^2}, \end{aligned} \tag{32}$$

for a certain constant $C_2 > 0$. Combining (30) and (32), we get the final condition on S as stated in the theorem.

E. Proof of Theorem 2

Theorem 2 *Assume that the $\widehat{\boldsymbol{R}}_{m,j}$'s are estimated using (3) with $S_{m,j} = |\mathcal{S}_{m,j}|$ for all $(m,j) \in \boldsymbol{\Omega}$. Also assume that each $\widehat{\boldsymbol{R}}_{m,j}$ is observed with the same probability. Let $\{\boldsymbol{U}_m^*, \boldsymbol{U}_j^*\}_{(m,j)\in\boldsymbol{\Omega}}$ be any optimal solution of (9). Define $L = M(M-1)/2$. Then we have*

$$\frac{1}{L}\sum_{m<j}\|\boldsymbol{U}_m^*(\boldsymbol{U}_j^*)^\top - \boldsymbol{R}_{m,j}\|_{\mathrm{F}} \leq C\sqrt{\frac{MK^2\log(M)}{|\boldsymbol{\Omega}|}} + \left(\frac{1}{|\boldsymbol{\Omega}|} + \frac{1}{L}\right)\sum_{(m,j)\in\boldsymbol{\Omega}}\frac{1+\sqrt{M}}{\sqrt{S_{m,j}}},$$

with probability of at least $1 - 3\exp(-M)$, where $C > 0$.

Let $\boldsymbol{R}^*_{m,j} = \boldsymbol{U}^*_m \boldsymbol{U}^{*\top}_j$, where $\{\boldsymbol{U}^*_m, \boldsymbol{U}^*_j\}_{(m,j)\in\boldsymbol{\Omega}}$ be any optimal solution of (9) and $\boldsymbol{N}_{m,j} = \widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{R}_{m,j}$ for every m, j. Note that we treat $\boldsymbol{N}_{m,j} = \boldsymbol{0}$ for $(m,j) \notin \boldsymbol{\Omega}$, since the co-occurrences are unobserved. We define the following quantity that will be useful in our proof:

$$\tau(\boldsymbol{\Omega}) = \left| \frac{1}{|\boldsymbol{\Omega}|} \sum_{(m,j)\in\boldsymbol{\Omega}} \|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{R}^*_{m,j}\|_{\mathrm{F}} - \frac{1}{L}\sum_{m<j} \|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{R}^*_{m,j}\|_{\mathrm{F}} \right|, \tag{33}$$

where $L = M(M-1)/2$. Then we have

$$\begin{aligned}
\frac{1}{L}\sum_{m<j}\|\boldsymbol{R}^*_{m,j} - \boldsymbol{R}_{m,j}\|_{\mathrm{F}} &= \frac{1}{L}\sum_{m<j}\|\boldsymbol{R}^*_{m,j} - \widehat{\boldsymbol{R}}_{m,j} + \boldsymbol{N}_{m,j}\|_{\mathrm{F}} \\
&\overset{(a)}{\leq} \frac{1}{L}\sum_{m<j}\|\boldsymbol{R}^*_{m,j} - \widehat{\boldsymbol{R}}_{m,j}\|_{\mathrm{F}} + \frac{1}{L}\sum_{m<j}\|\boldsymbol{N}_{m,j}\|_{\mathrm{F}} \\
&\overset{(b)}{\leq} \frac{1}{|\boldsymbol{\Omega}|}\sum_{(m,j)\in\boldsymbol{\Omega}}\|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{R}^*_{m,j}\|_{\mathrm{F}} + \tau(\boldsymbol{\Omega}) + \frac{1}{L}\sum_{m<j}\|\boldsymbol{N}_{m,j}\|_{\mathrm{F}} \\
&\overset{(c)}{\leq} \frac{1}{|\boldsymbol{\Omega}|}\sum_{(m,j)\in\boldsymbol{\Omega}}\|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{R}_{m,j}\|_{\mathrm{F}} + \tau(\boldsymbol{\Omega}) + \frac{1}{L}\sum_{m<j}\|\boldsymbol{N}_{m,j}\|_{\mathrm{F}} \\
&= \frac{1}{|\boldsymbol{\Omega}|}\sum_{(m,j)\in\boldsymbol{\Omega}}\|\boldsymbol{N}_{m,j}\|_{\mathrm{F}} + \tau(\boldsymbol{\Omega}) + \frac{1}{L}\sum_{m<j}\|\boldsymbol{N}_{m,j}\|_{\mathrm{F}},
\end{aligned} \tag{34}$$

where (a) is due to triangle inequality, (b) is due to the definition of $\tau(\boldsymbol{\Omega})$ and triangle inequality, and (c) is due to the fact that $\boldsymbol{R}^*_{m,j}$ is the optimal solution of (9).

Next, we will characterize $\tau(\boldsymbol{\Omega})$. For this, let us define the set

$$\mathcal{S}_K = \{\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top \in \mathbb{R}^{MK\times MK} : \mathrm{rank}(\boldsymbol{X}) \leq K,\ \|\boldsymbol{U}\|_{\mathrm{F}} \leq B,\ \|\boldsymbol{V}\|_{\mathrm{F}} \leq B\},$$

where the constant $B = \sqrt{M}D$ and D is the constant from Problem (9).

If $\|\boldsymbol{U}\|_{\mathrm{F}} \leq B$ and $\|\boldsymbol{V}\|_{\mathrm{F}} \leq B$, then $\|\boldsymbol{X}\|_F \leq \|\boldsymbol{U}\|_{\mathrm{F}}\|\boldsymbol{V}\|_{\mathrm{F}} = B^2$. Therefore, we can rewrite the definition of the set $\mathcal{S}_K$ as below:

$$\mathcal{S}_K = \{\boldsymbol{X} \in \mathbb{R}^{MK\times MK} : \mathrm{rank}(\boldsymbol{X}) \leq K,\ \|\boldsymbol{X}\|_{\mathrm{F}} \leq B^2\}. \tag{35}$$

We will invoke the following lemma to characterize the covering number of the set $\mathcal{S}_K$.

Lemma 4 *(Wang & Xu, 2012) Let $\mathcal{S}_r = \{\boldsymbol{X} \in \mathbb{R}^{n_1\times n_2} : \mathrm{rank}(\boldsymbol{X}) \leq r,\ \|\boldsymbol{X}\|_{\mathrm{F}} \leq C\}$. Then there exists an ϵ-net $\overline{\mathcal{S}}_r$ for the Frobenius norm obeying*

$$|\overline{\mathcal{S}}_r(\epsilon)| \leq (9C/\epsilon)^{(n_1+n_2+1)r}.$$

By denoting the ϵ-net of $\mathcal{S}_K$ defined in (35) as $\overline{\mathcal{S}}_K(\epsilon)$ and applying Lemma 4, we get that

$$|\overline{\mathcal{S}}_K(\epsilon)| \leq (9B^2/\epsilon)^{(2MK+1)K}. \tag{36}$$

Let $\widetilde{\boldsymbol{X}} \in \overline{\mathcal{S}}_K(\epsilon)$ and we can define the following:

$$\widehat{\mathcal{L}}(\widetilde{\boldsymbol{X}}) = \frac{1}{|\boldsymbol{\Omega}|}\sum_{(m,j)\in\boldsymbol{\Omega}}\|\widehat{\boldsymbol{R}}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} \tag{37a}$$

$$\mathcal{L}(\widetilde{\boldsymbol{X}}) = \frac{1}{L}\sum_{m<j}\|\widehat{\boldsymbol{R}}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}}, \tag{37b}$$

where $\widetilde{\boldsymbol{X}}_{m,j} \in \mathbb{R}^{K\times K}$ is the (m,j)th block of $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{MK\times MK}$.

To proceed, consider the below lemma:

Lemma 5 *(Serfling, 1974) Let $X = [X_1, \dots, X_n]$ be a set of samples taken without replacement from a set $\{x_1, \dots, x_N\}$ with mean u where $n \le N$. Denote $a := \max_i x_i$ and $b := \max_i b_i$. Then, we have*

$$\mathsf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - u\right| \ge t\right) \le 2\exp\left(-\frac{2nt^2}{\left(1-\frac{n-1}{N}\right)(b-a)^2}\right).$$

Notice that $\{\|\widehat{\boldsymbol{R}}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}}\}_{m<j}$ forms a set of L elements with $\mathcal{L}(\widetilde{\boldsymbol{X}})$ as its mean and $\widehat{\mathcal{L}}(\widetilde{\boldsymbol{X}})$ as the mean estimated from $|\boldsymbol{\Omega}|$ samples, drawn without replacement. Also, we have

$$\begin{aligned}
&\max_{m,j} \|\widehat{\boldsymbol{R}}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} \le \max_{m,j} \|\widehat{\boldsymbol{R}}_{m,j}\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} \le 2,\\
&\min_{m,j} \|\widehat{\boldsymbol{R}}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} = 0.
\end{aligned}$$

Therefore, by applying Lemma 6, we have

$$\mathsf{Pr}\left(\left|\widehat{\mathcal{L}}(\widetilde{\boldsymbol{X}}) - \mathcal{L}(\widetilde{\boldsymbol{X}})\right| \ge t\right) \le 2\exp\left(-\frac{2|\boldsymbol{\Omega}|t^2}{\left(1-\frac{|\boldsymbol{\Omega}|-1}{L}\right)4}\right).$$

Applying union bound over every $\widetilde{\boldsymbol{X}} \in \overline{\mathcal{S}}_K(\epsilon)$, we get

$$\mathsf{Pr}\left(\sup_{\widetilde{\boldsymbol{X}} \in \overline{\mathcal{S}}_K(\epsilon)} \left|\widehat{\mathcal{L}}(\widetilde{\boldsymbol{X}}) - \mathcal{L}(\widetilde{\boldsymbol{X}})\right| \ge t\right) \le 2|\overline{\mathcal{S}}_K(\epsilon)|\exp\left(-\frac{2|\boldsymbol{\Omega}|t^2}{\left(1-\frac{|\boldsymbol{\Omega}|-1}{L}\right)4}\right).$$

By letting $|\overline{\mathcal{S}}_K(\epsilon)|\exp\left(-\frac{2|\boldsymbol{\Omega}|t^2}{\left(1-\frac{|\boldsymbol{\Omega}|-1}{L}\right)4}\right) = \exp(-M)$, we get that

$$\begin{aligned}
\log\ |\overline{\mathcal{S}}_K(\epsilon)| - \frac{2L|\boldsymbol{\Omega}|t^2}{(L-|\boldsymbol{\Omega}|+1)\,4} &= -M\\
\implies M + \log\ |\overline{\mathcal{S}}_K(\epsilon)| &= \frac{2L|\boldsymbol{\Omega}|t^2}{(L-|\boldsymbol{\Omega}|+1)\,4}\\
\implies t &= \sqrt{\frac{(M+\log\ |\overline{\mathcal{S}}_K(\epsilon)|)(L-|\boldsymbol{\Omega}|+1)4}{2L|\boldsymbol{\Omega}|}}
\end{aligned}$$

Therefore, we get that with probability at least $1-2\exp(-M)$, we have

$$\sup_{\widetilde{\boldsymbol{X}} \in \overline{\mathcal{S}}_K(\epsilon)} \left|\widehat{\mathcal{L}}(\widetilde{\boldsymbol{X}}) - \mathcal{L}(\widetilde{\boldsymbol{X}})\right| \le \sqrt{\frac{(M+\log\ |\overline{\mathcal{S}}_K(\epsilon)|)(L-|\boldsymbol{\Omega}|+1)4}{2L|\boldsymbol{\Omega}|}}.$$

By applying (36), we have

$$\sup_{\widetilde{\boldsymbol{X}} \in \overline{\mathcal{S}}_K(\epsilon)} \left|\widehat{\mathcal{L}}(\widetilde{\boldsymbol{X}}) - \mathcal{L}(\widetilde{\boldsymbol{X}})\right| \le \sqrt{\frac{(M+(2MK+1)K\log(9B^2/\epsilon))(L-|\boldsymbol{\Omega}|+1)4}{2L|\boldsymbol{\Omega}|}} := \zeta. \tag{38}$$

With the above result, we proceed to relate $\mathcal{S}_K$ and $\overline{\mathcal{S}}_K(\epsilon)$. Let $\boldsymbol{X} \in \mathcal{S}_K$ and for every $\boldsymbol{X}$, there exists $\widetilde{\boldsymbol{X}} \in \overline{\mathcal{S}}_K(\epsilon)$ satisfying $\|\boldsymbol{X} - \widetilde{\boldsymbol{X}}\|_{\mathrm{F}} \leq \epsilon$. This implies that

$$\begin{aligned}
|\mathcal{L}(\boldsymbol{X}) - \mathcal{L}(\widetilde{\boldsymbol{X}})| &= \left| \frac{1}{L} \sum_{m<j} \|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{X}_{m,j}\|_{\mathrm{F}} - \frac{1}{L} \sum_{m<j} \|\widehat{\boldsymbol{R}}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} \right| \\
&= \left| \frac{1}{L} \sum_{m<j} \left(\|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{X}_{m,j}\|_{\mathrm{F}} - \|\widehat{\boldsymbol{R}}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} \right) \right| \\
&\leq \left| \frac{1}{L} \sum_{m<j} \left(\|\boldsymbol{X}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} \right) \right| \leq \epsilon
\end{aligned}$$

where we have used the relation that $\|\boldsymbol{X}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} \leq \|\boldsymbol{X} - \widetilde{\boldsymbol{X}}\|_{\mathrm{F}} \leq \epsilon$. Similarly, we have

$$\begin{aligned}
|\widehat{\mathcal{L}}(\boldsymbol{X}) - \widehat{\mathcal{L}}(\widetilde{\boldsymbol{X}})| &= \left| \frac{1}{|\boldsymbol{\Omega}|} \sum_{(m,j)\in\boldsymbol{\Omega}} \|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{X}_{m,j}\|_{\mathrm{F}} - \frac{1}{|\boldsymbol{\Omega}|} \sum_{(m,j)\in\boldsymbol{\Omega}} \|\widehat{\boldsymbol{R}}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} \right| \\
&= \left| \frac{1}{|\boldsymbol{\Omega}|} \sum_{(m,j)\in\boldsymbol{\Omega}} \left(\|\widehat{\boldsymbol{R}}_{m,j} - \boldsymbol{X}_{m,j}\|_{\mathrm{F}} - \|\widehat{\boldsymbol{R}}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} \right) \right| \\
&\leq \left| \frac{1}{|\boldsymbol{\Omega}|} \sum_{(m,j)\in\boldsymbol{\Omega}} \left(\|\boldsymbol{X}_{m,j} - \widetilde{\boldsymbol{X}}_{m,j}\|_{\mathrm{F}} \right) \right| \leq \epsilon.
\end{aligned}$$

From the above results, we further have

$$\begin{aligned}
\sup_{\boldsymbol{X}\in\mathcal{S}_K} \left| \widehat{\mathcal{L}}(\boldsymbol{X}) - \mathcal{L}(\boldsymbol{X}) \right| &\leq \sup_{\boldsymbol{X}\in\mathcal{S}_K} \left(\left| \widehat{\mathcal{L}}(\boldsymbol{X}) - \widehat{\mathcal{L}}(\widetilde{\boldsymbol{X}}) \right| + \left| \mathcal{L}(\widetilde{\boldsymbol{X}}) - \mathcal{L}(\boldsymbol{X}) \right| + \left| \widehat{\mathcal{L}}(\widetilde{\boldsymbol{X}}) - \mathcal{L}(\widetilde{\boldsymbol{X}}) \right| \right) \\
&\leq \epsilon + \epsilon + \sup_{\widetilde{\boldsymbol{X}}\in\overline{\mathcal{S}}_K(\epsilon)} \left| \widehat{\mathcal{L}}(\widetilde{\boldsymbol{X}}) - \mathcal{L}(\widetilde{\boldsymbol{X}}) \right| \\
&\leq 2\epsilon + \zeta,
\end{aligned}$$

where we have applied (38) in the last inequality.

Setting $\epsilon = 1/L$, we get the below with probability at least $1 - 2\exp(-M)$:

$$\begin{aligned}
\sup_{\boldsymbol{X}\in\mathcal{S}_K} \left| \widehat{\mathcal{L}}(\boldsymbol{X}) - \mathcal{L}(\boldsymbol{X}) \right| &\leq 2\frac{1}{L} + \sqrt{\frac{(M + (2MK+1)K\log(9LB^2))(L - |\boldsymbol{\Omega}| + 1)4}{2L|\boldsymbol{\Omega}|}} \\
&\leq 2\frac{1}{L} + \sqrt{\frac{(M + 3MK^2\log(9LB^2))(L - |\boldsymbol{\Omega}| + 1)4}{2L|\boldsymbol{\Omega}|}} \\
&\leq 2\frac{1}{|\boldsymbol{\Omega}|} + \sqrt{\frac{2(M + 3MK^2\log(9LB^2))}{|\boldsymbol{\Omega}|}} \\
&\leq 2\frac{1}{|\boldsymbol{\Omega}|} + \sqrt{\frac{2(M + 3MK^2\log(9M^2B^2))}{|\boldsymbol{\Omega}|}}
\end{aligned}$$

where we have used the relation that $L = M(M-1)/2$ in the last inequality. Note that B is defined such that $\|\boldsymbol{X}\|_{\mathrm{F}} \leq B^2$, where $\boldsymbol{X} \in \mathcal{S}_K$. In our case, we have $\boldsymbol{R}_{m,j} \geq \boldsymbol{0}$, $\sum_{p,q} \boldsymbol{R}_{m,j}(p,q) = 1$ and therefore we get $\|\boldsymbol{R}_{m,j}\|_{\mathrm{F}}^2 \leq 1$ for all m, j. It implies that all the elements $\boldsymbol{X}$ of the feasible set $\mathcal{S}_K$ can be set to have $\|\boldsymbol{X}\|_{\mathrm{F}}^2 \leq M^2$. Therefore, we can set $B^2 = M$.

Using the definition of $\tau(\boldsymbol{\Omega})$ given by (33), $\widehat{\mathcal{L}}(\boldsymbol{X})$ and $\mathcal{L}(\boldsymbol{X})$ given by (37) and setting $B^2 = M$, we can then see that, there exists a constant $C > 0$ such that

$$\tau(\boldsymbol{\Omega}) \leq C\sqrt{\frac{MK^2\log(M)}{|\boldsymbol{\Omega}|}}. \tag{39}$$

Substituting (39) in (34), we get that with probability at least $1 - 2\exp(-M)$,

$$\frac{1}{L}\sum_{m<j}\|\boldsymbol{R}^*_{m,j} - \boldsymbol{R}_{m,j}\|_{\rm F} \leq \frac{1}{|\boldsymbol{\Omega}|}\sum_{(m,j)\in\boldsymbol{\Omega}}\|\boldsymbol{N}_{m,j}\|_{\rm F} + \frac{1}{L}\sum_{m<j}\|\boldsymbol{N}_{m,j}\|_{\rm F} + C\sqrt{\frac{MK^2\log(M)}{|\boldsymbol{\Omega}|}}. \tag{40}$$

Using Lemma 13 from (Zhang et al., 2016), we get that with probability at least $1 - \delta$,

$$\|\boldsymbol{N}_{m,j}\|_{\rm F} \leq \frac{1 + \sqrt{\log(1/\delta)}}{\sqrt{S_{m,j}}}, \text{ if } (m,j) \in \boldsymbol{\Omega}, \tag{41}$$

where $S_{m,j}$ is the (nonzero) number of samples that the annotators m and j have co-labeled. Also, without loss of any generality, we can let $\widehat{\boldsymbol{R}}_{m,j} = \boldsymbol{R}_{m,j}$, for all $(m,j) \notin \boldsymbol{\Omega}$. Therefore, we have

$$\|\boldsymbol{N}_{m,j}\|_{\rm F} = 0,\ (m,j) \notin \boldsymbol{\Omega}. \tag{42}$$

By substituting $\delta = \exp(-M)$, combining (40)-(42) with union bound, we have the below with probability at least $1 - 3\exp(-M)$,

$$\frac{1}{L}\sum_{m<j}\|\boldsymbol{R}^*_{m,j} - \boldsymbol{R}_{m,j}\|_{\rm F} \leq \left(\frac{1}{|\boldsymbol{\Omega}|} + \frac{1}{L}\right)\sum_{(m,j)\in\boldsymbol{\Omega}}\frac{1+\sqrt{M}}{\sqrt{S_{m,j}}} + C\sqrt{\frac{MK^2\log(M)}{|\boldsymbol{\Omega}|}}, \tag{43}$$

where $L = M(M-1)/2$.

F. Proof of Theorem 3

We restate the assumptions and the convergence theorem here:

Assumption 1 *The nonnegative factor* $\boldsymbol{H} \in \mathbb{R}_+^{MK\times K}$ *satisfies: (i)* $\operatorname{rank}(\boldsymbol{H}) = K$ *and* $\|\boldsymbol{H}\|_{\rm F} = \sigma$*; (ii)* $\frac{\|\boldsymbol{H}(j,:)\boldsymbol{\Theta}\|_2^2}{\|\boldsymbol{H}\boldsymbol{\Theta}\|_{\rm F}^2} \leq \zeta,\ \forall j,\ \forall\boldsymbol{\Theta} \in \mathbb{R}^{K\times K}$*; (iii) the locations of the nonzero elements of* $\boldsymbol{H}$ *are uniformly distributed over* $[MK]\times[K]$*, and the set* $\boldsymbol{\Delta} = \{(j,k) : [\boldsymbol{H}]_{j,k} > 0\}$ *has the following cardinality bound*

$$|\boldsymbol{\Delta}| = O\left(\frac{MK\gamma_0^2}{(1+MK\zeta)\sigma^4}\right); \tag{44}$$

and (iv) $0 < \gamma_0 \leq \min_{1\leq k\leq K}\{\beta_k^2 - \beta_{k+1}^2\}$*, where* β_k *is the kth singular value of* $\boldsymbol{H}$ *and* $\beta_{K+1} = 0$*.*

Theorem 3 Under Assumption 1, consider $\widehat{\boldsymbol{U}} = \boldsymbol{H}\boldsymbol{Q}^\top + \boldsymbol{N}$, where $\boldsymbol{Q} \in \mathbb{R}^{K\times K}$ is orthogonal, and apply (13). Denote $\nu = \|\boldsymbol{N}\|_{\rm F}$, $h_{(t)} = \|\boldsymbol{H}_{(t)} - \boldsymbol{H}\boldsymbol{\Pi}\|_{\rm F}^2$ and $q_{(t)} = \|\boldsymbol{Q}_{(t)} - \boldsymbol{Q}\boldsymbol{\Pi}\|_{\rm F}^2$, where $\boldsymbol{\Pi}$ is any permutation matrix. Suppose that $\nu \leq \sigma\min\{(1-\rho)\sqrt{\eta}q_{(0)}, 1\}$ for $\rho := O({}^{K\eta\sigma^4}/_{\gamma_0^2}) \in (0,1)$, where $\eta = ({}^{|\boldsymbol{\Delta}|}/_{MK^2})(1+MK\zeta)$, and that

$$2\sigma q_{(0)} + 2\nu < \min_{(j,k)\in\boldsymbol{\Delta}}[\boldsymbol{H}]_{j,k}. \tag{45}$$

Then, there exists $\alpha_{(t)} = \alpha > 0$ such that with probability of at least $1 - \delta$ the following holds:

$$q_{(t)} \leq \rho q_{(t-1)} + O\left({}^{K\sigma^2\nu^2}/_{\gamma_0^2}\right), \tag{46a}$$

$$h_{(t)} \leq 2\eta\sigma^2 q_{(t-1)} + 2\nu^2, \tag{46b}$$

where $\delta = 2\exp\left(-2|\boldsymbol{\Delta}|/K^2(1-\frac{|\boldsymbol{\Delta}|-1}{MK^2})\right)$.

Let $\widehat{\boldsymbol{X}}$ be the estimated $\boldsymbol{X}$ in (6). Consider the rank-K square root decomposition of $\widehat{\boldsymbol{X}} \in \mathbb{R}^{MK\times MK}$:

$$\widehat{\boldsymbol{X}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top.$$

It can be shown that $\widehat{\boldsymbol{U}} = \boldsymbol{U} + \boldsymbol{N} = \boldsymbol{H}\boldsymbol{Q}^\top + \boldsymbol{N}$ with bounded noise $\boldsymbol{N}$, if $\widehat{\boldsymbol{X}}$ is a reasonable estimate for $\boldsymbol{X}$ (cf. Lemma 1).

Using $\widehat{\boldsymbol{U}} \in \mathbb{R}^{MK\times K}$, the proposed SymNMF algorithm has the following updates:

$$\boldsymbol{H}_{(t+1)} \leftarrow \mathsf{ReLU}_{\alpha_{(t)}}\left(\widehat{\boldsymbol{U}}\boldsymbol{Q}_{(t)}\right) \tag{47a}$$

$$\boldsymbol{W}_{(t+1)}\boldsymbol{\Sigma}_{(t+1)}\boldsymbol{V}_{(t+1)}^\top \leftarrow \mathsf{svd}\left(\boldsymbol{H}_{(t+1)}^\top\widehat{\boldsymbol{U}}\right) \tag{47b}$$

$$\boldsymbol{Q}_{(t+1)} \leftarrow \boldsymbol{V}_{(t+1)}\boldsymbol{W}_{(t+1)}^\top, \tag{47c}$$

where $\alpha_{(t)} > 0$.

In the proof, we omit the permutation notation $\boldsymbol{\Pi}$ for notation simplicity, since all the column-permuted version of $\boldsymbol{H}$ and $\boldsymbol{Q}$ are considered equally good—i.e., the column permutation ambiguity in NMF problems is intrinsic; see (Fu et al., 2019; Huang et al., 2014).

Suppose that $\boldsymbol{Q}^\top\boldsymbol{Q}_{(t)} = \boldsymbol{I} + \boldsymbol{E}_{\boldsymbol{Q}_{(t)}}$. Note that

$$\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}} = \|\boldsymbol{Q} - \boldsymbol{Q}_{(t)}\|_{\mathrm{F}}$$

per the orthogonality of $\boldsymbol{Q}$ and $\boldsymbol{Q}_{(t)}$.

Also define

$$\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}} := \boldsymbol{H}_{(t+1)} - \boldsymbol{H}.$$

F.1. The $\boldsymbol{H}$-update

From the update in (47a), the below set of relations can be obtained:

$$\begin{aligned}
||\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}||_{\mathrm{F}} &= \left\|\mathsf{ReLU}_{\alpha_{(t)}}\left(\widehat{\boldsymbol{U}}\boldsymbol{Q}_{(t)}\right) - \boldsymbol{H}\right\|_{\mathrm{F}} \\
&= \left\|\mathsf{ReLU}_{\alpha_{(t)}}\left((\boldsymbol{H}\boldsymbol{Q}^\top + \boldsymbol{N})\boldsymbol{Q}_{(t)}\right) - \boldsymbol{H}\right\|_{\mathrm{F}} \\
&= \left\|\mathsf{ReLU}_{\alpha_{(t)}}\left(\boldsymbol{H}(\boldsymbol{Q}^\top\boldsymbol{Q}_{(t)}) + \boldsymbol{N}\boldsymbol{Q}_{(t)}\right) - \boldsymbol{H}\right\|_{\mathrm{F}} \\
&= \left\|\mathsf{ReLU}_{\alpha_{(t)}}\left(\boldsymbol{H}(\boldsymbol{I} + \boldsymbol{E}_{\boldsymbol{Q}_{(t)}}) + \boldsymbol{N}\boldsymbol{Q}_{(t)}\right) - \boldsymbol{H}\right\|_{\mathrm{F}} \\
&= \left\|\mathsf{ReLU}_{\alpha_{(t)}}\left(\boldsymbol{H} + \boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}\right) - \boldsymbol{H}\right\|_{\mathrm{F}}.
\end{aligned} \tag{48}$$

Recall that $\boldsymbol{\Delta} := \{(j,k) : [\boldsymbol{H}]_{j,k} > 0\}$. Assume that the following conditions are satisfied for $\alpha_{(t)}$ (at the end of the proof, using Lemma 8, we will establish the feasibility of $\alpha_{(t)}$ satisfying the below conditions),

$$\alpha_{(t)} \leq [\boldsymbol{H} + \boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}, \quad \forall (j,k) \in \boldsymbol{\Delta}, \tag{49}$$

$$\alpha_{(t)} \geq [\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}, \quad \forall j,k. \tag{50}$$

Then, we have

$$\begin{aligned}
&\|\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}\|_{\mathrm{F}}^2 \\
&= \sum_{(j,k)\in\boldsymbol{\Delta}} \left|[\mathsf{ReLU}_{\alpha_{(t)}}\left(\boldsymbol{H} + \boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}\right)]_{j,k} - [\boldsymbol{H}]_{j,k}\right|^2 + \sum_{(j,k)\notin\boldsymbol{\Delta}} \left|[\mathsf{ReLU}_{\alpha_{(t)}}\left(\boldsymbol{H} + \boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}\right)]_{j,k} - [\boldsymbol{H}]_{j,k}\right|^2 \\
&= \sum_{(j,k)\in\boldsymbol{\Delta}} \left|[\mathsf{ReLU}_{\alpha_{(t)}}\left(\boldsymbol{H} + \boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}\right)]_{j,k} - [\boldsymbol{H}]_{j,k}\right|^2 + \sum_{(j,k)\notin\boldsymbol{\Delta}} \left|[\mathsf{ReLU}_{\alpha_{(t)}}\left(\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}\right)]_{j,k}\right|^2 \\
&= \sum_{(j,k)\in\boldsymbol{\Delta}} \left|[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}\right|^2,
\end{aligned} \tag{51}$$

where we used $[\boldsymbol{H}]_{j,k} = 0, \forall (j,k) \notin \boldsymbol{\Delta}$ to get the second equality and applied the conditions in (49) and (50) to obtain the last equality.

Note that the below holds:

$$
\begin{aligned}
\left|[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}\right|^2 &= |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}]_{j,k}|^2 + |[\boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}|^2 + 2[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}]_{j,k}[\boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k} \\
&\leq |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}]_{j,k}|^2 + |[\boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}|^2 + |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}|^2 + |[\boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}|^2 \\
&= 2|[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}]_{j,k}|^2 + 2|[\boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}|^2, \qquad (52)
\end{aligned}
$$

where we have applied the Young's inequality in the first inequality.

Combining (51) and (52), we get that

$$
\|\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}\|_{\mathrm{F}}^2 \leq 2 \sum_{(j,k)\in\boldsymbol{\Delta}} \left|[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}]_{j,k}\right|^2 + 2 \sum_{(j,k)\in\boldsymbol{\Delta}} \left|[\boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}\right|^2. \qquad (53)
$$

Next, we consider the following lemma to bound the first term in (53).

Lemma 6 *(Serfling, 1974) Let $X = [X_1, \ldots, X_n]$ be a set of samples taken without replacement from a set $\{x_1, \ldots, x_N\}$ with mean u where $n \leq N$. Denote $a := \min_i x_i$ and $b := \max_i x_i$. Then, we have*

$$
\mathsf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - u\right| \geq s\right) \leq 2\exp\left(-\frac{2ns^2}{\left(1 - \frac{n-1}{N}\right)(b-a)^2}\right).
$$

Applying Lemma 6, and by the assumption that nonzero elements of $\boldsymbol{H}$ are located over $[MK] \times [K]$ uniformly at random, we get

$$
\mathsf{Pr}\left(\frac{1}{|\boldsymbol{\Delta}|}\sum_{(j,k)\in\boldsymbol{\Delta}} |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}]_{j,k}|^2 - \frac{1}{JK}\|\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 \geq s\right) \leq 2\exp\left(-\frac{2|\boldsymbol{\Delta}|s^2}{(1 - \frac{|\boldsymbol{\Delta}|-1}{JK})(b-a)^2}\right),
$$

where $J = MK$. Using the assumption that $\frac{\|\boldsymbol{H}(j,:)\boldsymbol{\Theta}\|_2^2}{\|\boldsymbol{H}\boldsymbol{\Theta}\|_{\mathrm{F}}^2} \leq \zeta,\ \forall j,\ \forall \boldsymbol{\Theta} \in \mathbb{R}^{K\times K}$, we get

$$
b = \max_{j,k}\ |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}]_{j,k}|^2 \leq \max_{j}\ \|\boldsymbol{H}(j,:)\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_2^2 \leq \zeta\|\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 \qquad (54a)
$$

$$
a = \min_{j,k}\ |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}]_{j,k}|^2 \geq 0. \qquad (54b)
$$

Using the bounds (54) and by letting $s = \zeta\|\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2/K$, we get that

$$
\mathsf{Pr}\left(\frac{1}{|\boldsymbol{\Delta}|}\sum_{i,k\in\boldsymbol{\Delta}} |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}]_{j,k}|^2 - \frac{1}{JK}\|\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 \geq \frac{1}{K}\zeta\|\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2\right) \leq 2\exp\left(-\frac{2|\boldsymbol{\Delta}|}{K^2(1 - \frac{|\boldsymbol{\Delta}|-1}{JK})}\right).
$$

It implies that with probability at least $1 - 2\exp\left(-\frac{2|\boldsymbol{\Delta}|}{K^2(1-\frac{|\boldsymbol{\Delta}|-1}{JK})}\right)$, we get

$$
\sum_{(j,k)\in\boldsymbol{\Delta}} |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}]_{j,k}|^2 \leq \frac{|\boldsymbol{\Delta}|}{JK}(1 + J\zeta)\|\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 \leq \frac{|\boldsymbol{\Delta}|}{JK}(1 + J\zeta)\|\boldsymbol{H}\|_{\mathrm{F}}^2\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2. \qquad (55)
$$

Letting $\eta = \frac{|\boldsymbol{\Delta}|}{JK}(1 + J\zeta)$ and applying (55) in (53), we get that

$$
\begin{aligned}
\|\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}\|_{\mathrm{F}}^2 &\leq 2\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 + 2\sum_{(j,k)\in\boldsymbol{\Delta}} \left|[\boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}\right|^2 \\
&\leq 2\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 + 2\|\boldsymbol{N}\boldsymbol{Q}_{(t)}\|_{\mathrm{F}}^2 \\
&= 2\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 + 2\nu^2, \qquad (56)
\end{aligned}
$$

where we have used $\|\boldsymbol{N}\|_{\mathrm{F}} = \nu$ and the orthogonality of $\boldsymbol{Q}_{(t)}$ in the last equality.

F.2. The $\boldsymbol{Q}$-update

We will now consider the update in (47b):

$$\begin{aligned}\boldsymbol{H}_{(t+1)}^\top \widehat{\boldsymbol{U}} &= (\boldsymbol{H} + \boldsymbol{E}_{\boldsymbol{H}_{(t+1)}})^\top (\boldsymbol{U} + \boldsymbol{N}) \\ &= \boldsymbol{H}^\top \boldsymbol{U} + \boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}^\top \boldsymbol{U} + \boldsymbol{H}^\top \boldsymbol{N} + \boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}^\top \boldsymbol{N}.\end{aligned}$$

We bound the below:

$$\begin{aligned}\|\boldsymbol{H}_{(t+1)}^\top \widehat{\boldsymbol{U}} - \boldsymbol{H}^\top \boldsymbol{U}\|_{\rm F}^2 &= \|\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}^\top \boldsymbol{U} + \boldsymbol{H}^\top \boldsymbol{N} + \boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}^\top \boldsymbol{N}\|_{\rm F}^2 \\ &\leq 3\|\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}^\top \boldsymbol{U}\|_{\rm F}^2 + 3\|\boldsymbol{H}^\top \boldsymbol{N}\|_{\rm F}^2 + 3\|\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}^\top \boldsymbol{N}\|_{\rm F}^2 \\ &\leq 3\|\boldsymbol{H}\|_{\rm F}^2 \|\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}\|_{\rm F}^2 + 3\|\boldsymbol{H}\|_{\rm F}^2 \nu^2 + 3\nu^2 \|\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}\|_{\rm F}^2 \\ &= 3(\|\boldsymbol{H}\|_{\rm F}^2 + \nu^2)\|\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}\|_{\rm F}^2 + 3\|\boldsymbol{H}\|_{\rm F}^2 \nu^2 \\ &\leq 3(\|\boldsymbol{H}\|_{\rm F}^2 + \nu^2)\left(2\eta\|\boldsymbol{H}\|_{\rm F}^2 \|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\rm F}^2 + 2\nu^2\right) + 3\|\boldsymbol{H}\|_{\rm F}^2 \nu^2 \\ &= 6\eta(\|\boldsymbol{H}\|_{\rm F}^2 + \nu^2)\|\boldsymbol{H}\|_{\rm F}^2 \|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\rm F}^2 + 6(\|\boldsymbol{H}\|_{\rm F}^2 + \nu^2)\nu^2 + 3\|\boldsymbol{H}\|_{\rm F}^2 \nu^2 \\ &\leq 12\eta\|\boldsymbol{H}\|_{\rm F}^4 \|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\rm F}^2 + 15\|\boldsymbol{H}\|_{\rm F}^2 \nu^2, \end{aligned} \tag{57}$$

where we have used the Young's inequality for the first inequality, used the fact that $\|\boldsymbol{U}\|_{\rm F} = \|\boldsymbol{H}\|_{\rm F}$ for the second inequality, applied the result in (56) for the third inequality and used the assumption that $\|\boldsymbol{N}\|_{\rm F} = \nu \leq \|\boldsymbol{H}\|_{\rm F}$ for the last inequality.

Let us proceed to characterize the SVD operation in (47b). Denote the full SVD of $\boldsymbol{H}^\top \boldsymbol{U}$ using the following notation:

$$\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \mathsf{svd}\left(\boldsymbol{H}^\top \boldsymbol{U}\right).$$

We invoke the below lemma:

Lemma 7 *(Fan et al., 2018; Mirsky, 1960; Wedin, 1972) Let $\boldsymbol{C} \in \mathbb{R}^{m\times n}$ and $\widehat{\boldsymbol{C}} \in \mathbb{R}^{m\times n}$ have singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)}$ and $\widehat{\sigma}_1 \geq \widehat{\sigma}_2 \geq \ldots \widehat{\sigma}_{\min(m,n)}$, respectively. Let $r \leq \min\{m,n\}$. Denote $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r \in \mathbb{R}^m$ and $\widehat{\boldsymbol{w}}_1, \ldots, \widehat{\boldsymbol{w}}_r \in \mathbb{R}^m$ as the orthonormal columns satisfying $\boldsymbol{C}^\top \boldsymbol{w}_i = \sigma_i \boldsymbol{v}_i$ and $\widehat{\boldsymbol{C}}^\top \widehat{\boldsymbol{w}}_i = \widehat{\sigma}_i \widehat{\boldsymbol{v}}_i$ for $i = 1, \ldots, r$ and let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r \in \mathbb{R}^n$ and $\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_r \in \mathbb{R}^n$ are orthonormal columns satisfying $\boldsymbol{C}\boldsymbol{v}_i = \sigma_i \boldsymbol{w}_i$ and $\widehat{\boldsymbol{C}}\widehat{\boldsymbol{v}}_i = \widehat{\sigma}_i \widehat{\boldsymbol{w}}_i$ for $i = 1, \ldots, r$. Denote $\gamma_0 = \min\{\sigma_i - \sigma_{i+1} : i = 1, \ldots, r\}$ where $\sigma_{r+1} = 0$. Then, if $\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_2 \leq \gamma_0/2$, we have*

$$\max_{1\leq i\leq r}\{\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2 \vee \|\widehat{\boldsymbol{v}}_i - \boldsymbol{v}_i\|_2\} \leq \frac{2\sqrt{2}\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_2}{\gamma_0}, \tag{58}$$

where the operation $a \vee b = \max\{a, b\}$.

A short proof of how the bound in (58) is obtained from the classic result in (Wedin, 1972) is given in Section I.

By letting $\boldsymbol{C} := \boldsymbol{H}^\top \boldsymbol{U}$, $\widehat{\boldsymbol{C}} := \boldsymbol{H}_{(t+1)}^\top \widehat{\boldsymbol{U}}$ and applying Lemma 7, we have

$$\|\boldsymbol{W}_{(t+1)} - \boldsymbol{W}\|_{\rm F} \leq \frac{2\sqrt{2K}\|\boldsymbol{H}_{(t+1)}^\top \widehat{\boldsymbol{U}} - \boldsymbol{H}^\top \boldsymbol{U}\|_{\rm F}}{\gamma_0}, \tag{59}$$

$$\|\boldsymbol{V}_{(t+1)} - \boldsymbol{V}\|_{\rm F} \leq \frac{2\sqrt{2K}\|\boldsymbol{H}_{(t+1)}^\top \widehat{\boldsymbol{U}} - \boldsymbol{H}^\top \boldsymbol{U}\|_{\rm F}}{\gamma_0}, \tag{60}$$

where we have used the fact that for any matrix $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K]$, the equality $\|\boldsymbol{\Theta}\|_{\rm F} = \sqrt{\sum_{i=1}^K \|\boldsymbol{\theta}_i\|_2^2}$ holds. We have also applied matrix norm equivalence $\|\boldsymbol{\Theta}\|_2 \leq \|\boldsymbol{\Theta}\|_{\rm F}$. Note that since the singular values of $\boldsymbol{H}^\top \boldsymbol{U}$ are the same as that of $\boldsymbol{H}^\top \boldsymbol{H}$, we re-define γ_0 as

$$\gamma_0 = \min_{1\leq k\leq K}\{\beta_k^2 - \beta_{k+1}^2\},$$

where β_k's are the singular values of $\boldsymbol{H}$.

By squaring the term in the right hand side of (59), we get

$$\begin{aligned}\|\boldsymbol{W}_{(t+1)} - \boldsymbol{W}\|_{\mathrm{F}}^2 &\le \frac{8K\|\boldsymbol{H}_{(t+1)}^\top \widehat{\boldsymbol{U}} - \boldsymbol{H}^\top \boldsymbol{U}\|_{\mathrm{F}}^2}{\gamma_0^2} \\ &\le \frac{8K\left(12\eta\|\boldsymbol{H}\|_{\mathrm{F}}^4\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 + 15\|\boldsymbol{H}\|_{\mathrm{F}}^2\nu^2\right)}{\gamma_0^2}, \qquad (61)\end{aligned}$$

where we applied (57) to obtain the last inequality. We can similarly get that

$$\|\boldsymbol{V}_{(t+1)} - \boldsymbol{V}\|_{\mathrm{F}}^2 \le \frac{8K\left(12\eta\|\boldsymbol{H}\|_{\mathrm{F}}^4\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 + 15\|\boldsymbol{H}\|_{\mathrm{F}}^2\nu^2\right)}{\gamma_0^2}. \qquad (62)$$

Consider $\boldsymbol{E}_{\boldsymbol{Q}_{(t+1)}} = \boldsymbol{Q}_{(t+1)} - \boldsymbol{Q}$. Then,

$$\begin{aligned}\|\boldsymbol{E}_{\boldsymbol{Q}_{(t+1)}}\|_{\mathrm{F}}^2 = \|\boldsymbol{Q}_{(t+1)} - \boldsymbol{Q}\|_{\mathrm{F}}^2 &= \|\boldsymbol{V}_{(t+1)}\boldsymbol{W}_{(t+1)}^\top - \boldsymbol{V}\boldsymbol{W}^\top\|_{\mathrm{F}}^2 \\ &= \|\boldsymbol{V}_{(t+1)}(\boldsymbol{W}_{(t+1)}^\top - \boldsymbol{W}^\top) + (\boldsymbol{V}_{(t+1)} - \boldsymbol{V})\boldsymbol{W}^\top\|_{\mathrm{F}}^2 \\ &\le 2\|\boldsymbol{W}_{(t+1)} - \boldsymbol{W}\|_{\mathrm{F}}^2 + 2\|\boldsymbol{V}_{(t+1)} - \boldsymbol{V}\|_{\mathrm{F}}^2,\end{aligned}$$

where the last inequality is by the Young's inequality and the fact that $\|\boldsymbol{\Theta}\boldsymbol{\Phi}\|_{\mathrm{F}}^2 \le \|\boldsymbol{\Theta}\|_2^2\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2$ for two matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$; we have also used that $\|\boldsymbol{W}\|_2 = \|\boldsymbol{V}_{(t+1)}\|_2 = 1$. The above leads to

$$\|\boldsymbol{E}_{\boldsymbol{Q}_{(t+1)}}\|_{\mathrm{F}}^2 \le \frac{CK\left(\eta\|\boldsymbol{H}\|_{\mathrm{F}}^4\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 + \|\boldsymbol{H}\|_{\mathrm{F}}^2\nu^2\right)}{\gamma_0^2}. \qquad (63)$$

for a certain constant $C > 1$. Let us denote $\rho := \frac{CK\eta\|\boldsymbol{H}\|_{\mathrm{F}}^4}{\gamma_0^2}$. Then we have

$$\|\boldsymbol{E}_{\boldsymbol{Q}_{(t+1)}}\|_{\mathrm{F}}^2 \le \rho\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 + \frac{\rho\nu^2}{\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2}. \qquad (64)$$

We can see that if the below condition is satisfied, then $\rho < 1$:

$$\begin{aligned}\eta = \frac{|\boldsymbol{\Delta}|}{JK}(1 + J\zeta) &\le \frac{\gamma_0^2}{CK\|\boldsymbol{H}\|_{\mathrm{F}}^4}, \\ \implies |\boldsymbol{\Delta}| &\le \frac{J\gamma_0^2}{C(1 + J\zeta)\|\boldsymbol{H}\|_{\mathrm{F}}^4}. \qquad (65)\end{aligned}$$

Therefore, under the conditions of $\alpha_{(t)}$ in (49) and (50) and the condition on $|\boldsymbol{\Delta}|$ in (65), we get the bound for $\|\boldsymbol{E}_{\boldsymbol{Q}_{(t+1)}}\|_{\mathrm{F}}^2$ and $\|\boldsymbol{E}_{\boldsymbol{H}_{(t+1)}}\|_{\mathrm{F}}^2$ given by (64) and (56), respectively, with $\rho < 1$ and with probability greater than $1 - 2\exp\left(-\frac{2|\boldsymbol{\Delta}|}{K^2(1 - \frac{|\boldsymbol{\Delta}| - 1}{JK})}\right)$.

Regarding the feasibility of $\alpha_{(t)}$ satisfying the conditions (49) and (50), we have the following lemma:

Lemma 8 *Assume that the following conditions are satisfied:*

$$\nu \le (1 - \rho)\sqrt{\eta}\|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}, \qquad \min_{(j,k)\in\boldsymbol{\Delta}}[\boldsymbol{H}]_{j,k} > 2\|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}} + 2\nu.$$

Then there exists $\alpha_{(t)} = \alpha > 0$, for all t, specified as below such that the bounds given by (64) and (56) hold true:

$$\|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}} + \nu \le \alpha \le \min_{(j,k)\in\boldsymbol{\Delta}}[\boldsymbol{H}]_{j,k} - \|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}} - \nu.$$

The proof can be found in Sec. J.

G. Proof of Lemma 2

Consider the below:

$$\begin{aligned}\|(\boldsymbol{Y}+\boldsymbol{E})^{-1}-\boldsymbol{Y}^{-1}\|_2 &= \|(\boldsymbol{Y}+\boldsymbol{E})^{-1}(\boldsymbol{I}-(\boldsymbol{Y}+\boldsymbol{E})\boldsymbol{Y}^{-1}\|_2 \\ &= \|(\boldsymbol{Y}+\boldsymbol{E})^{-1}\boldsymbol{E}\boldsymbol{Y}^{-1}\|_2 \\ &\leq \frac{\|\boldsymbol{E}\|_2}{\sigma_{\min}(\boldsymbol{Y})\sigma_{\min}(\boldsymbol{Y}+\boldsymbol{E})}. \end{aligned} \tag{66}$$

Next, we consider the following relations for any vector $\boldsymbol{x} \in \mathbb{R}^K$ satisfying $\|\boldsymbol{x}\| = 1$:

$$\begin{aligned}\|(\boldsymbol{Y}+\boldsymbol{E})\boldsymbol{x}\|_2 &= \|\boldsymbol{Y}\boldsymbol{x}+\boldsymbol{E}\boldsymbol{x}\|_2 \\ &\geq \|\boldsymbol{Y}\boldsymbol{x}\|_2 - \|\boldsymbol{E}\boldsymbol{x}\|_2, \\ \implies \min_{\boldsymbol{x}} \|(\boldsymbol{Y}+\boldsymbol{E})\boldsymbol{x}\|_2 &\geq \min_{\boldsymbol{x}} \|\boldsymbol{Y}\boldsymbol{x}\|_2 - \max_{\boldsymbol{x}} \|\boldsymbol{E}\boldsymbol{x}\|_2, \\ \implies \sigma_{\min}(\boldsymbol{Y}+\boldsymbol{E}) &\geq \sigma_{\min}(\boldsymbol{Y}) - \|\boldsymbol{E}|_2, \end{aligned}$$

where the first inequality is by applying the triangle inequality. Using the assumption that $\|\boldsymbol{E}\|_2 \leq \sigma_{\min}(\boldsymbol{Y})/2$, we get $\sigma_{\min}(\boldsymbol{Y}+\boldsymbol{E}) \geq \sigma_{\min}(\boldsymbol{Y})/2$. Applying this relation in (66), we get the bound in the lemma.

H. Proof of Lemma 3

Recall the below relation:

$$\boldsymbol{C} = [\boldsymbol{R}_{m,r}^\top, \boldsymbol{R}_{\ell,r}^\top]^\top = [\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top \boldsymbol{D}\boldsymbol{A}_r. \tag{67}$$

The SVD of $\boldsymbol{C}$ results the below:

$$\boldsymbol{C} = [\boldsymbol{U}_m^\top, \boldsymbol{U}_\ell^\top]^\top \boldsymbol{\Sigma}_{m,\ell,r}\boldsymbol{V}_r^\top \tag{68}$$

From (67) and (68), we get that there exists a nonsingular matrix $\boldsymbol{\Theta} \in \mathbb{R}^{K\times K}$ such that

$$[\boldsymbol{U}_m^\top, \boldsymbol{U}_\ell^\top]^\top = [\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top \boldsymbol{\Theta}, \tag{69}$$

where the matrix $[\boldsymbol{U}_m^\top, \boldsymbol{U}_\ell^\top]^\top$ is semi-orthogonal. Therefore, we get

$$\sigma_{\max}(\boldsymbol{\Theta}) = \frac{1}{\sigma_{\min}([\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top)} \quad \text{and} \quad \sigma_{\min}(\boldsymbol{\Theta}) = \frac{1}{\sigma_{\max}([\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top)}. \tag{70}$$

Since $\boldsymbol{A}_m$ is full row-rank, we have

$$\begin{aligned}\sigma_{\min}(\boldsymbol{U}_m) &= \min_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}_m\boldsymbol{\Theta}\boldsymbol{x}\|_2 \\ &\geq \min_{\|\boldsymbol{x}\|_2=1} \sigma_{\min}(\boldsymbol{A}_m)\|\boldsymbol{\Theta}\boldsymbol{x}\|_2 = \sigma_{\min}(\boldsymbol{A}_m) \min_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{\Theta}\boldsymbol{x}\|_2 \\ &= \sigma_{\min}(\boldsymbol{A}_m)\sigma_{\min}(\boldsymbol{\Theta}) = \frac{\sigma_{\min}(\boldsymbol{A}_m)}{\sigma_{\max}([\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top)}. \end{aligned} \tag{71}$$

where we have applied (70) to obtain the last equality.

We proceed to bound $\sigma_{\max}([\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top)$. Under the assumption $\kappa(\boldsymbol{A}_m) \leq \gamma$, for all m, there exists a positive scalar $\omega_{\max}$ and $\omega_{\min}$, such that for all m,

$$\sigma_{\max}(\boldsymbol{A}_m) \leq \omega_{\max}, \quad \sigma_{\min}(\boldsymbol{A}_m) \geq \omega_{\min}, \quad \gamma := \frac{\omega_{\max}}{\omega_{\min}}.$$

Then we have,

$$\begin{aligned}\sigma_{\max}^2([\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top) &= \|[\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top\|_2^2 \leq \|[\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top\|_{\mathrm{F}}^2 \\ &= \|\boldsymbol{A}_m\|_{\mathrm{F}}^2 + \|\boldsymbol{A}_\ell\|_{\mathrm{F}}^2 \leq K\|\boldsymbol{A}_m\|_2^2 + K\|\boldsymbol{A}_\ell\|_2^2 \leq 2K\omega_{\max}^2, \end{aligned}$$

where we have utilized the norm equivalence for the first and second inequalities. Hence, we have

$$\sigma_{\max}([\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top) \le \sqrt{2K}\omega_{\max}.$$

Applying the above results in (71), we get

$$\sigma_{\min}(\boldsymbol{U}_m) \ge \frac{\omega_{min}}{\sqrt{2K}\omega_{\max}} = \frac{1}{\sqrt{2K}\gamma}.$$

Similarly, we can easily show the above lower bound for $\sigma_{\min}(\boldsymbol{U}_\ell)$.

Next, we consider upper bounding $\sigma_{\max}(\boldsymbol{U}_m)$ and $\sigma_{\max}(\boldsymbol{U}_\ell)$. From (69) and (70), we have

$$\begin{aligned}\sigma_{\max}(\boldsymbol{U}_m) \le \sigma_{\max}(\boldsymbol{\Theta})\sigma_{\max}(\boldsymbol{A}_m) &= \frac{\sigma_{\max}(\boldsymbol{A}_m)}{\sigma_{\min}([\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top)} \\ &\le \frac{\sigma_{\max}(\boldsymbol{A}_m)}{\sigma_{\min}(\boldsymbol{A}_m)} \le \frac{\omega_{\max}}{\omega_{\min}} = \gamma,\end{aligned}$$

where we have applied $\sigma_{\min}([\boldsymbol{A}_m^\top, \boldsymbol{A}_\ell^\top]^\top) \ge \sigma_{\min}(\boldsymbol{A}_m)$ for second inequality. Similarly, we can easily show the above upper bound for $\sigma_{\max}(\boldsymbol{U}_\ell)$.

I. Proof of Lemma 7

The perturbation theorem in (Wedin, 1972) gives the below bound if $\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_2 \le \widetilde{\gamma}_0/2$,

$$\sqrt{\sum_{i=1}^{r}(\sin^2\theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i) + \sin^2\theta(\widehat{\boldsymbol{v}}_i, \boldsymbol{v}_i))} \le \frac{2\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_2}{\widetilde{\gamma}_0}, \tag{72}$$

where $\theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i)$ is the canonical angle between the left singular vectors $\widehat{\boldsymbol{w}}_i$ and $\boldsymbol{w}_i$. We can easily see that

$$\max\{\sin\theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i), \sin\theta(\widehat{\boldsymbol{v}}_i, \boldsymbol{v}_i)\} \le \sqrt{\sin^2\theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i) + \sin^2\theta(\widehat{\boldsymbol{v}}_i, \boldsymbol{v}_i)} \le \sqrt{\sum_{i=1}^{r}(\sin^2\theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i) + \sin^2\theta(\widehat{\boldsymbol{v}}_i, \boldsymbol{v}_i))}. \tag{73}$$

Also, consider the below:

$$\begin{aligned}\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}\|_2^2 &= 2 - 2\widehat{\boldsymbol{w}}_i^\top \boldsymbol{w} \\ &\le 2(1 - \cos\theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i)) \\ &\le 2(1 - \cos^2\theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i)) \\ &= 2\sin^2\theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i) \\ \implies \|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}\|_2 &\le \sqrt{2}\sin\theta(\widehat{\boldsymbol{w}}_i, \boldsymbol{w}_i).\end{aligned}$$

The above inequality combined with (72) and (73) gives the bound in Lemma 7.

J. Proof of Lemma 8

The conditions on $\alpha_{(t)}$ given by (49) and (50) can be re-written as:

$$\max_{(j,k)} [\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k} \le \alpha_{(t)} \le \min_{(j,k)\in\boldsymbol{\varDelta}}[\boldsymbol{H}]_{j,k} + \min_{(j,k)\in\boldsymbol{\varDelta}}[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}. \tag{74}$$

We can bound the term $\min_{(j,k)\in\boldsymbol{\varDelta}}[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}$ as below:

$$\min_{(j,k)\in\boldsymbol{\varDelta}}[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k} \ge -\max_{(j,k)\in\boldsymbol{\varDelta}}\left|[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}\right| \ge -\max_{(j,k)}\left|[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}\right|. \tag{75}$$

Using (75), we can re-write the conditions on $\alpha_{(t)}$ in (74) as below:

$$\max_{(j,k)} |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}| \le \alpha_{(t)} \le \min_{(j,k)\in\boldsymbol{\Delta}} [\boldsymbol{H}]_{j,k} - \max_{(j,k)} |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}|. \tag{76}$$

To proceed, we bound the term $\max_{(j,k)} \ |[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}|$ as below:

$$\max_{(j,k)} \left|[\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}]_{j,k}\right| \le \|\boldsymbol{H}\boldsymbol{E}_{\boldsymbol{Q}_{(t)}} + \boldsymbol{N}\boldsymbol{Q}_{(t)}\|_{\mathrm{F}} \le \|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}} + \nu, \tag{77}$$

where we used $\|\boldsymbol{N}\|_{\mathrm{F}} = \nu$ and the orthogonality of $\boldsymbol{Q}_{(t)}$ to obtain the last inequality. Applying (77) in (76), we can further re-write the conditions as:

$$\|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}} + \nu \le \alpha_{(t)} \le \min_{(j,k)\in\boldsymbol{\Delta}} [\boldsymbol{H}]_{j,k} - \|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}} - \nu. \tag{78}$$

Next, we proceed to bound $\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}$ using $\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}$. To accomplish this, we can recursively apply the results in (64) to obtain the below relation for any $t > 1$:

$$\begin{aligned}\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 = \|\boldsymbol{Q}_{(t)} - \boldsymbol{Q}\|_{\mathrm{F}}^2 &\le \rho^t\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2 + \frac{\nu^2}{\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2}\sum_{q=1}^{t}\rho^q, \\ &= \rho^t\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2 + \frac{\nu^2(1-\rho^{t+1})}{\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2(1-\rho)}.\end{aligned} \tag{79}$$

With the above result, we consider the following:

$$\begin{aligned}\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 - \|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2 &\le \left(\rho^t\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2 + \frac{\nu^2}{\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2}\sum_{q=1}^{t}\rho^q\right) - \|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2 \\ &= \left((\rho^t - 1)\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2 + \frac{\nu^2}{\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2}\sum_{q=1}^{t}\rho^q\right),\end{aligned} \tag{80}$$

where we applied (79) to get the first inequality. If the R.H.S of (80) is smaller than zero, then we have $\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 \le \|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2$. The condition to make the R.H.S of (80) smaller than zero can be written as below:

$$\begin{aligned}&(\rho^t - 1)\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2 + \frac{\nu^2}{\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2}\sum_{q=1}^{t}\rho^q \le 0 \\ \implies\ & \frac{\nu^2}{\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2}\sum_{q=1}^{t}\rho^q \le (1-\rho^t)\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2 \\ \implies\ & \frac{\nu^2}{\eta\|\boldsymbol{H}\|_{\mathrm{F}}^2}\frac{1}{1-\rho} \le (1-\rho)\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2 \\ &\implies \nu \le (1-\rho)\sqrt{\eta}\|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}},\end{aligned} \tag{81}$$

where the third inequality is obtained using the facts that $\sum_{q=1}^{t}\rho^q \le \sum_{q=1}^{\infty}\rho^q \le \frac{1}{1-\rho}$ and $1-\rho^t \ge 1-\rho$ since $\rho < 1$. It implies that if the conditions on ν given by (81) is satisfied,

$$\|\boldsymbol{E}_{\boldsymbol{Q}_{(t)}}\|_{\mathrm{F}}^2 \le \|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}}^2, \quad \forall t. \tag{82}$$

Applying (82) in (78), the condition on $\alpha_{(t)}$ can be further re-written as:

$$\|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}} + \nu \le \alpha_{(t)} \le \min_{(j,k)\in\boldsymbol{\Delta}} [\boldsymbol{H}]_{j,k} - \|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}} - \nu. \tag{83}$$

From (83), it is clear that we can find $\alpha_{(t)} = \alpha$ for every iteration t as long as

$$\min_{(j,k)\in\boldsymbol{\Delta}} [\boldsymbol{H}]_{j,k} > 2\|\boldsymbol{H}\|_{\mathrm{F}}\|\boldsymbol{E}_{\boldsymbol{Q}_{(0)}}\|_{\mathrm{F}} + 2\nu.$$

References

Fan, J., Wang, W., and Zhong, Y. An l_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

Fu, X., Huang, K., Sidiropoulos, N. D., and Ma, W.-K. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *IEEE Signal Process. Mag.*, 36(2):59–80, 2019.

Huang, K., Sidiropoulos, N., and Swami, A. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Trans. Signal Process.*, 62(1):211–224, 2014.

Mirsky, L. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

Razaviyayn, M., Hong, M., and Luo, Z.-Q. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

Rodrigues, F. and Pereira, F. Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (1), 2018.

Serfling, R. J. Probability inequalities for the sum in sampling without replacement. *Annals of Statistics*, 2(1):39–48, 1974.

Wang, Y.-X. and Xu, H. Stability of matrix factorization for collaborative filtering. In *Proceedings of International Conference on Machine Learning*, pp. 163–170, 2012.

Wedin, P. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

Yu, Y., Wang, T., and Samworth, R. J. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2): 315–323, 2014.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102):1–44, 2016.