Spectral Methods Meet EM: A Provably Optimal Algorithm for Crowdsourcing

Yuchen Zhang[†] **Xi Chen**[♯] **Dengyong Zhou**[*] **Michael I. Jordan**[†]

[†]University of California, Berkeley, Berkeley, CA 94720
{yuczhang,jordan}@berkeley.edu

[♯]New York University, New York, NY 10012
xichen@nyu.edu

[*]Microsoft Research, 1 Microsoft Way, Redmond, WA 98052
dengyong.zhou@microsoft.com

Abstract

The Dawid-Skene estimator has been widely used for inferring the true labels from the noisy labels provided by non-expert crowdsourcing workers. However, since the estimator maximizes a non-convex log-likelihood function, it is hard to theoretically justify its performance. In this paper, we propose a two-stage efficient algorithm for multi-class crowd labeling problems. The first stage uses the spectral method to obtain an initial estimate of parameters. Then the second stage refines the estimation by optimizing the objective function of the Dawid-Skene estimator via the EM algorithm. We show that our algorithm achieves the optimal convergence rate up to a logarithmic factor. We conduct extensive experiments on synthetic and real datasets. Experimental results demonstrate that the proposed algorithm is comparable to the most accurate empirical approach, while outperforming several other recently proposed methods.

1 Introduction

With the advent of online crowdsourcing services such as Amazon Mechanical Turk, crowdsourcing has become an appealing way to collect labels for large-scale data. Although this approach has virtues in terms of scalability and immediate availability, labels collected from the crowd can be of low quality since crowdsourcing workers are often non-experts and can be unreliable. As a remedy, most crowdsourcing services resort to labeling redundancy, collecting multiple labels from different workers for each item. Such a strategy raises a fundamental problem in crowdsourcing: how to infer true labels from noisy but redundant worker labels?

For labeling tasks with k different categories, Dawid and Skene [8] propose a maximum likelihood approach based on the Expectation-Maximization (EM) algorithm. They assume that each worker is associated with a $k \times k$ confusion matrix, where the (l, c)-th entry represents the probability that a randomly chosen item in class l is labeled as class c by the worker. The true labels and worker confusion matrices are jointly estimated by maximizing the likelihood of the observed worker labels, where the unobserved true labels are treated as latent variables. Although this EM-based approach has had empirical success [21, 20, 19, 26, 6, 25], there is as yet no theoretical guarantee for its performance. A recent theoretical study [10] shows that the global optimal solutions of the Dawid-Skene estimator can achieve minimax rates of convergence in a simplified scenario, where the labeling task is binary and each worker has a single parameter to represent her labeling accuracy (referred to as a "one-coin model" in what follows). However, since the likelihood function is non-convex, this guarantee is not operational because the EM algorithm may get trapped in a local optimum. Several alternative approaches have been developed that aim to circumvent the theoretical deficiencies of the

EM algorithm, still in the context of the one-coin model [14, 15, 11, 7]. Unfortunately, they either fail to achieve the optimal rates or depend on restrictive assumptions which are hard to justify in practice.

We propose a computationally efficient and provably optimal algorithm to simultaneously estimate true labels and worker confusion matrices for multi-class labeling problems. Our approach is a two-stage procedure, in which we first compute an initial estimate of worker confusion matrices using the spectral method, and then in the second stage we turn to the EM algorithm. Under some mild conditions, we show that this two-stage procedure achieves minimax rates of convergence up to a logarithmic factor, even after only one iteration of EM. In particular, given any $\delta \in (0,1)$, we provide the bounds on the number of workers and the number of items so that our method can correctly estimate labels for all items with probability at least $1-\delta$. We also establish a lower bound to demonstrate the optimality of this approach. Further, we provide both upper and lower bounds for estimating the confusion matrix of each worker and show that our algorithm achieves the optimal accuracy.

This work not only provides an optimal algorithm for crowdsourcing but sheds light on understanding the general method of moments. Empirical studies show that when the spectral method is used as an initialization for the EM algorithm, it outperforms EM with random initialization [18, 5]. This work provides a concrete way to theoretically justify such observations. It is also known that starting from a root-n consistent estimator obtained by the spectral method, one Newton-Raphson step leads to an asymptotically optimal estimator [17]. However, obtaining a root-n consistent estimator and performing a Newton-Raphson step can be demanding computationally. In contrast, our initialization doesn't need to be root-n consistent, thus a small portion of data suffices to initialize. Moreover, performing one iteration of EM is computationally more attractive and numerically more robust than a Newton-Raphson step especially for high-dimensional problems.

2 Related Work

Many methods have been proposed to address the problem of estimating true labels in crowdsourcing [23, 20, 22, 11, 19, 26, 7, 15, 14, 25]. The methods in [20, 11, 15, 19, 14, 7] are based on the generative model proposed by Dawid and Skene [8]. In particular, Ghosh et al. [11] propose a method based on Singular Value Decomposition (SVD) which addresses binary labeling problems under the one-coin model. The analysis in [11] assumes that the labeling matrix is full, that is, each worker labels all items. To relax this assumption, Dalvi et al. [7] propose another SVD-based algorithm which explicitly considers the sparsity of the labeling matrix in both algorithm design and theoretical analysis. Karger et al. propose an iterative algorithm for binary labeling problems under the one-coin model [15] and extend it to multi-class labeling tasks by converting a k-class problem into $k-1$ binary problems [14]. This line of work assumes that tasks are assigned to workers according to a random regular graph, thus imposing specific constraints on the number of workers and the number of items. In Section 5, we compare our theoretical results with that of existing approaches [11, 7, 15, 14]. The methods in [20, 19, 6] incorporate Bayesian inference into the Dawid-Skene estimator by assuming a prior over confusion matrices. Zhou et al. [26, 25] propose a minimax entropy principle for crowdsourcing which leads to an exponential family model parameterized with worker ability and item difficulty. When all items have zero difficulty, the exponential family model reduces to the generative model suggested by Dawid and Skene [8].

Our method for initializing the EM algorithm in crowdsourcing is inspired by recent work using spectral methods to estimate latent variable models [3, 1, 4, 2, 5, 27, 12, 13]. The basic idea in this line of work is to compute third-order empirical moments from the data and then to estimate parameters by computing a certain orthogonal decomposition of a tensor derived from the moments. Given the special symmetric structure of the moments, the tensor factorization can be computed efficiently using the robust tensor power method [3]. A problem with this approach is that the estimation error can have a poor dependence on the condition number of the second-order moment matrix and thus empirically it sometimes performs worse than EM with multiple random initializations. Our method, by contrast, requires only a rough initialization from the moment of moments; we show that the estimation error does not depend on the condition number (see Theorem 2 (b)).

Algorithm 1: Estimating confusion matrices

```
Input: integer $k$, observed labels $z_{ij} \in \mathbb{R}^k$ for $i \in [m]$ and $j \in [n]$.
Output: confusion matrix estimates $\widehat{C}_i \in \mathbb{R}^{k\times k}$ for $i \in [m]$.
(1) Partition the workers into three disjoint and non-empty group $G_1$, $G_2$ and $G_3$. Compute the
    group aggregated labels $Z_{gj}$ by Eq. (1).
(2) For $(a,b,c) \in \{(2,3,1),(3,1,2),(1,2,3)\}$, compute the second and the third order moments
    $\widehat{M_2} \in \mathbb{R}^{k\times k}$, $\widehat{M_3} \in \mathbb{R}^{k\times k\times k}$ by Eq. (2a)-(2d), then compute $\widehat{C}_c^\diamond \in \mathbb{R}^{k\times k}$ and $\widehat{W} \in \mathbb{R}^{k\times k}$ by
    tensor decomposition:
    (a) Compute whitening matrix $\widehat{Q} \in \mathbb{R}^{k\times k}$ (such that $\widehat{Q}^T \widehat{M_2} \widehat{Q} = I$) using SVD.
    (b) Compute eigenvalue-eigenvector pairs $\{(\widehat{\alpha}_h, \widehat{v}_h)\}_{h=1}^k$ of the whitened tensor $\widehat{M_3}(\widehat{Q},\widehat{Q},\widehat{Q})$
        by using the robust tensor power method [3]. Then compute $\widehat{w}_h = \widehat{\alpha}_h^{-2}$ and
        $\widehat{\mu}_h^\diamond = (\widehat{Q}^T)^{-1}(\widehat{\alpha}_h \widehat{v}_h)$.
    (c) For $l = 1,\ldots,k$, set the $l$-th column of $\widehat{C}_c^\diamond$ by some $\widehat{\mu}_h^\diamond$ whose $l$-th coordinate has the
        greatest component, then set the $l$-th diagonal entry of $\widehat{W}$ by $\widehat{w}_h$.
(3) Compute $\widehat{C}_i$ by Eq. (3).
```

3 Problem Setup

Throughout this paper, $[a]$ denotes the integer set $\{1,2,\ldots,a\}$ and $\sigma_b(A)$ denotes the b-th largest singular value of the matrix A. Suppose that there are m workers, n items and k classes. The true label y_j of item $j \in [n]$ is assumed to be sampled from a probability distribution $\mathbb{P}[y_j = l] = w_l$ where $\{w_l : l \in [k]\}$ are positive values satisfying $\sum_{l=1}^k w_l = 1$. Denote by a vector $z_{ij} \in \mathbb{R}^k$ the label that worker i assigns to item j. When the assigned label is c, we write $z_{ij} = e_c$, where e_c represents the c-th canonical basis vector in $\mathbb{R}^k$ in which the c-th entry is 1 and all other entries are 0. A worker may not label every item. Let π_i indicate the probability that worker i labels a randomly chosen item. If item j is not labeled by worker i, we write $z_{ij} = 0$. Our goal is to estimate the true labels $\{y_j : j \in [n]\}$ from the observed labels $\{z_{ij} : i \in [m], j \in [n]\}$.

In order to obtain an estimator, we need to make assumptions on the process of generating observed labels. Following the work of Dawid and Skene [8], we assume that the probability that worker i labels an item in class l as class c is independent of any particular chosen item, that is, it is a constant over $j \in [n]$. Let us denote the constant probability by μ_{ilc}. Let $\mu_{il} = [\mu_{il1}\ \mu_{il2}\ \cdots\ \mu_{ilk}]^T$. The matrix $C_i = [\mu_{i1}\ \mu_{i2}\ \ldots\ \mu_{ik}] \in \mathbb{R}^{k\times k}$ is called the *confusion matrix* of worker i. Besides estimating the true labels, we also want to estimate the confusion matrix for each worker.

4 Our Algorithm

In this section, we present an algorithm to estimate confusion matrices and true labels. Our algorithm consists of two stages. In the first stage, we compute an initial estimate of confusion matrices via the method of moments. In the second stage, we perform the standard EM algorithm by taking the result of the Stage 1 as an initialization.

4.1 Stage 1: Estimating Confusion Matrices

Partitioning the workers into three disjoint and non-empty groups G_1, G_2 and G_3, the outline of this stage is the following: we use the spectral method to estimate the averaged confusion matrices for the three groups, then utilize this intermediate estimate to obtain the confusion matrix of each individual worker. In particular, for $g \in \{1,2,3\}$ and $j \in [n]$, we calculate the averaged labeling within each group by

$$Z_{gj} := \frac{1}{|G_g|} \sum_{i \in G_g} z_{ij}. \tag{1}$$

Denoting the aggregated confusion matrix columns by $\mu^\diamond_{gl} := \mathbb{E}(Z_{gj}|y_j = l) = \frac{1}{|G_g|}\sum_{i\in G_g} \pi_i\mu_{il}$, our first step is to estimate $C^\diamond_g := [\mu^\diamond_{g1}, \mu^\diamond_{g2}, \dots, \mu^\diamond_{gk}]$ and to estimate the distribution of true labels $W := \mathrm{diag}(w_1, w_2, \dots, w_k)$. The following proposition shows that we can solve for $C^\diamond_g$ and W from the moments of $\{Z_{gj}\}$.

Proposition 1 (Anandkumar et al. [3]). *Assume that the vectors $\{\mu^\diamond_{g1}, \mu^\diamond_{g2}, \dots, \mu^\diamond_{gk}\}$ are linearly independent for each $g \in \{1, 2, 3\}$. Let (a, b, c) be a permutation of $\{1, 2, 3\}$. Define*

$$\begin{aligned}
Z'_{aj} &:= \mathbb{E}[Z_{cj} \otimes Z_{bj}] \left(\mathbb{E}[Z_{aj} \otimes Z_{bj}]\right)^{-1} Z_{aj},\\
Z'_{bj} &:= \mathbb{E}[Z_{cj} \otimes Z_{aj}] \left(\mathbb{E}[Z_{bj} \otimes Z_{aj}]\right)^{-1} Z_{bj},\\
M_2 &:= \mathbb{E}[Z'_{aj} \otimes Z'_{bj}] \quad \textit{and} \quad M_3 := \mathbb{E}[Z'_{aj} \otimes Z'_{bj} \otimes Z_{cj}];
\end{aligned}$$

then we have $M_2 = \sum_{l=1}^k w_l\, \mu^\diamond_{cl} \otimes \mu^\diamond_{cl}$ and $M_3 = \sum_{l=1}^k w_l\, \mu^\diamond_{cl} \otimes \mu^\diamond_{cl} \otimes \mu^\diamond_{cl}$.

Since we only have finite samples, the expectations in Proposition 1 have to be approximated by empirical moments. In particular, they are computed by averaging over indices $j = 1, 2, \dots, n$. For each permutation $(a, b, c) \in \{(2, 3, 1), (3, 1, 2), (1, 2, 3)\}$, we compute

$$\widehat{Z}'_{aj} := \Big(\frac{1}{n}\sum_{j=1}^n Z_{cj} \otimes Z_{bj}\Big)\Big(\frac{1}{n}\sum_{j=1}^n Z_{aj} \otimes Z_{bj}\Big)^{-1} Z_{aj}, \tag{2a}$$

$$\widehat{Z}'_{bj} := \Big(\frac{1}{n}\sum_{j=1}^n Z_{cj} \otimes Z_{aj}\Big)\Big(\frac{1}{n}\sum_{j=1}^n Z_{bj} \otimes Z_{aj}\Big)^{-1} Z_{bj}, \tag{2b}$$

$$\widehat{M}_2 := \frac{1}{n}\sum_{j=1}^n \widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj}, \tag{2c}$$

$$\widehat{M}_3 := \frac{1}{n}\sum_{j=1}^n \widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} \otimes Z_{cj}. \tag{2d}$$

The statement of Proposition 1 suggests that we can recover the columns of $C^\diamond_c$ and the diagonal entries of W by operating on the moments $\widehat{M}_2$ and $\widehat{M}_3$. This is implemented by the tensor factorization method in Algorithm 1. In particular, the tensor factorization algorithm returns a set of vectors $\{(\widehat{\mu}^\diamond_h, \widehat{w}_h) : h = 1, \dots, k\}$, where each $(\widehat{\mu}^\diamond_h, \widehat{w}_h)$ estimates a particular column of $C^\diamond_c$ (for some $\mu^\diamond_{cl}$) and a particular diagonal entry of W (for some w_l). It is important to note that the tensor factorization algorithm doesn't provide a one-to-one correspondence between the recovered column and the true columns of $C^\diamond_c$. Thus, $\widehat{\mu}^\diamond_1, \dots, \widehat{\mu}^\diamond_k$ represents an arbitrary permutation of the true columns.

To discover the index correspondence, we take each $\widehat{\mu}^\diamond_h$ and examine its greatest component. We assume that within each group, the probability of assigning a correct label is always greater than the probability of assigning any specific incorrect label. This assumption will be made precise in the next section. As a consequence, if $\widehat{\mu}^\diamond_h$ corresponds to the l-th column of $C^\diamond_c$, then its l-th coordinate is expected to be greater than other coordinates. Thus, we set the l-th column of $\widehat{C}^\diamond_c$ to some vector $\widehat{\mu}^\diamond_h$ whose l-th coordinate has the greatest component (if there are multiple such vectors, then randomly select one of them; if there is no such vector, then randomly select a $\widehat{\mu}^\diamond_h$). Then, we set the l-th diagonal entry of $\widehat{W}$ to the scalar $\widehat{w}_h$ associated with $\widehat{\mu}^\diamond_h$. Note that by iterating over $(a, b, c) \in \{(2, 3, 1), (3, 1, 2), (1, 2, 3)\}$, we obtain $\widehat{C}^\diamond_c$ for $c = 1, 2, 3$ respectively. There will be three copies of $\widehat{W}$ estimating the same matrix W—we average them for the best accuracy.

In the second step, we estimate each individual confusion matrix C_i. The following proposition shows that we can recover C_i from the moments of $\{z_{ij}\}$. Its proof is deferred to Appendix B.

Proposition 2. *For any $g \in \{1, 2, 3\}$ and any $i \in G_g$, let $a \in \{1, 2, 3\}\backslash\{g\}$ be one of the remaining group index. Then*

$$\pi_i C_i W (C^\diamond_a)^T = \mathbb{E}[z_{ij} Z^T_{aj}].$$

Proposition 2 suggests a plug-in estimator for C_i. We compute $\widehat{C}_i$ using the empirical approximation of $\mathbb{E}[z_{ij} Z_{aj}^T]$ and using the matrices $\widehat{C}_a^\diamond$, $\widehat{C}_b^\diamond$, $\widehat{W}$ obtained in the first step. Concretely, we calculate

$$\widehat{C}_i := \text{normalize}\left\{ \Big(\frac{1}{n}\sum_{j=1}^{n} z_{ij} Z_{aj}^T\Big)\Big(\widehat{W}(\widehat{C}_a^\diamond)^T\Big)^{-1} \right\}, \tag{3}$$

where the normalization operator rescales the matrix columns, making sure that each column sums to one. The overall procedure for Stage 1 is summarized in Algorithm 1.

4.2 Stage 2: EM algorithm

The second stage is devoted to refining the initial estimate provided by Stage 1. The joint likelihood of true label y_j and observed labels z_{ij}, as a function of confusion matrices μ_i, can be written as

$$L(\mu; y, z) := \prod_{j=1}^{n}\prod_{i=1}^{m}\prod_{c=1}^{k} (\mu_{iy_jc})^{\mathbb{I}(z_{ij}=e_c)}.$$

By assuming a uniform prior over y, we maximize the marginal log-likelihood function $\ell(\mu) := \log(\sum_{y\in[k]^n} L(\mu; y, z))$. We refine the initial estimate of Stage 1 by maximizing the objective function, which is implemented by the Expectation Maximization (EM) algorithm. The EM algorithm takes the values $\{\widehat{\mu}_{ilc}\}$ provided as output by Stage 1 as initialization, then executes the following E-step and M-step *for at least one round.*

E-step Calculate the expected value of the log-likelihood function, with respect to the conditional distribution of y given z under the current estimate of μ:

$$Q(\mu) := \mathbb{E}_{y|zf,\widehat{\mu}}\left[\log(L(\mu; y, z))\right] = \sum_{j=1}^{n}\left\{\sum_{l=1}^{k} \widehat{q}_{jl} \log\left(\prod_{i=1}^{m}\prod_{c=1}^{k} (\mu_{ilc})^{\mathbb{I}(z_{ij}=e_c)}\right)\right\},$$

$$\text{where} \quad \widehat{q}_{jl} \leftarrow \frac{\exp\big(\sum_{i=1}^{m}\sum_{c=1}^{k} \mathbb{I}(z_{ij}=e_c)\log(\widehat{\mu}_{ilc})\big)}{\sum_{l'=1}^{k}\exp\big(\sum_{i=1}^{m}\sum_{c=1}^{k} \mathbb{I}(z_{ij}=e_c)\log(\widehat{\mu}_{il'c})\big)} \qquad \text{for } j\in[n],\, l\in[k]. \tag{4}$$

M-step Find the estimate $\widehat{\mu}$ that maximizes the function $Q(\mu)$:

$$\widehat{\mu}_{ilc} \leftarrow \frac{\sum_{j=1}^{n} \widehat{q}_{jl}\mathbb{I}(z_{ij}=e_c)}{\sum_{c'=1}^{k}\sum_{j=1}^{n} \widehat{q}_{jl}\mathbb{I}(z_{ij}=e_{c'})} \qquad \text{for } i\in[m],\, l\in[k],\, c\in[k]. \tag{5}$$

In practice, we alternatively execute the updates (4) and (5), for one iteration or until convergence. Each update increases the objective function $\ell(\mu)$. Since $\ell(\mu)$ is not concave, the EM update doesn't guarantee converging to the global maximum. It may converge to distinct local stationary points for different initializations. Nevertheless, as we prove in the next section, it is guaranteed that the EM algorithm will output statistically optimal estimates of true labels and worker confusion matrices if it is initialized by Algorithm 1.

5 Convergence Analysis

To state our main theoretical results, we first need to introduce some notation and assumptions. Let

$$w_{\min} := \min\{w_l\}_{l=1}^{k} \quad \text{and} \quad \pi_{\min} := \min\{\pi_i\}_{i=1}^{m}$$

be the smallest portion of true labels and the most extreme sparsity level of workers. Our first assumption assumes that both $w_{\min}$ and $\pi_{\min}$ are strictly positive, that is, every class and every worker contributes to the dataset.

Our second assumption assumes that the confusion matrices for each of the three groups, namely $C_1^\diamond$, $C_2^\diamond$ and $C_3^\diamond$, are nonsingular. As a consequence, if we define matrices S_{ab} and tensors T_{abc} for any $a, b, c \in \{1, 2, 3\}$ as

$$S_{ab} := \sum_{l=1}^{k} w_l\, \mu_{al}^\diamond \otimes \mu_{bl}^\diamond = C_a^\diamond W (C_b^\diamond)^T \quad \text{and} \quad T_{abc} := \sum_{l=1}^{k} w_l\, \mu_{al}^\diamond \otimes \mu_{bl}^\diamond \otimes \mu_{cl}^\diamond,$$

then there will be a positive scalar σ_L such that $\sigma_k(S_{ab}) \geq \sigma_L > 0$.

Our third assumption assumes that within each group, the average probability of assigning a correct label is always higher than the average probability of assigning any incorrect label. To make this statement rigorous, we define a quantity

$$\kappa := \min_{g \in \{1,2,3\}} \min_{l \in [k]} \min_{c \in [k] \setminus \{l\}} \{\mu^{\diamond}_{gll} - \mu^{\diamond}_{glc}\}$$

indicating the smallest gap between diagonal entries and non-diagonal entries in the same confusion matrix column. The assumption requires κ being strictly positive. Note that this assumption is group-based, thus does not assume the accuracy of any individual worker.

Finally, we introduce a quantity that measures the average ability of workers in identifying distinct labels. For two discrete distributions P and Q, let $\mathbb{D}_{\mathrm{KL}}\left(P, Q\right) := \sum_i P(i) \log(P(i)/Q(i))$ represent the KL-divergence between P and Q. Since each column of the confusion matrix represents a discrete distribution, we can define the following quantity:

$$\overline{D} = \min_{l \neq l'} \frac{1}{m} \sum_{i=1}^{m} \pi_i \mathbb{D}_{\mathrm{KL}}\left(\mu_{il}, \mu_{il'}\right). \tag{6}$$

The quantity $\overline{D}$ lower bounds the averaged KL-divergence between two columns. If $\overline{D}$ is strictly positive, it means that every pair of labels can be distinguished by at least one subset of workers. As the last assumption, we assume that $\overline{D}$ is strictly positive.

The following two theorems characterize the performance of our algorithm. We split the convergence analysis into two parts. Theorem 1 characterizes the performance of Algorithm 1, providing sufficient conditions for achieving an arbitrarily accurate initialization. See Appendix C for the proof.

Theorem 1. *For any scalar $\delta > 0$ and any scalar ϵ satisfying $\epsilon \leq \min\left\{\frac{36\kappa k}{\pi_{\min} w_{\min} \sigma_L}, 2\right\}$, if the number of items n satisfies*

$$n = \Omega\left(\frac{k^5 \log((k+m)/\delta)}{\epsilon^2 \pi_{\min}^2 w_{\min}^2 \sigma_L^{13}}\right),$$

then the confusion matrices returned by Algorithm 1 are bounded as

$$\|\widehat{C}_i - C_i\|_\infty \leq \epsilon \qquad \textit{for all } i \in [m],$$

with probability at least $1 - \delta$. Here, $\|\cdot\|_\infty$ denotes the element-wise ℓ_∞-norm of a matrix.

Theorem 2 characterizes the error rate in Stage 2. It states that when a sufficiently accurate initialization is taken, the updates (4) and (5) refine the estimates $\widehat{\mu}$ and $\widehat{y}$ to the optimal accuracy. See Appendix D for the proof.

Theorem 2. *Assume that there is a positive scalar ρ such that $\mu_{ilc} \geq \rho$ for all $(i, l, c) \in [m] \times [k]^2$. For any scalar $\delta > 0$, if confusion matrices $\widehat{C}_i$ are initialized in a manner such that*

$$\|\widehat{C}_i - C_i\|_\infty \leq \alpha := \min\left\{\frac{\rho}{2}, \frac{\rho \overline{D}}{16}\right\} \qquad \textit{for all } i \in [m], \tag{7}$$

and the number of workers m and the number of items n satisfy

$$m = \Omega\left(\frac{\log(1/\rho)\log(kn/\delta) + \log(mn)}{\overline{D}}\right) \quad \textit{and} \quad n = \Omega\left(\frac{\log(mk/\delta)}{\pi_{\min} w_{\min} \alpha^2}\right),$$

then, for $\widehat{\mu}$ and $\widehat{q}$ obtained by iterating (4) and (5) (for at least one round), with probability at least $1 - \delta$,

(a) Letting $\widehat{y}_j = \arg\max_{l \in [k]} \widehat{q}_{jl}$, we have that $\widehat{y}_j = y_j$ holds for all $j \in [n]$.

(b) $\|\widehat{\mu}_{il} - \mu_{il}\|_2^2 \leq \frac{48 \log(2mk/\delta)}{\pi_i w_l n}$ holds for all $(i, l) \in [m] \times [k]$.

In Theorem 2, the assumption that all confusion matrix entries are lower bounded by $\rho > 0$ is somewhat restrictive. For datasets violating this assumption, we enforce positive confusion matrix entries by adding random noise: Given any observed label z_{ij}, we replace it by a random label in $\{1, ..., k\}$ with probability $k\rho$. In this modified model, every entry of the confusion matrix is lower

Dataset name	# classes	# items	# workers	# worker labels
Bird	2	108	39	4,212
RTE	2	800	164	8,000
TREC	2	19,033	762	88,385
Dog	4	807	52	7,354
Web	5	2,665	177	15,567

Table 1: Summary of datasets used in the real data experiment.

bounded by ρ, so that Theorem 2 holds. The random noise makes the constant $\overline{D}$ smaller than its original value, but the change is minor for small ρ.

To see the consequence of the convergence analysis, we take error rate ϵ in Theorem 1 equal to the constant α defined in Theorem 2. Then we combine the statements of the two theorems. This shows that if we choose the number of workers m and the number of items n such that

$$m = \widetilde{\Omega}\left(\frac{1}{\overline{D}}\right) \quad \text{and} \quad n = \widetilde{\Omega}\left(\frac{k^5}{\pi_{\min}^2 w_{\min}^2 \sigma_L^{13} \min\{\rho^2, (\rho\overline{D})^2\}}\right); \tag{8}$$

that is, if both m and n are lower bounded by a problem-specific constant and logarithmic terms, then with high probability, the predictor $\widehat{y}$ will be perfectly accurate, and the estimator $\widehat{\mu}$ will be bounded as $\|\widehat{\mu}_{il} - \mu_{il}\|_2^2 \leq \widetilde{\mathcal{O}}(1/(\pi_i w_l n))$. To show the optimality of this convergence rate, we present the following minimax lower bounds. See Appendix E for the proof.

Theorem 3. *There are universal constants $c_1 > 0$ and $c_2 > 0$ such that:*

(a) For any $\{\mu_{ilc}\}$, $\{\pi_i\}$ and any number of items n, if the number of workers $m \leq 1/(4\overline{D})$, then

$$\inf_{\widehat{y}} \sup_{v \in [k]^n} \mathbb{E}\Big[\sum_{j=1}^{n} \mathbb{I}(\widehat{y}_j \neq y_j)\Big| \{\mu_{ilc}\}, \{\pi_i\}, y = v\Big] \geq c_1 n.$$

(b) For any $\{w_l\}$, $\{\pi_i\}$, any worker-item pair (m, n) and any pair of indices $(i, l) \in [m] \times [k]$, we have

$$\inf_{\widehat{\mu}} \sup_{\mu \in \mathbb{R}^{m \times k \times k}} \mathbb{E}\Big[\|\widehat{\mu}_{il} - \mu_{il}\|_2^2 \Big| \{w_l\}, \{\pi_i\}\Big] \geq c_2 \, \min\left\{1, \frac{1}{\pi_i w_l n}\right\}.$$

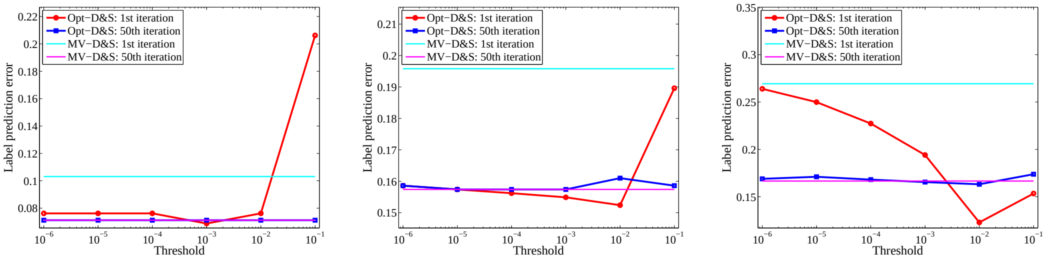
In part (a) of Theorem 3, we see that the number of workers should be at least $1/(4\overline{D})$, otherwise any predictor will make many mistakes. This lower bound matches our sufficient condition on the number of workers m (see Eq. (8)). In part (b), we see that the best possible estimate for μ_{il} has $\Omega(1/(\pi_i w_l n))$ mean-squared error. It verifies the optimality of our estimator $\widehat{\mu}_{il}$. It is worth noting that the constraint on the number of items n (see Eq. (8)) might be improvable. In real datasets we usually have $n \gg m$ so that the optimality for m is more important than for n.

It is worth contrasting our convergence rate with existing algorithms. Ghosh et al. [11] and Dalvi et al. [7] proposed consistent estimators for the binary one-coin model. To attain an error rate δ, their algorithms require m and n scaling with $1/\delta^2$, while our algorithm only requires m and n scaling with $\log(1/\delta)$. Karger et al. [15, 14] proposed algorithms for both binary and multi-class problems. Their algorithm assumes that workers are assigned by a random regular graph. Moreover, their analysis assumes that the limit of number of items goes to infinity, or that the number of workers is many times the number of items. Our algorithm no longer requires these assumptions.

We also compare our algorithm with the majority voting estimator, where the true label is simply estimated by a majority vote among workers. Gao and Zhou [10] showed that if there are many spammers and few experts, the majority voting estimator gives almost a random guess. In contrast, our algorithm only requires $m\overline{D} = \widetilde{\Omega}(1)$ to guarantee good performance. Since $m\overline{D}$ is the aggregated KL-divergence, a small number of experts are sufficient to ensure it is large enough.

6 Experiments

In this section, we report the results of empirical studies comparing the algorithm we propose in Section 4 (referred to as Opt-D&S) with a variety of existing methods which are also based on the



(a) RTE (b) Dog (c) Web

Figure 1: Comparing MV-D&S and Opt-D&S with different thresholding parameter Δ. The label prediction error is plotted after the 1st EM update and after convergence.

	Opt-D&S	MV-D&S	Majority Voting	KOS	Ghosh-SVD	EigenRatio
Bird	**10.09**	11.11	24.07	11.11	27.78	27.78
RTE	**7.12**	**7.12**	10.31	39.75	49.13	9.00
TREC	**29.80**	30.02	34.86	51.96	42.99	43.96
Dog	16.89	**16.66**	19.58	31.72	–	–
Web	15.86	**15.74**	26.93	42.93	–	–

Table 2: Error rate (%) in predicting true labels on real data.

generative model of Dawid and Skene. Specifically, we compare to the Dawid & Skene estimator initialized by majority voting (referred to as MV-D&S), the pure majority voting estimator, the multi-class labeling algorithm proposed by Karger et al. [14] (referred to as KOS), the SVD-based algorithm proposed by Ghosh et al. [11] (referred to as Ghost-SVD) and the "Eigenvalues of Ratio" algorithm proposed by Dalvi et al. [7] (referred to as EigenRatio). The evaluation is made on five real datasets. See also Appendix A for experiments on synthetic data, where we show that Opt-D&S converges faster than MV-D&S.

We compare the crowdsourcing algorithms on three binary tasks and two multi-class tasks. Binary tasks include labeling bird species [22] (Bird dataset), recognizing textual entailment [21] (RTE dataset) and assessing the quality of documents in the TREC 2011 crowdsourcing track [16] (TREC dataset). Multi-class tasks include labeling the breed of dogs from ImageNet [9] (Dog dataset) and judging the relevance of web search results [26] (Web dataset). The statistics for the five datasets are summarized in Table 1. Since the Ghost-SVD algorithm and the EigenRatio algorithm work on binary tasks, they are evaluated only on the Bird, RTE and TREC datasets. For the MV-D&S and the Opt-D&S methods, we iterate their EM steps until convergence.

Since entries of the confusion matrix are positive, we find it helpful to incorporate this prior knowledge into the initialization stage of the Opt-D&S algorithm. In particular, when estimating the confusion matrix entries by Eq. (3), we add an extra checking step before the normalization, examining if the matrix components are greater than or equal to a small threshold Δ. For components that are smaller than Δ, they are reset to Δ. The default choice of the thresholding parameter is $\Delta = 10^{-6}$. Later, we will compare the Opt-D&S algorithm with respect to different choices of Δ. It is important to note that this modification doesn't change our theoretical result, since the thresholding is not needed in case that the initialization error is bounded by Theorem 1.

Table 2 summarizes the performance of each method. The MV-D&S and the Opt-D&S algorithms consistently outperform the other methods in predicting the true label of items. The KOS algorithm, the Ghost-SVD algorithm and the EigenRatio algorithm yield poorer performance, presumably due to the fact that they rely on idealized assumptions that are not met by the real data. In Figure 1, we compare the Opt-D&S algorithm with respect to different thresholding parameters $\Delta \in \{10^{-i}\}_{i=1}^{6}$. We plot results for three datasets (RET, Dog, Web), where the performance of MV-D&S is equal to or slightly better than that of Opt-D&S. The plot shows that the performance of the Opt-D&S algorithm is stable after convergence. But at the first EM iterate, the error rates are more sensitive to the choice of Δ. A proper choice of Δ makes Opt-D&S outperform MV-D&S. The result suggests that a proper initialization combined with one EM iterate is good enough for the purposes of prediction. In practice, the best choice of Δ can be obtained by cross validation.

References

[1] A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y.-K. Liu. A spectral algorithm for latent Dirichlet allocation. *arXiv preprint: 1204.6703*, 2012.

[2] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A tensor spectral approach to learning mixed membership community models. In *Annual Conference on Learning Theory*, 2013.

[3] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *arXiv preprint:1210.7559*, 2012.

[4] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. In *Annual Conference on Learning Theory*, 2012.

[5] A. T. Chaganty and P. Liang. Spectral experts for estimating mixtures of linear regressions. *arXiv preprint: 1306.3729*, 2013.

[6] X. Chen, Q. Lin, and D. Zhou. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *Proceedings of the 30th International Conferences on Machine Learning*, 2013.

[7] N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of World Wide Web Conference*, 2013.

[8] 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*, pages 20–28, 1979.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, 2009.

[10] C. Gao and D. Zhou. Minimax optimal convergence rates for estimating ground truth from crowdsourced labels. *arXiv preprint arXiv:1310.5764*, 2014.

[11] A. Ghosh, S. Kale, and P. McAfee. Who moderates the moderators?: Crowdsourcing abuse detection in user-generated content. In *Proceedings of the ACM Conference on Electronic Commerce*, 2011.

[12] 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.

[13] P. Jain and S. Oh. Learning mixtures of discrete product distributions using spectral decompositions. *arXiv preprint:1311.2972*, 2013.

[14] D. R. Karger, S. Oh, and D. Shah. Efficient crowdsourcing for multi-class labeling. In *ACM SIGMETRICS*, 2013.

[15] D. R. Karger, S. Oh, and D. Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

[16] M. Lease and G. Kazai. Overview of the TREC 2011 crowdsourcing track. In *Proceedings of TREC 2011*, 2011.

[17] E. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 2nd edition, 2003.

[18] P. Liang. Partial information from spectral methods. NIPS Spectral Learning Workshop, 2013.

[19] Q. Liu, J. Peng, and A. T. Ihler. Variational inference for crowdsourcing. In *NIPS*, 2012.

[20] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[21] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, 2008.

[22] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *NIPS*, 2010.

[23] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, 2009.

[24] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

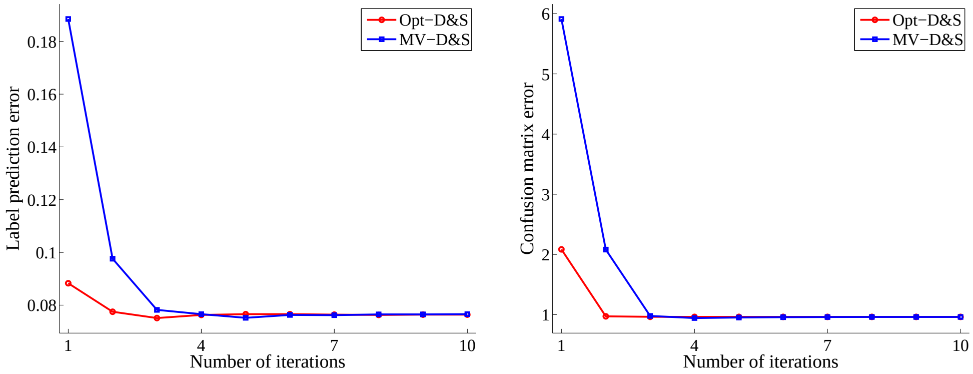
[25] D. Zhou, Q. Liu, J. C. Platt, and C. Meek. Aggregating ordinal labels from crowds by minimax conditional entropy. In *Proceedings of ICML*, 2014.

[26] D. Zhou, J. C. Platt, S. Basu, and Y. Mao. Learning from the wisdom of crowds by minimax entropy. In *NIPS*, 2012.

[27] J. Zou, D. Hsu, D. Parkes, and R. Adams. Contrastive learning using spectral methods. In *NIPS*, 2013.

	Opt-D&S	MV-D&S	Majority Voting	KOS	Ghosh-SVD	EigenRatio
$\pi = 0.2$	7.64	7.65	18.85	8.34	12.35	10.49
$\pi = 0.5$	0.84	0.84	7.97	1.04	4.52	4.52
$\pi = 1.0$	0.01	0.01	1.57	0.02	0.15	0.15

Table 3: Prediction error (%) on the synthetic dataset. The parameter π indicates the sparsity of data — it is the probability that the worker labels each task.



(a) (b)

Figure 2: Comparing the convergence rate of the Opt-D&S algorithm and the MV-D&S estimator on synthetic dataset with $\pi = 0.2$: (a) convergence of the prediction error. (b) convergence of the squared error $\sum_{i=1}^{m} \|\widehat{C}_i - C_i\|_F^2$ for estimating confusion matrices.

Appendix

A Experiments on synthetic data

For experiments on synthetic data, we generate $m = 100$ workers and $n = 1000$ binary tasks. The true label of each task is uniformly sampled from $\{1, 2\}$. For each worker, the 2-by-2 confusion matrix is generated as follow: the two diagonal entries are independently and uniformly sampled from the interval $[0.3, 0.9]$, then the non-diagonal entries are determined to make the confusion matrix columns sum to 1. To simulate a sparse dataset, we make each worker label a task with probability π. With the choice $\pi \in \{0.2, 0.5, 1.0\}$, we obtain three different datasets.

We execute every algorithm independently for 10 times and average the outcomes. For the Opt-D&S algorithm and the MV-D&S estimator, the estimation is outputted after 10 EM iterates. For the group partitioning step involved in the Opt-D&S algorithm, the workers are randomly and evenly partitioned into three groups.

The main evaluation metric is the error of predicting the true label of items. The performance of various methods are reported in Table 3. On all sparsity levels, the Opt-D&S algorithm achieves the best accuracy, followed by the MV-D&S estimator. All other methods are consistently worse. It is not surprising that the Opt-D&S algorithm and the MV-D&S estimator yield similar accuracies, since they optimize the same log-likelihood objective. It is also meaningful to look at the convergence speed of both methods, as they employ distinct initialization strategies. Figure 2 shows that the Opt-D&S algorithm converges faster than the MV-D&S estimator, both in estimating the true labels and in estimating confusion matrices. This is because that Opt-D&S starts from a provably consistent initialization (recall Theorem 1).

B Proof of Proposition 2

First, notice that

$$\mathbb{E}[z_{ij} Z_{aj}^T] = \mathbb{E}\left[\mathbb{E}[z_{ij} Z_{aj}^T | y_j]\right] = \sum_{l=1}^{k} w_l \mathbb{E}\left[z_{ij} Z_{aj}^T | y_j = l\right]. \tag{9}$$

Since z_{ij} for $1 \le i \le m$ are conditionally independent given y_j, we can write

$$\mathbb{E}\left[z_{ij} Z_{aj}^T | y_j = l\right] = \mathbb{E}\left[z_{ij} | y_j = l\right] \mathbb{E}\left[Z_{aj}^T | y_j = l\right] = (\pi_i \mu_{il})(\mu_{al}^\diamond)^T. \tag{10}$$

Combining (9) and (10) implies the desired result,

$$\mathbb{E}[z_{ij} Z_{aj}^T] = \pi_i \sum_{l=1}^{k} w_l \mu_{il} (\mu_{al}^\diamond)^T = \pi_i C_i W (C_a^\diamond)^T.$$

C Proof of Theorem 1

If $a \neq b$, it is easy to verify that $S_{ab} = C_a^\diamond W (C_b^\diamond)^T = \mathbb{E}[Z_{aj} \otimes Z_{bj}]$. Furthermore, we can upper bound the spectral norm of S_{ab}, namely

$$\|S_{ab}\|_{\text{op}} \le \sum_{l=1}^{k} w_l \, \|\mu_{al}^\diamond\|_2 \, \|\mu_{bl}^\diamond\|_2 \le \sum_{l=1}^{k} w_l \, \|\mu_{al}^\diamond\|_1 \, \|\mu_{bl}^\diamond\|_1 \le 1.$$

For the same reason, it can be shown that $\|T_{abc}\|_{\text{op}} \le 1$.

Our proof strategy is briefly described as follow: we upper bound the estimation error for computing empirical moments (2a)-(2d) in Lemma 1, and upper bound the estimation error for tensor decomposition in Lemma 2. Then, we combine both lemmas to upper bound the error of formula (3).

Lemma 1. *Given a permutation* (a, b, c) *of* $(1, 2, 3)$*, for any scalar* $\epsilon \le \sigma_L/2$*, the second and the third moments* $\widehat{M}_2$ *and* $\widehat{M}_3$ *computed by equation* (2c) *and* (2d) *are bounded as*

$$\max\{\|\widehat{M}_2 - M_2\|_{\text{op}}, \|\widehat{M}_3 - M_3\|_{\text{op}}\} \le 31\epsilon/\sigma_L^3 \tag{11}$$

with probability at least $1 - \delta$*, where* $\delta = 6\exp(-(\sqrt{n}\epsilon - 1)^2) + k\exp(-(\sqrt{n/k}\epsilon - 1)^2)$.

Lemma 2. *Suppose that* (a, b, c) *is permutation of* $(1, 2, 3)$*. For any scalar* $\epsilon \le \kappa/2$*, if the empirical moments* $\widehat{M}_2$ *and* $\widehat{M}_3$ *satisfy*

$$\max\{\|\widehat{M}_2 - M_2\|_{\text{op}}, \|\widehat{M}_3 - M_3\|_{\text{op}}\} \le \epsilon H \tag{12}$$

$$\textit{for} \quad H := \min\left\{\frac{1}{2}, \frac{2\sigma_L^{3/2}}{15k(24\sigma_L^{-1} + 2\sqrt{2})}, \frac{\sigma_L^{3/2}}{4\sqrt{3/2}\sigma_L^{1/2} + 8k(24/\sigma_L + 2\sqrt{2})}\right\}$$

then the estimates $\widehat{C}_c^\diamond$ *and* $\widehat{W}$ *are bounded as*

$$\|\widehat{C}_c^\diamond - C_c^\diamond\|_{\text{op}} \le \sqrt{k}\epsilon \qquad \textit{and} \qquad \|\widehat{W} - W\|_{\text{op}} \le \epsilon.$$

with probability at least $1 - \delta$*, where* δ *is defined in Lemma 1.*

Combining Lemma 1, Lemma 2, if we choose a scalar ϵ_1 satisfying

$$\epsilon_1 \le \min\{\kappa/2, \pi_{\min} w_{\min} \sigma_L/(36k)\}, \tag{13}$$

then the estimates $\widehat{C}_g^\diamond$ (for $g = 1, 2, 3$) and $\widehat{W}$ satisfy that

$$\|\widehat{C}_g^\diamond - C_g^\diamond\|_{\text{op}} \le \sqrt{k}\epsilon_1 \qquad \text{and} \qquad \|\widehat{W} - W\|_{\text{op}} \le \epsilon_1. \tag{14}$$

with probability at least $1 - 6\delta$, where

$$\delta = (6 + k)\exp\Big(-(\sqrt{n/k}\epsilon_1 H \sigma_L^3/31 - 1)^2\Big).$$

To be more precise, we obtain the bound (14) by plugging $\epsilon := \epsilon_1 H\sigma_L^3/31$ into Lemma 1, then plugging $\epsilon := \epsilon_1$ into Lemma 2. The high probability statement is obtained by apply union bound.

Assuming inequality (14), for any $a \in \{1,2,3\}$, since $\|C_a^\diamond\|_{\mathrm{op}} \le \sqrt{k}$, $\|\widehat{C}_a^\diamond - C_a^\diamond\|_{\mathrm{op}} \le \sqrt{k}\epsilon_1$ and $\|W\|_{\mathrm{op}} \le 1$, $\left\|\widehat{W}\right\|_{\mathrm{op}} \le \epsilon_1$, Lemma 8 (the preconditions are satisfied by inequality (13)) implies that

$$\left\|\widehat{W}\widehat{C}_a^\diamond - WC_a^\diamond\right\|_{\mathrm{op}} \le 4\sqrt{k}\epsilon_1,$$

Since condition (13) implies

$$\|\widehat{W}\widehat{C}_a^\diamond - WC_a^\diamond\|_{\mathrm{op}} \le 4\sqrt{k}\epsilon_1 \le \sqrt{w_{\min}\sigma_L}/2 \le \sigma_k(WC_a^\diamond)/2$$

Lemma 7 yields that

$$\left\|\left(\widehat{W}\widehat{C}_a^\diamond\right)^{-1} - (WC_a^\diamond)^{-1}\right\|_{\mathrm{op}} \le \frac{8\sqrt{k}\epsilon_1}{w_{\min}\sigma_L}.$$

By Lemma 9, for any $i \in [m]$, the concentration bound

$$\left\|\frac{1}{n}\sum_{j=1}^{n} z_{ij}Z_{aj}^T - \mathbb{E}[z_{ij}Z_{aj}^T]\right\|_{\mathrm{op}} \le \epsilon_1$$

holds with probability at least $1 - m\exp(-(\sqrt{n}\epsilon_1 - 1)^2)$. Combining the above two inequalities with Proposition 2, then applying Lemma 8 with preconditions

$$\left\|(WC_a^\diamond)^{-1}\right\|_{\mathrm{op}} \le \frac{1}{w_{\min}\sigma_L} \quad \text{and} \quad \left\|\mathbb{E}\left[z_{ij}Z_{aj}^T\right]\right\|_{\mathrm{op}} \le 1,$$

we have

$$\Big\| \underbrace{\Big(\frac{1}{n}\sum_{j=1}^{n} z_{ij}Z_{aj}^T\Big)\Big(\widehat{W}\widehat{C}_a^\diamond\Big)^{-1}}_{\widehat{G}} - \pi_i C_i \Big\|_{\mathrm{op}} \le \frac{18\sqrt{k}\epsilon_1}{w_{\min}\sigma_L}. \tag{15}$$

Let $\widehat{G} \in \mathbb{R}^{k\times k}$ be the first term on the left hand side of inequality (15). Each column of $\widehat{G}$, denoted by $\widehat{G}_l$, is an estimate of $\pi_i\mu_{il}$. The ℓ_2-norm estimation error is bounded by $\frac{18\sqrt{k}\epsilon_1}{w_{\min}\sigma_L}$. Hence, we have

$$\|\widehat{G}_l - \pi_i\mu_{il}\|_1 \le \sqrt{k}\|\widehat{G}_l - \pi_i\mu_{il}\|_2 \le \sqrt{k}\|\widehat{G} - \pi_i C_i\|_{\mathrm{op}} \le \frac{18k\epsilon_1}{w_{\min}\sigma_L}, \tag{16}$$

and consequently, using the fact that $\sum_{c=1}^{k}\mu_{ilc} = 1$, we have

$$\begin{aligned}\left\|\mathrm{normalize}(\widehat{G}_l) - \mu_{il}\right\|_2 &= \left\|\frac{\widehat{G}_l}{\pi_i + \sum_{c=1}^{k}\left(\widehat{G}_{lc} - \pi_i\mu_{ilc}\right)} - \mu_{il}\right\|_2 \\ &\le \frac{\|\widehat{G}_l - \pi_i\mu_{il}\|_2 + \|\widehat{G}_l - \pi_i\mu_{il}\|_1\|\mu_{il}\|_2}{\pi_i - \|\widehat{G}_l - \pi_i\mu_{il}\|_1} \\ &\le \frac{72k\epsilon_1}{\pi_{\min}w_{\min}\sigma_L} \end{aligned} \tag{17}$$

where the last step combines inequalities (15), (16) with the bound $\frac{18k\epsilon_1}{w_{\min}\sigma_L} \le \pi_i/2$ from condition (13), and uses the fact that $\|\mu_{il}\|_2 \le 1$.

Note that inequality (17) holds with probability at least

$$1 - (36 + 6k)\exp\Big(-(\sqrt{n/k}\epsilon_1 H\sigma_L^3/31 - 1)^2\Big) - m\exp(-(\sqrt{n}\epsilon_1 - 1)^2).$$

It can be verified that $H \geq \frac{\sigma_L^{5/2}}{230k}$. Thus, the above expression is lower bounded by

$$1 - (36 + 6k + m) \exp\Big(- \Big(\frac{\sqrt{n}\epsilon_1 \sigma_L^{11/2}}{31 \times 230 \cdot k^{3/2}} - 1\Big)^2\Big),$$

If we represent this probability in the form of $1 - \delta$, then

$$\epsilon_1 = \frac{31 \times 230 \cdot k^{3/2}}{\sqrt{n}\sigma_L^{11/2}} \left(1 + \sqrt{\log((36 + 6k + m)/\delta)}\right). \tag{18}$$

Combining condition (13) and inequality (17), we find that to make $\|\widehat{C} - C\|_\infty$ bounded by ϵ, it is sufficient to choose ϵ_1 such that

$$\epsilon_1 \leq \min\Big\{ \frac{\epsilon \pi_{\min} w_{\min} \sigma_L}{72k}, \frac{\kappa}{2}, \frac{\pi_{\min} w_{\min} \sigma_L}{36k} \Big\}$$

This condition can be further simplified to

$$\epsilon_1 \leq \frac{\epsilon \pi_{\min} w_{\min} \sigma_L}{72k} \tag{19}$$

for small ϵ, that is $\epsilon \leq \min\left\{ \frac{36\kappa k}{\pi_{\min} w_{\min} \sigma_L}, 2 \right\}$. According to equation (18), the condition (19) will be satisfied if

$$\sqrt{n} \geq \frac{72 \times 31 \times 230 \cdot k^{5/2}}{\epsilon \pi_{\min} w_{\min} \sigma_L^{13/2}} \left(1 + \sqrt{\log((36 + 6k + m)/\delta)}\right).$$

Taking square over both sides of the inequality completes the proof.

C.1 Proof of Lemma 1

Throughout the proof, we assume that the following concentration bound holds: for any distinct indices $(a', b') \in \{1, 2, 3\}$, we have

$$\left\| \frac{1}{n} \sum_{j=1}^{n} Z_{a'j} \otimes Z_{b'j} - \mathbb{E}[Z_{a'j} \otimes Z_{b'j}] \right\|_{\mathrm{op}} \leq \epsilon \tag{20}$$

By Lemma 9 and the union bound, this event happens with probability at least $1 - 6\exp(-(\sqrt{n}\epsilon - 1)^2)$. By the assumption that $\epsilon \leq \sigma_L/2 \leq \sigma_k(S_{ab})/2$ and Lemma 7, we have

$$\left\| \frac{1}{n} \sum_{j=1}^{n} Z_{cj} \otimes Z_{bj} - \mathbb{E}[Z_{cj} \otimes Z_{bj}] \right\|_{\mathrm{op}} \leq \epsilon \quad \text{and}$$

$$\left\| \left(\frac{1}{n} \sum_{j=1}^{n} Z_{aj} \otimes Z_{bj} \right)^{-1} - (\mathbb{E}[Z_{aj} \otimes Z_{bj}])^{-1} \right\|_{\mathrm{op}} \leq \frac{2\epsilon}{\sigma_k^2(S_{ab})}$$

Under the preconditions

$$\|\mathbb{E}[Z_{cj} \otimes Z_{bj}]\|_{\mathrm{op}} \leq 1 \quad \text{and} \quad \left\|(\mathbb{E}[Z_{aj} \otimes Z_{bj}])^{-1}\right\|_{\mathrm{op}} \leq \frac{1}{\sigma_k(S_{ab})},$$

Lemma 8 implies that

$$\left\| \left(\frac{1}{n} \sum_{j=1}^{n} Z_{cj} \otimes Z_{bj} \right) \left(\frac{1}{n} Z_{aj} \otimes Z_{bj} \right)^{-1} - \mathbb{E}[Z_{cj} \otimes Z_{bj}](\mathbb{E}[Z_{aj} \otimes Z_{bj}])^{-1} \right\|_{\mathrm{op}}$$

$$\leq 2 \left(\frac{\epsilon}{\sigma_k(S_{ab})} + \frac{2\epsilon}{\sigma_k^2(S_{ab})} \right) \leq 6\epsilon/\sigma_L^2 \tag{21}$$

and for the same reason, we have

$$\left\| \left(\frac{1}{n} \sum_{j=1}^{n} Z_{cj} \otimes Z_{aj} \right) \left(\frac{1}{n} Z_{bj} \otimes Z_{aj} \right)^{-1} - \mathbb{E}[Z_{cj} \otimes Z_{aj}](\mathbb{E}[Z_{bj} \otimes Z_{aj}])^{-1} \right\|_{\mathrm{op}} \leq 6\epsilon/\sigma_L^2 \quad (22)$$

Now, let matrices F_2 and F_3 be defined as

$$F_2 := \mathbb{E}[Z_{cj} \otimes Z_{bj}](\mathbb{E}[Z_{aj} \otimes Z_{bj}])^{-1},$$
$$F_3 := \mathbb{E}[Z_{cj} \otimes Z_{aj}](\mathbb{E}[Z_{bj} \otimes Z_{aj}])^{-1},$$

and let the matrix on the left hand side of inequalities (21) and (22) be denoted by Δ_2 and Δ_3, we have

$$\left\| \widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \right\|_{\mathrm{op}} = \left\| \left(F_2 + \Delta_2 \right) \left(Z_{aj} \otimes Z_{bj} \right) \left(F_3 + \Delta_3 \right)^T - F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \right\|_{\mathrm{op}}$$
$$\leq \left\| Z_{aj} \otimes Z_{bj} \right\|_{\mathrm{op}} \left(\left\| \Delta_2 \right\|_{\mathrm{op}} \left\| F_3 + \Delta_2 \right\|_{\mathrm{op}} + \left\| F_2 \right\|_{\mathrm{op}} \left\| \Delta_3 \right\|_{\mathrm{op}} \right) \leq 30\epsilon \left\| Z_{aj} \otimes Z_{bj} \right\|_{\mathrm{op}} / \sigma_L^3.$$

where the last steps uses inequality (21), (22) and the fact that $\max\{\left\| F_2 \right\|_{\mathrm{op}}, \left\| F_3 \right\|_{\mathrm{op}}\} \leq 1/\sigma_L$ and

$$\left\| F_3 + \Delta_2 \right\|_{\mathrm{op}} \leq \| F_3 \|_{\mathrm{op}} + \| \Delta_2 \|_{\mathrm{op}} \leq 1/\sigma_L + 6\epsilon/\sigma_L^2 \leq 4/\sigma_L.$$

To upper bound the norm $\left\| Z_{aj} \otimes Z_{bj} \right\|_{\mathrm{op}}$, notice that

$$\left\| Z_{aj} \otimes Z_{bj} \right\|_{\mathrm{op}} \leq \left\| Z_{aj} \right\|_2 \left\| Z_{bj} \right\|_2 \leq \left\| Z_{aj} \right\|_1 \left\| Z_{bj} \right\|_1 \leq 1.$$

Consequently, we have

$$\left\| \widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \right\|_{\mathrm{op}} \leq 30\epsilon/\sigma_L^3. \quad (23)$$

For the rest of the proof, we use inequality (23) to bound $\widehat{M_2}$ and $\widehat{M_3}$. For the second moment, we have

$$\left\| \widehat{M_2} - M_2 \right\|_{\mathrm{op}} \leq \frac{1}{n} \sum_{j=1}^{n} \left\| \widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \right\|_{\mathrm{op}} + \left\| F_2 \left(\frac{1}{n} \sum_{j=1}^{n} Z_{aj} \otimes Z_{bj} \right) F_3^T - M_2 \right\|_{\mathrm{op}}$$
$$\leq 30\epsilon/\sigma_L^3 + \left\| F_2 \left(\frac{1}{n} \sum_{j=1}^{n} Z_{aj} \otimes Z_{bj} - \mathbb{E}[Z_{aj} \otimes Z_{bj}] \right) F_3^T \right\|_{\mathrm{op}}$$
$$\leq 30\epsilon/\sigma_L^3 + \epsilon/\sigma_L^2 \leq 31\epsilon/\sigma_L^3.$$

For the third moment, we have

$$\widehat{M_3} - M_3 = \frac{1}{n} \sum_{j=1}^{n} \left(\widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \right) \otimes Z_{cj}$$
$$+ \left(\frac{1}{n} \sum_{j=1}^{n} F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \otimes Z_{cj} - \mathbb{E} \left[F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \otimes Z_{cj} \right] \right). \quad (24)$$

We examine the right hand side of equation (24). The first term is bounded as

$$\left\| \left(\widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \right) \otimes Z_{cj} \right\|_{\mathrm{op}} \leq \left\| \widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \right\|_{\mathrm{op}} \left\| Z_{cj} \right\|_2$$
$$\leq 30\epsilon/\sigma_L^3. \quad (25)$$

For the second term, since $\| F_2 Z_{aj} \|_2 \leq 1/\sigma_L$, $\| F_3 Z_{bj} \|_2 \leq 1/\sigma_L$ and $\| Z_{cj} \|_2 \leq 1$, Lemma 9 implies that

$$\left\| \frac{1}{n} \sum_{j=1}^{n} F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \otimes Z_{cj} - \mathbb{E} \left[F_2 \left(Z_{aj} \otimes Z_{bj} \right) F_3^T \otimes Z_{cj} \right] \right\|_{\mathrm{op}} \leq \epsilon/\sigma_L^2 \quad (26)$$

with probability at least $1 - k\exp(-(\sqrt{n/k}\epsilon - 1)^2)$. Combining inequalities (25) and (26), we have

$$\left\|\widehat{M}_3 - M_3\right\|_{\mathrm{op}} \le 30\epsilon/\sigma_L^3 + \epsilon/\sigma_L^2 \le 31\epsilon/\sigma_L^3.$$

Applying union bound to all high-probability events completes the proof.

C.2 Proof of Lemma 2

Chaganty and Liang (Lemma 4 in [5]) have proved that when condition (12) holds, the tensor decomposition method of Algorithm 1 outputs $\{\widehat{\mu}^\diamond_h, \widehat{w}_h\}_{h=1}^k$, such that with probability at least $1 - \delta$, a permutation π satisfies

$$\|\widehat{\mu}^\diamond_h - \mu^\diamond_{c\pi(h)}\|_2 \le \epsilon \qquad \text{and} \qquad \left\|\widehat{w}_h - w_{\pi(h)}\right\|_\infty \le \epsilon.$$

Note that the constant H in Lemma 2 is obtained by plugging upper bounds $\|M_2\|_{\mathrm{op}} \le 1$ and $\|M_3\|_{\mathrm{op}} \le 1$ into Lemma 4 of Chaganty and Liang [5].

The $\pi(h)$-th component of $\mu^\diamond_{c\pi(h)}$ is greater than other components of $\mu^\diamond_{c\pi(h)}$, by a margin of κ. Assuming $\epsilon \le \kappa/2$, the greatest component of $\widehat{\mu}^\diamond_h$ is its $\pi(h)$-th component. Thus, Algorithm 1 is able to correctly estimate the $\pi(h)$-th column of $\widehat{C}^\diamond_c$ by the vector $\widehat{\mu}^\diamond_h$. Consequently, for every column of $\widehat{C}^\diamond_c$, the ℓ_2-norm error is bounded by ϵ. Thus, the spectral-norm error of $\widehat{C}^\diamond_c$ is bounded by $\sqrt{k}\epsilon$. Since W is a diagonal matrix and $\left\|\widehat{w}_h - w_{\pi(h)}\right\|_\infty \le \epsilon$, we have $\|\widehat{W} - W\|_{\mathrm{op}} \le \epsilon$.

D Proof of Theorem 2

We define two random events that will be shown holding with high probability:

$$\mathcal{E}_1: \ \sum_{i=1}^m \sum_{c=1}^k \mathbb{I}(z_{ij} = e_c)\log(\mu_{iy_jc}/\mu_{ilc}) \ge m\overline{D}/2 \qquad \text{for all } j \in [n] \text{ and } l \in [k]\backslash\{y_j\}.$$

$$\mathcal{E}_2: \ \Big|\sum_{j=1}^n \mathbb{I}(y_j = l)\mathbb{I}(z_{ij} = e_c) - nw_l\pi_i\mu_{ilc}\Big| \le nt_{ilc} \qquad \text{for all } (i,l,c) \in [m] \times [k]^2.$$

where $t_{ilc} > 0$ are scalars to be specified later. We define $t_{\min}$ to be the smallest element among $\{t_{ilc}\}$. Assuming that $\mathcal{E}_1 \cap \mathcal{E}_2$ holds, the following lemma shows that performing updates (4) and (5) attains the desired level of accuracy. See Section D.1 for the proof.

Lemma 3. *Assume that $\mathcal{E}_1 \cap \mathcal{E}_2$ holds. Also assume that $\mu_{ilc} \ge \rho$ for all $(i,l,c) \in [m] \times [k]^2$. If $\widehat{C}$ is initialized such that inequality* (7) *holds, and scalars t_{ilc} satisfy*

$$2\exp\Big(-m\overline{D}/4 + \log(m)\Big) \le t_{ilc} \le \pi_{\min}w_{\min}\min\left\{\frac{\rho}{8}, \frac{\rho\overline{D}}{64}\right\} \tag{27}$$

Then by alternating updates (4) *and* (5) *for at least one round, the estimates $\widehat{C}$ and $\widehat{q}$ are bounded as*

$$|\widehat{\mu}_{il} - \mu_{ilc}| \le 4t_{ilc}/(\pi_i w_l). \qquad \textit{for all } i \in [m],\ l \in [k],\ c \in [k].$$

$$\max_{l\in[k]}\{|\widehat{q}_{jl} - \mathbb{I}(y_j = l)|\} \le \exp\left(-m\overline{D}/4 + \log(m)\right) \qquad \textit{for all } j \in [n].$$

Next, we characterize the probability that events $\mathcal{E}_1$ and $\mathcal{E}_2$ hold. For measuring $\mathbb{P}[\mathcal{E}_1]$, we define auxiliary variable $s_i := \sum_{c=1}^k \mathbb{I}(z_{ij} = e_c)\log(\mu_{iy_jc}/\mu_{ilc})$. It is straightforward to see that $s_1, s_2, \ldots, s_m$ are mutually independent on any value of y_j, and each s_i belongs to the interval $[0, \log(1/\rho)]$. it is easy to verify that

$$\mathbb{E}\left[\sum_{i=1}^m s_i \Big| y_i\right] = \sum_{i=1}^m \pi_i \mathbb{D}_{\mathrm{KL}}\left(\mu_{iy_j}, \mu_{il}\right).$$

We denote the right hand side of the above equation by D. The following lemma shows that the second moment of s_i is bounded by the KL-divergence between labels.

Lemma 4. *Conditioning on any value of y_j, we have*

$$\mathbb{E}[s_i^2|y_i] \le \frac{2\log(1/\rho)}{1-\rho}\pi_i \mathbb{D}_{\mathrm{KL}}\left(\mu_{iy_j}, \mu_{il}\right).$$

According to Lemma 4, the aggregated second moment of s_i is bounded by

$$\mathbb{E}\left[\sum_{i=1}^m s_i^2 \Big| y_i\right] \le \frac{2\log(1/\rho)}{1-\rho}\sum_{i=1}^m \pi_i \mathbb{D}_{\mathrm{KL}}\left(\mu_{iy_jc}, \mu_{ilc}\right) = \frac{2\log(1/\rho)}{1-\rho}D$$

Thus, applying the Bernstein inequality, we have

$$\mathbb{P}\Big[\sum_{i=1} s_i \ge D/2|y_i\Big] \ge 1 - \exp\left(-\frac{\frac{1}{2}(D/2)^2}{\frac{2\log(1/\rho)}{1-\rho}D + \frac{1}{3}(2\log(1/\rho))(D/2)}\right),$$

Since $\rho \le 1/2$ and $D \ge m\overline{D}$, combining the above inequality with the union bound, we have

$$\mathbb{P}[\mathcal{E}_1] \ge 1 - kn\exp\left(-\frac{m\overline{D}}{33\log(1/\rho)}\right). \tag{28}$$

For measuring $\mathbb{P}[\mathcal{E}_2]$, we observe that $\sum_{j=1}^n \mathbb{I}(y_j = l)\mathbb{I}(z_{ij} = e_c)$ is the sum of n i.i.d. Bernoulli random variables with mean $p := \pi_i w_l \mu_{ilc}$. Since $t_{ilc} \le \pi_{\min} w_{\min}\rho/8 \le p$, applying the Chernoff bound implies

$$\mathbb{P}\left[\Big|\sum_{j=1}^n \mathbb{I}(y_j = l)\mathbb{I}(z_{ij} = e_c) - np\Big| \ge nt_{ilc}\right] \le 2\exp(-nt_{ilc}^2/(3p)) = 2\exp\left(-\frac{nt_{ilc}^2}{3\pi_i w_l \mu_{ilc}}\right),$$

Summarizing the probability bounds on $\mathcal{E}_1$ and $\mathcal{E}_2$, we conclude that $\mathcal{E}_1 \cap \mathcal{E}_2$ holds with probability at least

$$1 - kn\exp\left(-\frac{m\overline{D}}{33\log(1/\rho)}\right) - \sum_{i=1}^m\sum_{l=1}^k 2\exp\left(-\frac{nt_{ilc}^2}{3\pi_i w_l \mu_{ilc}}\right). \tag{29}$$

Proof of Part (a) According to Lemma 3, for $\widehat{y}_j = y_j$ being true, it sufficient to have $\exp(-m\overline{D}/4 + \log(m)) < 1/2$, or equivalently

$$m > 4\log(2m)/\overline{D}. \tag{30}$$

To ensure that this bound holds with probability at least $1-\delta$, expression (29) needs to be lower bounded by δ. It is achieved if we have

$$m \ge \frac{33\log(1/\rho)\log(2kn/\delta)}{\overline{D}} \quad \text{and} \quad n \ge \frac{3\pi_i w_l \mu_{ilc}\log(2mk/\delta)}{t_{ilc}^2} \tag{31}$$

If we choose

$$t_{ilc} := \sqrt{\frac{3\pi_i w_l \mu_{ilc}\log(2mk/\delta)}{n}}. \tag{32}$$

then the second part of condition (31) is guaranteed. To ensure that t_{ilc} satisfies condition (27). We need to have

$$\sqrt{\frac{3\pi_i w_l \mu_{ilc}\log(2mk/\delta)}{n}} \ge 2\exp\Big(-m\overline{D}/4 + \log(m)\Big) \quad \text{and}$$
$$\sqrt{\frac{3\pi_i w_l \mu_{ilc}\log(2mk/\delta)}{n}} \le \pi_{\min} w_{\min}\alpha/4.$$

The above two conditions requires that m and n satisfy

$$m \geq \frac{4\log(m\sqrt{2n/(3\pi_{\min}w_{\min}\log(2mk/\delta))})}{\overline{D}} \tag{33}$$

$$n \geq \frac{48\log(2mk/\delta)}{\pi_{\min}w_{\min}\alpha^2} \tag{34}$$

The four conditions (30), (31), (33) and (34) are simultaneously satisfied if we have

$$m \geq \frac{\max\{33\log(1/\rho)\log(2kn/\delta), 4\log(2mn)\}}{\overline{D}} \quad \text{and}$$
$$n \geq \frac{48\log(2mk/\delta)}{\pi_{\min}w_{\min}\alpha^2}.$$

Under this setup, $\widehat{y}_j = y_j$ holds for all $j \in [n]$ with probability at least $1-\delta$.

Proof of Part (b) If t_{ilc} is set by equation (32), combining Lemma 3 with this assignment, we have

$$(\widehat{\mu}_{ilc} - \mu_{ilc})^2 \leq \frac{48\mu_{ilc}\log(2mk/\delta)}{\pi_i w_l n}$$

with probability at least $1-\delta$. Summing both sides of the inequality over $c = 1, 2, \ldots, k$ completes the proof.

D.1 Proof of Lemma 3

To prove Lemma 3, we look into the consequences of update (4) and update (5). We prove two important lemmas, which show that both updates provide good estimates if they are properly initialized.

Lemma 5. *Assume that event $\mathcal{E}_1$ holds. If μ and its estimate $\widehat{\mu}$ satisfies*

$$\mu_{ilc} \geq \rho \quad \textit{and} \quad |\widehat{\mu}_{ilc} - \mu_{ilc}| \leq \delta_1 \qquad \textit{for all } i \in [m],\ l \in [k],\ c \in [k], \tag{35}$$

and $\widehat{q}$ is updated by formula (4)*, then $\widehat{q}$ is bounded as:*

$$\max_{l\in[k]}\{|\widehat{q}_{jl} - \mathbb{I}(y_j = l)|\} \leq \exp\left(-m\left(\frac{\overline{D}}{2} - \frac{2\delta_1}{\rho-\delta_1}\right) + \log(m)\right) \qquad \textit{for all } j \in [n]. \tag{36}$$

Proof. For an arbitrary index $l \neq y_j$, we consider the quantity

$$A_l := \sum_{i=1}^{m}\sum_{c=1}^{k}\mathbb{I}(z_{ij} = e_c)\log(\widehat{\mu}_{iy_jc}/\widehat{\mu}_{ilc})$$

By the assumption that $\mathcal{E}_1$ and inequality (35) holds, we obtain that

$$\begin{aligned} A_l &= \sum_{i=1}^{m}\sum_{c=1}^{k}\mathbb{I}(z_{ij} = e_c)\log(\mu_{iy_jc}/\mu_{ilc}) + \sum_{i=1}^{m}\sum_{c=1}^{k}\mathbb{I}(z_{ij} = e_c)\left[\log\left(\frac{\widehat{\mu}_{iy_jc}}{\mu_{iy_jc}}\right) - \log\left(\frac{\widehat{\mu}_{ilc}}{\mu_{ilc}}\right)\right] \\ &\geq \left(\sum_{i=1}^{m}\frac{\pi_i\mathbb{D}_{\mathrm{KL}}\left(\mu_{iy_j},\mu_{il}\right)}{2}\right) - 2m\log\left(\frac{\rho}{\rho-\delta_1}\right) \geq m\left(\frac{\overline{D}}{2} - \frac{2\delta_1}{\rho-\delta_1}\right). \end{aligned} \tag{37}$$

Thus, for every index $l \neq y_j$, combining formula (4) and inequality (37) implies that

$$\widehat{q}_{jl} \leq \frac{1}{\exp(A_l)} \leq \exp\left(-m\left(\frac{\overline{D}}{2} - \frac{2\delta_1}{\rho-\delta_1}\right)\right).$$

Consequently, we have

$$\widehat{q}_{jy_j} \geq 1 - \sum_{l\neq y_j}\widehat{q}_{jl} \geq 1 - \exp\left(-m\left(\frac{\overline{D}}{2} - \frac{2\delta_1}{\rho-\delta_1}\right) + \log(m)\right).$$

Combining the above two inequalities completes the proof. □

Lemma 6. *Assume that event $\mathcal{E}_2$ holds. If $\widehat{q}$ satisfies*

$$\max_{l\in[k]}\{|\widehat{q}_{jl} - \mathbb{I}(y_j = l)|\} \leq \delta_2 \qquad \textit{for all } j \in [n], \tag{38}$$

and $\widehat{\mu}$ is updated by formula (5)*, then $\widehat{\mu}$ is bounded as:*

$$|\widehat{\mu}_{ilc} - \mu_{ilc}| \leq \frac{2nt_{ilc} + 2n\delta_2}{(7/8)n\pi_i w_l - n\delta_2}. \qquad \textit{for all } i \in [m],\ l \in [k],\ c \in [k]. \tag{39}$$

Proof. By formula (5), we can write $\widehat{\mu}_{il} = A/B$, where

$$A := \sum_{j=1}^{n} \widehat{q}_{jl}\mathbb{I}(z_{ij} = e_c) \quad \text{and} \quad B := \sum_{c'=1}^{k}\sum_{j=1}^{n} \widehat{q}_{jl}\mathbb{I}(z_{ij} = e_{c'}).$$

Combining this definition with inequality (38), we find that

$$\begin{aligned}|A - n\pi_i w_l \mu_{ilc}\mu_{ilc}| &\leq \Big|\sum_{j=1}^{n} \mathbb{I}(q_{jl} = y_j)\mathbb{I}(z_{ij} = e_c) - n\pi_i w_l \mu_{ilc}\mu_{ilc}\Big| + \Big|\sum_{j=1}^{n} \widehat{q}_{jl}\mathbb{I}(z_{ij} = e_c) - \sum_{j=1}^{n} \mathbb{I}(q_{jl} = y_j)\mathbb{I}(z_{ij} = e_c)\Big| \\ &\leq nt_{ilc} + n\delta_2.\end{aligned}$$

By the same argument, we have

$$|B - n\pi_i w_l \mu_{ilc}| \leq \left(\sum_{c=1}^{k} nt_{ilc}\right) + n\delta_2.$$

Combining the bound for A and B, we obtain that

$$\begin{aligned}|\widehat{\mu}_{il} - \mu_{ilc}| &= \left|\frac{n\pi_i w_l \mu_{ilc} + (A - n\pi_i w_l \mu_{ilc})}{n\pi_i w_l + (B - n\pi_i w_l)} - \mu_{ilc}\right| = \left|\frac{(A - n\pi_i w_l \mu_{ilc}) + \mu_{ilc}(B - n\pi_i w_l)}{n\pi_i w_l + (B - n\pi_i w_l)}\right| \\ &\leq \frac{2nt_{ilc} + 2n\delta_2}{n\pi_i w_l - n\sum_{c=1}^{k} t_{ilc} - n\delta_2}\end{aligned}$$

Condition (27) implies that $\sum_{c=1}^{k} t_{ilc} \leq \pi_{\min} w_{\min} \sum_{c=1}^{k} \rho/8 \leq \pi_{\min} w_{\min}/8$, where the last step follow from $k\rho \leq 1$. Plugging this upper bound into the above inequality completes the proof. □

To proceed with the proof, we assign specific values to δ_1 and δ_2. Let

$$\delta_1 := \min\left\{\frac{\rho}{2}, \frac{\rho\overline{D}}{16}\right\} \quad \text{and} \quad \delta_2 := t_{\min}/2. \tag{40}$$

We claim that at any step in the update, the preconditions (35) and (38) always hold.

We prove the claim by induction. Before the iteration begins, $\widehat{\mu}$ is initialized such that the accuracy bound (7) holds. Thus, condition (35) is satisfied at the beginning. We assume by induction that condition (35) is satisfied at time $1, 2, \ldots, \tau - 1$ and condition (38) is satisfied at time $2, 3, \ldots, \tau - 1$. At time τ, either update (4) or update (5) is performed. If update (4) is performed, then by the inductive hypothesis, condition (35) holds before the update. Thus, Lemma 5 implies that

$$\max_{l\in[k]}\{|\widehat{q}_{jl} - \mathbb{I}(y_j = l)|\} \leq \exp\left(-m\left(\frac{\overline{D}}{2} - \frac{2\delta_1}{\rho - \delta_1}\right) + \log(m)\right).$$

The assignment (40) implies $\frac{\overline{D}}{2} - \frac{2\delta_1}{\rho - \delta_1} \geq \frac{\overline{D}}{4}$, which yields that

$$\max_{l\in[k]}\{|\widehat{q}_{jl} - \mathbb{I}(y_j = l)|\} \leq \exp(-m\overline{D}/4 + \log(m)) \leq t_{\min}/2 = \delta_2,$$

where the last inequality follows from condition (27). It suggests that condition (38) holds after the update.

On the other hand, we assume that update (5) is performed at time τ. Since update (5) follows update (4), we have $\tau \geq 2$. By the inductive hypothesis, condition (38) holds before the update, so

Lemma 6 implies

$$|\widehat{\mu}_{il} - \mu_{ilc}| \le \frac{2nt_{ilc} + 2n\delta_2}{(7/8)n\pi_i w_l - n\delta_2} = \frac{2nt_{ilc} + nt_{\min}}{(7/8)n\pi_i w_l - nt_{\min}/2} \le \frac{3nt_{ilc}}{(7/8)n\pi_i w_l - nt_{\min}/2},$$

where the last step follows since $t_{\min} \le t_{ilc}$. Noticing $\rho \le 1$, condition (27) implies that $t_{\min} \le \pi_{\min} w_{\min}/8$. Thus, the right hand side of the above inequality is bounded by $4t_{ilc}/(\pi_i w_l)$. Using condition (27) again, we find

$$\frac{4t_{ilc}}{\pi_i w_l} \le \frac{4t_{ilc}}{\pi_{\min} w_{\min}} \le \min\left\{\frac{\rho}{2}, \frac{\rho\overline{D}}{16}\right\} = \delta_1,$$

which verifies that condition (35) holds after the update. This completes the induction.

Since preconditions (35) and (38) hold for any time $\tau \ge 2$, Lemma 5 and Lemma 6 implies that the concentration bounds (36) and (39) always hold. These two concentration bounds establish the lemma's conclusion.

D.2 Proof of Lemma 4

By the definition of s_i, we have

$$\mathbb{E}[s_i^2] = \pi_i \sum_{c=1}^{k} \mu_{iy_jc}(\log(\mu_{iy_jc}/\mu_{ilc}))^2 = \pi_i \sum_{c=1}^{k} \mu_{iy_jc}(\log(\mu_{ilc}/\mu_{iy_jc}))^2$$

We claim that for any $x \ge \rho$ and $\rho < 1$, the following inequality holds:

$$\log^2(x) \le \frac{2\log(1/\rho)}{1-\rho}(x - 1 - \log(x)) \tag{41}$$

We defer the proof of inequality (41), focusing on its consequence. Let $x := \mu_{ilc}/\mu_{iy_jc}$, then inequality (41) yields that

$$\mathbb{E}[s_i^2] \le \frac{2\log(1/\rho)}{1-\rho}\pi_i \left(\sum_{c=1}^{k} \mu_{ilc} - \mu_{iy_jc} - \mu_{iy_jc}\log(\mu_{ilc}/\mu_{iy_jc})\right) = \frac{2\log(1/\rho)}{1-\rho}\pi_i \mathbb{D}_{\mathrm{KL}}\left(\mu_{iy_j}, \mu_{il}\right).$$

It remains to prove the claim (41). Let $f(x) := \log^2(x) - \frac{2\log(1/\rho)}{1-\rho}(x - 1 - \log(x))$. It suffices to show that $f(x) \le 0$ for $x \ge \rho$. First, we have $f(1) = 0$ and

$$f'(x) = \frac{2(\log(x) - \frac{\log(1/\rho)}{1-\rho}(x-1))}{x}.$$

For any $x > 1$, we have

$$\log(x) < x - 1 \le \frac{\log(1/\rho)}{1-\rho}(x-1)$$

where the last inequality holds since $\log(1/\rho) \ge 1 - \rho$. Hence, we have $f'(x) < 0$ and consequently $f(x) < 0$ for $x > 1$.

For any $\rho \le x < 1$, notice that $\log(x) - \frac{\log(1/\rho)}{1-\rho}(x-1)$ is a concave function of x, and equals zero at two points $x = 1$ and $x = \rho$. Thus, $f'(x) \ge 0$ at any point $x \in [\rho, 1)$, which implies $f(x) \le 0$.

E Proof of Theorem 3

In this section we prove Theorem 3. The proof separates into two parts.

E.1 Proof of Part (a)

Throughout the proof, probabilities are implicitly conditioning on $\{\pi_i\}$ and $\{\mu_{ilc}\}$. We assume that (l, l') are the pair of labels such that

$$\overline{D} = \frac{1}{m} \sum_{i=1}^{m} \pi_i \mathbb{D}_{\mathrm{KL}}\left(\mu_{il}, \mu_{il'}\right).$$

Let $\mathbb{Q}$ be a uniform distribution over the set $\{l, l'\}^n$. For any predictor $\widehat{y}$, we have

$$\begin{aligned} \max_{v \in [k]^n} \mathbb{E}\Big[\sum_{j=1}^{n} \mathbb{I}(\widehat{y}_j \neq y_j)\Big| y = v\Big] &\geq \sum_{v \in \{l,l'\}^n} \mathbb{Q}(v)\ \mathbb{E}\Big[\sum_{j=1}^{n} \mathbb{I}(\widehat{y}_j \neq y_j)\Big| y = v\Big] \\ &= \sum_{j=1}^{n} \sum_{v \in \{l,l'\}^n} \mathbb{Q}(v)\ \mathbb{E}\Big[\mathbb{I}(\widehat{y}_j \neq y_j)\Big| y = v\Big]. \end{aligned} \tag{42}$$

Thus, it is sufficient to lower bound the right hand side of inequality (42).

For the rest of the proof, we lower bound the quantity $\sum_{y \in \{l,l'\}^n} \mathbb{Q}(v)\ \mathbb{E}[\mathbb{I}(\widehat{y}_j \neq y_j)|y]$ for every item j. Let $Z := \{z_{ij} : i \in [m],\ j \in [n]\}$ be the set of all observations. We define two probability measures $\mathbb{P}_0$ and $\mathbb{P}_1$, such that $\mathbb{P}_0$ is the measure of Z conditioning on $y_j = l$, while $\mathbb{P}_1$ is the measure of Z conditioning on $y_j = l'$. By applying Le Cam's method [24] and Pinsker's inequality, we have

$$\begin{aligned} \sum_{v \in \{l,l'\}^n} \mathbb{Q}(v)\ \mathbb{E}\Big[\mathbb{I}(\widehat{y}_j \neq y_j)\Big| y = v\Big] &= \mathbb{Q}(y_j = l)\mathbb{P}_0(\widehat{y}_j \neq l) + \mathbb{Q}(y_j = l')\mathbb{P}_1(\widehat{y}_j \neq l') \\ &\geq \frac{1}{2} - \frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_1\|_{\mathrm{TV}} \\ &\geq \frac{1}{2} - \frac{1}{4}\sqrt{\mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0, \mathbb{P}_1\right)}. \end{aligned} \tag{43}$$

The remaining arguments upper bound the KL-divergence between $\mathbb{P}_0$ and $\mathbb{P}_1$. Conditioning on y_j, the set of random variables $Z_j := \{z_{ij} : i \in [m]\}$ are independent of $Z \backslash Z_j$ for both $\mathbb{P}_0$ and $\mathbb{P}_1$. Letting the distribution of X with respect to probability measure $\mathbb{P}$ be denoted by $\mathbb{P}(X)$, we have

$$\mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0, \mathbb{P}_1\right) = \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_j), \mathbb{P}_1(Z_j)\right) + \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z \backslash Z_j), \mathbb{P}_1(Z \backslash Z_j)\right) = \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_j), \mathbb{P}_1(Z_j)\right), \tag{44}$$

where the last step follows since $\mathbb{P}_0(Z \backslash Z_j) = \mathbb{P}_1(Z \backslash Z_j)$. Next, we observe that $z_{1j}, z_{2j}, \ldots, z_{mj}$ are mutually independent given y_j, which implies

$$\begin{aligned} \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_j), \mathbb{P}_1(Z_j)\right) &= \sum_{i=1}^{m} \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(z_{ij}), \mathbb{P}_1(z_{ij})\right) \\ &= (1 - \pi_i)\log\left(\frac{1 - \pi_i}{1 - \pi_i}\right) + \sum_{c=1}^{k} \pi_i \mu_{ilc} \log\left(\frac{\pi_i \mu_{ilc}}{\pi_i \mu_{il'c}}\right) \\ &= \sum_{c=1}^{k} \pi_i \mathbb{D}_{\mathrm{KL}}\left(\mu_{ilc}, \mu_{il'c}\right) = m\overline{D}. \end{aligned} \tag{45}$$

Combining inequality (43) with equations (44) and (45), we have

$$\sum_{v \in \{l,l'\}^n} \mathbb{Q}(v)\ \mathbb{E}\Big[\mathbb{I}(\widehat{y}_j \neq y_j)\Big| y = v\Big] \geq \frac{1}{2} - \frac{1}{4}\sqrt{m\overline{D}}.$$

Thus, if $m \leq 1/(4\overline{D})$, then the above inequality is lower bounded by $3/8$. Plugging this lower bound into inequality (42) completes the proof.

E.2 Proof of Part (b)

Throughout the proof, probabilities are implicitly conditioning on $\{\pi_i\}$ and $\{w_l\}$. We define two vectors

$$u_0 := \left(\frac{1}{2}, \frac{1}{2}, 0, \dots, 0\right)^T \in \mathbb{R}^k \qquad \text{and} \qquad u_1 := \left(\frac{1}{2}+\delta, \frac{1}{2}-\delta, 0, \dots, 0\right)^T \in \mathbb{R}^k$$

where $\delta \leq 1/4$ is a scalar to be specified. Consider a m-by-k random matrix V whose entries are uniformly sampled from $\{0, 1\}$. We define a random tensor $u_V \in \mathbb{R}^{m\times k\times k}$, such that $(u_V)_{il} := u_{V_{il}}$ for all $(i,l) \in [m]\times[k]$. Givan an estimator $\widehat{\mu}$ and a pair of indices $(\bar{i}, \bar{l})$, we have

$$\sup_{\mu\in\mathbb{R}^{m\times k\times k}} \mathbb{E}\left[\|\widehat{\mu}_{\bar{i}\bar{l}} - \mu_{\bar{i}\bar{l}}\|_2^2\right] \geq \sum_{v\in[k]^n} \mathbb{P}(y=v)\left(\sum_V \mathbb{P}(V)\,\mathbb{E}\left[\|\widehat{\mu}_{\bar{i}\bar{l}} - \mu_{\bar{i}\bar{l}}\|_2^2 \middle| \mu = u_V, y = v\right]\right). \tag{46}$$

For the rest of the proof, we lower bound the term $\sum_V \mathbb{P}(V)\ \mathbb{E}[\|\widehat{\mu}_{\bar{i}\bar{l}} - \mu_{\bar{i}\bar{l}}\|_2^2 | \mu = u_V, y = v]$ for every $v \in [k]^n$. Let $\widehat{V}$ be an estimator defined as

$$\widehat{V} = \begin{cases} 0 & \text{if } \|\widehat{\mu}_{\bar{i}\bar{l}} - u_0\|_2 \leq \|\widehat{\mu}_{\bar{i}\bar{l}} - u_1\|_2. \\ 1 & \text{otherwise.} \end{cases}$$

If $\mu = u_V$, then $\widehat{V} \neq V_{\bar{i}\bar{l}} \Rightarrow \|\widehat{\mu}_{\bar{i}\bar{l}} - \mu_{\bar{i}\bar{l}}\|_2 \geq \frac{\sqrt{2}}{2}\delta$. Consequently, we have

$$\sum_V \mathbb{P}(V)\ \mathbb{E}[\|\widehat{\mu}_{\bar{i}\bar{l}} - \mu_{\bar{i}\bar{l}}\|_2^2 | \mu = u_V, y = v] \geq \frac{\delta^2}{2} \mathrm{P}[\widehat{V} \neq V_{\bar{i}\bar{l}} | y = v]. \tag{47}$$

Let $Z := \{z_{ij} : i \in [m],\ j \in [n]\}$ be the set of all observations. We define two probability measures $\mathbb{P}_0$ and $\mathbb{P}_1$, such that $\mathbb{P}_0$ is the measure of Z conditioning on $y = v$ and $\mu_{\bar{i}\bar{l}} = u_0$, and $\mathbb{P}_1$ is the measure of Z conditioning on $y = v$ and $\mu_{\bar{i}\bar{l}} = u_1$. For any other pair of indices $(i,l) \neq (\bar{i}, \bar{l})$, $\mu_{il} = u_{V_{il}}$ for both $\mathbb{P}_0$ and $\mathbb{P}_1$. By this definition, the distribution of Z conditioning on $y = v$ and $\mu = u_V$ is a mixture of distributions $\mathbb{Q} := \frac{1}{2}\mathbb{P}_0 + \frac{1}{2}\mathbb{P}_1$. By applying Le Cam's method [24] and Pinsker's inequality, we have

$$\begin{aligned} \mathrm{P}[\widehat{V} \neq V_{\bar{i}\bar{l}} | y = v] &\geq \frac{1}{2} - \frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_1\|_{\mathrm{TV}} \\ &\geq \frac{1}{2} - \frac{1}{4}\sqrt{\mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0, \mathbb{P}_1\right)}. \end{aligned} \tag{48}$$

Conditioning on $y = v$, the set of random variables $Z_i := \{z_{ij} : j \in [n]\}$ are mutually independent for both $\mathbb{P}_0$ and $\mathbb{P}_1$. Letting the distribution of X with respect to probability measure $\mathbb{P}$ be denoted by $\mathbb{P}(X)$, we have

$$\mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0, \mathbb{P}_1\right) = \sum_{i=1}^m \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_i), \mathbb{P}_1(Z_i)\right) = \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}}), \mathbb{P}_1(Z_{\bar{i}})\right) \tag{49}$$

where the last step follows since $\mathbb{P}_0(Z_i) = \mathbb{P}_1(Z_i)$ for all $i \neq \bar{i}$. Next, we let $J := \{j : v_j = \bar{l}\}$ and define a set of random variables $Z_{iJ} := \{z_{ij} : j \in J\}$. It is straightforward to see that Z_{iJ} is independent of $Z_i \backslash Z_{iJ}$ for both $\mathbb{P}_0$ and $\mathbb{P}_1$. Hence, we have

$$\begin{aligned} \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}}), \mathbb{P}_1(Z_{\bar{i}})\right) &= \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}J}), \mathbb{P}_1(Z_{\bar{i}J})\right) + \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}} \backslash Z_{\bar{i}J}), \mathbb{P}_1(Z_{\bar{i}} \backslash Z_{\bar{i}J})\right) \\ &= \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}J}), \mathbb{P}_1(Z_{\bar{i}J})\right) \end{aligned} \tag{50}$$

where the last step follows since $\mathbb{P}_0(Z_{\bar{i}} \backslash Z_{\bar{i}J}) = \mathbb{P}_1(Z_{\bar{i}} \backslash Z_{\bar{i}J})$. Finally, since $\mu_{\bar{i}\bar{l}}$ is explicitly given in both $\mathbb{P}_0$ and $\mathbb{P}_1$, the random variables contained in $Z_{\bar{i}J}$ are mutually independent. Consequently, we have

$$\begin{aligned} \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}J}), \mathbb{P}_1(Z_{\bar{i}J})\right) &= \sum_{j\in J} \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(z_{\bar{i}j}), \mathbb{P}_1(z_{\bar{i}j})\right) = |J|\ \pi_{\bar{i}}\ \frac{1}{2}\log\left(\frac{1}{1-4\delta^2}\right) \\ &\leq \frac{5}{2}|J|\ \pi_{\bar{i}}\delta^2. \end{aligned} \tag{51}$$

Here, we have used the fact that $\log(1/(1-4x^2)) \leq 5x^2$ holds for any $x \in [0, 1/4]$.

Combining the lower bound (48) with upper bounds (49), (50) and (51), we find

$$\mathbb{P}[\widehat{V}_{il} \neq V_{il} | y = v] \geq \frac{3}{8} \mathbb{I} \left(\frac{5}{2} |J| \; \pi_{\bar{i}} \delta^2 \leq \frac{1}{4} \right).$$

Plugging the above lower bound into inequalities (46) and (47) implies that

$$\sup_{\mu \in \mathbb{R}^{m \times k \times k}} \mathbb{E}\Big[\|\widehat{\mu}_{\bar{i}\bar{l}} - \mu_{\bar{i}\bar{l}}\|_2^2 \Big] \geq \frac{3\delta^2}{16} \mathbb{P}\left[|\{j : y_j = \bar{l}\}| \leq \frac{1}{10\pi_{\bar{i}}\delta^2} \right].$$

Note than $|\{j : y_j = \bar{l}\}| \sim \text{Binomial}(n, w_{\bar{l}})$. Thus, if we set

$$\delta^2 := \min \left\{ \frac{1}{16}, \frac{1}{10\pi_{\bar{i}} w_{\bar{l}} n} \right\},$$

then $\frac{1}{10\pi_{\bar{i}}\delta^2}$ is greater than or equal to the median of $|\{j : y_j = \bar{l}\}|$, and consequently,

$$\sup_{\mu \in \mathbb{R}^{m \times k \times k}} \mathbb{E}\Big[\|\widehat{\mu}_{\bar{i}\bar{l}} - \mu_{\bar{i}\bar{l}}\|_2^2 \Big] \geq \min \left\{ \frac{3}{512}, \frac{3}{320\pi_{\bar{i}} w_{\bar{l}} n} \right\},$$

which establishes the theorem.

F Basic Lemmas

In this section, we prove some standard lemmas that we use for proving technical results.

Lemma 7 (Matrix Inversion)**.** *Let $A, E \in \mathbb{R}^{k \times k}$ be given, where A is invertible and E satisfies that $\|E\|_{\text{op}} \leq \sigma_k(A)/2$. Then*

$$\|(A+E)^{-1} - A^{-1}\|_{\text{op}} \leq \frac{2\|E\|_{\text{op}}}{\sigma_k^2(A)}.$$

Proof. A little bit of algebra reveals that

$$(A+E)^{-1} - A^{-1} = (A+E)^{-1} E A^{-1}.$$

Thus, we have

$$\|(A+E)^{-1} - A^{-1}\|_{\text{op}} \leq \frac{\|E\|_{\text{op}}}{\sigma_k(A)\sigma_k(A+E)}$$

We can lower bound the eigenvalues of $A + E$ by $\sigma_k(A)$ and $\|E\|_{\text{op}}$. More concretely, since

$$\|(A+E)\theta\|_2 \geq \|A\theta\|_2 - \|E\theta\|_2 \geq \sigma_k(A) - \|E\|_{\text{op}}$$

holds for any $\|\theta\|_2 = 1$, we have $\sigma_k(A+E) \geq \sigma_k(A) - \|E\|_{\text{op}}$. By the assumption that $\|E\|_{\text{op}} \leq \sigma_k(A)/2$, we have $\sigma_k(A+E) \geq \sigma_k(A)/2$. Then the desired bound follows. □

Lemma 8 (Matrix Multiplication)**.** *Let $A_i, E_i \in \mathbb{R}^{k \times k}$ be given for $i = 1, \ldots, n$, where the matrix A_i and the perturbation matrix E_i satisfy $\|A_i\|_{\text{op}} \leq K_i$, $\|E_i\|_{\text{op}} \leq K_i$. Then*

$$\left\| \prod_{i=1}^{n} (A_i + E_i) - \prod_{i=1}^{n} A_i \right\|_{\text{op}} \leq 2^{n-1} \left(\sum_{i=1}^{n} \frac{\|E_i\|_{\text{op}}}{K_i} \right) \prod_{i=1}^{n} K_i$$

Proof. By triangular inequality, we have

$$\begin{aligned}
\left\|\prod_{i=1}^{n}(A_i+E_i)-\prod_{i=1}^{n}A_i\right\|_{\text{op}} &= \left\|\sum_{i=1}^{n}\left(\prod_{j=1}^{i-1}A_j\right)\left(\prod_{k=i+1}^{n}(A_k+E_k)\right)E_i\right\|_{\text{op}} \\
&\le \sum_{i=1}^{n}\|E_i\|_{\text{op}}\left(\prod_{j=1}^{i-1}\|A_j\|_{\text{op}}\right)\left(\prod_{k=i+1}^{n}\|A_k+E_k\|_{\text{op}}\right) \\
&\le \sum_{i=1}^{n}2^{n-i}\frac{\|E_i\|_{\text{op}}}{K_i}\prod_{i=1}^{n}K_i \\
&= 2^{n-1}\left(\sum_{i=1}^{n}\frac{\|E_i\|_{\text{op}}}{K_i}\right)\prod_{i=1}^{n}K_i
\end{aligned}$$

which completes the proof. □

Lemma 9 (Matrix and Tensor Concentration)**.** *Let $\{X_j\}_{j=1}^n$, $\{Y_j\}_{j=1}^n$ and $\{Z_j\}_{j=1}^n$ be i.i.k. samples from some distribution over $\mathbb{R}^k$ with bounded support ($\|X\|_2\le 1$, $\|Y\|_2\le 1$ and $\|Z\|_2\le 1$ with probability 1). Then with probability at least $1-\delta$,*

$$\left\|\frac{1}{n}\sum_{j=1}^{n}X_j\otimes Y_j-\mathbb{E}[X_1\otimes Y_1]\right\|_F \le \frac{1+\sqrt{\log(1/\delta)}}{\sqrt{n}}. \tag{52}$$

$$\left\|\frac{1}{n}\sum_{j=1}^{n}X_j\otimes Y_j\otimes Z_j-\mathbb{E}[X_1\otimes Y_1\otimes Z_1]\right\|_F \le \frac{1+\sqrt{\log(k/\delta)}}{\sqrt{n/k}}. \tag{53}$$

Proof. Inequality (52) is proved in Lemma D.1 of [1]. To prove inequality (53), we note that for any tensor $T\in\mathbb{R}^{k\times k\times k}$, we can define k-by-k matrices $T_1,\dots,T_k$ such that $(T_i)_{jk}:=T_{ijk}$. As a result, we have $\|T\|_F^2=\sum_{i=1}^{k}\|T_i\|_F^2$. If we set T to be the tensor on the left hand side of inequality (53), then

$$T_i=\frac{1}{n}\sum_{j=1}^{n}(Z_j^{(i)}\ X_j)\otimes Y_j-\mathbb{E}[(Z_j^{(i)}\ X_1)\otimes Y_1]$$

By applying the result of inequality (52), we find that with probability at least $1-k\delta'$, we have

$$\left\|\frac{1}{n}\sum_{j=1}^{n}X_j\otimes Y_j\otimes Z_j-\mathbb{E}[X_1\otimes Y_1\otimes Z_1]\right\|_F^2 \le k\left(\frac{1+\sqrt{\log(1/\delta')}}{\sqrt{n}}\right)^2.$$

Setting $\delta'=\delta/k$ completes the proof. □