JMLR: Workshop and Conference Proceedings vol 35:1–33, 2014

# Learning Mixtures of Discrete Product Distributions using Spectral Decompositions

**Prateek Jain** PRAJAIN@MICROSOFT.COM
**Microsoft Research India, Bangalore**

**Sewoong Oh** SWOH@ILLINOIS.EDU
**Department of Industrial and Enterprise Systems Engineering**
**University of Illinois at Urbana-Champaign**

## Abstract

We study the problem of learning a distribution from samples, when the underlying distribution is a mixture of product distributions over discrete domains. This problem is motivated by several practical applications such as crowdsourcing, recommendation systems, and learning Boolean functions. The existing solutions either heavily rely on the fact that the number of mixtures is finite or have sample/time complexity that is exponential in the number of mixtures. In this paper, we introduce a polynomial time/sample complexity method for learning a mixture of $r$ discrete product distributions over $\{1, 2, \dots, \ell\}^n$, for general $\ell$ and $r$. We show that our approach is consistent and further provide finite sample guarantees.

We use recently developed techniques from tensor decompositions for moment matching. A crucial step in these approaches is to construct certain tensors with low-rank spectral decompositions. These tensors are typically estimated from the sample moments. The main challenge in learning mixtures of discrete product distributions is that the corresponding low-rank tensors cannot be obtained directly from the sample moments. Instead, we need to estimate a low-rank matrix using only off-diagonal entries, and estimate a tensor using a few linear measurements. We give an alternating minimization based method to estimate the low-rank matrix, and formulate the tensor estimation problem as a least-squares problem.

## 1. Introduction

Consider the following generative model for sampling from a mixture of product distributions over discrete domains. We use $r$ to denote the number of components in the mixture, $\ell$ to denote the size of the discrete output alphabet in each coordinate, and $n$ to denote the total number of coordinates. Each sample belongs to one of $r$ components, and conditioned on its component $q \in \{1, \dots, r\}$ the $n$ dimensional discrete sample $y \in \{1, \dots, \ell\}^n$ is drawn from some distribution $\pi_q$. Precisely, the model is represented by the non-negative weights of the components $w = [w_1 \dots w_r] \in \mathbb{R}^r$ that sum to one, and the $r$ distributions $\Pi = [\pi_1 \dots \pi_r] \in \mathbb{R}^{\ell n \times r}$. We use an $\ell n$ dimensional binary random vector $x$ to represent a sample $y$. For $x = [x_1 \dots x_n] \in \{0,1\}^{\ell n}$, the $i$-th coordinate $x_i \in \{0,1\}^\ell$ is an $\ell$ dimensional binary random vector such that

$$x_i = e_j \text{ if and only if } y_i = j \ ,$$

where $e_j$ for some $j \in \{1, \dots, \ell\}$ is the standard coordinate basis vector.

When a sample is drawn, the *type* of the sample is drawn from $w = [w_1 \dots w_r]$ such that it has type $q$ with probability $w_q$. Conditioned on this type, the sample is distributed according to

$\pi_q \in \mathbb{R}^{\ell n}$, such that $y_i$'s are independent, hence it is a product distribution, and distributed according to

$$(\pi_q)_{(i;j)} = \mathbb{P}(y_i = j \mid y \text{ belong to component } q) \ ,$$

where $(\pi_q)_{(i;j)}$ is the $\big((i-1)\ell + j\big)$-th entry of the vector $\pi^{(q)}$. Note that using the binary encoding, $\mathbb{E}[x|\text{its type is } q] = \pi_q$, and $\mathbb{E}[x] = \sum_q w_q \pi_q$. Also, we let $\pi^{(i)} \in \mathbb{R}^{\ell \times r}$ represent the distribution in the $i$-th coordinate such that $\pi^{(i)}_{j;q} = (\pi_q)_{(i;j)} = \mathbb{P}(y_i = j | y \text{ belongs to component } q)$. Then, the discrete distribution can be represented by the matrix $\Pi \in \mathbb{R}^{\ell n \times r} = [\pi^{(1)}; \pi^{(2)}; \ldots; \pi^{(n)}]$ and the weights $w = [w_1, \ldots, w_r]$.

This mixture distribution (of $\ell$-wise discrete distributions over product spaces) captures as special cases the models used in several problems in domains such as crowdsourcing (Dawid and Skene, 1979), genetics (Sridhar et al., 2007), and recommendation systems (Tomozei and Massoulié, 2010). For example, in the crowdsourcing application, this model is same as the popular Dawid and Skene (Dawid and Skene, 1979) model: $x_i$ represents answer of the $i$-th worker to a multiple choice question (or task) of type $q \in [r]$. Given the ground truth label $q$, each of the worker is assumed to answer independently. The goal is to find out the "quality" of the workers (i.e. learn $\Pi$) and/or to learn the type of each question (clustering).

We are interested in the following two closely related problems:

- Learn mixture parameters $\{\pi_q\}_{q\in\{1,:::,r\}}$ and $\{w_q\}_{q\in\{1,:::,r\}}$ accurately and efficiently.
- Cluster the samples accurately and efficiently?

Historically, however, different algorithms have been proposed depending on which question is addressed. Also, for each of the problems, distinct measures of performances have been used to evaluate the proposed solution. In this paper, we propose an efficient method to address both questions.

The first question of estimating the underlying parameters of the mixture components has been addressed in (Kearns et al., 1994; Freund and Mansour, 1999; Feldman et al., 2008), where the error of a given algorithm is measured as the KL-divergence between the true distribution and the estimated distribution. More precisely, a mixture learning algorithm is said to be an *accurate learning algorithm*, if it outputs a mixture of product distribution such that the following holds with probability at least $1 - \delta$:

$$D_{\mathrm{KL}}\big(X \,||\, \widehat{X}\big) \equiv \sum_x \mathbb{P}(X = x) \log(\mathbb{P}(X = x)/\mathbb{P}(\widehat{X} = x)) \leq \varepsilon,$$

where $\epsilon, \delta \in (0, 1)$ are any given constants, and $X, \widehat{X} \in \{0, 1\}^{\ell n}$ denote the random vectors distributed according to the true and the estimated mixture distribution, respectively. Furthermore, the algorithm is said to *efficient* if its time complexity is polynomial in $n$, $r$, $\ell$, $1/\varepsilon$, and $\log(1/\delta)$.

This Probably Approximately Correct (PAC) style framework was first introduced by Kearns et al. (Kearns et al., 1994), where they provided the first analytical result for a simpler problem of learning mixtures of Hamming balls, which is a special case of our model with $\ell = 2$. However, the running time of the proposed algorithm is super-polynomial $O((n/\delta)^{\log r})$ and also assumes that one can obtain the exact probability of a sample $y$. Freund and Mansour (Freund and Mansour, 1999) were the first to addressed the sample complexity, but for the restrictive case of $r = 2$ and

$\ell = 2$. For this case, their method has running time $O(n^{3.5}\log^3(1/\delta)/\varepsilon^5)$ and sample complexity $O(n^2\log(1/\delta)/\varepsilon^2)$. Feldman, O'Donnell, and Servedio in Feldman et al. (2008) generalized approach of Freund and Mansour (1999) to arbitrary number of types $r$ and arbitrary number of output labels $\ell$. For general $\ell$, their algorithm requires running time scaling as $O((n\ell/\varepsilon)^{r^3})$. Hence, the proposed algorithm is an *efficient learning algorithm* only for finite values of $r = O(1)$ and $\ell = O(1)$.

A breakthrough in Feldman et al.'s result is that their result holds for all problem instances, with no dependence on the minimum weight $w_{\min}$ or the condition number $\sigma_1(\Pi W^{1=2})/\sigma_r(\Pi W^{1=2})$, where $\sigma_i(\Pi W^{1=2})$ is the $i$-th singular value of $\Pi W^{1=2}$, and $W$ is a $r \times r$ diagonal matrix with the weights $w$ in the diagonals. However, this comes at a cost of running time scaling exponentially in both $r^3$ and $\ell$, which is unacceptable in practice for any value of $r$ beyond two. Further, the running time is exponential for all problem instances, even when the problem parameters are *well-behaved*, with finite condition number.

In this paper, we alleviate this issue by proposing an efficient algorithm for *well-behave* mixture distributions. In particular, we give an algorithm with polynomial running time, and prove that it gives $\varepsilon$-accurate estimate for any problem instance that satisfy the following two conditions: $a$) the weight $w_q$ is strictly positive for all $q$; and $b$) the condition number $\sigma_1(\Pi W^{1=2})/\sigma_r(\Pi W^{1=2})$ is bounded as per hypotheses in Theorem 3.

The existence of an efficient learning algorithm for all problem instances and parameters still remains an open problem, especially in the PAC learning setting.

| | $r, \ell = O(1)$ | General $r$ and $\ell$ |
|---|---|---|
| $\sigma_1(\Pi W^{1=2})/\sigma_r(\Pi W^{1=2}) = poly(\ell, r, n)$ | WAM(Feldman et al., 2008), Algorithm 1 | Algorithm 1 |
| General cond. number | WAM (Feldman et al., 2008) | Open |

Table 1: Landscape of efficient learning algorithms

The second question finding the clusters has been addressed in (Chaudhuri et al., 2007; Chaudhuri and Rao, 2008). Chaudhuri et al. in (Chaudhuri et al., 2007) introduced an iterative clustering algorithm but their method is restricted to the case of a mixture of two product distributions with binary outputs, i.e. $r = 2$ and $\ell = 2$. Chaudhuri and Rao in Chaudhuri and Rao (2008) proposed a spectral method for general $r, \ell$. However, for the algorithm to correctly recover cluster of each sample w.h.p, the underlying mixture distribution should satisfy a certain 'spreading' condition. Moreover, the algorithm need to know the parameters characterizing the 'spread' of the distribution, which typically is not available apriori. Although it is possible to estimate the mixture distribution, once the samples are clustered, Chaudhuri et al. provides no guarantees for estimating the distribution. As is the case for the first problem, for clustering also, we provide an efficient algorithm for general $\ell, r$, under the assumption that the condition number of $\Pi W^{1=2}$ to be bounded. This condition is not directly comparable with the spreading condition assumed in previous work. Our algorithm first estimates the mixture parameters and then uses the distance based clustering method of Arora and Kannan (2001).

Our method for estimating the mixture parameters is based on the moment matching technique from Anandkumar et al. (2012a), Arora et al. (2012b). Typically, second and third (and sometimes fourth) moments of the true distribution are estimated using the given samples. Then, using the spectral decomposition of the second moment one develops certain whitening operators that reduce the

higher-order moment tensors to orthogonal tensors. Such higher order tensors are then decomposed using a power-method based method (Anandkumar et al., 2012b) to obtain the required distribution parameters.

While such a technique is generic and applies to several popular models (Hsu and Kakade, 2013; Anandkumar et al., 2012b), for many of the models the moments themselves constitute the "correct" intermediate quantity that can be used for whitening and tensor decomposition. However, because there are dependencies in the $\ell$-wise model (for example, $x_1$ to $x_\ell$ are correlated), the higher-order moments are "incomplete" versions of the intermediate quantities that we require (see (1), (2)). Hence, we need to complete these moments so as to use them for estimating distribution parameters $\Pi, W$.

Completion of the "incomplete" second moment, can be posed as a low-rank matrix completion problem where the *block-diagonal* elements are missing. For this problem, we propose an alternating minimization based method and, borrowing techniques from the recent work of Jain et al. (2013), we prove that alternating minimization is able to complete the second moment exactly. We would like to note that our alternating minimization result also solves a generalization of the low-rank+diagonal decomposition problem of Saunderson et al. (2012). Moreover, unlike trace-norm based method of Saunderson et al. (2012), which in practice is computationally expensive, our method is efficient, requires only one Singular Value Decomposition (SVD) step, and is robust to noise as well.

We reduce the completion of the "incomplete" third moment to a simple least squares problem that is robust as well. Using techniques from our second moment completion method, we can analyze an alternating minimization method also for the third moment case as well. However, for the mixture problem we can exploit the structure to reduce the problem to an efficient least squares problem with closed form solution.

Next, we present our method (see Algorithm 1) that combines the estimates from the above mentioned steps to estimate the distribution parameters $\Pi, W$ (see Theorem 2, Theorem 3). After estimating the model parameters $\Pi$, and $W$, we also show that the KL-divergence measure and the clustering error measure can also be shown to be small. In fact the excess error vanishes as the number of samples grow (see Corollary 3.1, Corollary 3.2).

## 2. Related Work

Learning mixtures of distributions is an important problem with several applications such as clustering, crowdsourcing, community detection etc. One of the most well studied problems in this domain is that of learning a mixture of Gaussians. There is a long list of interesting recent results, and discussing the literature in detail is out side of the scope of this paper. Our approach is inspired by both spectral and moment-matching based techniques that have been successfully applied in learning a mixture of Gaussians (Vempala and Wang, 2004; Arora and Kannan, 2001; Moitra and Valiant, 2010; Hsu and Kakade, 2013).

Another popular mixture distribution arises in topic models, where each word $x_i$ is selected from a $\ell$-sized dictionary. Several recent results show that such a model can also be learned efficiently using spectral as well as moments based methods (Rabani et al., 2012; Anandkumar et al., 2012a; Arora et al., 2012a). However, there is a crucial difference between the general mixture of product distribution that we consider and the topic model distribution. Given a topic (or question) $q$, each

of the words $x_i$ in the topic model have exactly the same probability. That is, $\pi^{(i)} = \pi$ for all $i \in \{1, \ldots, n\}$. In contrast, for our problem, $\pi^{(i)} \neq \pi^{(j)}, i \neq j$, in general.

Learning mixtures of discrete distribution over product spaces has several practical applications such as crowdsourcing, recommendation systems, etc. However, as discussed in the previous section, most of the existing results for this problem are designed for the case of small alphabet size $\ell$ or the number of mixture components $r$. For several practical problems (Karger et al., 2013), $\ell$ can be large and hence existing methods either do not apply or are very inefficient. In this work, we propose first provably efficient method for learning mixture of discrete distributions for general $\ell$ and $r$.

Our method is based on tensor decomposition methods for moment matching that have recently been made popular for learning mixture distributions. For example, Hsu and Kakade (2013) provided a method to learn mixture of Gaussians without any separation assumption. Similarly, Anandkumar et al. (2012a) introduced a method for learning mixture of HMMs, and also for topic models. Using similar techniques, another interesting result has been obtained for the problem of independent component analysis (ICA) (Arora et al., 2012b; Goyal and Rademacher, 2012; Hsu and Kakade, 2013).

Typically, tensor decomposition methods proceed in two steps. First, obtain a whitening operator using the second moment estimates. Then, use this whitening operator to construct a tensor with orthogonal decomposition, which reveals the true parameters of the distribution. However, in a mixture of $\ell$-way distribution that we consider, the second or the third moment do not reveal all the "required" entries, making it difficult to find the standard whitening operator. We handle this problem by posing it as a matrix completion problem and using an alternating minimization method to complete the second moment. Our proof for the alternating minimization method closely follows the analysis of Jain et al. (2013). However, Jain et al. (2013) handled a matrix completion problem where the entries are missing uniformly at random, while in our case the block diagonal elements are missing.

### 2.1. Notation

Typically, we denote a matrix or a tensor by an upper-case letter (e.g. $M$) while a vector is denoted by a small-case letter (e.g. $v$). $M_i$ denotes the $i$-th column of matrix $M$. $M_{ij}$ denotes the $(i, j)$-th entry of matrix $M$ and $M_{ijk}$ denotes the $(i, j, k)$-th entry of the third order *tensor* $M$. $A^T$ denotes the transpose of matrix $A$, i.e., $A^T_{ij} = A_{ji}$. $[k] = \{1, \ldots, k\}$ denotes the set of first $k$ integers. $\mathbf{e}_i$ denotes the $i$-th standard basis vector.

If $M \in \mathbb{R}^{\ell n \times d}$, then $M^{(m)}$ $(1 \leq m \leq n)$ denotes the $m$-th block of $M$, i.e., $(m-1)\ell + 1$ to $m\ell$-th rows of $M$. The operator $\otimes$ denotes the outer product. For example, $H = v_1 \otimes v_2 \otimes v_3$ denote a rank-one tensor such that $H_{abc} = (v_1)_a \cdot (v_2)_b \cdot (v_3)_c$. For a symmetric third-order tensor $T \in \mathbb{R}^{d \times d \times d}$, define an $r \times r \times r$ dimensional operation with respect to a matrix $R \in \mathbb{R}^{d \times r}$ as

$$T[R, R, R] \equiv \sum_{i_1, i_2, i_3 \in [d]} T_{i_1, i_2, i_3} R_{i_1, j_1} R_{i_2, j_2} R_{i_3, j_3} (e_{j_1} \otimes e_{j_2} \otimes e_{j_3}).$$

$\|A\| = \|A\|_2$ denotes the spectral norm of a tensor $A$. That is, $\|A\|_2 = \max_{x; \|x\|=1} A[x, \ldots, x]$. $\|A\|_F$ denotes the Frobenius norm of $A$, i.e., $\|A\|_F = \sqrt{\sum_{i_1, i_2, \ldots, i_p} A^2_{i_1 i_2 \ldots i_p}}$. We use $M = U \Sigma V^T$ to denote the singular value decomposition (SVD) of $M$, where $\sigma_r(M)$ denotes the $r$-th singular value of $M$. Also, wlog, assume that $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_r$.

## 3. Main results

In this section, we present our main results for estimating the mixture weights $w_q, 1 \leq q \leq r$ and the probability matrix $\Pi$ of the mixture distribution. Our estimation method is based on the moment-matching technique that has been popularized by several recent results (Anandkumar et al., 2012a; Hsu et al., 2012; Hsu and Kakade, 2013; Anandkumar et al., 2012b). However, our method differs from the existing methods in the following crucial aspects: we propose $(a)$ a matrix completion approach to estimate the second moments from samples (Algorithm 2); and $(b)$ a least squares approach with an appropriate change of basis to estimate the third moments from samples (Algorithm 3). These approaches provide robust algorithms to estimating the moments and might be of independent interest to a broad range of applications in the domain of learning mixture distributions.

The key step in our method is estimation of the following two quantities:

$$M_2 \equiv \sum_{q\in[r]} w_q \left(\pi_q \otimes \pi_q\right) = \Pi W \Pi^T \in \mathbb{R}^{\grave{n}\times\grave{n}}, \tag{1}$$

$$M_3 \equiv \sum_{q\in[r]} w_q \left(\pi_q \otimes \pi_q \otimes \pi_q\right) \in \mathbb{R}^{\grave{n}\times\grave{n}\times\grave{n}}, \tag{2}$$

where $W$ is a diagonal matrix s.t. $W_{qq} = w_q$.

Now, as is standard in the moment based methods, we exploit spectral structure of $M_2, M_3$ to recover the latent parameters $\Pi$ and $W$. The following theorem presents a method for estimating $\Pi, W$, assuming $M_2, M_3$ are estimated *exactly*:

**Theorem 1** *Let $M_2, M_3$ be as defined in (1), (2). Also, let $M_2 = U_{M_2}\Sigma_{M_2}U_{M_2}^T$ be the eigenvalue decomposition of $M_2$. Now, define $G = M_3[U_{M_2}\Sigma_{M_2}^{-1=2}, U_{M_2}\Sigma_{M_2}^{-1=2}, U_{M_2}\Sigma_{M_2}^{-1=2}]$. Let $V^G = [v_1^G\, v_2^G\, \ldots v_r^G] \in \mathbb{R}^{r\times r}$, $\lambda_q^G, 1 \leq q \leq r$ be the eigenvectors and eigenvalues obtained by the orthogonal tensor decomposition of $G$ (see (Anandkumar et al., 2012b)), i.e., $G = \sum_{q=1}^r \lambda_q^G (v_q^G \otimes v_q^G \otimes v_q^G)$. Then,*

$$\Pi = U_{M_2}\Sigma_{M_2}^{1=2} V^G \Lambda^G, \quad \textit{and} \qquad W = (\Lambda^G)^{-2},$$

*where $\Lambda^G \in \mathbb{R}^{r\times r}$ is a diagonal matrix with $\Lambda_{qq}^G = \lambda_q^G$.*

The above theorem reduces the problem of estimation of mixture parameters $\Pi, W$ to that of estimating $M_2$ and $M_3$. Typically, in moment based methods, tensors corresponding to $M_2$ and $M_3$ can be estimated directly using the second moment or third moment of the distribution, which can be estimated efficiently using the provided data samples. In our problem, however, the block-diagonal entries of $M_2$ and $M_3$ cannot be directly computed from these sample moments. For example, the expected value of a diagonal entry at $j$-th coordinate is $\mathbb{E}[xx^T]_{j;j} = \mathbb{E}[x_j] = \sum_{q\in[r]} w_q \Pi_{j;q}$, where as the corresponding entry for $M_2$ is $(M_2)_{j;j} = \sum_{q\in[r]} w_q (\Pi_{j;q})^2$.

To recover these unknown $\ell \times \ell$ block-diagonal entries of $M_2$, we use an alternating minimization algorithm. Our algorithm writes $M_2$ in a bi-linear form and solves for each factor of the bi-linear form using the computed off-diagonal blocks of $M_2$. We then prove that this algorithm exactly recovers the missing entries when we are given the exact second moment. For estimating $M_3$, we reduce the problem of estimating unknown block-diagonal entries of $M_3$ to a least squares problem that can be solved efficiently.

Concretely, to get a consistent estimate of $M_2$, we pose it as a matrix completion problem, where we use the off-block-diagonal entries of the second moment, which we know are consistent, to estimate the missing entries. Precisely, let

$$\Omega_2 \equiv \Big\{ (i,j) \subseteq [\ell n] \times [\ell n] \,|\, \lceil \frac{i}{\ell} \rceil \neq \lceil \frac{j}{\ell} \rceil \Big\},$$

be the indices of the off-block-diagonal entries, and define a masking operator as:

$$\mathcal{P}_{\Omega_2}(A)_{i;j} \equiv \begin{cases} A_{i;j} \,, & \text{if } (i,j) \in \Omega_2 \,, \\ 0 \,, & \text{otherwise} \,. \end{cases} \tag{3}$$

Now, using the fact that $M_2$ has rank at most $r$, we find a rank-$r$ estimate that explains the off-block-diagonal entries using an alternating minimization algorithm defined in Section 4.

$$\widehat{M_2} \equiv \textsc{MatrixAltMin}\left( \frac{2}{|\mathcal{S}|} \sum_{t \in [|\mathcal{S}|=2]} x_t x_t^T, \Omega_2, r, T \right) \,, \tag{4}$$

where $\{x_1, \ldots, x_{|\mathcal{S}|}\}$ is the set of observed samples, and $T$ is the number of iterations. We use the first half of the samples to estimate $M_2$ and the rest to estimate the third-order tensor.

Similarly for the tensor $M_3$, the sample third moment does not converge to $M_3$. However, the off-block diagonal entries do converge to the corresponding entries of $M_3$. That is, let

$$\Omega_3 \equiv \Big\{ (i,j,k) \subseteq [\ell n] \times [\ell n] \times [\ell n] \,|\, \lceil \frac{i}{\ell} \rceil \neq \lceil \frac{j}{\ell} \rceil \neq \lceil \frac{k}{\ell} \rceil \neq \lceil \frac{i}{\ell} \rceil \Big\},$$

be the indices of the off-block-diagonal entries, and define the following masking operator:

$$\mathcal{P}_{\Omega_3}(A)_{i;j;k} \equiv \begin{cases} A_{i;j;k} \,, & \text{if } (i,j,k) \in \Omega_3 \,, \\ 0 \,, & \text{otherwise} \,. \end{cases} \tag{5}$$

Then, we have consistent estimates for $\mathcal{P}_{\Omega_3}(M_3)$ from the sample third moment.

Now, in the case of $M_3$, we do not explicitly compute $M_3$. Instead, we estimate a $r \times r \times r$ dimensional tensor $\widetilde{G} \equiv M_3[\widehat{U}_{M_2} \widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2} \widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2} \widehat{\Sigma}_{M_2}^{-1=2}]$ (cf. Theorem 1), using a least squares formulation that uses only off-diagonal blocks of $P_\Omega(M_3)$. That is,

$$\widehat{G} \equiv \textsc{TensorLS}\Big( \frac{2}{|\mathcal{S}|} \sum_{t=1+|\mathcal{S}|=2}^{|\mathcal{S}|} x_t \otimes x_t \otimes x_t, \Omega_3, \widehat{U}_{M_2}, \widehat{\Sigma}_{M_2} \Big) \,,$$

where $\widehat{M_2} = \widehat{U}_{M_2} \widehat{S}_{M_2} \widehat{U}_{M_2}^T$ is the singular value decomposition of the rank-$r$ matrix $\widehat{M_2}$. After estimation of $\widehat{G}$, similar to Theorem 1, we use the whitening and tensor decomposition to estimate $\Pi, W$. See Algorithm 1 for a pseudo-code of our approach.

**Remark**: Note that we use a new set of $|\mathcal{S}|/2$ samples to estimate the third moment. This sub-sampling helps us in our analysis, as it ensures independence of the samples $x_{|\mathcal{S}|=2+1}, \ldots, x_{|\mathcal{S}|}$ from the output of the alternating minimization step (4).

The next theorem shows that the moment matching approach (Algorithm 1) is consistent. Let $\widehat{W} = \text{diag}([\hat{w}_1, \ldots, \hat{w}_r])$ and $\widehat{\Pi} = [\widehat{\pi}_1, \ldots, \widehat{\pi}_r]$ denote the estimates obtained using Algorithm 1. Also, let $\mu$ denote the block-incoherence of $M_2 = \Pi W \Pi^T$ as defined in (7).

**Algorithm 1 Spectral-Dist:** Moment method for Mixture of Discrete Distribution

1: Input: Samples $\{x_t\}_{t\in\mathcal{S}}$
2: $\widehat{M}_2 \leftarrow \textsc{MatrixAltMin}\left(\left(\frac{2}{|\mathcal{S}|}\sum_{t\in[|\mathcal{S}|=2]} x_t x_t^T\right), \Omega_2, r, T\right)$ (see Algorithm 2)
3: Compute eigenvalue decomposition of $\widehat{M}_2 = \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}\widehat{U}_{M_2}^T$
4: $\widehat{G} \leftarrow \textsc{TensorLS}\left(\left(\frac{2}{|\mathcal{S}|}\sum_{t=|\mathcal{S}|=2+1}^{|\mathcal{S}|} x_t\otimes x_t\otimes x_t\right), \Omega_3, \widehat{U}_{M_2}, \widehat{\Sigma}_{M_2}\right)$ (see Algorithm 3)
5: Compute a rank-$r$ orthogonal tensor decomposition $\sum_{q\in[r]}\widehat{\lambda}_q^G(\hat{v}_q^G\otimes\hat{v}_q^G\otimes\hat{v}_q^G)$ of $\widehat{G}$, using Robust Power-method of (Anandkumar et al., 2012b)
6: Output: $\widehat{\Pi} = \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}\widehat{V}^G\widehat{\Lambda}^G$, $\widehat{W} = (\widehat{\Lambda}_q^G)^{-2}$, where $(\widehat{V}^G)^T = [\hat{v}_1^G\ \dots\ \hat{v}_r^G]$

**Theorem 2** *Assume that the sample second and the third moments are exact, i.e.,* $\mathcal{P}_{\Omega_2}(\frac{2}{|\mathcal{S}|}\sum_{t\in[|\mathcal{S}|=2]} x_t x_t^T) = \mathcal{P}_{\Omega_2}(M_2)$ *and* $\mathcal{P}_{\Omega_3}(\frac{2}{|\mathcal{S}|}\sum_{t=|\mathcal{S}|=2+1}^{|\mathcal{S}|} x_t\otimes x_t\otimes x_t) = \mathcal{P}_{\Omega_3}(M_3)$. *Also, let* $T=\infty$ *for the* MATRIXALTMIN*procedure and let* $n \geq C\,\sigma_1(M_2)^5\mu^5 r^{3:5}/\sigma_r(M_2)^5$, *for a global constant* $C>0$. *Then, there exists a permutation* $P$ *over* $[r]$ *such that, for all* $q\in[r]$,

$$\pi_q = \widehat{\pi}_{P(q)} \quad \textit{and} \quad w_q = \hat{w}_{P(q)}\ .$$

We now provide a finite sample version of the above theorem.

**Theorem 3 (Finite sample bound)** *There exists positive constants* $C_0$, $C_1$, $C_2$, $C_3$ *and a permutation* $P$ *on* $[r]$ *such that if* $n \geq C_0\,\sigma_1(M_2)^{4:5}\mu^4 r^{3:5}/\sigma_r(M_2)^{4:5}$ *then for any* $\varepsilon_M \leq \frac{C_1}{\sqrt{r+}}$ *and for a large enough sample size:*

$$|\mathcal{S}| \quad \geq \quad C_2\,\frac{\mu^6\,r^6}{w_{\min}}\,\frac{\sigma_1(M_2)^6 n^3}{\sigma_r(M_2)^9}\,\frac{\log(n/\delta)}{\varepsilon_M^2}\ ,$$

*the following holds for all* $q\in[r]$, *with probability at least* $1-\delta$:

$$|\hat{w}_{P(q)} - w_q| \quad \leq \quad \varepsilon_M\ ,$$

$$\|\widehat{\pi}_{P(q)} - \pi_q\| \quad \leq \quad \varepsilon_M\sqrt{\frac{r\,w_{\max}\,\sigma_1(M_2)}{w_{min}}}\ .$$

*Further, Algorithm 1 runs in time* $\mathrm{poly}\big(n,\ell,r,1/\varepsilon,\log(1/\delta),1/w_{\min},\sigma_1(M_2)/\sigma_r(M_2)\big)$.

Note that, the estimated $\hat{\pi}_i$'s and $\hat{w}_i$'s using Algorithm 1 do not necessarily define a *valid* probability measure: they can take negative values and might not sum to one. We can process the estimates further to get a valid probability distribution, and show that the estimated mixture distribution is close in Kullback-Leibler divergence to the original one. Let $\varepsilon_w = C_3\varepsilon_M/\sqrt{w_{\min}}$. We first set

$$\tilde{w}'_q \quad = \quad \begin{cases} \hat{w}_q & \text{if } \hat{w}_q \geq \varepsilon_w\ , \\ \varepsilon_w & \text{if } \hat{w}_q < \varepsilon_w\ , \end{cases}$$

and set mixture weights $\tilde{w}_q = \tilde{w}'_q/\sum_{q'}\tilde{w}'_{q'}$. Similarly, let $\varepsilon \ = C_3\varepsilon_M\sqrt{\frac{_1(M_2)\,r(1+"_M\ \ _r(M_2))}{w_{min}}}$ and set

$$\tilde{\pi}'^{(j)}_{q;p} \quad = \quad \begin{cases} \hat{\pi}^{(j)}_{q;p} & \text{if } \hat{\pi}^{(j)}_{q;p} \geq \varepsilon \quad , \\ \varepsilon & \text{if } \hat{\pi}^{(j)}_{q;p} < \varepsilon \quad , \end{cases}$$

**Algorithm 2** MATRIXALTMIN: Alternating Minimization for Matrix Completion

1: Input: $S_2 = \frac{2}{|\mathcal{S}|}\sum_{t\in\{1,\ldots,|\mathcal{S}|/2\}} x_t x_t^T$, $\Omega_2$, $r$, $T$
2: Initialize $\ell n \times r$ dimensional matrix $U_0 \leftarrow$ top-$r$ eigenvectors of $\mathcal{P}_{\Omega_2}(S_2)$
3: **for all** $\tau = 1$ to $T-1$ **do**
4: $\quad \widehat{U}_{\tau+1} = \arg\min_U \|\mathcal{P}_{\Omega_2}(S_2) - \mathcal{P}_{\Omega_2}(UU^T)\|_F^2$
5: $\quad [U_{\tau+1} R_{\tau+1}] = \mathrm{QR}(\widehat{U}_{t+1})$ (standard QR decomposition)
6: **end for**
7: Output: $\widehat{M}_2 = (\widehat{U}_T)(U_{T-1})^T$

**Algorithm 3** TENSORLS: Least Squares method for Tensor Estimation

1: Input: $S_3 = \frac{2}{|\mathcal{S}|}\sum_{t\in\{|\mathcal{S}|/2+1,\ldots,|\mathcal{S}|\}}(x_t \otimes x_t \otimes x_t)$, $\Omega_3$, $\widehat{U}_{M_2}$, $\widehat{\Sigma}_{M_2}$
2: Define operator $\widehat{\nu} : \mathbb{R}^{r\times r\times r} \to \mathbb{R}^{\ell n\times \ell n\times \ell n}$ as follows

$$\widehat{\nu}_{ijk}(Z) = \begin{cases} \sum_{abc} Z_{abc}(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1/2})_{ia}(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1/2})_{jb}(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1/2})_{kc}, & \text{if } \lceil \frac{i}{\ell}\rceil \neq \lceil \frac{j}{\ell}\rceil \neq \lceil \frac{k}{\ell}\rceil \neq \lceil \frac{i}{\ell}\rceil, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

3: Define $\widehat{A} : \mathbb{R}^{r\times r\times r} \to \mathbb{R}^{r\times r\times r}$ s.t. $\widehat{A}(Z) = \widehat{\nu}(Z)[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1/2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1/2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1/2}]$
4: Output: $\widehat{G} = \arg\min_Z \|\widehat{A}(Z) - \mathcal{P}_{\Omega_3}(S_3)[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1/2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1/2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1/2}]\|_F^2$

for all $q \in [r]$, $p \in [\ell]$, and $j \in [n]$, and normalize it to get valid distributions $\tilde{\pi}_{q;p}^{(j)} = \tilde{\pi}_{q;p}^{\prime(j)} / \sum_{p'} \tilde{\pi}_{q;p'}^{\prime(j)}$. Let $\widehat{X}$ denote a random vector in $\{0,1\}^{\ell n}$ obtained by first selecting a random type $q$ with probability $\tilde{w}_q$ and then drawing from a random vector according to $\tilde{\pi}_q$.

**Corollary 3.1 (KL-divergence bound)** *Under the hypotheses of Theorem 3, there exists a positive constant $C$ such that if $|\mathcal{S}| \geq Cn^7 r^7 \mu^6 \sigma_1(M_2)^7 \ell^{12} w_{\max} \log(n/\delta)/(\sigma_r(M_2)^9 \eta^6 w_{\min}^2)$, then Algorithm 1 with the above post-processing produces a $r$-mixture distribution $\widehat{X}$ that, with probability at least $1-\delta$, satisfies : $D_{KL}(X||\widehat{X}) \leq \eta$.*

Moreover, we can show that the "type" of each data point can also be recovered accurately.

**Corollary 3.2 (Clustering bound)** *Define:*

$$\tilde{\varepsilon} \equiv \max_{i,j\in[r]} \left\{ \frac{\|\pi_i - \pi_j\|^2 - 2\|\Pi\|_F\sqrt{2\log(r/\delta)}}{(\|\pi_i - \pi_j\| + 2\sqrt{2\log(r/\delta)})r^{1/2}} \right\}.$$

*Under the hypotheses of Theorem 3, there exists a positive numerical constant $C$ such that if $\tilde{\varepsilon} > 0$ and $|\mathcal{S}| \geq C\mu^6 r^7 n^3 \sigma_1(M_2)^7 w_{\max} \log(n/\delta)/(w_{\min}^2 \sigma_r(M_2)^9 \tilde{\varepsilon}^2)$, then with probability at least $1-\delta$, the distance based clustering algorithm of (Arora and Kannan, 2001) computes a correct clustering of the samples.*

## 4. Algorithm

In this section, we describe the proposed approach in detail and provide finite sample performance guarantees for each components: MATRIXALTMIN and TENSORLS. These results are crucial in proving the finite sample bound in Theorem 3. As mentioned in the previous section, the algorithm first estimates $M_2$ using the alternating minimization procedure. Recall that the second moment of the data given by $S_2$ cannot estimate the block-diagonal entries of $M_2$. That is, even in the case of infinite samples, we only have consistency in the off-block-diagonal entries: $\mathcal{P}_{\Omega_2}(S_2) = \mathcal{P}_{\Omega_2}(M_2)$. However, to apply the "whitening" operator to the third order tensor (see Theorem 1) we need to estimate $M_2$.

In general it is not possible to estimate $M_2$ from $\mathcal{P}_{\Omega_2}(M_2)$ as one can fill any entries in the block-diagonal entries. Fortunately, we can avoid such a case since $M_2$ is guaranteed to be of rank $r \ll \ell n$. However, even a low-rank assumption is not enough to recover back $M_2$. For example, if $M_2 = \mathbf{e}_1\mathbf{e}_1^T$, then $\mathcal{P}_{\Omega_2}(M_2) = 0$ and one cannot recover back $M_2$. Hence, we make an additional standard assumption that $M_2$ is $\mu$-block-incoherent, where a symmetric rank-$r$ matrix $A$ with singular value decomposition $A = USV^T$ is $\mu$-block-incoherent if the operator norm of all $\ell \times r$ blocks of $U$ are upper bounded by

$$\left\|U^{(i)}\right\|_2 \;\leq\; \mu\sqrt{\frac{r}{n}}\;, \text{ for all } i \in [n]\;, \tag{7}$$

where $U^{(i)}$ is an $\ell \times r$ sub matrix of $U$ which is defined by the block from the $((i-1)\ell+1)$-th row to the $(i\ell)$-th row. For a given matrix $M$, the smallest value of $\mu$ that satisfy the above condition is referred to as the block-incoherence of $M$.

Now, assuming that $M_2$ satisfies two assumptions, $r \ll \ell n$ and $M_2$ is $\mu$-block incoherent, we provide an alternating minimization method that provably recovers $M_2$. In particular, we model $M_2$ explicitly using a bi-linear form $M_2 = \widehat{U}^{(t+1)}(U^{(t)})^T$ with variables $\widehat{U}^{(t+1)} \in \mathbb{R}^{\grave{}n\times r}$ and $U^{(t)} \in \mathbb{R}^{\grave{}n\times r}$. We iteratively solve for $\widehat{U}^{(t+1)}$ for fixed $U^{(t)}$, and use QR decomposition to orthonormalize $\widehat{U}^{(t+1)}$ to get $U^{(t+1)}$. Note that the QR-decomposition is *not required* for our method but we use it only for ease of analysis. Below, we give the precise recovery guarantee for the alternating minimization method (Algorithm 2).

**Theorem 4 (Matrix completion using alternating minimization)** *For an $\ell n \times \ell n$ symmetric rank-$r$ matrix $M$ with block-incoherence $\mu$, we observe off-block-diagonal entries corrupted by noise:*

$$\widehat{M}_{ij} \;=\; \begin{cases} M_{ij} + E_{ij} & \text{if } \lceil \frac{i}{\ell} \rceil \neq \lceil \frac{j}{\ell} \rceil\;, \\ 0 & \text{otherwise.} \end{cases}$$

*Let $\widehat{M}^{(\tau)}$ denote the output after $\tau$ iterations of* MATRIXALTMIN. *If $\mu \leq (\sigma_r(M)/\sigma_1(M))\sqrt{n/(32\, r^{1.5})}$, the noise is bounded by $\|\mathcal{P}_{\Omega_2}(E)\|_2 \leq \sigma_r(M)/32\sqrt{r}$, and each column of the noise is bounded by $\|\mathcal{P}_{\Omega_2}(E)_i\| \leq \sigma_1(M)\mu\sqrt{3r/(8\, n\, \ell)}$, $\forall i \in n\ell$, then after $\tau \geq (1/2)\log\left(2\|M\|_F/\varepsilon\right)$ iterations of* MATRIXALTMIN, *the estimate $\widehat{M}^{(\tau)}$ satisfies:*

$$\|M - \widehat{M}^{(\tau)}\|_2 \;\leq\; \varepsilon + \frac{9\,\|M\|_F\,\sqrt{r}}{\sigma_r(M)}\|\mathcal{P}_{\Omega_2}(E)\|_2\;,$$

*for any $\varepsilon \in (0,1)$. Further, $\widehat{M}^{(\tau)}$ is $\mu_1$-incoherent with $\mu_1 = 6\mu\sigma_1(M_2)/\sigma_r(M_2)$.*

For estimating $M_2$, the noise $E$ in the off-block-diagonal entries are due to insufficient sample size. We can precisely bound how large the sampling noise is in the following lemma.

**Lemma 5** *Let* $S_2 = \frac{2}{|\mathcal{S}|}\sum_{t\in\{1,:::,|\mathcal{S}|=2\}} x_t x_t^T$ *be the sample co-variance matrix. Also, let* $E = \|\mathcal{P}_{\Omega_2}(S_2) - \mathcal{P}_{\Omega_2}(M_2)\|_2$. *Then,*

$$\|E\|_2 \;\leq\; 8\sqrt{\frac{n^2\log(n\ell/\delta)}{|\mathcal{S}|}}.$$

*Moreover,* $\|E_i\|_2 \leq 8\sqrt{n\log(1/\delta)/|\mathcal{S}|}$, *for all* $i \in [n\ell]$.

The above theorem shows that $M_2$ can be recovered exactly from infinite many samples, if $n \geq \frac{^2{}_1(M)^2 r^{1.5}}{r(M)^2}$. Furthermore, using Lemma 5, $M_2$ can be recovered approximately, with sample size $|\mathcal{S}| = O(n^2(\ell+r)/\sigma_r(M)^2)$. Now, recovering $M_2 = \Pi W \Pi^T$ recovers the left-singular space of $\Pi$, i.e., range($U$). However, we still need to recover $W$ and the right-singular space of $\Pi$, i.e., range($V$).

To this end, we can estimate the tensor $M_3$, "whiten" the tensor using $\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}$ (recall that, $\widehat{M_2} = \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}\widehat{U}_{M_2}^T$), and then use tensor decomposition techniques to solve for $V, W$. However, we show that estimating $M_3$ is not necessary, we can directly estimate the "whitened" tensor by solving a system of linear equations. In particular, we design an operator $\widehat{A} : \mathbb{R}^{r\times r\times r} \to \mathbb{R}^{r\times r\times r}$ such that $\widehat{A}(\widetilde{G}) \approx \mathcal{P}_{\Omega_3}(S_3)[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}]$, where

$$\widetilde{G} \;\equiv\; \sum_{q\in[r]} \frac{1}{\sqrt{w_q}}(R_3\mathbf{e}_q \otimes R_3\mathbf{e}_q \otimes R_3\mathbf{e}_q), \text{ and } \quad R_3 \;\equiv\; \widehat{\Sigma}_{M_2}^{-1=2}\widehat{U}_{M_2}^T \Pi W^{1=2}. \tag{8}$$

Moreover, we show that $\widehat{A}$ is nearly-isometric. Hence, we can efficiently estimate $\widetilde{G}$, using the following system of equations:

$$\widehat{G} \;=\; \arg\min_Z \|\widehat{A}(Z) - \mathcal{P}_{\Omega_3}(S_3)[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}]\|_F^2. \tag{9}$$

Let $\mu$ and $\mu_1$ denote the block-incoherence of $M_2$ and $\widehat{M_2}$ respectively, as defined in (7).

**Theorem 6** *Let* $\widetilde{G}$, $\widehat{G}$ *be as defined in* (8), (9), *respectively. If* $n \geq 144r^3\sigma_1(M_2)^2/\sigma_r(M_2)^2$, *then the following holds with probability at least* $1-\delta$:

$$\|\widehat{G} - \widetilde{G}\|_F \;\leq\; \frac{24\mu_1^3\mu r^{3:5}\sigma_1(M_2)^{3=2}}{n\sqrt{w_{\min}}\sigma_r(M_2)^{3=2}}\varepsilon_{M_2} + 2\left\|\mathcal{P}_{\Omega_3}(M_3 - S_3)[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}]\right\|_F,$$

*for* $\varepsilon_{M_2} \equiv (1/\sigma_r(M_2))\|\widehat{M_2} - M_2\|_2$.

We can also prove a bound on the sampling noise for the third order tensor in the following lemma.

**Lemma 7** *Let* $S_3 = \frac{2}{|\mathcal{S}|}\sum_{t\in\{|\mathcal{S}|=2+1,:::,|\mathcal{S}|\}}(x_t \otimes x_t \otimes x_t)$. *Then, there exists a positive numerical constant* $C$ *such that, with probability at least* $1-\delta$,

$$\left\|\mathcal{P}_{\Omega_3}(M_3 - S_3)[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}]\right\|_F \;\leq\; \frac{C\, r^3\, \mu_1^3\, n^{3=2}}{\sigma_r(M_2)^{3=2}}\sqrt{\frac{\log(1/\delta)}{|\mathcal{S}|}}\,.$$

Next, we apply the tensor decomposition method of (Anandkumar et al., 2012b) to decompose obtained tensor, $\widehat{G}$, and obtain $\widehat{R}_3, \widehat{W}$ that approximates $R_3$ and $W$. We then use the obtained estimate $\widehat{R}_3, \widehat{W}$ to estimate $\Pi$; see Algorithm 1 for the details. In particular, using Theorem 4 and Theorem 6, Algorithm 1 provides the following estimate for $\Pi$:

$$\widehat{\Pi} = \widehat{U}_{M_2} \widehat{\Sigma}_{M_2}^{1=2} \widehat{R}_3 \widehat{W}^{-1=2} \approx \widehat{U}_{M_2} \widehat{U}_{M_2}^{T} \Pi.$$

Now, $\|\widehat{\Pi} - \Pi\|_2$ can be bounded by using the above equation along with the fact that $\mathrm{range}(\widehat{U}_{M_2}) \approx \mathrm{range}(\Pi)$. See Section A.6 for a detailed proof.

## 5. Applications in Crowdsourcing

Crowdsourcing has emerged as an effective paradigm for solving large-scale data-processing tasks in domains where humans have an advantage over computers. Examples include image classification, video annotation, data entry, optical character recognition, and translation. For tasks with discrete choice outputs, one of the most widely used model is the Dawid-Skene model introduced in Dawid and Skene (1979): each expert $j$ is modeled through a $r \times r$ *confusion matrix* $\pi^{(j)}$ where $\pi_{pq}^{(j)}$ is the probability that the expert answers $q$ when the true label is $p$. This model was developed to study how different clinicians give different diagnosis, even when they are presented with the same medical chart. This is a special case, with $\ell = r$, of the mixture model studied in this paper.

Historically, a greedy algorithm based on Expectation-Maximization has been widely used for inference (Dawid and Skene, 1979; Smyth et al., 1995; Hui and Zhou, 1998; Sheng et al., 2008), but with no understanding of how the performance changes with the problem parameters and sample size. Recently, spectral approaches were proposed and analyzed with provable guarantees. For a simple case when there are only two labels, i.e. $r = \ell = 2$, Ghosh et al. in Ghosh et al. (2011) and Karger et al. in Karger et al. (2011b) analyzed a spectral approach of using the top singular vector for clustering under Dawid-Skene model. The model studied in these work is a special case of our model with $r = \ell = 2$ and $w = [1/2, 1/2]$, and $\pi^{(j)} = \begin{bmatrix} p_j & 1-p_j \\ 1-p_j & p_j \end{bmatrix}$. Let $q = (1/n) \sum_{j \in [n]} 2(p_j - 1)^2$, then it follows that $\sigma_1(M_2) = (1/2)n$ and $\sigma_2(M_2) = (1/2)nq$. It was proved in Ghosh et al. (2011); Karger et al. (2011b) that if we project each data point $x_i$ onto the second singular vector of $S_2$ the empirical second moment, and make a decision based on the sign of this projection, we get good estimates with the probability of misclassification scales as $O(1/\sigma_r(M_2))$.

More recently, Karger et al. in Karger et al. (2011a) proposed a new approach based on a message-passing algorithm for computing the top singular vectors, and improved this misclassification bound to an exponentially decaying $O(e^{-C\ r(M_2)})$ for some positive numerical constant $C$. However, these approaches highly rely on the fact that there are only two ground truth labels, and the algorithm and analysis cannot be generalized. These spectral approaches has been extended to general $r$ in Karger et al. (2013) with misclassification probability scaling as $O(r/\sigma_r(M_2))$, but this approach still uses the existing binary classification algorithms as a black box and tries to solve a series of binary classification tasks.

Furthermore, existing spectral approaches use $S_2$ directly for inference. This is not consistent, since even if infinite number of samples are provided, this empirical second moment does not converge to $M_2$. Instead, we use recent developments in matrix completion to recover $M_2$ from

samples, thus providing a consistent estimator. Hence, we provide a robust clustering algorithm for crowdsourcing and provide estimates for the mixture distribution with provable guarantees. Corollary 3.2 shows that with large enough samples, the misclassification probability of our approach scales as $O(re^{-C(r\ \ r(M_2)^2=n)})$ for some positive constant $C$. This is an exponential decay and is a significant improvement over the known error bound of $O(r/\sigma_r(M_2))$.

## 6. Conclusion

We presented a method for learning a mixture of $\ell$-wise discrete distribution with distribution parameters $\Pi, W$. Our method shows that assuming $n \geq Cr^3\kappa^{4.5}$ and the number of samples to be $|S| \geq C_1(n\, r^7\, \kappa^9 \log(n/\delta))/(w_{min}^2 \varepsilon_\Pi^2)$, we have $\|\widehat{\Pi} - \Pi\|_2 \leq \varepsilon_\Pi$ where $\kappa = \sigma_1(M_2)/\sigma_r(M_2)$, and $M_2 = \Pi W \Pi^T$.

Note that our algorithm does not require any separability condition on the distribution, is consistent for infinite samples, and is robust to noise as well. That is, our analysis can be easily extended to the noisy case, where there is a small amount of noise in each sample.

Our sample complexity bounds include the condition number of the distribution $\kappa$ which implies that our method requires $\kappa$ to be at most $poly(\ell, r)$. This makes our method unsuitable for the problem of learning Boolean functions (Feldman et al., 2008). However, it is not clear if is possible to design an efficient algorithm with sample complexity independent of the condition number. We leave further study of the dependence of sample complexity on the condition number as a topic for future research.

Another drawback of our method is that $n$ is required to be $n = \Omega(r^3)$. We believe that this condition is natural, as one cannot recover the distribution for $n = 1$. However, establishing tight information theoretic lower bound on $n$ (w.r.t. $\ell, r$) is still an open problem.

For the crowdsourcing application, the current error bound for clustering translates into $O(e^{-Cnq^2})$ when $r = 2$. This is not as strong as the best known error bound of $O(e^{-Cnq})$, since $q$ is always less than one. The current analysis and algorithm for clustering needs to be improved to get an error bound of $O(re^{-Cr\ \ r(M_2)})$ for general $r$ such that it gives optimal error rate for the special case of $r = 2$.

The sample complexity also depends on $1/w_{\min}$, which we believe is unnecessary. If there is a component with small mixing weight, we should be able to ignore such component smaller than the sample noise level and still guarantee the same level accuracy. To this end, we need an adaptive algorithm that detects the number of components that are non-trivial and this is a subject of future research.

More fundamentally, all of the moment matching methods based on the spectral decompositions suffer from the same restrictions. It is required that the underlying tensors have rank equal to the number of components, and the condition number needs to be small. However, the problem itself is not necessarily more difficult when the condition number is larger.

Finally, we believe that our technique of completion of the second and the higher order moments should have application to several other mixture models that involve $\ell$-wise distributions, e.g., mixed membership stochastic block model with $\ell$-wise connections between nodes.

## References

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *Learning Theory*, pages 458–469. Springer, 2005.

A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. *arXiv preprint arXiv:1203.0683*, 2012a.

Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012b.

S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *STOC*, pages 247–257, 2001.

Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models - going beyond SVD. In *FOCS*, pages 1–10, 2012a.

Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ICA with unknown Gaussian noise, with implications for Gaussian mixtures and autoencoders. In *NIPS*, pages 2384–2392, 2012b.

K. Chaudhuri and S. Rao. Learning mixtures of product distributions using correlations and independence. In *COLT*, pages 9–20, 2008.

Kamalika Chaudhuri, Eran Halperin, Satish Rao, and Shuheng Zhou. A rigorous analysis of population stratification with limited data. In *SODA*, pages 1046–1055, 2007.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

J. Feldman, R. O'Donnell, and R. A Servedio. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008.

Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *COLT*, pages 53–62, 1999.

A. Ghosh, S. Kale, and P. McAfee. Who moderates the moderators?: crowdsourcing abuse detection in user-generated content. In *EC*, pages 167–176, 2011.

Navin Goyal and Luis Rademacher. Efficient learning of simplices. *CoRR*, abs/1211.2227, 2012.

Suriya Gunasekar, Ayan Acharya, Neeraj Gaur, and Joydeep Ghosh. Noisy matrix completion using alternating minimization. In *Machine Learning and Knowledge Discovery in Databases*, pages 194–209. Springer, 2013.

D. Hsu and S. M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *ITCS*, pages 11–20, 2013.

D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Siu L Hui and Xiao H Zhou. Evaluation of diagnostic tests without gold standards. *Statistical methods in medical research*, 7(4):354–370, 1998.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, pages 665–674, 2013.

D. R. Karger, S. Oh, and D. Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *arXiv preprint arXiv:1110.3564*, 2011a.

D. R. Karger, S. Oh, and D. Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Allerton*, 2011b.

D. R. Karger, S. Oh, and D. Shah. Efficient crowdsourcing for multi-class labeling. In *Proceedings of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems*, pages 81–92, 2013.

M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie. On the learnability of discrete distributions. In *STOC*, pages 273–282, 1994.

Frank McSherry. Spectral partitioning of random graphs. In *FOCS*, pages 529–537, 2001.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. Learning mixtures of arbitrary distributions over large discrete domains. *CoRR*, abs/1212.1527, 2012.

James Saunderson, Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM J. Matrix Analysis Applications*, 33(4):1395–1416, 2012.

V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD*, pages 614–622, 2008.

P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. In *NIPS*, pages 1085–1092, 1995.

S. Sridhar, S. Rao, and E. Halperin. An efficient and accurate graph-based approach to detect population substructure. In *Research in Computational Molecular Biology*, pages 503–517, 2007.

Dan-Cristian Tomozei and Laurent Massoulié. Distributed user profiling via spectral methods. In *ACM SIGMETRICS Performance Evaluation Review*, volume 38, pages 383–384, 2010.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci.*, 68(4):841–860, 2004.

# Appendix

## Appendix A. Proofs

In this section, we give detailed proofs for all the key theorems/lemmata that we require to prove our main result (Theorem 4, Theorem 6).

### A.1. Proof of Theorem 4

We analyze each iteration and show that we get closer to the optimal solution up to a certain noise level at each step. To make the block structures explicit, we use index $(i, a)$ for some $i \in [n]$ and $a \in [\ell]$ to denote $(i-1)\ell + a \in [\ell n]$. The least squares update gives:

$$U^{(t+1)} \;=\; \arg\min_{V \in \mathbb{R}^{\ell n \times \ell n}} \sum_{i;j \in [n];a;b \in [\grave{}];i \neq j} \Big( \widehat{M}_{(i;a);(j;b)} - \big( V (\widehat{U}^{(t)})^{T} \big)_{(i;a);(j;b)} \Big)^2 .$$

Setting the gradient to zero, we get:

$$-2 \sum_{j \neq i; b \in [\grave{}]} \Big( M_{(i;a);(j;b)} + E_{(i;a);(j;b)} - \big\langle U^{(t+1)}_{(i;a)}, \widehat{U}^{(t)}_{(j;b)} \big\rangle \Big) \widehat{U}^{(t)}_{(j;b)} \;=\; 0 \,,$$

for all $i \in [n]$ and $a \in [\ell]$. Here, $U^{(t)}_{(j;b)}$ is a $r$-dimensional column vector representing the $((j-1)\ell + b)$-th row of $U^{(t)}$. Let $M = USU^{T}$ be the singular value decomposition of $M$. The $r$-dimensional column vector $U^{(t+1)}_{(i;a)}$ can be written as:

$$\begin{aligned} U^{(t+1)}_{(i;a)} \;&=\; (B^{(i;a)})^{-1} C^{(i;a)} S U_{(i;a)} \;+\; (B^{(i;a)})^{-1} N_{(i;a)} \\ &=\; \underbrace{D S U_{(i;a)}}_{\text{power iteration}} \;-\; \underbrace{(B^{(i;a)})^{-1} \big( B^{(i;a)} D \;-\; C^{(i;a)} \big) S U_{(i;a)}}_{\text{error due to missing entries}} \;+\; \underbrace{(B^{(i;a)})^{-1} N_{(i;a)}}_{\text{error due to noise}} \,, \end{aligned} \tag{10}$$

where,

$$\begin{aligned} B^{(i;a)} \;&=\; \sum_{j \neq i; j \in [n]; b \in [\grave{}]} \widehat{U}^{(t)}_{(j;b)} (\widehat{U}^{(t)})^{T}_{(j;b)} \;\in\; \mathbb{R}^{r \times r} \\ C^{(i;a)} \;&=\; \sum_{j \neq i; j \in [n]; b \in [\grave{}]} \widehat{U}^{(t)}_{(j;b)} U^{T}_{(j;b)} \;\in\; \mathbb{R}^{r \times r} \\ D \;&=\; \sum_{j \in [n]; b \in [\grave{}]} \widehat{U}^{(t)}_{(j;b)} U^{T}_{(j;b)} \;\in\; \mathbb{R}^{r \times r} \\ N_{(i;a)} \;&=\; \sum_{j \neq i; j \in [n]; b \in [\grave{}]} E_{(i;a);(j;b)} \widehat{U}^{(t)}_{(j;b)} \;\in\; \mathbb{R}^{r \times 1} \,. \end{aligned}$$

Note that, the above quantities are independent of index $a$, but we carry the index for uniformity of notation.

In a matrix form of dimension $\ell n \times r$, we use $F_{\text{miss}} \in \mathbb{R}^{\grave{} n \times r}$ to denote the error due to missing entries and $F_{\text{noise}} \in \mathbb{R}^{\grave{} n \times r}$ to denote the error due to the noise such that

$$\begin{aligned} U^{(t+1)} \;&=\; M \widehat{U}^{(t)} - F^{(t+1)}_{\text{miss}} + F^{(t+1)}_{\text{noise}} \,, \text{ and} \\ \widehat{U}^{(t+1)} \;&=\; \Big( M \widehat{U}^{(t)} - F^{(t+1)}_{\text{miss}} + F^{(t+1)}_{\text{noise}} \Big) \big( R^{(t+1)}_{U} \big)^{-1} \,, \end{aligned} \tag{11}$$

where we define $R_U^{(t+1)}$ to be the upper triangular matrix obtained by QR decomposition of $U^{(t+1)} = \widehat{U}^{(t+1)} R_U^{(t+1)}$. The explicit formula for $F_{\text{miss}}$ and $F_{\text{noise}}$ is given in (14) and (18). Then, the error after $t$ iterations of the alternating minimization is bounded by

$$\left\| M - \widehat{U}^{(t)} \left(U^{(t+1)}\right)^T \right\|_F \;\leq\; \left\| \left(\mathbb{I} - \widehat{U}^{(t)} \left(\widehat{U}^{(t)}\right)^T\right) U\, S \right\|_F + \left\| F_{\text{miss}}^{(t+1)} \right\|_F + \left\| F_{\text{noise}}^{(t+1)} \right\|_F . \quad (12)$$

Let $U_\perp \in \mathbb{R}^{\binom{n}{1} \times \binom{n}{1}-r}$ be an orthogonal matrix spanning the subspace orthogonal to $U$. We use the following definition of distance between two $r$-dimensional subspaces in $\mathbb{R}^{\binom{n}{1}}$.

$$d(\widehat{U}, U) = \left\| U_\perp^T \widehat{U} \right\|_2 .$$

The following key technical lemma provides upper bounds on each of the error terms in (12).

**Lemma 8** *For any $\mu_1$-incoherent orthogonal matrix $U^{(t)} \in \mathbb{R}^{\binom{n}{1} \times r}$ and $\mu$-incoherent matrix $M \in \mathbb{R}^{\binom{n}{1} \times \binom{n}{1}}$, the error after one step of alternating minimization is upper bounded by*

$$\begin{aligned} \|F_{\text{miss}}^{(t+1)}\|_F &\leq \frac{\sigma_1(M) r^{1.5} \mu \mu_1}{n(1 - \frac{2r}{n})} \, d(\widehat{U}^{(t)}, U) \, , \\ \|F_{\text{noise}}^{(t+1)}\|_F &\leq \frac{1}{1 - \frac{2r}{n}} \sqrt{r} \, \|\mathcal{P}_\Omega(E)\|_2 \, , \end{aligned}$$

*where $\sigma_i(M)$ is the $i$-th singular value of $M$.*

We show in Lemma 10 that the incoherence assumption is satisfied for all $t$ with $\mu_1 = 6(\sigma_1(M)/\sigma_r(M))\mu$. For $\mu_1 \leq \sqrt{n/2r}$ as per our assumption and substituting these bounds into (12), we get

$$\left\| M - \widehat{U}^{(t)} \left(U^{(t+1)}\right)^T \right\|_F \;\leq\; \|M\|_F \, d(\widehat{U}^{(t)}, U) \, + \, \frac{12\, \sigma_1(M)^2\, r^{1.5}\, \mu^2}{n\, \sigma_r(M)} \, d(\widehat{U}^{(t)}, U) \, + \, 2\sqrt{r}\, \|\mathcal{P}_\Omega(E)\|_2 \, ,$$

where the first term follows from the fact that $\|(\mathbb{I} - \widehat{U}^{(t)}(\widehat{U}^{(t)})^T)U\|_2 = \|\widehat{U}_\perp^{(t)}(\widehat{U}_\perp^{(t)})^T U\|_2 = d(\widehat{U}^{(t)}, U)$. To further bound the distance $d(\widehat{U}^{(t)}, U)$, we first claim that after $t$ iterations of the alternating minimization algorithm, the estimates satisfy

$$d(\widehat{U}^{(t)}, U) \;\leq\; \frac{\varepsilon}{2\|M\|_F} + \frac{2\sqrt{3\, r}}{\sigma_r(M)} \, \|\mathcal{P}_\Omega(E)\|_2 \, , \quad (13)$$

for $t \geq (1/2) \log\left(2\|M\|_F / \varepsilon\right)$. For $\mu \leq \sqrt{n\, \sigma_r(M)/(12 r \sigma_1(M))}$ as per our assumption, this gives

$$\left\| M - \widehat{U}^{(t)} \left(U^{(t+1)}\right)^T \right\|_F \;\leq\; \varepsilon \, + \, \frac{9\, \|M\|_F \, \sqrt{r}}{\sigma_r(M)} \, \|\mathcal{P}_\Omega(E)\|_2 \, .$$

This proves the desired error bound of Theorem 4.

Now, we are left to prove (13) for $t \geq (1/2) \log\left(2\|M\|_F / \varepsilon\right)$. This follows from the analysis of each step of the algorithm, which shows that we improve at each step up to a certain noise

level. Define $R_U^{(t+1)}$ to be the upper triangular matrix obtained by QR decomposition of $U^{(t+1)} = \widehat{U}^{(t+1)} R_U^{(t+1)}$. Then we can represent the distance using (11) as:

$$\begin{aligned} d(\widehat{U}^{(t+1)}, U) &= \left\| U_\perp^T \left( U S U^T U^{(t)} - F_{\text{miss}}^{(t+1)} + F_{\text{noise}}^{(t+1)} \right) \left( R_U^{(t+1)} \right)^{-1} \right\|_2, \\ &\leq \left( \|F_{\text{miss}}^{(t+1)}\|_2 + \|F_{\text{noise}}^{(t+1)}\|_2 \right) \left\| \left( R_U^{(t+1)} \right)^{-1} \right\|_2, \\ &\leq \frac{12\sqrt{3}\,\sigma_1(M)^2\, r^{1.5}\,\mu^2}{\sigma_r(M)^2\, n}\, d(\widehat{U}^{(t)}, U) + \frac{2\sqrt{3r}}{\sigma_r(M)} \|\mathcal{P}_\Omega(E)\|_2 \,, \end{aligned}$$

where we used Lemma 9 to bound $\left\| \left( R_U^{(t+1)} \right)^{-1} \right\|_2$, Lemma 8 to bound $\|F_{\text{miss}}^{(t+1)}\|_2$ and $\|F_{\text{noise}}^{(t+1)}\|_2$, and Lemma 10 to bound $\mu_1$. For $\mu \leq \sqrt{n\sigma_r(M)/(10\, r^{1.5}\sigma_1(M))}$ as per our assumption, it follows that

$$d(\widehat{U}^{(t)}, U) = \left(\frac{1}{4}\right)^t d(\widehat{U}^{(0)}, U) + \frac{2\sqrt{3r}}{\sigma_r(M)} \|\mathcal{P}_\Omega(E)\|_2 \,,$$

Taking $t \geq (1/2)\log\left(2\|M\|_F/\varepsilon\right)$, this finishes the proof of the desired bound in (13).

Now we are left to prove that starting from a good initial guess we obtain using a simple Singular Value Decomposition(SVD), the estimates at every iterate $t$ is incoherent with bounded $\|(R_U^{(t+1)})^{-1}\|_2$. We first state the following two lemmas upper bounding $\mu_1$ and $\|(R_U^{(t+1)})^{-1}\|_2$. Then we prove that the hypotheses of the lemmas are satisfied, if we start from a good initialization.

**Lemma 9** *Assume that $U$ is $\mu$-incoherent with $\mu \leq (\sigma_r(M)/\sigma_1(M))\sqrt{n/(32r^{1.5})}$, $d(\widehat{U}^{(t)}, U) \leq 1/2$, and $\|\mathcal{P}_\Omega(E)\|_2 \leq \sigma_r(M)/(16\sqrt{r})$. Then,*

$$\|(R_U^{(t+1)})^{-1}\|_2 \leq \frac{\sqrt{3}}{\sigma_r(M)} \,.$$

**Lemma 10 (Incoherence of the estimates)** *Assume that $\widehat{U}^{(t)}$ is $\tilde{\mu}$-incoherent with $\tilde{\mu} \leq \sqrt{n/(2r)}$, and $U$ is $\mu$-incoherent with $\mu \leq (\sigma_r(M)/\sigma_1(M))\sqrt{n/(32r)}$, and the noise $E$ satisfy $\|\mathcal{P}_\Omega(E)_{(i,\mathsf{a})}\| \leq \sigma_1(M)\mu\sqrt{3r/(8n\ell)}$ for all $i \in [n]$ and $a \in [\ell]$. Then, $\widehat{U}^{(t+1)}$ is $\mu_1$-incoherent with*

$$\mu_1 = \frac{6\,\mu\,\sigma_1(M)}{\sigma_r(M)} \,.$$

For the above two lemmas to hold, we need a good initial guess $\widehat{U}^{(0)}$ with incoherence less than $4\mu$ and error upper bounded by $d(\widehat{U}^{(0)}, U) \leq 1/2$. Next lemmas shows that we can get such a good initial guess by singular value decomposition and truncation. And this finishes the proof of Theorem 4.

**Lemma 11 (Bound on the initial guess)** *Let $\widehat{U}^{(0)}$ be the output of step 3 in the alternating minimization algorithm, and let $\mu_0$ be the incoherence of $\widehat{U}^{(0)}$. Assuming $\mu \leq \sqrt{\sigma_r(M)\, n/(32\,\sigma_1(M)\, r^{1.5})}$ and $\|\mathcal{P}_\Omega(E)\|_2 \leq \sigma_r(M)/(32\sqrt{r})$, we have the following upper bound on the error and the incoherence:*

$$\begin{aligned} d(\widehat{U}^{(0)}, U) &\leq \frac{1}{2}, \\ \mu_0 &\leq 4\,\mu. \end{aligned}$$

### A.1.1. PROOFS OF LEMMAS 8, 9, 10, 11

**Proof** [Proof of Lemma 8] First, we prove the following upper bound for $\mu_1$-incoherent $\widehat{U}^{(t+1)}$.

$$\|F_{\text{miss}}^{(t+1)}\|_F \;\leq\; \frac{\sigma_1(M) r^{1.5} \mu \mu_1}{n(1-\frac{2r}{n})} d(\widehat{U}^{(t)}, U) \;.$$

We drop the time index $(t+1)$ whenever it is clear from the context, to simplify notations. Let $F_{(i,a)} \in \mathbb{R}^r$ be a column vector representing the $(\ell(i-1)+a)$-th row of $F_{\text{miss}} \in \mathbb{R}^{\ell n \times r}$. We know from (10) that

$$F_{(i,a)} \;=\; (B^{(i,a)})^{-1} \underbrace{\left(B^{(i,a)} D - C^{(i,a)}\right)}_{\equiv H^{(i)}} S\, U_{(i,a)} \;, \tag{14}$$

where we define $H^{(i)} \equiv B^{(i,a)} D - C^{(i,a)}$. Notice that we dropped $a$ from the index to emphasize that $B^{(i,a)}$ and $C^{(i,a)}$ do not depend on $a$.

$$\begin{aligned}
\|F_{\text{miss}}\|_F \;&\leq\; \sqrt{\sum_{i,a} \|(B^{(i,a)})^{-1}\|_2^2 \, \|H^{(i)}\, S\, U_{(i,a)}\|^2} \\
&=\; \max_{j,b} \|(B^{(j,b)})^{-1}\|_2 \max_{x \in \mathbb{R}^{\ell n \times r}; \|x\|_F = 1} \sum_{i \in [n]; a \in [\ell]; q \in [r]} x_{(i,a),q}\, e_q^T\, H^{(i)}\, S\, U_{(i,a)} \;.
\end{aligned}$$

To upper bound the first term, notice that $\|(B^{(j,b)})^{-1}\|_2 \leq 1/\sigma_r(B^{(j,b)})$. Since $B^{(j,b)} = \mathbb{I}_{r\times r} - \sum_{a \in [\ell]} \widehat{U}_{(j,a)} (\widehat{U}_{(j,a)})^T$, and by incoherence property from Lemma 10, we have

$$\|(B^{(j,b)})^{-1}\|_2 \leq \frac{1}{1-\frac{2r}{n}} \;, \tag{15}$$

for all $(j,b)$.

The second term can be bounded using Cauchy-Schwarz inequality:

$$\begin{aligned}
\sum_{i \in [n]; a \in [\ell]; q \in [r]} x_{(i,a),q}\, e_q^T\, H^{(i)}\, S\, U_{(i,a)} \;&=\; \sum_{i \in [n]; q,p \in [r]} \Big(\sum_{a \in [\ell]} S_p\, U_{(i,a),p} x_{(i,a),q}\Big) \left(e_q^T\, H^{(i)} e_p\right) \\
&\leq\; \sqrt{\sum_{i,p,q} \Big(\sum_{a \in [\ell]} S_p\, U_{(i,a),p} x_{(i,a),q}\Big)^2} \sqrt{\sum_{i,p,q} (e_q^T\, H^{(i)} e_p)^2} \;,
\end{aligned}$$

where $S_p$ is the $p$-th eigenvalue of $M$. Applying Cauchy-Schwarz again, and by the incoherence of $U$ is and $\|x\| = 1$,

$$\begin{aligned}
\sum_{i,p,q} \Big(\sum_{a \in [\ell]} S_p\, U_{(i,a),p} x_{(i,a),q}\Big)^2 \;&\leq\; \sum_{i,p,q} S_p^2 \Big(\sum_{a \in [\ell]} U_{(i,a),p}^2 \sum_{b \in [\ell]} x_{(i,b),q}^2\Big) \\
&\leq\; \sigma_1(M)^2\, \frac{\mu^2 r}{n} \;.
\end{aligned} \tag{16}$$

$$\begin{aligned}
\sum_{i;p;q}(e_q^T H^{(i)} e_p)^2 &= \sum_{i;p;q}\Big(\sum_a \widehat{U}_{(i;a);q}\big(U_{(i;a);p} - \widehat{U}_{(i;a)}^T \widehat{U}^T U_p\big)\Big)^2 \\
&\le \sum_i \Big\{\sum_{a;q}\widehat{U}_{(i;a);q}^2 \sum_{b;p}\big(U_{(i;b);p} - \widehat{U}_{(i;b)}^T\widehat{U}^T U_p\big)^2\Big\} \\
&\le \frac{\mu_1^2 r}{n}\sum_{i;b;p}\big(U_{(i;b);p} - \widehat{U}_{(i;b)}^T\widehat{U}^T U_p\big)^2 \\
&\le \frac{\mu_1^2 r}{n}\big(r - \mathrm{Tr}(U^T\widehat{U}\widehat{U}^T U)\big)^2 \\
&\le \frac{\mu_1^2 r^2\, d(\widehat{U},U)^2}{n}\,, \qquad (17)
\end{aligned}$$

where the last inequality follows from the fact that $d(\widehat{U},U)^2 = \|\widehat{U}_\perp^T U\|_2^2 = \|U^T\widehat{U}_\perp\widehat{U}_\perp^T U\|_2 = \|\mathbb{I}_{r\times r} - U^T\widehat{U}\widehat{U}^T U\|_2 = 1 - \sigma_r(\widehat{U}^T U)^2 \ge 1 - (1/r)\sum_p \sigma_p(\widehat{U}^T U)^2$.

Now, we prove an upper bound on $\|F_{\mathrm{noise}}^{(t+1)}\|_F$. Again, we drop the time index $(t+1)$ or $(t)$ whenever it is clear from the context. Let $\widetilde{F}_{(i;a)} \in \mathbb{R}^r$ denote a column vector representing the $(\ell(i-1)+a)$-th row of $F_{\mathrm{noise}}$. We know from (10) that

$$\widetilde{F}_{(i;a)} = (B^{(i;a)})^{-1}\Big(\widehat{U}^T E_{(i;a)} - \sum_{b\in[\grave{}]} E_{(i;a)(i;b)}\widehat{U}_{i;b}\Big)\,, \qquad (18)$$

where $E_{(i;a)} \in \mathbb{R}^{\grave{}n}$ is an column vector representing the $(\ell(i-1)+a)$-th row of $E$. Then,

$$\begin{aligned}
\|F_{\mathrm{noise}}\|_F &\le \sqrt{\sum_{i\in[n];a\in[\grave{}]}\big\|(B^{(i;a)})^{-1}\big\|_2^2\Big\|\widehat{U}^T E_{(i;a)} - \sum_{b\in[\grave{}]}E_{(i;a)(i;b)}\widehat{U}_{i;b}\Big\|^2} \\
&\le \max_{i;a}\big\|(B^{(i;a)})^{-1}\big\|_2\,\|\mathcal{P}_\Omega(E)\widehat{U}\|_F \\
&\le \frac{1}{1-\frac{2}{1}\frac{r}{n}}\sqrt{r}\,\|\mathcal{P}_\Omega(E)\|_2\,,
\end{aligned}$$

where $\mathcal{P}_\Omega$ is the projection onto the sampled entries defined in (3), and we used (15) to bound $\|(B^{(i;a)})^{-1}\|_2$. □

**Proof** [Proof of Lemma 9] From Lemma 7 in (Gunasekar et al., 2013), we know that

$$\|(R_U^{(t+1)})^{-1}\|_2 \le \frac{1}{\sigma_r(M)\sqrt{1-d^2(U^{(t)},U)} - \|F_{\mathrm{miss}}^{(t+1)}\|_2 - \|F_{\mathrm{noise}}^{(t+1)}\|_2}\,.$$

From Lemma 8 with $\mu \le (\sigma_r(M)/(6\sigma_1(M)))\sqrt{n/(2r^{1.5})}$ and $\|\mathcal{P}_\Omega(E)\|_2 \le \sigma_r(M)/(16\sqrt{r})$, we have $\|F_{\mathrm{noise}}^{(t+1)}\|_2 \le \sigma_r(M)/8$ and $\|F_{\mathrm{miss}}^{(t+1)}\|_2 \le (1/6)\sigma_r(M)\,d(\widehat{U}^{(t)},U)$. Assuming $d(\widehat{U}^{(t)},U) \le 1/2$, this proves the desired claim. □

**Proof** [Proof of Lemma 10] Assuming that $\widehat{U}^{(t)}$ is $\tilde{\mu}$-incoherent, we make use of the following set of inequalities:

$$\begin{aligned}
\|(B^{(i;a)})^{-1}\|_2 &\geq 1-(\tilde{\mu}^2 r/n) \\
\|B^{(i;a)}\|_2 &= \|\mathbb{I}_{r\times r} - \widehat{U}_{(i)}\widehat{U}_{(i)}^T\|_2 \leq 1 \\
\|D\|_2 &= \|\widehat{U}^T U\|_2 \leq 1 \\
\|C^{(i;a)}\|_2 &= \|\widehat{U}^T U - \widehat{U}_{(i)}U_{(i)}^T\|_2 \leq 1+\mu\tilde{\mu}\, r/n\ .
\end{aligned}$$

Also, from Lemma 9, we know that if $\tilde{\mu} \leq \sqrt{n/2r}$ as per our assumption, then $\|(R_U^{(t+1)})^{-1}\|_2 \leq \sqrt{3}/\sigma_r(M)$. Then, by (10) and the triangular inequality,

$$\begin{aligned}
\sum_{a\in[\ell]} \|\widehat{U}_{(i;a)}^{(t+1)}\|^2 &\leq \sum_{a\in[\ell]} \left\|(B^{(i;a)})^{-1} C^{(i;a)} S\, U_{(i;a)} + (B^{(i;a)})^{-1} N_{(i;a)}\right\|^2 \left\|\left(R_U^{(t+1)}\right)^{-1}\right\|_2^2 \\
&\leq \sum_{a\in[\ell]} 2\left\|\left(R_U^{(t+1)}\right)^{-1}\right\|_2^2 \|(B^{(i;a)})^{-1}\|_2^2 \left\{ \|C^{(i;a)}\|_2^2 \|S\|_2^2 \|U_{(i;a)}\|^2 + \|N_{(i;a)}\|^2 \right\} \\
&\leq \frac{6}{\sigma_r(M)^2\,(1-(\tilde{\mu}^2\, r/n))^2} \sum_{a\in[\ell]} \left\{ \sigma_1(M)^2 \Big(1+\mu\tilde{\mu} r/n\Big) \|U_{(i;a)}\|^2 + \left\|\widehat{U}^T\, \mathcal{P}_\Omega(E)_{(i;a)}\right\|^2 \right\} \\
&\leq \frac{6}{\sigma_r(M)^2\,(1-(\tilde{\mu}^2\, r/n))^2} \left\{ \sigma_1(M)^2 \Big(1+\frac{\mu\tilde{\mu} r}{n}\Big) \frac{\mu^2 r}{n} + \|\mathcal{P}_\Omega(E)_{(i;a)}\|^2 \right\} \\
&\leq \frac{36\sigma_1(M)^2}{\sigma_r(M)^2} \frac{\mu^2 r}{n}\ ,
\end{aligned}$$

where the last inequality follows from our assumption that $\tilde{\mu} \leq \sqrt{n/(2r)}$, $\mu \leq (\sigma_r(M)/\sigma_1(M))\sqrt{n/(32r)}$, and $\|\mathcal{P}_\Omega(E)_{(i;a)}\| \leq \sigma_1(M)\mu\sqrt{3r/(8n\ell)}$. This proves that $\widehat{U}^{(t+1)}$ is $\mu_1$-incoherent for $\mu_1 = 6\mu(\sigma_1(M)/\sigma_r(M))$. □

**Proof** [Proof of Lemma 11] Let $\mathcal{P}_r(\widehat{M}) = \widetilde{U}\widetilde{S}\widetilde{U}^T$ denote the best rank-$r$ approximation of the observed matrix $\widehat{M}$ and $\mathcal{P}_\Omega$ is the sampling mask operator defined in (3) such that $M - \widehat{M} = \mathcal{P}_\Omega(E) + M - \mathcal{P}_\Omega(M)$. Then,

$$\begin{aligned}
\|\, M - \mathcal{P}_r(\widehat{M})\,\|_2 &\leq \|\, M-\widehat{M}\,\|_2 + \|\,\widehat{M} - \mathcal{P}_r(\widehat{M})\,\|_2 \\
&\leq 2\,\|\, M-\widehat{M}\,\|_2 \\
&\leq 2\left(\|\mathcal{P}_\Omega(E)\|_2 + \|M-\mathcal{P}_\Omega(M)\|_2\right) \\
&\leq 2\left(\|\mathcal{P}_\Omega(E)\|_2 + \sigma_1(M)\,\mu^2 r/n\right)\ , \qquad (19)
\end{aligned}$$

where we used the fact that $\mathcal{P}_r(\widehat{M})$ is the best rank-$r$ approximation such that $\|\widehat{M} - \mathcal{P}_r(\widehat{M})\|_2 \leq \|\widehat{M} - A\|_2$ for any rank-$r$ matrix $A$, and $\|M - \mathcal{P}_\Omega(M)\|_2 = \max_i \|U_{(i)}SU_{(i)}^T\|_2 \leq (\mu^2 r/n)\sigma_1(M)$.

The next series of inequalities provide an upper bound on $d(\widetilde{U}, U)$ in terms of the spectral norm:

$$\begin{aligned}
\|\, M - \mathcal{P}_r(\widehat{M})\,\|_2 &= \|\,(\widetilde{U}\widetilde{U}^T)(USU^T - \widetilde{U}\widetilde{S}\widetilde{U}^T) + (\widetilde{U}_\perp\widetilde{U}_\perp^T)(USU^T - \widetilde{U}\widetilde{S}\widetilde{U}^T)\,\|_2 \\
&\geq \|(\widetilde{U}_\perp\widetilde{U}_\perp^T)(USU^T - \widetilde{U}\widetilde{S}\widetilde{U}^T)\,\|_2 \\
&= \|\widetilde{U}_\perp^T USU^T\,\|_2 \\
&\geq \sigma_r(S)\,\|\widetilde{U}_\perp^T U\|_2 \\
&\geq \sigma_r(S)\, d(\widetilde{U}, U)\ ,
\end{aligned}$$

Together with (19), this implies that

$$d(\widetilde{U}, U) \;\leq\; \frac{2}{\sigma_r(M)} \big( \|\mathcal{P}_\Omega(E)\|_2 + \sigma_1(M) \mu^2 r/n \big) \,.$$

For $\|\mathcal{P}_\Omega(E)\|_2 \leq \sigma_r(M)/(32\sqrt{r})$ and $\mu \leq \sqrt{\sigma_r(M)\, n/(32\, \sigma_1(M)\, r^{1.5})}$ as per our assumptions, we have

$$d(\widetilde{U}, U) \;\leq\; \frac{1}{8\sqrt{r}} \,.$$

Next, we show that by truncating large components of $\widetilde{U}$, we can get an incoherent matrix $\widehat{U}^{(0)}$ which is also close to $U$. Consider a sub-matrix of $U$ which consists of the rows from $\ell(i-1)+1$ to $\ell i$. We denote this block by $U_{(i)} \in \mathbb{R}^{\ell \times r}$. Let $\overline{U}$ denote an $\ell n \times r$ matrix obtained from $\widetilde{U}$ by setting to zero all blocks that have Frobenius norm greater than $2\mu\sqrt{r/n}$. Let $\widehat{U}^{(0)}$ be the orthonormal basis of $\overline{U}$. We use the following lemma to bound the error and incoherence of the resulting $\widehat{U}^{(0)}$. A similar lemma has been proven in (Jain et al., 2013, Lemma C.2), and we provide a tighter bound in the following lemma. For $\delta \leq 1/(8\sqrt{r})$, this lemma proves that we get the desired bound of $d(\widehat{U}^{(0)}, U) \leq 1/2$ and $\mu_0 \leq 4\mu$. □

**Lemma 12** *Let $\mu_0$ denote the incoherence of $\overline{U}$, and define $\delta \equiv d(\widetilde{U}, U)$. Then*

$$d(\widehat{U}^{(0)}, U) \;\leq\; \frac{3\sqrt{r}\,\delta}{1 - 2\sqrt{r}\,\delta} \,, \qquad \textit{and} \qquad \mu_0 \;\leq\; \frac{2\mu}{1 - 2\sqrt{r}\,\delta} \,.$$

**Proof** Denote the QR decomposition of $\overline{U}$ by $\overline{U} = \widehat{U}^{(0)} R$ and let $\delta \equiv d(\widetilde{U}, U)$. Then,

$$\begin{aligned} d(\widehat{U}^{(0)}, U) &= \|U_\perp^T \widehat{U}^{(0)}\|_2 \\ &\leq \|U_\perp^T \overline{U}\|_2 \, \|R^{-1}\|_2 \\ &\leq \big( \|U_\perp^T (\overline{U} - \widetilde{U})\|_2 + \|U_\perp^T \widetilde{U}\|_2 \big) \, \|R^{-1}\|_2 \\ &= \big( \|\overline{U} - \widetilde{U}\|_2 + \delta \big) \, \|R^{-1}\|_2 \,. \end{aligned} \tag{20}$$

First, we upper bound $\|\overline{U} - \widetilde{U}\|_F$ as follows. Let $\mathcal{P}()$ denote a projection operator that sets to zero those blocks whose Frobenius norm is smaller than $2\mu\sqrt{r/n}$ such that $\mathcal{P}(\widetilde{U}) = \widetilde{U} - \overline{U}$. Then,

$$\|\mathcal{P}(\widetilde{U})\|_F \;\leq\; \|\mathcal{P}(\widetilde{U} - U(U^T \widetilde{U}))\|_F \;+\; \|\mathcal{P}(U(U^T \widetilde{U}))\|_F \,. \tag{21}$$

The first term can be bounded by $\|\mathcal{P}(\widetilde{U} - U(U^T \widetilde{U}))\|_F \leq \|\widetilde{U} - U(U^T \widetilde{U})\|_F \leq \sqrt{r} \|\widetilde{U} - U(U^T \widetilde{U})\|_2 = \sqrt{r}\,\delta$. The second term can be bounded by $\|\mathcal{P}(U(U^T \widetilde{U}))\|_F = \|\mathcal{P}(U)\,(U^T \widetilde{U})\|_F \leq \|\mathcal{P}(U)\|_F$. By incoherence of $U$, we have that $\|\mathcal{P}(U)\|_F \leq \sqrt{N}\mu\sqrt{r/n}$, where $N$ is the number of $\ell \times r$ block matrices that is not set to zero by $\mathcal{P}(\cdot)$.

To provide an upper bound on $N$, notice that the incoherence of an $\ell n \times r$ matrix $U(U^T \widetilde{U})$ is $\mu$. This follows from the fact that $\|U^T \widetilde{U}\|_2 \leq 1$. Then,

$$\begin{aligned} \|U(U^T \widetilde{U}) - \widetilde{U}\|_F &\geq \|\mathcal{P}(U(U^T \widetilde{U}) - \widetilde{U})\|_F \\ &\geq \sqrt{N}\,\mu\sqrt{\frac{r}{n}} \,, \end{aligned}$$

where the last line follows from the fact that there are $N$ blocks where the Frobenius norm of $U(U^T\widetilde{U}$ in that block is at most $\mu\sqrt{r/n}$ and the Frobenius norm of $\widetilde{U}$ is at least $2\mu\sqrt{r/n}$. On the other hand, we have $\|U(U^T\widetilde{U}) - \widetilde{U}\|_F \leq \sqrt{r}\delta$. Putting these inequalities together, we get that

$$\sqrt{N} \leq \frac{\delta\sqrt{n}}{\mu} \quad \text{and } \|\mathcal{P}(U(U^T\widetilde{U}))\|_F \leq \sqrt{r}\,\delta\,.$$

Substituting these bounds in (21) gives

$$\|\widetilde{U} - \overline{U}\|_F \quad \leq \quad 2\,\delta\,\sqrt{r}\,. \tag{22}$$

Next, we show that

$$\|R^{-1}\|_2 \quad \leq \quad \frac{1}{1 - 2\delta\sqrt{r}}\,. \tag{23}$$

By the definition of $R$, we know that $\|R^{-1}\|_2 = 1/\sigma_r(R) = 1/\sigma_r(\widehat{U}^{(0)} = 1/\sigma_r(\overline{U})$. Using Weyl's inequality, we can lower bound $\sigma_r(\overline{U}) = \sigma_r(\overline{U} - \widetilde{U} + \widetilde{U}) \geq \sigma_r(\widetilde{U}) - \sigma_1(\overline{U} - \widetilde{U})$. Since $\widetilde{U}$ is an orthogonal matrix and using (22), this proves (23). Substituting (22) and (23) into (20), we get

$$d(\widehat{U}^{(0)}, U) \quad \leq \quad \frac{(2\sqrt{r}+1)\delta}{1 - 2\delta\sqrt{r}}\,.$$

For $\delta \leq$, this gives the desired bound.

To provide an upper bound on the incoherence $\mu_0$ of $\widehat{U}^{(0)}$, recall that the incoherence is defined as $\mu_0\sqrt{r/n} = \max_i \|\widehat{U}^{(0)}_{(i)}\|_F = \max_i \|\overline{U}_{(i)}R^{-1}\|_F$. By construction, $\|\overline{U}_{(i)}\|_F \leq 2\mu\sqrt{r/n}$, and from (23) we know that $\|R^{-1}\|_2 \leq 1/(1 - 2\delta\sqrt{r})$. Together, this gives

$$\mu_0 \quad \leq \quad \frac{2\mu}{1 - 2\delta\sqrt{r}}\,.$$

This finishes the proof of the desired bounds. □

### A.2. Proof of Theorem 6

In this section, we provide a detailed proof of Theorem 6. To this end, we first provide an infinite sample version of the proof, i.e., when $P_{\Omega_3}(S_3) = P_{\Omega_3}(M_3)$. Then, in the next subsection, we bound each element of $P_{\Omega_3}(S_3) - P_{\Omega_3}(M_3)$ and extend the infinite sample version of the proof to the finite sample case.

Recall that $\widehat{M}_2 = \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}\widehat{U}^T_{M_2}$, $\varepsilon = \|\widehat{M}_2 - M_2\|_2\|/\sigma_r(M_2)$, $M_2$ is $\mu$-incoherent and $\widehat{M}_2$ is $\mu_1$-incoherent. Incoherence of a matrix is defined as in (7). Then, the following two remarks can be easily proved using standard matrix perturbation results (for example, see (Anandkumar et al., 2012a)).

**Remark 13** *Suppose* $\|\widehat{M}_2 - M_2\|_2 \leq \varepsilon\sigma_r(M_2)$, *then*

$$1 - 4\frac{\varepsilon^2}{(1-\varepsilon)^2} \leq \sigma_r(U^T\widehat{U}_{M_2}) \leq \sigma_1(\widehat{U}^T_{M_2}U) \leq 1.$$

*That is,*

$$\|(I - \widehat{U}_{M_2}\widehat{U}_{M_2}^T)U\|_2 \leq \varepsilon, \text{ and },$$
$$\|(U^T\widehat{U}_{M_2})^T(U^T\widehat{U}_{M_2}) - I\| \leq 8\frac{\varepsilon^2}{(1-\varepsilon)^2}.$$

**Remark 14** *Suppose* $\|\widehat{M_2} - M_2\|_2 \leq \varepsilon\sigma_r(M_2)$, *then*

$$\|\mathbb{I} - \widehat{\Sigma}_{M_2}^{-1=2}\widehat{U}_{M_2}^T M_2 \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\|_2 \leq 2\varepsilon.$$

**Proof**

$$\begin{aligned}\|I - \widehat{\Sigma}_{M_2}^{-1=2}\widehat{U}_{M_2}^T M_2 \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\|_2 &= \|\widehat{\Sigma}_{M_2}^{-1=2}\widehat{U}_{M_2}^T(\widehat{M_2} - M_2)\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\|_2 \\ &\leq \|\widehat{\Sigma}_{M_2}^{-1=2}\widehat{U}_{M_2}^T\|_2^2\,\|\widehat{M_2} - M_2\|_2 \\ &\leq \frac{1}{\sigma_r(M_2)(1-\varepsilon)}\sigma_r(M_2)\varepsilon\ ,\end{aligned}$$

where we used the fact that $\|\widehat{\Sigma}_{M_2}^{-1=2}\|_2^2 \geq 1/\sigma_r(\widehat{M_2})$ and $\sigma_r(\widehat{M_2}) \geq \sigma_r(M_2)(1-\varepsilon)$ by Weyl's inequality. For $\varepsilon < 1/2$ we have the desired bound.

□

We now define the following operators: $\widehat{\nu}$ and $\widehat{A}$. Define $\widehat{\nu} : \mathbb{R}^{r\times r\times r} \to \mathbb{R}^{`n\times `n\times `n}$ as:

$$\widehat{\nu}_{ijk}(Z) = \begin{cases}\sum_{abc} Z_{abc}(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2})_{ia}(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2})_{jb}(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2})_{kc}, & \text{if } \lceil \frac{i}{`} \rceil \neq \lceil \frac{j}{`} \rceil \neq \lceil \frac{k}{`} \rceil \neq \lceil \frac{i}{`} \rceil, \\ 0, & \text{otherwise.}\end{cases} \tag{24}$$

Define $\widehat{A} : \mathbb{R}^{`n\times `n\times `n} \to \mathbb{R}^{r\times r\times r}$ as:

$$\widehat{A}(Z) = \widehat{\nu}(Z)\left[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\right]. \tag{25}$$

Now, let $R_3$ be defined as: $R_3 = \widehat{\Sigma}_{M_2}^{-1=2}\widehat{U}_{M_2}^T U\Sigma V^T W^{1=2}$. Note that, using Remark 14,

$$\|R_3 R_3^T - I\| \leq 2\varepsilon.$$

Also, define the following tensor:

$$\widetilde{G} = \sum_{q\in[r]} \frac{1}{\sqrt{w_q}}(R_3\mathbf{e}_q \otimes R_3\mathbf{e}_q \otimes R_3\mathbf{e}_q). \tag{26}$$

Note that, as $R_3$ is nearly orthonormal, $\widetilde{G}$ is a *nearly* orthogonally decomposable tensor.

We now present a lemma that shows that $P_{\Omega_3}(M_3)\left[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\right]$ and $\widehat{A}(\widetilde{G})$ are "close".

**Lemma 15**

$$P_{\Omega_3}(M_3)\left[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2},\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2},\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\right]=\widehat{A}(\widetilde{G})+E,$$

*where*

$$\|E\|_F \;\leq\; \frac{12\,\mu_1^3\,\mu\,r^{3:5}\,\sigma_1(M_2)^{3=2}\,\varepsilon}{n\sqrt{w_{\min}}\sigma_r(M_2)^{3=2}}\;,$$

*and we denote the Frobenius norm of a tensor as* $\|E\|_F=\{\sum_{i;j;k}E_{i;j;k}^2\}^{1=2}$

**Proof** Define $H=\widehat{A}(G)$ and $F=P_{\Omega_3}(M_3)\left[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2},\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2},\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\right]$. Also, let $Q=U\Sigma V^T W^{1=2}$ and $\widehat{Q}=\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}$.

Note that, $F_{abc}=\sum_{ijk}\delta_{ijk}M_3(i,j,k)\widehat{Q}_{ia}\widehat{Q}_{jb}\widehat{Q}_{kc}$, where $\delta_{ijk}=1$, if $(i,j,k)\in\Omega_3$ and 0 otherwise. Also, $M_3(i,j,k)=\sum_{q\in[r]}\frac{1}{\sqrt{w_q}}Q_{iq}\cdot Q_{jq}\cdot Q_{kq}$. Hence,

$$F_{abc}=\sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\sum_{ijk}\delta_{ijk}Q_{iq}\cdot Q_{jq}\cdot Q_{kq}\cdot\widehat{Q}_{ia}\cdot\widehat{Q}_{jb}\cdot\widehat{Q}_{kc}. \tag{27}$$

Note that, $\sum_i\widehat{Q}_{ia}Q_{iq}=\langle\widehat{Q}_a,Q_q\rangle=\mathbf{e}_a^T\widehat{\Sigma}_{M_2}^{-1=2}\widehat{U}_{M_2}^T U\Sigma V^T W^{1=2}\mathbf{e}_q=\mathbf{e}_a^T R_3\mathbf{e}_q$. That is,

$$\begin{aligned}F_{abc}=G_{abc}&-\sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\sum_{m\in[n]}\langle\widehat{Q}_a^{(m)},Q_q^{(m)}\rangle\cdot\langle\widehat{Q}_b^{(m)},Q_q^{(m)}\rangle\cdot\langle\widehat{Q}_c^{(m)},Q_q^{(m)}\rangle\\
-\sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\mathbf{e}_a^T R_3\mathbf{e}_q\sum_{m\in[n]}\langle\widehat{Q}_b^{(m)},Q_q^{(m)}\rangle\cdot\langle\widehat{Q}_c^{(m)},Q_q^{(m)}\rangle&-\sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\mathbf{e}_b^T R_3\mathbf{e}_q\sum_{m\in[n]}\langle\widehat{Q}_a^{(m)},Q_q^{(m)}\rangle\cdot\langle\widehat{Q}_c^{(m)},Q_q^{(m)}\rangle\\
&-\sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\mathbf{e}_c^T R_3\mathbf{e}_q\sum_{m\in[n]}\langle\widehat{Q}_a^{(m)},Q_q^{(m)}\rangle\cdot\langle\widehat{Q}_b^{(m)},Q_q^{(m)}\rangle.\end{aligned} \tag{28}$$

On the other hand,

$$\widehat{\nu}(G)_{ijk}=\begin{cases}\sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\mathbf{e}_i^T(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}R_3)\mathbf{e}_q\cdot\mathbf{e}_j^T(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}R_3)\mathbf{e}_q\cdot\mathbf{e}_k^T(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}R_3)\mathbf{e}_q, & \text{if } \lceil\tfrac{i}{\epsilon}\rceil\neq\lceil\tfrac{j}{\epsilon}\rceil\neq\lceil\tfrac{k}{\epsilon}\rceil\neq\lceil\tfrac{i}{\epsilon}\rceil,\\ 0, & \text{otherwise.}\end{cases} \tag{29}$$

That is,

$$H_{abc}=\sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\sum_{ijk}\delta_{ijk}\mathbf{e}_i^T(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}R_3)\mathbf{e}_q\cdot\mathbf{e}_j^T(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}R_3)\mathbf{e}_q\cdot\mathbf{e}_k^T(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}R_3)\mathbf{e}_q\cdot\widehat{Q}_{ia}\cdot\widehat{Q}_{jb}\cdot\widehat{Q}_{kc}. \tag{30}$$

Now, note that $\sum_i \widehat{Q}_{ia}\mathbf{e}_i^T(\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}R_3)\mathbf{e}_q = \langle \widehat{Q}_a, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}R_3\mathbf{e}_q\rangle = \mathbf{e}_a^T\widehat{\Sigma}_{M_2}^{-1=2}\widehat{U}_{M_2}^T\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}R_3\mathbf{e}_q = \mathbf{e}_a^T R_3\mathbf{e}_q$. Also, let $\widetilde{Q} = \widehat{U}_{M_2}\widehat{U}_{M_2}^T Q$. That is,

$$H_{abc} = G_{abc} - \sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\sum_{m\in[n]}\langle \widehat{Q}_a^{(m)}, \widetilde{Q}_q^{(m)}\rangle\cdot\langle \widehat{Q}_b^{(m)}, \widetilde{Q}_q^{(m)}\rangle\cdot\langle \widehat{Q}_c^{(m)}, \widetilde{Q}_q^{(m)}\rangle$$
$$-\sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\mathbf{e}_a^T R_3\mathbf{e}_q\sum_{m\in[n]}\langle \widehat{Q}_b^{(m)}, \widetilde{Q}_q^{(m)}\rangle\cdot\langle \widehat{Q}_c^{(m)}, \widetilde{Q}_q^{(m)}\rangle - \sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\mathbf{e}_b^T R_3\mathbf{e}_q\sum_{m\in[n]}\langle \widehat{Q}_a^{(m)}, \widetilde{Q}_q^{(m)}\rangle\cdot\langle \widehat{Q}_c^{(m)}, \widetilde{Q}_q^{(m)}\rangle$$
$$-\sum_{q\in[r]}\frac{1}{\sqrt{w_q}}\mathbf{e}_c^T R_3\mathbf{e}_q\sum_{m\in[n]}\langle \widehat{Q}_a^{(m)}, \widetilde{Q}_q^{(m)}\rangle\cdot\langle \widehat{Q}_b^{(m)}, \widetilde{Q}_q^{(m)}\rangle. \quad (31)$$

Now,

$$\begin{aligned}\left|\langle \widehat{Q}_c^{(m)}, \widetilde{Q}_q^{(m)}\rangle - \langle \widehat{Q}_c^{(m)}, Q_q^{(m)}\rangle\right| &\le \|\widehat{Q}_c^{(m)}\|\,\|\widetilde{Q}_q^{(m)} - Q_q^{(m)}\| \\ &\le \|\widehat{Q}_c^{(m)}\|\,\|(I-\widehat{U}_{M_2}\widehat{U}_{M_2}^T)U\|_2\,\|\Sigma V^T W^{1=2}\|_2 \\ &\le \frac{\mu_1\sqrt{r}}{\sqrt{n(1-\varepsilon)\sigma_r(M_2)}}\,\varepsilon\,\sqrt{\sigma_1(M_2)},\end{aligned}$$

where we used $\|(I-\widehat{U}_{M_2}\widehat{U}_{M_2}^T)U\|_2 \le \varepsilon$ from Remark 13, and the following remark to bound $\|\widehat{Q}_c^{(m)}\|$. Then, from Remark 16,

$$\begin{aligned}&|\langle \widehat{Q}_a^{(m)}, \widetilde{Q}_q^{(m)}\rangle\langle \widehat{Q}_c^{(m)}, \widetilde{Q}_q^{(m)}\rangle - \langle \widehat{Q}_a^{(m)}, Q_q^{(m)}\rangle\langle \widehat{Q}_c^{(m)}, Q_q^{(m)}\rangle| \\ &\le |(\langle \widehat{Q}_a^{(m)}, \widetilde{Q}_q^{(m)}\rangle - \langle \widehat{Q}_a^{(m)}, Q_q^{(m)}\rangle)\langle \widehat{Q}_c^{(m)}, \widetilde{Q}_q^{(m)}\rangle| + |\langle \widehat{Q}_a^{(m)}, Q_q^{(m)}\rangle(\langle \widehat{Q}_c^{(m)}, \widetilde{Q}_q^{(m)}\rangle - \langle \widehat{Q}_c^{(m)}, Q_q^{(m)}\rangle)| \\ &\le \frac{\mu_1\sqrt{r\,\sigma_1(M_2)}}{\sqrt{n(1-\varepsilon)\sigma_r(M_2)}}\,\varepsilon\,\frac{\mu_1(\mu+\mu_1)\,r}{n\,(1-\varepsilon)\,\sigma_r(M_2)}\end{aligned}$$

Further, $|e_a^T R_3 e_q| \le \mu_1\sqrt{(r\sigma_1(M_2))/(n(1-\varepsilon)\sigma_r(M_2))}$. The desired bound now follows by using the above inequalities to bound $\|E\|_F = \|H-F\|_F$. □

**Remark 16** *For $\widetilde{Q} = \widehat{U}_{M_2}\widehat{U}_{M_2}^T Q$, $Q = U\Sigma V^T W^{1=2}$, and $\widehat{Q} = \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}$, suppose $M_2$ is $\mu$-incoherent and $\widehat{M_2}$ is $\mu_1$-incoherent. Then,*

$$\|\widehat{Q}_c^{(m)}\| \le \frac{\mu_1 r^{1=2}}{\sqrt{(1-\varepsilon)\,n\,\sigma_r(M_2)}}, \quad \|\widetilde{Q}_c^{(m)}\| \le \mu_1\sqrt{\frac{r\sigma_1(M_2)}{n}}, \quad \textit{and} \quad \|Q_c^{(m)}\| \le \mu\sqrt{\frac{r\sigma_1(M_2)}{n}}.$$

**Proof**

$$\|\widehat{Q}_c^{(m)}\| = \frac{1}{\sqrt{\widehat{\Sigma}_{cc}}}\Big\{\sum_{a\in[\grave{}]}(\widehat{U}_{M_2})_{\grave{}(m-1)+a;c}\Big\}^{1=2} \le \frac{\mu_1\sqrt{r/n}}{\sqrt{\sigma_r(M_2)(1-\varepsilon)}}.$$

The rest of the remark follows similarly. □

Next, we now show that $\|\widehat{A}^{-1}\|_2$ is small.

**Lemma 17** *$\sigma_{\min}(\widehat{A}) \geq 1 - 8r^3\sigma_1(M_2)^2(1+\varepsilon)^2/(n\sigma_r(M_2)^2(1-\varepsilon)^2)$ and hence,*

$$\|\widehat{A}^{-1}\|_2 \;\leq\; \frac{1}{1 - 72r^3\sigma_1(M_2)^2/(n\sigma_r(M_2)^2)}\,.$$

**Proof** Let $\widehat{Q} = \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}$, $\widetilde{Q} = \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}$, and $H = \widehat{A}(Z)$. Then,

$$H_{abc} = \sum_{ijk}\delta_{ijk}\sum_{a'b'c'} Z_{a'b'c'}\widetilde{Q}_{ia'}\cdot\widetilde{Q}_{jb'}\cdot\widetilde{Q}_{kc'}\cdot\widehat{Q}_{ia}\cdot\widehat{Q}_{jb}\cdot\widehat{Q}_{kc},$$

where $\delta_{ijk} = 1,$ if $(i,j,k) \in \Omega_3$ and 0 otherwise. That is,

$$\begin{aligned} H_{abc} = Z_{abc} &- \sum_{a'b'c'} Z_{a'b'c'}\sum_{m\in[n]}\langle\widehat{Q}_a^{(m)},\widetilde{Q}_{a'}^{(m)}\rangle\cdot\langle\widehat{Q}_b^{(m)},\widetilde{Q}_{b'}^{(m)}\rangle\cdot\langle\widehat{Q}_c^{(m)},\widetilde{Q}_{c'}^{(m)}\rangle \\ &- \sum_{b'c'} Z_{ab'c'}\sum_{m\in[n]}\langle\widehat{Q}_b^{(m)},\widetilde{Q}_{b'}^{(m)}\rangle\cdot\langle\widehat{Q}_c^{(m)},\widetilde{Q}_{c'}^{(m)}\rangle - \sum_{a'c'} Z_{a'bc'}\sum_{m\in[n]}\langle\widehat{Q}_a^{(m)},\widetilde{Q}_{a'}^{(m)}\rangle\cdot\langle\widehat{Q}_c^{(m)},\widetilde{Q}_{c'}^{(m)}\rangle \\ &- \sum_{a'b'} Z_{a'b'c}\sum_{m\in[n]}\langle\widehat{Q}_a^{(m)},\widetilde{Q}_{a'}^{(m)}\rangle\cdot\langle\widehat{Q}_c^{(m)},\widetilde{Q}_{c'}^{(m)}\rangle. \quad (32) \end{aligned}$$

Let $\mathrm{vec}(H) = B\cdot\mathrm{vec}(Z)$. We know that $|\langle\widetilde{Q}_a^{(m)},\widehat{Q}_a^{(m)}\rangle| \leq \mu_1^2 r/n$ and $|\langle\widetilde{Q}_a^{(m)},\widetilde{Q}_{a'}^{(m)}\rangle| \leq \mu_1^2 r\sigma_1(M_2)(1+\varepsilon)/(n\sigma_r(m_2)(1-\varepsilon))$ for $a \neq a'$. Now, using the above equation and using incoherence:

$$1 - 4r^2\mu_1^4/n \leq B_{pp} \leq 1 + 4r^2\mu_1^4/n, \forall 1 \leq p \leq r.$$

Similarly, $|B_{pq}| \leq 4r^2\mu_1^4\sigma_1(M_2)^2(1+\varepsilon)^2/(n\sigma_r(M_2)^2(1-\varepsilon)^2), \forall p \neq q$. Theorem now follows using Gershgorin's theorem. □

Finally, we combine the above two lemmas to show that the least squares procedure approximately recovers $\widetilde{G}$.

**Lemma 18** *Let $G$ be as defined in (26). Also, let $\widehat{G}$ be obtained by solving the following least squares problem:*

$$\widehat{G} = \arg\min_{Z}\|\widehat{A}(Z) - P_{\Omega_3}(M_3)\left[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2},\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2},\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\right]\|_F^2.$$

*Then, for $n \geq 144r^3\sigma_1(M_2)^2/\sigma_r(M_2)^2$ such that $\|\widehat{A}^{-1}\|_2 \leq 2$,*

$$\|\widehat{G} - \widetilde{G}\|_F \;\leq\; \frac{24\mu_1^3\,\mu\, r^{3:5}\sigma_1(M_2)^{3=2}\varepsilon}{n\sqrt{w_{\min}}\sigma_r(M_2)}\,.$$

**Proof** Note that $\widehat{A} : \mathbb{R}^{r\times r\times r} \to \mathbb{R}^{r\times r\times r}$ is a square operator. Moreover, using Lemma 15:

$$P_{\Omega_3}(M_3)\left[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2},\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2},\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\right] = \widehat{A}(\widetilde{G}) + E.$$

Hence, $\|\widehat{G} - \widetilde{G}\|_F = \|\widehat{A}^{-1}(\widehat{A}(\widehat{G}) - \widehat{A}(\widetilde{G}))\|_2 \leq \|\widehat{A}^{-1}\|_2\,\|E\|_F$. Together with Lemma 15 and 17, we get the desired bound. □

**Proof** [Proof of Theorem 6] Note that $A : \mathbb{R}^{r\times r\times r} \to \mathbb{R}^{r\times r\times r}$ is a square operator. Moreover, using Lemma 15:

$$P_{\Omega_3}(M_3)\left[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\right] = \widehat{A}(\widetilde{G}) + E.$$

In the case of finite many samples, we use $S_3 = \frac{1}{|\mathcal{S}|}\sum_{t=1+|\mathcal{S}|=2}^{|\mathcal{S}|} x_t \otimes x_t \otimes x_t$ for estimating the low-dimensional tensor $\widetilde{G}$. In particular, we compute the following quantity:

$$\widehat{H} = P_{\Omega_3}(S_3)\left[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\right]. \tag{33}$$

We then use this quantity to solve the least squares problem. That is, we find $\widehat{G}$ as:

$$\widehat{G} = \arg\min_{Z} \|\widehat{A}(Z) - \widehat{H}\|_F^2.$$

Now, we show that such a procedure gives $\widehat{G}$ that is close to $\widetilde{G}$ (see (26)).

$$\begin{aligned}
\|\widehat{G} - \widetilde{G}\|_F &= \|\widehat{A}^{-1}(\widehat{A}(\widehat{G})) - \widehat{A}^{-1}(\widehat{A}(\widetilde{G}))\|_F \\
&= \|\widehat{A}^{-1}(\mathcal{P}_{\Omega_3}(S_3)[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}]) - \widehat{A}^{-1}(\widehat{A}(\widetilde{G}))\|_F \\
&= \|\widehat{A}^{-1}(\mathcal{P}_{\Omega_3}(S_3 - M_3 + M_3)[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}]) - \widehat{A}^{-1}(\widehat{A}(\widetilde{G}))\|_F \\
&\le \|\widehat{A}^{-1}\|_2 \left( \|E\|_F + \|\mathcal{P}_{\Omega_3}(S_3 - M_3)[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}]\,\|_F \right) \\
&\le \|A^{-1}\|_2 \left( \frac{12\mu_1^3\mu r^{3:5}\sigma_1(M_2)^{3=2}\varepsilon}{n\sqrt{w_{\min}}\sigma_r(M_2)^{3=2}} + \|\mathcal{P}_{\Omega_3}(S_3 - M_3)[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}]\,\|_F \right).
\end{aligned}$$

This finishes the proof of the desired claim. □

### A.3. Proof of Lemma 5

Let $E = E^{(1)} - E^{(2)}$ where $E^{(1)} \equiv S_2 - \mathbb{E}[S_2]$, $E^{(2)} \equiv \mathcal{P}_{\Omega_2^c}(S_2 - \mathbb{E}[S_2])$, and $\Omega_2^c$ is the complement of $\Omega_2$. We first note that $\|x_t\|^2 = n$. Hence, applying Matrix Hoeffding bound (see Theorem 1.3 of (Tropp, 2012)), we get with probability at least $1-\delta$:

$$\|E^{(1)}\|_2 = \Big\|\frac{2}{|\mathcal{S}|}\sum_{t\in\{1;:::;|\mathcal{S}|=2\}}(x_t x_t^T) - \mathbb{E}\Big[\frac{2}{|\mathcal{S}|}\sum_{t\in\{1;:::;|\mathcal{S}|=2\}}(x_t x_t^T)\Big]\Big\|_2 \le \sqrt{\frac{32n^2\log(n\ell/\delta)}{|\mathcal{S}|}}.$$

The second term $E^{(2)}$ is a diagonal matrix, with each diagonal entry $E_{ii}^{(2)}$ distributed as a binomial distribution. Applying standard Hoeffding's bound, we get that with probability at least $1-\delta$,

$$\|E^{(2)}\|_2 = \max_{i\in[\grave{}n]}|E_{ii}^{(2)}| \le \sqrt{\frac{2\log(2/\delta)}{|\mathcal{S}|}}.$$

This gives the desired bound on $\|E^{(1)} + E^{(2)}\|_2$.

Similarly, $x_{t;i}\|x_t\|_2 \le \sqrt{n}, \forall i$. Hence, using standard Hoeffding Bound, we get with probability at least $1-\delta$,

$$\Big\|\frac{2}{|\mathcal{S}|}\sum_{t\in[|\mathcal{S}|=2]}(x_t x_t)_i - \mathbb{E}[S_2]_i\Big\|_2 \le \sqrt{\frac{16n\log(2/\delta)}{|\mathcal{S}|}}.$$

### A.4. Proof of Lemma 7

The claim follows form the following lemma.

**Lemma 19** *Let* $H = P_{\Omega_3}(M_3)\left[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}\right]$ *and* $\widehat{H}$ *be as defined above. Then, with probability larger than* $1-\delta$*, we have:*

$$|H_{abc} - \widehat{H}_{abc}| \;\leq\; 2\Big(\frac{2\,r\,n}{\sigma_r(M_2)}\Big)^{3=2}\mu_1^3\sqrt{\frac{\log(1/\delta)}{|\mathcal{S}|}}\,.$$

**Proof** Let $\widehat{H}_{abc} = \frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}} Y^t_{a;b;c}$, where $Y^t_{a;b;c} = \sum_{(i;j;k)\in\Omega_3} x_{t;i}x_{t;j}x_{t;k}\widehat{Q}_{ia}\widehat{Q}_{jb}\widehat{Q}_{kc}$, where $\widehat{Q} = \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{-1=2}$. Then $\mathbb{E}[Y^t] = H$. That is,

$$\begin{aligned}
Y^t_{a;b;c} &= \langle\widehat{Q}_a, x^t\rangle\cdot\langle\widehat{Q}_b, x^t\rangle\cdot\langle\widehat{Q}_c, x^t\rangle - \sum_{m\in[n]}\langle\widehat{Q}_a^{(m)}, (x^t)^{(m)}\rangle\langle\widehat{Q}_b^{(m)}, (x^t)^{(m)}\rangle\langle\widehat{Q}_c^{(m)}, (x^t)^{(m)}\rangle \\
&- \langle\widehat{Q}_a, x^t\rangle\cdot\sum_{m\in[n]}\langle\widehat{Q}_b^{(m)}, (x^t)^{(m)}\rangle\langle\widehat{Q}_c^{(m)}, (x^t)^{(m)}\rangle - \langle\widehat{Q}_b, x^t\rangle\cdot\sum_{m\in[n]}\langle\widehat{Q}_a^{(m)}, (x^t)^{(m)}\rangle\langle\widehat{Q}_c^{(m)}, (x^t)^{(m)}\rangle \\
&- \langle\widehat{Q}_c, x^t\rangle\cdot\sum_{m\in[n]}\langle\widehat{Q}_a^{(m)}, (x^t)^{(m)}\rangle\langle\widehat{Q}_b^{(m)}, (x^t)^{(m)}\rangle. \quad (34)
\end{aligned}$$

Note that, $|\langle\widehat{Q}_b^{(m)}, x_t^{(m)}\rangle| \leq \frac{1\sqrt{r}}{\sqrt{n(1-")\ r(M_2)}}$. Hence, for all $a\in[r]$, $|\langle\widehat{Q}_a, x^t\rangle| \leq \frac{1\sqrt{r\,n}}{\sqrt{(1-")\ r(M_2)}}$.

Using the above inequality with (34), we get: $|Y^t_{a;b;c}| \leq \big(r\,n/((1-\varepsilon)\sigma_r(M_2))\big)^{3=2}\mu_1^3$. Lemma now follows by using Hoeffding's inequality. □

### A.5. Proof of Theorem 1

We first observe that as $U_{M_2} = UR_1$, where $R_1\in\mathbb{R}^{r\times r}$ is an orthonormal matrix. Also, $\Sigma_{M_2} = R_1^T\Sigma V^T W V\Sigma R_1$. Hence, $\Sigma_{M_2}^{1=2} = R_1^T\Sigma V^T W^{1=2}R_3$, where $R_3$ is an orthonormal matrix. Moreover, $\Sigma_{M_2}^{-1=2} = R_3^T W^{-1=2}V\Sigma^{-1}R_1$. Hence,

$$\begin{aligned}
G = M_3[U_{M_2}\Sigma_{M_2}^{-1=2}, U_{M_2}\Sigma_{M_2}^{-1=2}, U_{M_2}\Sigma_{M_2}^{-1=2}] &= \sum_{q=1}^{k} w_q(R_3^T W^{-1=2}\mathbf{e}_q)\otimes(R_3^T W^{-1=2}\mathbf{e}_q)\otimes(R_3^T W^{-1=2}\mathbf{e}_q) \\
&= \sum_{q=1}^{k}\frac{1}{\sqrt{w_q}}(R_3^T\mathbf{e}_q)\otimes(R_3^T\mathbf{e}_q)(R_3^T\mathbf{e}_q). \quad (35)
\end{aligned}$$

Now, using orthogonal tensor decomposition method of (Anandkumar et al., 2012b), we get: $\Lambda^G = W^{-1=2}$ as the eigenvalues and $V^G = R_3^T$ as the eigenvectors. Theorem now follows by observing:

$$U_{M_2}\cdot\Sigma_{M_2}^{1=2}\cdot V^G\cdot\Lambda^G = U_{M_2}\cdot R_1^T\Sigma V^T W^{1=2}R_3\cdot R_3^T\cdot W^{-1=2} = U\Sigma V^T = \Pi.$$

### A.6. Proof of Theorem 2 and Theorem 3

**Proof** [Proof of Theorem 2] Recall that in this case, the number of samples are infinite, i.e., $|\mathcal{S}| = \infty$. Hence, $P_{\Omega_2}(S_2) = P_{\Omega_2}(M_2)$. That is, $E = 0$. Furthermore, $T = \infty$. Hence, using Theorem 4, Algorithm 2 exactly recovers $M_2$, i.e., $\widehat{M}_2^{(T)} = M_2$.

Furthermore, using Theorem 6, we have $\widehat{G} = G$; as, $\varepsilon = \|M_2 - \widehat{M}_2\|_2 = 0$ and $|\mathcal{S}| = \infty$. Now, consider $R_3 R_3^T = \widehat{\Sigma}_{M_2}^{-1=2} \widehat{U}_{M_2}^T \Pi W^{1=2} \cdot W^{1=2} \Pi^T \widehat{U}_{M_2} \widehat{\Sigma}_{M_2}^{-1=2} = \widehat{\Sigma}_{M_2}^{-1=2} \widehat{U}_{M_2}^T M_2 \widehat{U}_{M_2} \widehat{\Sigma}_{M_2}^{-1=2} = I$. That is, $R_3$ is orthonormal. Hence, using orthogonal decomposition method of (Anandkumar et al., 2012b) (see Theorem 20), we get $V^G = R_3$ and $\Lambda^G = W^{-1=2}$. Now, using step 6 of Algorithm 1, $\widehat{\Pi} = \widehat{U}_{M_2} \widehat{U}_{M_2}^T \Pi$. Theorem now follows as $\widehat{U}_{M_2} \widehat{U}_{M_2}^T U = U$ using Remark 13.

Also note that from Theorem 4, $\widehat{M}_2$ is $\mu_1$ incoherent with $\mu_1 = 6\mu\sigma_1(M_2)/\sigma_r(M_2)$.

□

**Proof** [Proof of Theorem 3]

To simplify the notations, we will assume that the permutation that matches the output of our algorithm to the actual types is the identity permutation. Let's define

$$\varepsilon_M \equiv \frac{\|\widehat{M}_2 - M_2\|_2}{\sigma_r(M_2)} \quad \text{and} \quad \varepsilon_G \equiv \|\widehat{G} - \widetilde{G}\|_2 , \tag{36}$$

where $\widehat{G}$ is the output of the TENSORLS and $\widetilde{G} = M_3[\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}, \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}]$.

The spectral algorithm outputs $\widehat{\Pi} = \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}\widehat{V}^G\widehat{\Lambda}^G$, and we know that $\Pi = U_{M_2}\Sigma_{M_2}^{1=2}V^G W^{-1=2}$. In order to show that these two matrices are close, now might hope to prove that each of the terms are close. For example we want $\|U_{M_2} - \widehat{U}_{M_2}\|_2$ to be small. However, even if $U_{M_2}$ and $\widehat{U}_{M_2}$ span the same subspaces the distance might be quite large. Hence, we project $P$ onto the subspace spanned by $\widehat{U}_{M_2}$ to prove the bound we want. Define

$$\widetilde{V} \equiv \widehat{\Sigma}_{M_2}^{-1=2}\widehat{U}_{M_2}^T \Pi W^{1=2} , \tag{37}$$

such that

$$\widetilde{G} = \sum_{i=1}^{r} \frac{1}{\sqrt{w_i}} (\tilde{v}_i \otimes \tilde{v}_i \otimes \tilde{v}_i) , \tag{38}$$

where $\widetilde{V} = [\tilde{v}_1, \ldots, \tilde{v}_r]$. Then, we have $\widehat{U}_{M_2}\Pi = \widehat{\Sigma}_{M_2}^{1=2}\widetilde{V}Q^{-1=2}$. Then,

$$\begin{aligned}
\|\Pi - \widehat{\Pi}\|_2 &\leq \|\widehat{U}_{M_2}\widehat{U}_{M_2}^T\Pi - \Pi\|_2 + \|\widehat{\Pi} - \widehat{U}_{M_2}\widehat{U}_{M_2}^T\Pi\|_2 \\
&= \|(\widehat{U}_{M_2}\widehat{U}_{M_2}^T - \mathbb{I})\Pi\|_2 + \|\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}\widehat{V}^G\widehat{\Lambda}^G - \widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}\widetilde{V}^G W^{-1=2}\|_2 \\
&\leq \|(\widehat{U}_{M_2}\widehat{U}_{M_2}^T - \mathbb{I})\Pi\|_2 + \|\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}(\widehat{V}^G - \widetilde{V})W^{-1=2}\|_2 + \|\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}\widehat{V}^G(\widehat{\Lambda}^G - W^{-1=2})\|_2
\end{aligned} \tag{39}$$

To bound the first term, denote the SVD of $\Pi$ as $\Pi = U\Sigma V^T$. Using Remark 13, $\|\widehat{U}_{M_2}\widehat{U}_{M_2}^T\Pi - \Pi\|_2 \leq \|\widehat{U}_{M_2}\widehat{U}_{M_2}^T U - U\|_2\|\Sigma\|_2 \leq \varepsilon_M \sigma_1(\Pi)$.

Note that $\|\widehat{\Sigma}_{M_2}\|_2 \leq \|\widehat{M}_2 - M_2\| + \|M_2\|_2 \leq \varepsilon_M \sigma_r(M_2) + \|M_2\|_2 \leq 2\|M_2\|_2$, when $\varepsilon_M \leq 1/2$. To prove that the second term is bounded by $C\sqrt{\|M_2\|_2 r w_{\max}/w_{\min}}(\varepsilon_G + (1/\sqrt{w_{\min}})\varepsilon_M)$, we

claim that

$$\begin{aligned}\|\widetilde{V} - \widehat{V}^{G}\|_2 &\leq C\sqrt{r w_{\max}}\Big(\varepsilon_G + \frac{1}{\sqrt{w_{\min}}}\varepsilon_M\Big)\,, \text{ and}\\ \|W^{-1=2} - \widehat{\Lambda}^{G}\|_2 &\leq C\left(\varepsilon_G + \frac{1}{\sqrt{w_{\min}}}\varepsilon_M\right).\end{aligned}$$

Now recall that, $R_3 = \widehat{\Sigma}_{M_2}^{-1=2}\widehat{U}_{M_2}^{T}\Pi W^{1=2}$. Let the SVD of $\widetilde{V}$ be $\widetilde{V} = U_1\Sigma_1 V_1^{T}$. Define an orthogonal matrix $R = U_1 V_1^{T}$, such that $RR^{T} = R^{T}R = \mathbb{I}$. Using Remark 14 we have $\|\widetilde{V} - R\|_2 \leq 2\varepsilon_M$. Moreover, $\widetilde{G} = \sum_{q\in[r]}\frac{1}{\sqrt{w_q}}(R\mathbf{e}_q \otimes R\mathbf{e}_q \otimes R\mathbf{e}_q) + E_G$, where

$$\|E_G\|_2 \leq 2\frac{\varepsilon_M(1+\varepsilon_M)^2}{\sqrt{w_{min}}} \leq \frac{8\varepsilon_M}{\sqrt{w_{min}}}, \tag{40}$$

where, last inequality follows by $\varepsilon_M \leq 1$.

Hence, using (36), (40), we have (w.p. $\geq 1-2\delta$):

$$\|\widehat{G} - \sum_{q\in[r]}\frac{1}{\sqrt{w_q}}(R\mathbf{e}_q \otimes R\mathbf{e}_q \otimes R\mathbf{e}_q)\|_2 \leq \varepsilon_G + \|E_G\|_2 \leq \varepsilon_G + (8/\sqrt{w_{\min}})\varepsilon_M\,. \tag{41}$$

Since $R$ is orthogonal by construction, we can apply Theorem 20 to bound the distance between $\widehat{V}^{G}$ and $R$, i.e. $\|\widehat{V}^{G} - R\|_2 \leq 8\sqrt{r\,w_{\max}}(\varepsilon_G + (8/\sqrt{w_{\min}})\varepsilon_M)$. By triangular inequality, we get that

$$\begin{aligned}\|\widehat{V}^{G} - \widetilde{V}\|_2 &\leq \|\widehat{V}^{G} - R\|_2 + \|R - \widetilde{V}\|_2\\ &\leq 8\sqrt{r\,w_{\max}}\left(\varepsilon_G + \frac{8}{\sqrt{w_{\min}}}\varepsilon_M\right) + 2\varepsilon_M\\ &\leq C\sqrt{r\,w_{\max}}\left(\varepsilon_G + \frac{1}{\sqrt{w_{\min}}}\varepsilon_M\right).\end{aligned}$$

Similarly,

$$\|W^{-1=2} - \widehat{\Lambda}^{G}\|_2 \leq 5\left(\varepsilon_G + \frac{8}{\sqrt{w_{\min}}}\varepsilon_M\right).$$

This implies that the third term in (39) is bounded by $\|\widehat{U}_{M_2}\widehat{\Sigma}_{M_2}^{1=2}\widehat{V}^{G}(\widehat{\Lambda}^{G} - W^{-1=2})\|_2 \leq C\sqrt{\|M_2\|_2}(\varepsilon_G + \varepsilon_M/\sqrt{w_{\min}})$, using the assumption on $|\mathcal{S}|$ such that $(\sqrt{r w_{\max}})\varepsilon_G \leq C$ and $(\sqrt{r w_{\max}/w_{\min}})\varepsilon_M \leq C$.

Putting these bounds together, we get that

$$\|\widehat{\Pi} - \Pi\|_2 \leq C\sqrt{\frac{r\,w_{\max}\,\|M_2\|_2}{w_{\min}}}\left(\varepsilon_G + \frac{1}{\sqrt{w_{\min}}}\varepsilon_M\right),$$

where we used the fact that $\|\Pi\|_2 \leq (1/\sqrt{w_{\min}})\|M_2\|_2^{1=2}$.

From Theorems 4 and 6 and Lemmas 5 and 7, we get that

$$\begin{aligned}\varepsilon_M &\leq C\frac{n\,\|M_2\|_F\,r^{1=2}}{\sigma_r(M_2)^2}\sqrt{\frac{\log(n/\delta)}{|\mathcal{S}|}}\,, \text{ and}\\ \varepsilon_G &\leq C\frac{\mu^4 r^{3:5}}{\sqrt{w_{\min}}}\Big(\frac{\sigma_1(M_2)}{\sigma_r(M_2)}\Big)^{4:5}\frac{1}{n}\varepsilon_M + Cr^3\mu^3\frac{\sigma_1(M_2)^3\,n^{1:5}}{\sigma_r(M_2)^{4:5}}\sqrt{\frac{\log(n/\delta)}{|\mathcal{S}|}}\,,\end{aligned}$$

when $|\mathcal{S}| \geq C'(\ell + r)(n^2/\sigma_r(M_2)^2)\log(n/\delta)$ and $n \geq C'(r^3 + r^{1:5}\mu^2)(\sigma_1(M_2)/\sigma_r(M_2))^2$. Further, if $n \geq C'\mu^4 r^{3:5}(\sigma_1(M_2)/\sigma_r(M_2))^{4:5}$, then

$$\varepsilon_G \leq C\frac{1}{\sqrt{w_{\min}}}\varepsilon_M + Cr^3\mu^3\frac{\sigma_1(M_2)^3\, n^{1:5}}{\sigma_r(M_2)^{4:5}}\sqrt{\frac{\log(n/\delta)}{|\mathcal{S}|}}\,.$$

□

**Theorem 20 (Restatement of Theorem 5.1 by (Anandkumar et al., 2012b))** *Let $G = \sum_{i\in[r]}\lambda_i(v_i \otimes v_i \otimes v_i) + E$, where $\|E\|_2 \leq C_1\frac{w_{\min}}{r}$. Then the tensor power-method after $N \geq C_2(\log r + \log\log\left(\frac{w_{\max}}{\|E\|_2}\right)$, generates vectors $\hat{v}_i, 1 \leq i \leq r$, and $\widehat{\lambda}_i, 1 \leq i \leq r$, s.t.,*

$$\|v_i - \hat{v}_{P(i)}\|_2 \leq 8\|E\|_2/\lambda_{P(i)}, \quad |\lambda_i - \widehat{\lambda}_{P(i)}| \leq 5\|E\|_2. \tag{42}$$

*where $P$ is some permutation on $[r]$.*

### A.7. Proof of Corollary 3.1

Feldman et al. proved that if we have a good estimate of $w_i$'s and $\pi_i$'s in absolute difference, then the thresholding and normalization defined in Section 3 gives a good estimate in KL-divergence.

**Theorem A.1 ((Feldman et al., 2008, Theorem 12))** *Assume $Z$ is a mixture of $r$ product distributions on $\{1,\ldots,\ell\}^n$ with mixing weights $w_1\ldots,w_r$ and probabilities $\pi^j_{i;a}$, and the following are satisfied:*

- *for all $i \in [r]$ we have $|w_i - \hat{w}_i| \leq \varepsilon_w$, and*
- *for all $i \in [r]$ such that $w_i \geq \varepsilon_{\min}$ we have $|\pi^{(j)}_{i;a} - \hat{\pi}^{(j)}_{i;a}| \leq \varepsilon$ for all $j \in [n]$ and $a \in [\ell]$.*

*Then, for sufficiently small $\varepsilon_w$ and $\varepsilon$ , the mixture $\widehat{Z}$ satisfies*

$$D_{\mathrm{KL}}(Z||\widehat{Z}) \leq 12n\ell^3\varepsilon^{1=2} + nk\varepsilon_{\min}\log(\ell/\varepsilon\ ) + \varepsilon_w^{1=3}\,. \tag{43}$$

For the right-hand-side of (43) to be less than $\eta$, it suffices to have $\varepsilon_w = O(\eta^3)$, $\varepsilon = O(\eta^2/n^2\ell^6)$, and $\varepsilon_{\min} = O(\eta/nk\log(\ell/\varepsilon\ ))$.

From Theorem 3, $|\hat{w}_i - w_i| = O(\varepsilon_M)$. Then $\varepsilon_M \leq C\eta^3$ for some positive constant $C$ ensures that the condition is satisfied with $\varepsilon_w = O(\eta^3)$. From Theorem 3, we know that that $|\hat{\pi}^{(j)}_{i;a} - \pi^{(j)}_{i;a}| = O(\varepsilon_M\sqrt{\sigma_1(M_2)w_{\max}r/w_{\min}})$. Then $\varepsilon_M \leq C(\eta^2 w_{\min}^{1=2}/(n^2\ell^6(\sigma_1(M_2)w_{\max}r)^{1=2}))$ for some positive constant $C$ ensures that the condition is satisfied with $\varepsilon = O(\eta^2/n^2\ell^6)$.

These results are true for any values of $w_{\min}$, as long as it is positive. Hence, we have $\varepsilon_{\min} = 0$. It follows that for a choice of

$$\varepsilon_M \leq C\,\eta^2\,\min\left\{\frac{w_{\min}^{1=2}}{n^2\ell^6(\sigma_1(M_2)\,w_{\max}\,r)^{1=2}}, \eta\right\},$$

we have the desired bound on the KL-divergence.

### A.8. Proof of Corollary 3.2

We use a technique similar to those used to analyze distance based clustering algorithms in (Arora and Kannan, 2001; Achlioptas and McSherry, 2005; McSherry, 2001). The clustering algorithm of (Arora and Kannan, 2001) uses $\widehat{\Pi}$ obtained in Algorithm 1 to reduce the dimension of the samples and apply distance based clustering algorithm of (Arora and Kannan, 2001).

Following the analysis of (Arora and Kannan, 2001), we want to identify the conditions such that two samples from the same type are closer than the distance between two samples from two different types. In order to get a large enough gap, we apply $\widehat{\Pi}$ and show that

$$\|\widehat{\Pi}^T(x_i - x_j)\| \;<\; \|\widehat{\Pi}^T(x_i - x_k)\|\,,$$

for all $x_i$ and $x_j$ that belong to the same type and for all $x_k$ with a different type. Then, it is sufficient to show that $\|\widehat{\Pi}^T(\pi_a - \pi_b)\| \geq 4\max_{i\in\mathcal{S}}\|x_i - \mathbb{E}[x_i]\|$ for all $a \neq b \in [r]$. From Theorem 3, we know that for $|\mathcal{S}| \geq C\mu^6 r^7 n^3 \sigma_1(M_2)^7 w_{\max}\log(n/\delta)/(w_{\min}^2\sigma_r(M_2)^9\tilde{\varepsilon}^2)$, $\|\pi_a - \hat{\pi}_a\| \leq \varepsilon_M\sqrt{r w_{\max}\sigma_1(M_2)/w_{\min}} \leq \tilde{\varepsilon}$ for all $a \in [r]$. Then,

$$\begin{aligned}
\|\widehat{\Pi}^T(\pi_a - \pi_b)\| &\geq \|\Pi^T(\pi_a - \pi_b)\| - \|(\Pi - \widehat{\Pi})^T(\pi_a - \pi_b)\| \\
&\geq \sqrt{(\pi_a^T(\pi_a - \pi_b))^2 + (\pi_b^T(\pi_a - \pi_b))^2} - \|\Pi - \widehat{\Pi}\|_2\,\|\pi_a - \pi_b\| \\
&\geq \|\pi_a - \pi_b\|^2 - \sqrt{r}\tilde{\varepsilon}\|\pi_a - \pi_b\|
\end{aligned}$$

On the other hand, applying a concentration of measure inequality gives

$$\mathbb{P}\Big(|\hat{\pi}_a^T(x_i - \mathbb{E}[x_i])| \geq \|\hat{\pi}_a\|\sqrt{2\log(r/\delta)}\Big) \;\leq\; \frac{\delta}{r}\,.$$

Applying union bound, $\|\widehat{\Pi}^T(x_i - \mathbb{E}[x_i])\| \leq \|\widehat{\Pi}\|_F\sqrt{2\log(r/\delta)} \leq (\sqrt{2}\,\|\Pi\|_F + \sqrt{2r}\tilde{\varepsilon})\sqrt{4\log(r/\delta)}$ with probability at least $1-\delta$, where we used the fact that $\|\widehat{\Pi}\|_F^2 \leq \sum_a(\|\pi_a\| + \tilde{\varepsilon})^2 \leq 2\sum_a(\|\pi_a\|^2 + \tilde{\varepsilon}^2) \leq 2(\|\Pi\|_F + \sqrt{r}\,\tilde{\varepsilon})^2$.

For $\tilde{\varepsilon} \geq (\|\pi_a - \pi_b\|^2 - \|\Pi\|_F\sqrt{8\log(r/\delta)})/(\sqrt{r}\|\pi_a - \pi_b\| + \sqrt{8r\log(r/\delta)})$, it follows that $\|\widehat{\Pi}^T(\pi_a - \pi_b)\| \geq 4\max_{i\in\mathcal{S}}\|x_i - \mathbb{E}[x_i]\|$, and this proves that the distance based algorithm of (Arora and Kannan, 2001) will succeed in finding the right clusters for all samples.