Journal of Machine Learning Research 17 (2016) 1-44 Submitted 12/14; Revised 4/15; Published 7/16

Spectral Methods Meet EM: A Provably Optimal Algorithm for Crowdsourcing

Yuchen Zhang YUCZHANG@EECS.BERKELEY.EDU
Department of Electrical Engineering and Computer Science
University of California, Berkeley, Berkeley, CA 94720, USA

Xi Chen XCHEN3@STERN.NYU.EDU
Stern School of Business
New York University, New York, NY 10012, USA

Dengyong Zhou DENGYONG.ZHOU@MICROSOFT.COM
Microsoft Research
1 Microsoft Way, Redmond, WA 98052, USA

Michael I. Jordan JORDAN@STAT.BERKELEY.EDU
Department of Electrical Engineering and Computer Science and Department of Statistics
University of California, Berkeley, Berkeley, CA 94720, USA

Editor: Alexander Ihler

Abstract

Crowdsourcing is a popular paradigm for effectively collecting labels at low cost. 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.

Keywords: crowdsourcing, spectral methods, EM, Dawid-Skene model, non-convex optimization, minimax rate

1. Introduction

With the advent of online services such as Amazon Mechanical Turk, crowdsourcing has become an efficient and inexpensive way to collect labels for large-scale data. However, labels collected from the crowd can be of low quality since crowdsourcing workers are often non-experts and sometimes 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?

©2016 Yuchen Zhang and Xi Chen and Dengyong Zhou and Michael I. Jordan.

For labeling tasks with k different categories, Dawid and Skene (1979) develop a maximum likelihood approach to this problem based on the 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 random 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 labels, where the unobserved true labels are treated as latent variables.

Although this EM-based approach has had empirical success (Snow et al., 2008; Raykar et al., 2010; Liu et al., 2012; Zhou et al., 2012; Chen et al., 2013; Zhou et al., 2014), there is as yet no theoretical guarantee for its performance. A recent theoretical study (Gao and Zhou, 2014) 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 the "one-coin" model in what follows). However, since the likelihood function is nonconvex, this guarantee is not operational because the EM algorithm can get trapped in a local optimum. Several alternative approaches have been developed that aim to circumvent the theoretical deficiencies of the EM algorithm, still the context of the one-coin model (Karger et al., 2013, 2014; Ghosh et al., 2011; Dalvi et al., 2013), but, as we survey in Section 2, they either fail to achieve an optimal rate or make restrictive assumptions that can be 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 a 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 an upper bound 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 matching lower bound. 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 provides new general insight into the 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 (Liang, 2013; Chaganty and Liang, 2013). This work provides a concrete way to justify such observations theoretically. 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 (Lehmann and Casella, 2003). 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.

The paper is organized as follows. In Section 2, we provide background on crowdsourcing and the method of moments for latent variables models. In Section 3, we describe our crowdsourcing problem. Our provably optimal algorithm is presented in Section 4. Section

5 is devoted to theoretical analysis (with the proofs gathered in the Appendix). In Section 6, we consider the special case of the one-coin model. A simpler algorithm is introduced together with a sharper rate. Numerical results on both synthetic and real datasets are reported in Section 7, followed by our conclusions in Section 8.

2. Related Work

Many methods have been proposed to address the problem of estimating true labels in crowdsourcing (Whitehill et al., 2009; Raykar et al., 2010; Welinder et al., 2010; Ghosh et al., 2011; Liu et al., 2012; Zhou et al., 2012; Dalvi et al., 2013; Karger et al., 2014, 2013; Parisi et al., 2014; Zhou et al., 2014). The methods in Raykar et al. (2010); Ghosh et al. (2011); Karger et al. (2014); Liu et al. (2012); Karger et al. (2013); Dalvi et al. (2013) are based on the generative model proposed by Dawid and Skene (1979). In particular, Ghosh et al. (2011) propose a method based on Singular Value Decomposition (SVD) which addresses binary labeling problems under the one-coin model. The analysis in Ghosh et al. (2011) assumes that the labeling matrix is full, that is, each worker labels all items. To relax this assumption, Dalvi et al. (2013) propose another SVD-based algorithm which explicitly considers the sparsity of the labeling matrix in both algorithm design and theoretical analysis. Karger et al. (2014) propose an iterative algorithm for binary labeling problems under the one-coin model and extended it to multi-class labeling tasks by converting a k-class problem into $k-1$ binary problems (Karger et al., 2013). This line of work assumes that tasks are assigned to workers according to a random regular graph, thus imposes 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 (Ghosh et al., 2011; Dalvi et al., 2013; Karger et al., 2014, 2013). The methods in Raykar et al. (2010); Liu et al. (2012); Chen et al. (2013) incorporate Bayesian inference into the Dawid-Skene estimator by assuming a prior over confusion matrices. Zhou et al. (2012, 2014) 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 (1979).

Our method for initializing the EM algorithm in crowdsourcing is inspired by recent work using spectral methods to estimate latent variable models (Anandkumar et al., 2014, 2015, 2012, 2013; Chaganty and Liang, 2013; Zou et al., 2013; Hsu et al., 2012; Jain and Oh, 2014). 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 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 (Anandkumar et al., 2014). 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 4 (b)).

Recently, Balakrishnan et al. (2016) study the convergence rate of EM algorithm under a good initialization, which belongs to a ball centered at the true parameter. They show that

when the radius of the ball is small enough to satisfy certain gradient stability condition and sample deviation condition, EM has a geometric convergence rate. Although this is an insightful theoretical result, Balakrishnan et al. (2016) fail to provide a practical approach to constructing such an initialization.

Other related work is by Dasgupta and Schulman (2007), who study an EM algorithm for learning mixture of Gaussians under certain initialization. They establish a nearly optimal estimation precision when using a two-round EM algorithm to learn the parameters from a mixture of k well-separated spherical Gaussians in the high-dimensional space. The space dimension d is assumed to be much greater than $\log(k)$. Although the high-level idea is quite similar to ours (i.e., constructing a good initializer and running one or two steps of EM), our work is different from this work in several respects. First, while Dasgupta and Schulman (2007) require the Gaussian means to be well separated, we do not assume the workers' labeling distributions for different true labels to be separated. In fact, for any worker, we allow his/her labeling distributions for two classes to be arbitrarily close or even identical. We only assume that there is a partitioning of the workers into three groups, such that the averaged confusion matrix of each group has full rank. Second, Dasgupta and Schulman (2007) consider the high-dimensional case where the dimension of the parameter space is $d \gg \log(k)$. In crowdsourcing, the labeling distribution lies in a k-dimensional space where k is the number of classes. Typically k is a small integer (below 10). Finally, in terms of analysis technique, Dasgupta and Schulman (2007) heavily rely on Gaussian concentration results. Our work uses a variety of techniques for different theorems. In particular, for Theorem 3, we repeatedly use matrix perturbation and matrix concentration results to establish the sample complexity for the spectral method. For Theorem 4, we create three random events in equation (33) (that holds with high probability), then establish both a prediction error bound and a confusion matrix estimation error bound for the EM algorithm. We use Le Cam's method (Yu, 1997) to establish the minimax lower bound result in Theorem 5.

3. Problem Setting

Throughout this paper, $[a]$ denotes the integer set $\{1, 2, \ldots, a\}$ and $\sigma_b(A)$ denotes the b-th largest singular value of 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 $\{z_{ij} : i \in [m], j \in [n]\}$.

For this estimation purpose, we need to make assumptions on the process of generating observed labels. Following the work of Dawid and Skene (1979), 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*

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. 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. (5).

matrix of worker i. In the special case of the one-coin model, all the diagonal elements of C_i are equal to a constant while all the off-diagonal elements are equal to another constant such that each column of C_i sums to 1.

4. Our Algorithm

In this section, we present an algorithm to estimate the confusion matrices and true labels. Our algorithm consists of two stages. In the first stage, we compute an initial estimate for the 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 method of moments 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

$$\mu_{gl}^{\diamond} := \mathbb{E}(Z_{gj}|y_j = l) = \frac{1}{|G_g|}\sum_{i\in G_g}\pi_i\mu_{il},$$

our first step is to estimate $C_g^{\diamond} := [\mu_{g1}^{\diamond}, \mu_{g2}^{\diamond}, \ldots, \mu_{gk}^{\diamond}]$ and to estimate the distribution of true labels $W := \mathrm{diag}(w_1, w_2, \ldots, w_k)$. The following proposition shows that we can solve for $C_g^{\diamond}$ and W from the moments of $\{Z_{gj}\}$.

Proposition 1 (Anandkumar et al. (2015)) *Assume that the vectors $\{\mu_{g1}^{\diamond}, \mu_{g2}^{\diamond}, \ldots, \mu_{gk}^{\diamond}\}$ are linearly independent for each $g \in \{1, 2, 3\}$. Let (a, b, c) be a permutation of $\{1, 2, 3\}$. Define*

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

Then,

$$M_2 = \sum_{l=1}^{k} w_l\ \mu_{cl}^{\diamond} \otimes \mu_{cl}^{\diamond} \quad \textit{and} \quad M_3 = \sum_{l=1}^{k} w_l\ \mu_{cl}^{\diamond} \otimes \mu_{cl}^{\diamond} \otimes \mu_{cl}^{\diamond}.$$

Since we only have finite samples, the expectations in Proposition 1 must be approximated by empirical moments. In particular, they are computed by averaging over indices $j = 1, 2, \ldots, n$. 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 $C_c^{\diamond}$ 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 $\{(\widehat{\mu}_h^{\diamond}, \widehat{w}_h) : h = 1, \ldots, k\}$, where each $(\widehat{\mu}_h^{\diamond}, \widehat{w}_h)$ estimates a particular column of $C_c^{\diamond}$ (for some $\mu_{cl}^{\diamond}$) 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 $C_c^\diamond$. Thus, $\widehat{\mu}_1^\diamond, \ldots, \widehat{\mu}_k^\diamond$ represents an arbitrary permutation of the true columns.

To discover the index correspondence, we take each $\widehat{\mu}_h^\diamond$ 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 $\widehat{\mu}_h^\diamond$ corresponds to the l-th column of $C_c^\diamond$, then its l-th coordinate is expected to be greater than other coordinates. Thus, we set the l-th column of $\widehat{C}_c^\diamond$ to some vector $\widehat{\mu}_h^\diamond$ 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 $\widehat{\mu}_h^\diamond$). Then, we set the l-th diagonal entry of $\widehat{W}$ to the scalar $\widehat{w}_h$ associated with $\widehat{\mu}_h^\diamond$. Note that by iterating over $(a, b, c) \in \{(2, 3, 1), (3, 1, 2), (1, 2, 3)\}$, we obtain $\widehat{C}_c^\diamond$ 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 $\{z_{ij}\}$.

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*

$$\pi_i C_i W (C_a^\diamond)^T = \mathbb{E}[z_{ij} Z_{aj}^T].$$

Proof First, notice that

$$\mathbb{E}[z_{ij} Z_{aj}^T] = \mathbb{E}\left[\mathbb{E}[z_{ij} Z_{aj}^T | y_j]\right] = \sum_{l=1}^{k} w_l \mathbb{E}\left[z_{ij} Z_{aj}^T | y_j = l\right]. \tag{3}$$

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

$$\mathbb{E}\left[z_{ij} Z_{aj}^T | y_j = l\right] = \mathbb{E}\left[z_{ij} | y_j = l\right] \mathbb{E}\left[Z_{aj}^T | y_j = l\right] = (\pi_i \mu_{il})(\mu_{al}^\diamond)^T. \tag{4}$$

Combining (3) and (4) implies the desired result,

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

■

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

$$\widehat{C}_i := \text{normalize}\left\{ \left(\frac{1}{n} \sum_{j=1}^{n} z_{ij} Z_{aj}^T \right) \left(\widehat{W} (\widehat{C}_a^\diamond)^T \right)^{-1} \right\}, \tag{5}$$

where the normalization operator rescales the matrix columns, making sure that each column sums to 1. 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

$$\ell(\mu) := \log \left(\sum_{y \in [k]^n} L(\mu; y, z) \right). \tag{6}$$

We refine the initial estimate of Stage 1 by maximizing the objective function (6), which is implemented by the Expectation Maximization (EM) algorithm. The EM algorithm takes as initialization the values $\{\widehat{\mu}_{ilc}\}$ provided as output by Stage 1, and 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:

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

$$\text{where} \quad \widehat{q}_{jl} \leftarrow \frac{\exp\left(\sum_{i=1}^{m} \sum_{c=1}^{k} \mathbb{I}(z_{ij} = e_c) \log(\widehat{\mu}_{ilc}) \right)}{\sum_{l'=1}^{k} \exp\left(\sum_{i=1}^{m} \sum_{c=1}^{k} \mathbb{I}(z_{ij} = e_c) \log(\widehat{\mu}_{il'c}) \right)} \qquad \text{for } j \in [n],\ l \in [k]. \tag{7}$$

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

$$\widehat{\mu}_{ilc} \leftarrow \frac{\sum_{j=1}^{n} \widehat{q}_{jl} \mathbb{I}(z_{ij} = e_c)}{\sum_{c'=1}^{k} \sum_{j=1}^{n} \widehat{q}_{jl} \mathbb{I}(z_{ij} = e_{c'})} \qquad \text{for } i \in [m],\ l \in [k],\ c \in [k]. \tag{8}$$

In practice, we alternatively execute the updates (7) and (8), 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

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

indicating the smallest gap between diagonal entries and non-diagonal entries in the confusion matrix. The assumption requires that κ is strictly positive. Note that this assumption is group-based, thus doesn't 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 $\mathbb{D}_{\mathrm{KL}}(P, Q) := \sum_i P(i) \log(P(i)/Q(i))$ 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:

$$\overline{D} = \min_{l \neq l'} \frac{1}{m} \sum_{i=1}^{m} \pi_i \mathbb{D}_{\mathrm{KL}}(\mu_{il}, \mu_{il'}). \tag{9}$$

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 3 characterizes the performance of Algorithm 1, providing sufficient conditions for achieving an arbitrarily accurate initialization. We provide the proof of Theorem 3 in Appendix A.

Theorem 3 *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

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

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

Theorem 4 characterizes the error rate in Stage 2. It states that when a sufficiently accurate initialization is taken, the updates (7) and (8) refine the estimates $\widehat{\mu}$ and $\widehat{y}$ to the optimal accuracy. See Appendix B for the proof.

Theorem 4 *Assume that $\mu_{ilc} \geq \rho$ holds 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 way such that*

$$\|\widehat{C}_i - C_i\|_\infty \leq \alpha := \min\left\{\frac{\rho}{2}, \frac{\rho\overline{D}}{16}\right\} \qquad \text{for all } i \in [m] \tag{10}$$

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

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

then, for $\widehat{\mu}$ and $\widehat{q}$ obtained by iterating (7) and (8) (for at least one round), with probability at least $1-\delta$,

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

(b) $\|\widehat{\mu}_{il} - \mu_{il}\|_2^2 \leq \frac{48\log(8mk/\delta)}{\pi_i w_l n}$ holds for all $(i,l) \in [m] \times [k]$.

In Theorem 4, 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 bounded by ρ, so that Theorem 4 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 3 equal to the constant α defined in Theorem 4. 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

$$m = \widetilde{\Omega}\left(\frac{1}{\overline{D}}\right) \quad \text{and} \quad n = \widetilde{\Omega}\left(\frac{k^5}{\pi_{\min}^2 w_{\min}^2 \sigma_L^{13} \min\{\rho^2, (\rho\overline{D})^2\}}\right); \tag{11}$$

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 C for the proof.

Theorem 5 *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* $m \le 1/(4\overline{D})$, *then*

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

(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*

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

In part (a) of Theorem 5, we see that the number of workers should be at least $1/\overline{D}$, otherwise any predictor will make many mistakes. This lower bound matches our sufficient condition on the number of workers m (see Eq. (11)). In part (b), we see that the best possible estimate for μ_{il} has $1/(\pi_i w_l n)$ mean-squared error. It verifies the optimality of our estimator $\widehat{\mu}_{il}$. It is also worth noting that the constraint on the number of items n (see Eq. (11)) depends on problem-specific constants, which might be improvable. Nevertheless, the constraint scales logarithmically with m and $1/\delta$, thus is easy to satisfy for reasonably large datasets.

5.1 Discussion of theoretical results

In this section, we present a discussion of the foregoing theoretical results. In particular, we compare Theorem 3 and Theorem 4 to existing theoretical results on crowdsourcing and the EM method.

5.1.1 Sparse sampling regime vs. dense sampling regime

In our theoretical analysis, we make the assumption that the minimum labeling frequency $\pi_{\min}$ is bounded away from zero, i.e., $\pi_{\min} = \Omega(1)$. This corresponds to the *dense sampling regime* where the average number of labels for each item should be $\Theta(m)$. According to Theorem 4, to guarantee perfect label prediction with probability $1-\delta$, we require the average number of samples for each item to be $\Omega\left(\frac{\log(1/\delta)\pi_{\min}}{D}\right)$. Two related works in the dense sampling regime include Ghosh et al. (2011) and Gao and Zhou (2014). Ghosh et al. (2011) studied the one-coin model for binary labeling. To attain a δ prediction error, their algorithm requires m and n to scale with $1/\delta^2$, while our algorithm allows m and n to scale with $\log(1/\delta)$. Gao and Zhou (2014) studied the minimax rate of the prediction error and showed that maximizing likelihood is statistically optimal. However, they didn't provide a polynomial-time algorithm to achieve the minimax rate.

Another sampling regime is the *sparse sampling regime*, where the labeling frequency $\pi_{\min}$ goes to zero as the number of items n goes to infinity. In fact, this is a practical regime for large-scale datasets when workers complete only a vanishing fraction of the total tasks. As can be seen in condition (11), our theoretical result doesn't apply to the very sparse regime when $\pi = o(1/\sqrt{n})$. Several work have been devoted to investigate the sparse sampling regime (Karger et al., 2014, 2013; Dalvi et al., 2013). Under the sparse sampling

regime, the high-probability recovery of *all* true labels might be impossible. However, it is still of great interest to establish the upper bound on the prediction error as a function of the average number of labels per item. Karger et al. (2014, 2013) show that if worker labels are organized by a random regular bipartite graph, then the number of labels on each item should scale as $\log(1/\delta)$, where δ is the label prediction error. Their analysis assumes that the limit of number of items goes to infinity, or that the number of workers is many times of the number of items. Dalvi et al. (2013) provide algorithms that improve the theoretical guarantee for the one-coin model. Their algorithms succeed without the regular bipartite graph assumption, and without the requirement that the limit of number of items goes to infinity.

Although our theoretical analysis does not fully cover the sparse sampling regime, the algorithm sill applies to the sparse regime and achieves reasonably good performance (see, e.g., the experiment with TREC data in Section 7.2). Empirically, spectral-initialized EM is rather robust for both dense and sparse sampling regimes.[1]

5.1.2 Lower bound on the number of items

In both Theorem 3 and 4, we require lower bounds on the number of times. It is interesting to see whether these bounds can be improved. One idea is to improve those lower bounds using the technique from Balakrishnan et al. (2016) discussed in Section 2. Nevertheless, as we explain below, such improvement is up to a multiplicative factor $1/\alpha$, which doesn't depend on σ_L, $w_{\min}$ and $\pi_{\min}$. We recall that there are two lower bounds on the number of items n in our theoretical results.

1. The lower bound in the condition of Theorem 3 (denoted by n_{spectral}). It relies on the target error ϵ, the minimum singular value σ_L and the minimum prior probabilities $\pi_{\min}, w_{\min}$.

2. The lower bound in the condition of Theorem 4 (denoted by n_{EM}). It establishes the performance guarantee for EM. This lower bound relies on the initialization accuracy α and the minimum prior probabilities $\pi_{\min}, w_{\min}$.

Theorem 4 shows that it is sufficient to set the target error of Theorem 3 equal to $\epsilon := \alpha$. Thus, the constant n_{spectral} also depends on α. Due to the additional dependence on σ_L, the condition on n_{spectral} is more restrictive than that on n_{EM}.

The result of Balakrishnan et al. (2016) provides conditions (e.g., gradient stability condition and certain sample deviation condition) under which the EM method has geometric convergence rate. Assuming that these conditions are weaker than ours in Theorem 4, then instead of requiring the initialization condition $\|\widehat{C}_i - C_i\|_\infty \leq \alpha$ for all $i \in [m]$ as in equation (10), we may have another constant $\alpha' > \alpha$, such that $\|C_i - \widehat{C}_i\|_\infty \leq \alpha'$ ensures the linear convergence of EM. (It might be difficult to find the largest α' that makes the conditions of Balakrishnan et al. (2016) hold). The best possible value of α' is 1 since any entry of $\widehat{C}_i$ and C_i is bounded by 1. This means that we can potentially improve n_{spectral} by

1. An anonymous reviewer pointed out that if the system designer has the control over how to assign tasks to the workers, our result can be applied to sparse sampling regime via a grouping technique. In particular, one can partition the items into groups of small size and assign different workers to each groups of items so that the each subgroups of items and workers form a dense sub-matrix.

a factor of $1/\alpha$. Note that α is a constant that doesn't depend on σ_L. Thus, this potential improvement doesn't affect the lower bound's dependence on the condition number of the confusion matrix.

The result of Balakrishnan et al. (2016) provides conditions for EM to work but it doesn't show how to initialize EM to satisfy these conditions. Our paper uses the spectral method to do the initialization. As a consequence, the restriction on n_{spectral} is a premise for the spectral method to work. In order to improve the dependence on the condition number, one has to invent a better initialization scheme, which is out of the scope of this paper.

5.1.3 Discussion of simple majority voting estimator

It is also interesting to compare our algorithm with the majority voting estimator, where the true label is simply estimated by a majority vote among workers. Gao and Zhou (2014) show that if there are many spammers and few experts, the majority voting estimator gives almost a random guess. In contrast, our algorithm requires a sufficiently large $m\overline{D}$ to guarantee good performance. Since $m\overline{D}$ is the aggregated KL-divergence, a small number of experts suffices to ensure it is large enough.

6. One-Coin Model

In this section, we consider a simpler crowdsourcing model that is usually referred to as the "one-coin model." For the one-coin model, the confusion matrix C_i is parameterized by a single parameter p_i. More concretely, its entries are defined as

$$\mu_{ilc} = \begin{cases} p_i & \text{if } l = c, \\ \frac{1-p_i}{k-1} & \text{if } l \neq c. \end{cases} \tag{12}$$

In other words, the worker i uses a single coin flip to decide her assignment. No matter what the true label is, the worker has p_i probability to assign the correct label, and has $1 - p_i$ probability to randomly assign an incorrect label. For the one-coin model, it suffices to estimate p_i for every worker i and estimate y_j for every item j. Because of its simplicity, the one-coin model is easier to estimate and enjoys better convergence properties.

To simplify our presentation, we consider the case where $\pi_i \equiv 1$; noting that with proper normalization, the algorithm can be easily adapted to the case where $\pi_i < 1$. The statement of the algorithm relies on the following notation: For every two workers a and b, let the quantity N_{ab} be defined as

$$N_{ab} := \frac{k-1}{k}\left(\frac{\sum_{j=1}^{n}\mathbb{I}(z_{aj} = z_{bj})}{n} - \frac{1}{k}\right).$$

For every worker i, let workers a_i, b_i be defined as

$$(a_i, b_i) = \arg\max_{(a,b)}\{|N_{ab}| : \ a \neq b \neq i\}.$$

The algorithm contains two separate stages. First, we initialize $\widehat{p}_i$ by an estimator based on the method of moments. In contrast with the algorithm for the general model, the estimator

Algorithm 2: Estimating one-coin model

Input: integer k, observed labels $z_{ij} \in \mathbb{R}^k$ for $i \in [m]$ and $j \in [n]$.
Output: Estimator $\widehat{p}_i$ for $i \in [m]$ and $\widehat{y}_j$ for $j \in [n]$.

(1) Initialize $\widehat{p}_i$ by

$$\widehat{p}_i \leftarrow \frac{1}{k} + \operatorname{sign}(N_{ia_1}) \sqrt{\frac{N_{ia_i} N_{ib_i}}{N_{a_i b_i}}} \tag{13}$$

(2) If $\frac{1}{m}\sum_{i=1}^m \widehat{p}_i \geq \frac{1}{k}$ does not hold, then set $\widehat{p}_i \leftarrow \frac{2}{k} - \widehat{p}_i$ for all $i \in [m]$.

(3) Iteratively execute the following two steps for at least one round:

$$\widehat{q}_{jl} \propto \exp\left(\sum_{i=1}^m \mathbb{I}(z_{ij} = e_l) \log(\widehat{p}_i) + \mathbb{I}(z_{ij} \neq e_l) \log\left(\frac{1-\widehat{p}_i}{k-1}\right)\right) \qquad \text{for } j \in [n],\ l \in [k], \tag{14}$$

$$\widehat{p}_i \leftarrow \frac{1}{n}\sum_{j=1}^n \sum_{l=1}^k \widehat{q}_{jl} \mathbb{I}(z_{ij} = e_l) \qquad \text{for } i \in [m], \tag{15}$$

where update (14) normalizes $\widehat{q}_{jl}$, making $\sum_{l=1}^k \widehat{q}_{jl} = 1$ hold for all $j \in [n]$.

(4) Output $\{\widehat{p}_i\}$ and $\widehat{y}_j := \arg\max_{l \in [k]} \{\widehat{q}_{jl}\}$.

for the one-coin model doesn't need third-order moments. Instead, it only relies on pairwise statistics N_{ab}. Second, an EM algorithm is employed to iteratively maximize the objective function (6). See Algorithm 2 for a detailed description.

To theoretically characterize the performance of Algorithm 2, we need some additional notation. Let κ_i be the i-th largest element in $\{|p_i - 1/k|\}_{i=1}^m$. In addition, let $\overline{\kappa} := \frac{1}{m}\sum_{i=1}^m (p_i - 1/k)$ be the average gap between all accuracies and $1/k$. We assume that $\overline{\kappa}$ is strictly positive. We follow the definition of $\overline{D}$ in Eq. (9). The following theorem is proved in Appendix D.

Theorem 6 *Assume that $\rho \leq p_i \leq 1 - \rho$ holds for all $i \in [m]$. For any scalar $\delta > 0$, if the number of workers m and the number of items n satisfy*

$$m = \Omega\left(\frac{\log(1/\rho)\log(kn/\delta)}{\overline{D}}\right) \quad \text{and} \quad n = \Omega\left(\frac{\log(mk/\delta)}{\kappa_3^6 \min\{\overline{\kappa}^2, \rho^2, (\rho\overline{D})^2\}}\right), \tag{16}$$

then, for $\widehat{p}$ and $\widehat{y}$ returned by Algorithm 2, with probability at least $1 - \delta$,

(a) $\widehat{y}_j = y_j$ holds for all $j \in [n]$;

(b) $|\widehat{p}_i - p_i| \leq 2\sqrt{\frac{3\log(6m/\delta)}{n}}$ holds for all $i \in [m]$.

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

It is worth contrasting condition (11) with condition (16), namely the sufficient conditions for the general model and for the one-coin model. It turns out that the one-coin model requires much milder conditions on the number of items. In particular, κ_3 will be close to 1 if among all the workers there are three experts giving high-quality answers. As a consequence, the one-coin model is more robust than the general model. By contrasting the convergence rate of $\widehat{\mu}_{il}$ (by Theorem 4) and $\widehat{p}_i$ (by Theorem 6), the convergence rate of $\widehat{p}_i$ does not depend on $\{w_l\}_{l=1}^k$. This is additional evidence that the one-coin model enjoys a better convergence rate because of its simplicity.

7. 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 other methods. We compare to the Dawid & Skene estimator initialized by majority voting (refereed to as MV-D&S), the pure majority voting estimator, the multi-class labeling algorithm proposed by Karger et al. (2013) (referred to as KOS), the SVD-based algorithm proposed by Ghosh et al. (2011) (referred to as Ghost-SVD) and the "Eigenvalues of Ratio" algorithm proposed by Dalvi et al. (2013) (referred to as EigenRatio). The evaluation is made on three synthetic datasets and five real datasets.

7.1 Synthetic data

For 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 ten times and average the outcomes. For the Opt-D&S algorithm and the MV-D&S estimator, the estimation is outputted after ten 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 1. On all sparsity levels, the Opt-D&S algorithm achieves the best accuracy, followed by the MV-D&S estimator. All other

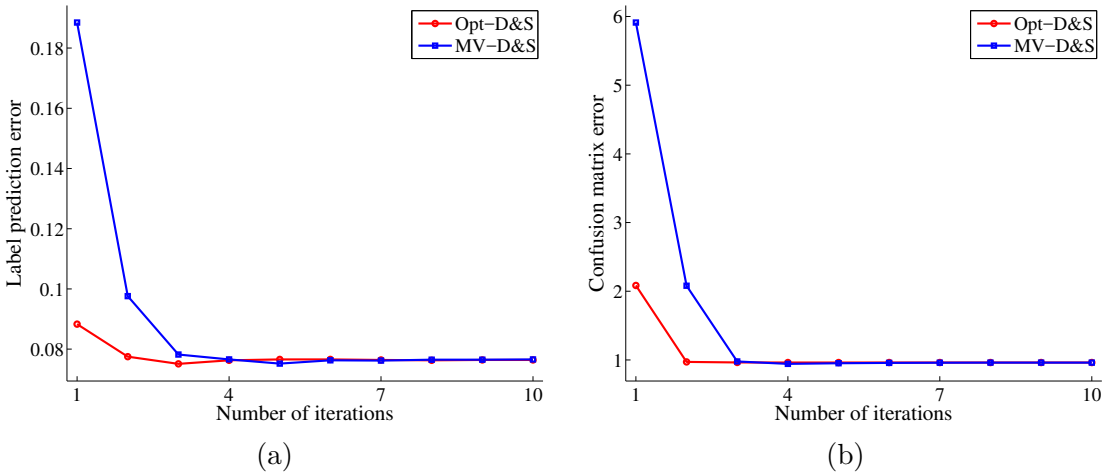


Figure 1: Comparing the convergence rate of the Opt-D&S algorithm and the MV-D&S estimator on synthetic dataset with $\pi = 0.2$: (a) convergence of the prediction error. (b) convergence of the squared error $\sum_{i=1}^{m} \|\widehat{C}_i - C_i\|_F^2$ for estimating confusion matrices.

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 1 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 can be explained by the general theoretical guarantee associated with Opt-D&S (recall Theorem 3).

The first and the second iteration of the Opt-D&S curve in Figure 1(b) correspond to the confusion matrix estimation error achieved by 1) the spectral method and 2) the spectral method with one iteration of EM, respectively. In Table 2, we provide a more detailed comparison, where the errors are compared with $\pi \in \{0.2, 0.5, 1.0\}$ and at different stages of the iteration. We found that for dense data ($\pi = 1$), the spectral method alone provides a reasonably good estimate on the confusion matrix, but its performance degenerates as the data becomes sparser. For both dense and sparse data, the spectral method with one iteration of EM achieves the nearly optimal error rate, which coincides with our theoretical prediction. In contrast, if the algorithm is initialized by majority voting, then one iteration of EM fails to provide a good estimate. Table 2 shows that its error rate is an order-of-magnitude higher than the spectral method initialized EM. This supports our concern that majority-voting initialization can be far from optimal—a principal motivation for this paper. Nevertheless, both initializations converge to the same error rate after ten iterations. Given the robustness of the EM method, deriving a sufficient and necessary condition under which the majority-voting initialization converges to an optimal solution remains an open problem.

	$\pi = 0.2$	$\pi = 0.5$	$\pi = 1.0$
Spectral Method	2.084	0.583	0.164
Opt-D&S (1st iteration)	0.972	0.352	0.143
Opt-D&S (10th iteration)	0.962	0.343	0.143
MV-D&S (1st iteration)	5.912	6.191	6.618
MV-D&S (10st iteration)	0.962	0.343	0.143

Table 2: Squared error for estimating the confusion matrix. The table compares (1) spectral method; (2) spectral initialization + one iteration of EM; (3) spectral initialization + 10 iterations of EM; (4) majority voting initialization + one iteration of EM; and (5) majority voting initialization + 10 iterations of EM.

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 3: The summary of datasets used in the real data experiment.

7.2 Real data

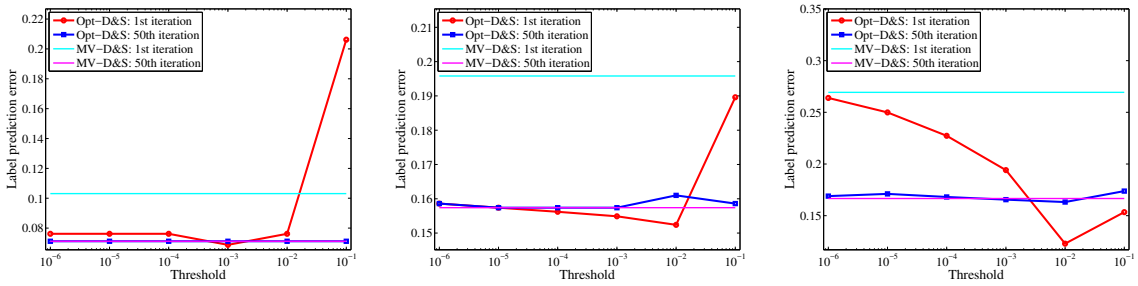
For real data experiments, we compare crowdsourcing algorithms on five datasets: three binary tasks and two multi-class tasks. Binary tasks include labeling bird species (Welinder et al., 2010) (Bird dataset), recognizing textual entailment (Snow et al., 2008) (RTE dataset) and assessing the quality of documents in TREC 2011 crowdsourcing track (Lease and Kazai, 2011) (TREC dataset). Multi-class tasks include labeling the bread of dogs from ImageNet (Deng et al., 2009) (Dog dataset) and judging the relevance of web search results (Zhou et al., 2012) (Web dataset). The statistics for the five datasets are summarized in Table 3. Since the Ghost-SVD algorithm and the EigenRatio algorithm work on binary tasks, they are evaluated on the Bird, RTE and TREC dataset. For the MV-D&S estimator and the Opt-D&S algorithm, 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 equation (5), 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 step doesn't take effect if the initialization error is bounded by Theorem 3.

Table 4 summarizes the performance of each method. The MV-D&S estimator and the Opt-D&S algorithm consistently outperform the other methods in predicting the true label of items. The KOS algorithm, the Ghost-SVD algorithm and the EigenRatio algorithm yield

	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 4: Error rate (%) in predicting the true labels on real data.



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

Figure 2: Comparing the MV-D&S estimator the Opt-D&S algorithm with different thresholding parameter Δ. The predict error is plotted after the 1st EM update and after convergence.

poorer performance, presumably due to the fact that they rely on idealized assumptions that are not met by the real data. In Figure 2, 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 the MV-D&S estimator 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 when only using the spectral method with just one E-step for the label prediction (i.e., the red curve Opt-D&S: 1st iteration in Figure 2), the error rates are more sensitive to the choice of Δ. A proper choice of Δ makes the Opt-D&S algorithm perform better than MV-D&S. The result suggests that a proper spectral initialization with just an E-step is good enough for the purposes of prediction. In practice, the best choice of Δ can be obtained by cross validation.

8. Conclusions

Under the generative model proposed by Dawid and Skene (1979), we propose an optimal algorithm for inferring true labels in the multi-class crowd labeling setting. Our approach utilizes the method of moments to construct an initial estimator for the EM algorithm. We proved that our method achieves the optimal rate with only one iteration of the EM algorithm.

To the best of our knowledge, this work provides the first instance of provable convergence for a latent variable model in which EM is initialized with the method of moments.

One-step EM initialized by the method of moments not only leads to better estimation error in terms of the dependence on the condition number of the second-order moment matrix but it also computationally more attractive than the standard one-step estimator obtained via a Newton-Raphson step. It is interesting to explore whether a properly initialized one-step EM algorithm can achieve the optimal rate for other latent variable models such as latent Dirichlet allocation or other mixed membership models.

Acknowledgments

We would like to thank anonymous reviewers and the associate editor for their constructive comments on improving the quality of the paper. Xi Chen would like to thank the support from Google Faculty Research Awards. This research was supported in part by DHS Award HSHQDC-16-3-00083, NSF CISE Expeditions Award CCF-1139158, DOE Award SN10040 DE-SC0012463, and DARPA XData Award FA8750-12-2-0331, and gifts from Amazon Web Services, Google, IBM, SAP, The Thomas and Stacey Siebel Foundation, Apple Inc., Arimo, Blue Goji, Bosch, Cisco, Cray, Cloudera, Ericsson, Facebook, Fujitsu, HP, Huawei, Intel, Microsoft, Pivotal, Samsung, Schlumberger, Splunk, State Farm and VMware.

Appendix A. Proof of Theorem 3

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

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

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

Our proof strategy is briefly described as follows: we upper bound the estimation error for computing empirical moments (2a)-(2d) in Lemma 7, and upper bound the estimation error for tensor decomposition in Lemma 8. Then, we combine both lemmas to upper bound the error of formula (5).

Lemma 7 *Given a permutation (a, b, c) of $(1, 2, 3)$, for any scalar $\epsilon \leq \sigma_L/2$, the second and the third moments $\widehat{M}_2$ and $\widehat{M}_3$ computed by equation (2c) and (2d) are bounded as*

$$\max\{\|\widehat{M}_2 - M_2\|_{\text{op}}, \|\widehat{M}_3 - M_3\|_{\text{op}}\} \leq 31\epsilon/\sigma_L^3 \tag{17}$$

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

Lemma 8 *Suppose that (a, b, c) is permutation of $(1, 2, 3)$. For any scalar $\epsilon \leq \kappa/2$, if the empirical moments $\widehat{M}_2$ and $\widehat{M}_3$ satisfy*

$$\max\{\|\widehat{M}_2 - M_2\|_{\text{op}}, \|\widehat{M}_3 - M_3\|_{\text{op}}\} \leq \epsilon H \tag{18}$$

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

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

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

with probability at least $1 - \delta$, where δ is defined in Lemma 7.

Combining Lemma 7, Lemma 8, if we choose a scalar ϵ_1 satisfying

$$\epsilon_1 \leq \min\{\kappa/2, \pi_{\min} w_{\min} \sigma_L/(36k)\}, \tag{19}$$

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

$$\|\widehat{C}_g^{\diamond} - C_g^{\diamond}\|_{\text{op}} \leq \sqrt{k}\epsilon_1 \qquad \text{and} \qquad \|\widehat{W} - W\|_{\text{op}} \leq \epsilon_1. \tag{20}$$

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 (20) by plugging $\epsilon := \epsilon_1 H \sigma_L^3/31$ into Lemma 7, then plugging $\epsilon := \epsilon_1$ into Lemma 8. The high probability statement is obtained by applying a union bound.

Assuming inequality (20), for any $a \in \{1,2,3\}$, since $\|C_a^\diamond\|_{\text{op}} \le \sqrt{k}, \|\widehat{C}_a^\diamond - C_a^\diamond\|_{\text{op}} \le \sqrt{k}\epsilon_1$ and $\|W\|_{\text{op}} \le 1, \left\|\widehat{W} - W\right\|_{\text{op}} \le \epsilon_1$, Lemma 18 (the preconditions are satisfied by inequality (19)) implies that

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

Since condition (19) implies

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

Lemma 17 yields that

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

By Lemma 19, for any $i \in [m]$, the concentration bound

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

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 18 with preconditions

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

we have

$$\Bigg\| \underbrace{\left(\frac{1}{n}\sum_{j=1}^{n} z_{ij}Z_{aj}^T\right)\left(\widehat{W}\widehat{C}_a^\diamond\right)^{-1}}_{\widehat{G}} - \pi_i C_i \Bigg\|_{\text{op}} \le \frac{18\sqrt{k}\epsilon_1}{w_{\min}\sigma_L}. \tag{21}$$

Let $\widehat{G} \in \mathbb{R}^{k\times k}$ be the first term on the left hand side of inequality (21). 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

$$\|\widehat{G}_l - \pi_i\mu_{il}\|_1 \le \sqrt{k}\|\widehat{G}_l - \pi_i\mu_{il}\|_2 \le \sqrt{k}\|\widehat{G} - \pi_i C_i\|_{\text{op}} \le \frac{18k\epsilon_1}{w_{\min}\sigma_L}, \tag{22}$$

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

$$\begin{aligned}\left\|\text{normalize}(\widehat{G}_l) - \mu_{il}\right\|_2 &= \left\| \frac{\widehat{G}_l}{\pi_i + \sum_{c=1}^{k}\left(\widehat{G}_{lc} - \pi_i \mu_{ilc}\right)} - \mu_{il} \right\|_2 \\ &\leq \frac{\|\widehat{G}_l - \pi_i\mu_{il}\|_2 + \|\widehat{G}_l - \pi_i\mu_{il}\|_1 \|\mu_{il}\|_2}{\pi_i - \|\widehat{G}_l - \pi_i\mu_{il}\|_1} \\ &\leq \frac{72k\epsilon_1}{\pi_{\min} w_{\min} \sigma_L} \end{aligned} \tag{23}$$

where the last step combines inequalities (21), (22) with the bound $\frac{18k\epsilon_1}{w_{\min}\sigma_L} \leq \pi_i/2$ from condition (19), and uses the fact that $\|\mu_{il}\|_2 \leq 1$.

Note that inequality (23) 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

$$\epsilon_1 = \frac{31 \times 230 \cdot k^{3/2}}{\sqrt{n}\sigma_L^{11/2}} \left(1 + \sqrt{\log((36 + 6k + m)/\delta)}\right). \tag{24}$$

Combining condition (19) and inequality (23), we find that to make $\|\widehat{C} - C\|_\infty$ bounded by ϵ, it is sufficient to choose ϵ_1 such that

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

This condition can be further simplified to

$$\epsilon_1 \leq \frac{\epsilon \pi_{\min} w_{\min} \sigma_L}{72k} \tag{25}$$

for small ϵ, that is $\epsilon \leq \min\left\{ \frac{36\kappa k}{\pi_{\min} w_{\min} \sigma_L}, 2 \right\}$. According to equation (24), the condition (25) will be satisfied if

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

Squaring both sides of the inequality completes the proof.

A.1 Proof of Lemma 7

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

$$\left\| \frac{1}{n} \sum_{j=1}^{n} Z_{a'j} \otimes Z_{b'j} - \mathbb{E}[Z_{a'j} \otimes Z_{b'j}] \right\|_{\mathrm{op}} \leq \epsilon. \tag{26}$$

By Lemma 19 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 17, 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 18 implies that

$$\begin{aligned} &\left\| \left(\frac{1}{n} \sum_{j=1}^{n} Z_{cj} \otimes Z_{bj} \right) \left(\frac{1}{n} \sum_{j=1}^{n} Z_{aj} \otimes Z_{bj} \right)^{-1} - \mathbb{E}[Z_{cj} \otimes Z_{bj}](\mathbb{E}[Z_{aj} \otimes Z_{bj}])^{-1} \right\|_{\mathrm{op}} \\ &\quad \leq 2 \left(\frac{\epsilon}{\sigma_k(S_{ab})} + \frac{2\epsilon}{\sigma_k^2(S_{ab})} \right) \leq 6\epsilon/\sigma_L^2 \end{aligned} \tag{27}$$

and for the same reason, we have

$$\left\| \left(\frac{1}{n} \sum_{j=1}^{n} Z_{cj} \otimes Z_{aj} \right) \left(\frac{1}{n} \sum_{j=1}^{n} Z_{bj} \otimes Z_{aj} \right)^{-1} - \mathbb{E}[Z_{cj} \otimes Z_{aj}](\mathbb{E}[Z_{bj} \otimes Z_{aj}])^{-1} \right\|_{\mathrm{op}} \leq 6\epsilon/\sigma_L^2 \tag{28}$$

Now, let matrices F_2 and F_3 be defined as

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

and let the matrix on the left hand side of inequalities (27) and (28) be denoted by Δ_2 and Δ_3, we have

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

where the last steps uses inequality (27), (28) and the fact that $\max\{\|F_2\|_{\text{op}}, \|F_3\|_{\text{op}}\} \leq 1/\sigma_L$ and

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

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

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

Consequently, we have

$$\left\|\widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 (Z_{aj} \otimes Z_{bj}) F_3^T\right\|_{\text{op}} \leq 30\epsilon/\sigma_L^3. \tag{29}$$

For the rest of the proof, we use inequality (29) to bound $\widehat{M}_2$ and $\widehat{M}_3$. For the second moment, we have

$$\begin{aligned}
\left\|\widehat{M}_2 - M_2\right\|_{\text{op}} &\leq \frac{1}{n}\sum_{j=1}^n \left\|\widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 (Z_{aj} \otimes Z_{bj}) F_3^T\right\|_{\text{op}} + \left\|F_2 \left(\frac{1}{n}\sum_{j=1}^n Z_{aj} \otimes Z_{bj}\right) F_3^T - M_2\right\|_{\text{op}} \\
&\leq 30\epsilon/\sigma_L^3 + \left\|F_2 \left(\frac{1}{n}\sum_{j=1}^n Z_{aj} \otimes Z_{bj} - \mathbb{E}[Z_{aj} \otimes Z_{bj}]\right) F_3^T\right\|_{\text{op}} \\
&\leq 30\epsilon/\sigma_L^3 + \epsilon/\sigma_L^2 \leq 31\epsilon/\sigma_L^3.
\end{aligned}$$

For the third moment, we have

$$\begin{aligned}
\widehat{M}_3 - M_3 = \frac{1}{n}\sum_{j=1}^n &\left(\widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 (Z_{aj} \otimes Z_{bj}) F_3^T\right) \otimes Z_{cj} \\
&+ \left(\frac{1}{n}\sum_{j=1}^n F_2 (Z_{aj} \otimes Z_{bj}) F_3^T \otimes Z_{cj} - \mathbb{E}\left[F_2 (Z_{aj} \otimes Z_{bj}) F_3^T \otimes Z_{cj}\right]\right).
\end{aligned} \tag{30}$$

We examine the right hand side of equation (30). The first term is bounded as

$$\begin{aligned}
\left\|\left(\widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 (Z_{aj} \otimes Z_{bj}) F_3^T\right) \otimes Z_{cj}\right\|_{\text{op}} &\leq \left\|\widehat{Z}'_{aj} \otimes \widehat{Z}'_{bj} - F_2 (Z_{aj} \otimes Z_{bj}) F_3^T\right\|_{\text{op}} \|Z_{cj}\|_2 \\
&\leq 30\epsilon/\sigma_L^3.
\end{aligned} \tag{31}$$

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 19 implies that

$$\left\|\frac{1}{n}\sum_{j=1}^n F_2 (Z_{aj} \otimes Z_{bj}) F_3^T \otimes Z_{cj} - \mathbb{E}\left[F_2 (Z_{aj} \otimes Z_{bj}) F_3^T \otimes Z_{cj}\right]\right\|_{\text{op}} \leq \epsilon/\sigma_L^2 \tag{32}$$

with probability at least $1 - k\exp(-(\sqrt{n/k}\epsilon - 1)^2)$. Combining inequalities (31) and (32), we have

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

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

A.2 Proof of Lemma 8

Chaganty and Liang (2013) (Lemma 4) prove that when condition (18) holds, the tensor decomposition method of Algorithm 1 outputs $\{\widehat{\mu}_h^{\diamond}, \widehat{w}_h\}_{h=1}^k$, such that with probability at least $1-\delta$, a permutation π satisfies

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

Note that the constant H in Lemma 8 is obtained by plugging upper bounds $\|M_2\|_{\text{op}} \le 1$ and $\|M_3\|_{\text{op}} \le 1$ into Lemma 4 of Chaganty and Liang (2013).

The $\pi(h)$-th component of $\mu_{c\pi(h)}^{\diamond}$ is greater than other components of $\mu_{c\pi(h)}^{\diamond}$, by a margin of κ. Assuming $\epsilon \le \kappa/2$, the greatest component of $\widehat{\mu}_h^{\diamond}$ is its $\pi(h)$-th component. Thus, Algorithm 1 is able to correctly estimate the $\pi(h)$-th column of $\widehat{C}_c^{\diamond}$ by the vector $\widehat{\mu}_h^{\diamond}$. Consequently, for every column of $\widehat{C}_c^{\diamond}$, the ℓ_2-norm error is bounded by ϵ. Thus, the spectral-norm error of $\widehat{C}_c^{\diamond}$ 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 $\|\widehat{W} - W\|_{\text{op}} \le \epsilon$.

Appendix B. Proof of Theorem 4

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

$$\begin{aligned}
\mathcal{E}_1 &: \sum_{i=1}^{m}\sum_{c=1}^{k} \mathbb{I}(z_{ij} = e_c)\log(\mu_{iy_jc}/\mu_{ilc}) \ge m\overline{D}/2 \qquad \text{for all } j \in [n] \text{ and } l \in [k]\backslash\{y_j\}. \\
\mathcal{E}_2 &: \left|\sum_{j=1}^{n} \mathbb{I}(y_j = l)\mathbb{I}(z_{ij} = e_c) - nw_l\pi_i\mu_{ilc}\right| \le nt_{ilc} \qquad \text{for all } (i,l,c) \in [m]\times[k]^2. \\
\mathcal{E}_3 &: \left|\sum_{j=1}^{n} \mathbb{I}(y_j = l)\mathbb{I}(z_{ij} \ne 0) - nw_l\pi_i\right| \le \frac{nt_{ilc}}{\mu_{ilc}} \qquad \text{for all } (i,l,c) \in [m]\times[k]^2. \qquad (33)
\end{aligned}$$

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 (7) and (8) attains the desired level of accuracy. See Section B.1 for the proof.

Lemma 9 *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 (10) holds, and scalars t_{ilc} satisfy*

$$2\exp\Big(-m\overline{D}/4 + \log(k)\Big) \le t_{ilc} \le \pi_{\min}w_{\min}\min\left\{\frac{\rho}{8}, \frac{\rho\overline{D}}{64}\right\}. \qquad (34)$$

Then by alternating updates (7) and (8) for at least one round, the estimates $\widehat{C}$ and $\widehat{q}$ are bounded as

$$\begin{aligned}
|\widehat{\mu}_{ilc} - \mu_{ilc}| &\le 4t_{ilc}/(\pi_i w_l). && \text{for all } i \in [m],\ l \in [k],\ c \in [k]. \\
\max_{l\in[k]}\{|\widehat{q}_{jl} - \mathbb{I}(y_j = l)|\} &\le \exp\big(-m\overline{D}/4 + \log(k)\big) && \text{for all } j \in [n].
\end{aligned}$$

Next, we characterize the probability that events $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ 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 10 *Conditioning on any value of y_j, we have*

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

According to Lemma 10, the aggregated second moment of s_i is bounded by

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

Thus, applying the Bernstein inequality, we have

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

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

$$\mathbb{P}[\mathcal{E}_1] \geq 1 - kn \exp\left(-\frac{m\overline{D}}{33\log(1/\rho)}\right). \tag{35}$$

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 $t_{ilc} \leq \pi_{\min} w_{\min} \rho/8 \leq p$, applying the Chernoff bound implies

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

For measuring $\mathbb{P}[\mathcal{E}_3]$, note that $\sum_{j=1}^{n} \mathbb{I}(y_j = l)\mathbb{I}(z_{ij} \neq e_c)$ is the sum of n i.i.d. Bernoulli random variables with mean $q := \pi_i w_l$. Since $\frac{t_{ilc}}{\mu_{ilc}} \leq \frac{\pi_{\min} w_{\min} \rho/8}{\rho} \leq q$, using a Chernoff bound yields

$$\mathbb{P}\left[\Big|\sum_{j=1}^{n} \mathbb{I}(y_j = l)\mathbb{I}(z_{ij} \neq 0) - nq\Big| \geq n\frac{t_{ilc}}{\mu_{ilc}}\right] \leq 2\exp(-\frac{nt_{ilc}^2}{3q\mu_{ilc}^2}) \leq 2\exp\left(-\frac{nt_{ilc}^2}{3\pi_i w_l \mu_{ilc}}\right),$$

Summarizing the probability bounds on $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$, we conclude that $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ holds with probability at least

$$1 - kn\exp\left(-\frac{m\overline{D}}{33\log(1/\rho)}\right) - \sum_{i=1}^{m}\sum_{l=1}^{k} 4\exp\left(-\frac{nt_{ilc}^2}{3\pi_i w_l \mu_{ilc}}\right). \tag{36}$$

Proof of Part (a) According to Lemma 9, for $\widehat{y}_j = y_j$ being true, it sufficient to have $\exp(-m\overline{D}/4 + \log(k)) < 1/2$, or equivalently

$$m > 4\log(2k)/\overline{D}. \tag{37}$$

To ensure that this bound holds with probability at least $1-\delta$, expression (36) needs to be lower bounded by δ. It is achieved if we have

$$m \geq \frac{33\log(1/\rho)\log(2kn/\delta)}{\overline{D}} \quad \text{and} \quad n \geq \frac{3\pi_i w_l \mu_{ilc}\log(8mk/\delta)}{t_{ilc}^2} \tag{38}$$

If we choose

$$t_{ilc} := \sqrt{\frac{3\pi_i w_l \mu_{ilc}\log(8mk/\delta)}{n}}. \tag{39}$$

then the second part of condition (38) is guaranteed. To ensure that t_{ilc} satisfies condition (34). We need to have

$$\sqrt{\frac{3\pi_i w_l \mu_{ilc}\log(8mk/\delta)}{n}} \geq 2\exp\Big(-m\overline{D}/4 + \log(k)\Big) \quad \text{and}$$
$$\sqrt{\frac{3\pi_i w_l \mu_{ilc}\log(8mk/\delta)}{n}} \leq \pi_{\min} w_{\min}\alpha/4.$$

The above two conditions require that m and n satisfy

$$m \geq \frac{4\log(2k\sqrt{n/(3\pi_{\min} w_{\min}\rho\log(8mk/\delta))})}{\overline{D}} \tag{40}$$
$$n \geq \frac{48\log(8mk/\delta)}{\pi_{\min} w_{\min}\alpha^2} \tag{41}$$

The four conditions (37), (38), (40) and (41) are simultaneously satisfied if we have

$$m \geq \frac{33\log(1/\rho)\log(2kn/\delta)}{\overline{D}} \quad \text{and}$$
$$n \geq \frac{48\log(8mk/\delta)}{\pi_{\min} w_{\min}\alpha^2}.$$

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 (39), combining Lemma 9 with this assignment, we have

$$(\widehat{\mu}_{ilc} - \mu_{ilc})^2 \leq \frac{48\mu_{ilc}\log(8mk/\delta)}{\pi_i w_l n}$$

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

B.1 Proof of Lemma 9

To prove Lemma 9, we study the consequences of update (7) and update (8). We prove two important lemmas, which show that both updates provide good estimates if they are properly initialized.

Lemma 11 *Assume that event $\mathcal{E}_1$ holds. If μ and its estimate $\widehat{\mu}$ satisfy*

$$\mu_{ilc} \geq \rho \quad \text{and} \quad |\widehat{\mu}_{ilc} - \mu_{ilc}| \leq \delta_1 \qquad \text{for all } i \in [m],\, l \in [k],\, c \in [k], \tag{42}$$

and $\widehat{q}$ is updated by formula (7), then $\widehat{q}$ is bounded as:

$$\max_{l \in [k]} \{|\widehat{q}_{jl} - \mathbb{I}(y_j = l)|\} \leq \exp\left(-m\left(\frac{\overline{D}}{2} - \frac{2\delta_1}{\rho - \delta_1}\right) + \log(k)\right) \qquad \text{for all } j \in [n]. \tag{43}$$

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 (42) holds, we obtain that

$$\begin{aligned} A_l &= \sum_{i=1}^{m} \sum_{c=1}^{k} \mathbb{I}(z_{ij} = e_c) \log(\mu_{iy_jc} / \mu_{ilc}) + \sum_{i=1}^{m} \sum_{c=1}^{k} \mathbb{I}(z_{ij} = e_c) \left[\log\left(\frac{\widehat{\mu}_{iy_jc}}{\mu_{iy_jc}}\right) - \log\left(\frac{\widehat{\mu}_{ilc}}{\mu_{ilc}}\right)\right] \\ &\geq \left(\sum_{i=1}^{m} \frac{\pi_i \mathbb{D}_{\mathrm{KL}}\left(\mu_{iy_j}, \mu_{il}\right)}{2}\right) - 2m \log\left(\frac{\rho}{\rho - \delta_1}\right) \geq m\left(\frac{\overline{D}}{2} - \frac{2\delta_1}{\rho - \delta_1}\right). \end{aligned} \tag{44}$$

Thus, for every index $l \neq y_j$, combining formula (7) and inequality (44) 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

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

Combining the above two inequalities completes the proof. ■

Lemma 12 *Assume that event $\mathcal{E}_2$ holds. If $\widehat{q}$ satisfies*

$$\max_{l \in [k]} \{|\widehat{q}_{jl} - \mathbb{I}(y_j = l)|\} \leq \delta_2 \qquad \text{for all } j \in [n], \tag{45}$$

and $\widehat{\mu}$ is updated by formula (8), then $\widehat{\mu}$ is bounded as:

$$|\widehat{\mu}_{ilc} - \mu_{ilc}| \leq \frac{2nt_{ilc} + 2n\delta_2}{(7/8)n\pi_i w_l - n\delta_2}. \qquad \text{for all } i \in [m],\, l \in [k],\, c \in [k]. \tag{46}$$

Proof By formula (8), we can write $\widehat{\mu}_{ilc} = A/B$, where

$$A := \sum_{j=1}^{n} \widehat{q}_{jl} \mathbb{I}(z_{ij} = e_c) \quad \text{and} \quad B := \sum_{c'=1}^{k} \sum_{j=1}^{n} \widehat{q}_{jl} \mathbb{I}(z_{ij} = e_{c'}) = \sum_{j=1}^{n} \widehat{q}_{jl} \mathbb{I}(z_{ij} \neq 0).$$

Combining this definition with the assumption that event $\mathcal{E}_2$ and inequality (45) hold, we find that

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

Similarly, using the assumption that event $\mathcal{E}_3$ and inequality (45) hold, we have

$$\begin{aligned} |B - n\pi_i w_l| &\leq \left| \sum_{j=1}^{n} \mathbb{I}(q_{jl} = y_j) \mathbb{I}(z_{ij} \neq 0) - n\pi_i w_l \right| + \left| \sum_{j=1}^{n} \widehat{q}_{jl} \mathbb{I}(z_{ij} \neq 0) - \sum_{j=1}^{n} \mathbb{I}(q_{jl} = y_j) \mathbb{I}(z_{ij} \neq 0) \right| \\ &\leq \frac{n t_{ilc}}{\mu_{ilc}} + n\delta_2. \end{aligned}$$

Combining the bound for A and B, we obtain that

$$\begin{aligned} |\widehat{\mu}_{ilc} - \mu_{ilc}| &= \left| \frac{n\pi_i w_l \mu_{ilc} + (A - n\pi_i w_l \mu_{ilc})}{n\pi_i w_l + (B - n\pi_i w_l)} - \mu_{ilc} \right| = \left| \frac{(A - n\pi_i w_l \mu_{ilc}) - \mu_{ilc}(B - n\pi_i w_l)}{n\pi_i w_l + (B - n\pi_i w_l)} \right| \\ &\leq \frac{2n t_{ilc} + 2n\delta_2}{n\pi_i w_l - n(t_{ilc}/\mu_{ilc}) - n\delta_2}. \end{aligned}$$

Condition (34) implies

$$\frac{t_{ilc}}{\mu_{ilc}} \leq \frac{t_{ilc}}{\rho} \leq \frac{\pi_{\min} w_{\min} \rho}{8\rho} = \frac{\pi_{\min} w_{\min}}{8},$$

lower bounding the denominator. Plugging in this bound completes the proof. ■

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

$$\delta_1 := \min\left\{ \frac{\rho}{2}, \frac{\rho \overline{D}}{16} \right\} \quad \text{and} \quad \delta_2 := t_{\min}/2. \tag{47}$$

We claim that at any step in the update, the preconditions (42) and (45) always hold.

We prove the claim by induction. Before the iteration begins, $\widehat{\mu}$ is initialized such that the accuracy bound (10) holds. Thus, condition (42) is satisfied at the beginning. We assume by induction that condition (42) is satisfied at time $1, 2, \ldots, \tau - 1$ and condition (45) is satisfied at time $2, 3, \ldots, \tau - 1$. At time τ, either update (7) or update (8) is performed. If update (7) is performed, then by the inductive hypothesis, condition (42) holds before the update. Thus, Lemma 11 implies that

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

The assignment (47) implies $\frac{\overline{D}}{2} - \frac{2\delta_1}{\rho - \delta_1} \geq \frac{\overline{D}}{4}$, which yields that

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

where the last inequality follows from condition (34). It suggests that condition (45) holds after the update.

On the other hand, we assume that update (8) is performed at time τ. Since update (8) follows update (7), we have $\tau \geq 2$. By the inductive hypothesis, condition (45) holds before the update, so Lemma 12 implies

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

where the last step follows since $t_{\min} \leq t_{ilc}$. Noticing $\rho \leq 1$, condition (34) implies that $t_{\min} \leq \pi_{\min} w_{\min}/8$. Thus, the right hand side of the above inequality is bounded by $4t_{ilc}/(\pi_i w_l)$. Using condition (34) again, we find

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

which verifies that condition (42) holds after the update. This completes the induction.

Since preconditions (42) and (45) hold for any time $\tau \geq 2$, Lemma 11 and Lemma 12 implies that the concentration bounds (43) and (46) always hold. These two concentration bounds establish the lemma's conclusion.

B.2 Proof of Lemma 10

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 $x \geq \rho$ and $\rho < 1$, the following inequality holds:

$$\log^2(x) \leq \frac{2\log(1/\rho)}{1-\rho}(x - 1 - \log(x)). \tag{48}$$

We defer the proof of inequality (48), focusing on its consequence. Let $x := \mu_{ilc}/\mu_{iy_jc}$, then inequality (48) yields that

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

It remains to prove the claim (48). Let $f(x) := \log^2(x) - \frac{2\log(1/\rho)}{1-\rho}(x - 1 - \log(x))$. It suffices to show that $f(x) \leq 0$ for $x \geq \rho$. 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

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

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

For any $\rho \leq x < 1$, 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, $f'(x) \geq 0$ at any point $x \in [\rho, 1)$, which implies $f(x) \leq 0$.

Appendix C. Proof of Theorem 5

In this section we prove Theorem 5. The proof separates into two parts.

C.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

$$\begin{aligned}\max_{v\in[k]^n} \mathbb{E}\Big[\sum_{j=1}^{n} \mathbb{I}(\widehat{y}_j \neq y_j)\Big| y = v\Big] &\geq \sum_{v\in\{l,l'\}^n} \mathbb{Q}(v)\, \mathbb{E}\Big[\sum_{j=1}^{n} \mathbb{I}(\widehat{y}_j \neq y_j)\Big| y = v\Big] \\ &= \sum_{j=1}^{n}\sum_{v\in\{l,l'\}^n} \mathbb{Q}(v)\, \mathbb{E}\Big[\mathbb{I}(\widehat{y}_j \neq y_j)\Big| y = v\Big]. \end{aligned} \tag{49}$$

Thus, it is sufficient to lower bound the right hand side of inequality (49).

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 (Yu, 1997) and Pinsker's inequality, we have

$$\begin{aligned}\sum_{v\in\{l,l'\}^n} \mathbb{Q}(v)\, \mathbb{E}\Big[\mathbb{I}(\widehat{y}_j \neq y_j)\Big| y = v\Big] &= \mathbb{Q}(y_j = l)\mathbb{P}_0(\widehat{y}_j \neq l) + \mathbb{Q}(y_j = l')\mathbb{P}_1(\widehat{y}_j \neq l') \\ &\geq \frac{1}{2} - \frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_1\|_{\mathrm{TV}} \\ &\geq \frac{1}{2} - \frac{1}{4}\sqrt{\mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0, \mathbb{P}_1\right)}. \end{aligned} \tag{50}$$

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

$$\mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0, \mathbb{P}_1\right) = \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_j), \mathbb{P}_1(Z_j)\right) + \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z \backslash Z_j), \mathbb{P}_1(Z \backslash Z_j)\right) = \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_j), \mathbb{P}_1(Z_j)\right), \tag{51}$$

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

$$\begin{aligned}
\mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_j), \mathbb{P}_1(Z_j)\right) &= \sum_{i=1}^{m} \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(z_{ij}), \mathbb{P}_1(z_{ij})\right) \\
&= \sum_{i=1}^{m} \left[(1-\pi_i) \log\left(\frac{1-\pi_i}{1-\pi_i}\right) + \sum_{c=1}^{k} \pi_i \mu_{ilc} \log\left(\frac{\pi_i \mu_{ilc}}{\pi_i \mu_{il'c}}\right) \right] \\
&= \sum_{i=1}^{m} \sum_{c=1}^{k} \pi_i \mathbb{D}_{\mathrm{KL}}\left(\mu_{ilc}, \mu_{il'c}\right) = m\overline{D}.
\end{aligned} \tag{52}$$

Combining inequality (50) with equations (51) and (52), 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 (49) completes the proof.

C.2 Proof of Part (b)

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

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

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

$$\sup_{\mu \in \mathbb{R}^{m \times k \times k}} \mathbb{E}\Big[\|\widehat{\mu}_{\bar{i}\bar{l}} - \mu_{\bar{i}\bar{l}}\|_2^2\Big] \geq \sum_{v \in [k]^n} \mathbb{P}(y = v) \left(\sum_{V} \mathbb{P}(V)\, \mathbb{E}\Big[\|\widehat{\mu}_{\bar{i}\bar{l}} - \mu_{\bar{i}\bar{l}}\|_2^2 \Big| \mu = u_V, y = v\Big] \right). \tag{53}$$

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

$$\sum_V \mathbb{P}(V)\, \mathbb{E}[\|\widehat{\mu}_{\bar{i}\bar{l}} - \mu_{\bar{i}\bar{l}}\|_2^2 | \mu = u_V, y = v] \geq \frac{\delta^2}{2} \mathrm{P}[\widehat{V} \neq V_{\bar{i}\bar{l}} | y = v]. \tag{54}$$

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 (Yu, 1997) and Pinsker's inequality, we have

$$\begin{aligned} \mathrm{P}[\widehat{V} \neq V_{\bar{i}\bar{l}} | y = v] &\geq \frac{1}{2} - \frac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_1\|_{\mathrm{TV}} \\ &\geq \frac{1}{2} - \frac{1}{4}\sqrt{\mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0, \mathbb{P}_1\right)}. \end{aligned} \tag{55}$$

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

$$\mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0, \mathbb{P}_1\right) = \sum_{i=1}^{m} \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_i), \mathbb{P}_1(Z_i)\right) = \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}}), \mathbb{P}_1(Z_{\bar{i}})\right), \tag{56}$$

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

$$\begin{aligned} \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}}), \mathbb{P}_1(Z_{\bar{i}})\right) &= \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}J}), \mathbb{P}_1(Z_{\bar{i}J})\right) + \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}} \backslash Z_{\bar{i}J}), \mathbb{P}_1(Z_{\bar{i}} \backslash Z_{\bar{i}J})\right) \\ &= \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}J}), \mathbb{P}_1(Z_{\bar{i}J})\right), \end{aligned} \tag{57}$$

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

$$\begin{aligned} \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(Z_{\bar{i}J}), \mathbb{P}_1(Z_{\bar{i}J})\right) &= \sum_{j \in J} \mathbb{D}_{\mathrm{KL}}\left(\mathbb{P}_0(z_{\bar{i}j}), \mathbb{P}_1(z_{\bar{i}j})\right) = |J|\, \pi_{\bar{i}}\, \frac{1}{2} \log\left(\frac{1}{1 - 4\delta^2}\right) \\ &\leq \frac{5}{2}|J|\, \pi_{\bar{i}}\delta^2. \end{aligned} \tag{58}$$

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 (55) with upper bounds (56), (57) and (58), 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 (53) and (54) 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.

Appendix D. Proof of Theorem 6

Our proof strategy is briefly described as follow: We first upper bound the error of Step (1)-(2) in Algorithm 2. This upper bound is presented as lemma 13. Then, we analyze the performance of Step (3), taking the guarantee obtained from the previous two steps.

Lemma 13 *Assume that $\kappa_3 > 0$. Let $\widehat{p}_i$ be initialized by Step (1)-(2). For any scalar $0 < t < \frac{\overline{\kappa}\kappa_3^3}{18}$, the upper bound*

$$\max_{i\in[m]}\{|\widehat{p}_i - p_i|\} \leq \frac{18t}{\kappa_3^3} \tag{59}$$

holds with probability at least $1 - m^2\exp(-nt^2/2)$.

The rest of the proof upper bounds the error of Step (3). The proof follows very similar steps as in the proof of Theorem 4. We first define two events that will be shown holding with high probability.

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

$$\mathcal{E}_2: \quad \Big|\sum_{j=1}^{n}\mathbb{I}(z_{ij} = e_{y_j}) - np_i\Big| \leq nt_i \qquad \text{for all } i\in[m].$$

Lemma 14 *Assume that $\mathcal{E}_1\cap\mathcal{E}_2$ holds. Also assume that $\rho \leq p_i \leq 1-\rho$ for all $i\in[m]$. If $\widehat{p}$ is initialized such that*

$$|\widehat{p}_i - p_i| \leq \alpha := \min\left\{\frac{\overline{\kappa}}{2}, \frac{\rho}{2}, \frac{\rho\overline{D}}{16}\right\} \qquad \textit{for all } i\in[m] \tag{60}$$

and scalars t_i satisfy

$$\exp\Big(-m\overline{D}/4+\log(k)\Big)\le t_i\le \min\left\{\frac{\rho}{4},\frac{\rho\overline{D}}{32}\right\} \tag{61}$$

Then the estimates $\widehat{p}$ and $\widehat{q}$ obtained by alternating updates (14) and (15) satisfy:

$$\begin{aligned} |\widehat{p}_i-p_i|&\le 2t_i. && \text{for all } i\in[m].\\ \max_{l\in[k]}\{|\widehat{q}_{jl}-\mathbb{I}(y_j=l)|\}&\le \exp\big(-m\overline{D}/4+\log(k)\big) && \text{for all } j\in[n].\end{aligned}$$

As in the proof of Theorem 4, we can lower bound the probability of the event $\mathcal{E}_1\cap\mathcal{E}_2$ by applying Bernstein's inequality and the Chernoff bound. In particular, the following bound holds:

$$\mathbb{P}[\mathcal{E}_1\cap\mathcal{E}_2]\ge 1-kn\exp\left(-\frac{m\overline{D}}{33\log(1/\rho)}\right)-\sum_{i=1}^{m}2\exp\left(-\frac{nt_i^2}{3p_i}\right). \tag{62}$$

The proof of inequality (62) precisely follows the proof of Theorem 4.

Proof of upper bounds (a) and (b) in Theorem 6 To apply Lemma 14, we need to ensure that condition (60) holds. If we assign $t:=\alpha\kappa_3^3/18$ in Lemma 13, then condition (60) holds with probability at least $1-m^2\exp(-n\alpha^2\kappa_3^6/648)$. To ensure that this event holds with probability at least $1-\delta/3$, we need to have

$$n\ge\frac{648\log(3m^2/\delta)}{\alpha^2\kappa_3^6}. \tag{63}$$

By Lemma 14, for $\widehat{y}_j=y_j$ being true, it suffices to have

$$m>4\log(2k)/\overline{D} \tag{64}$$

To ensure that $\mathcal{E}_1\cap\mathcal{E}_2$ holds with probability at least $1-2\delta/3$, expression (62) needs to be lower bounded by $1-2\delta/3$. It is achieved by

$$m\ge\frac{33\log(1/\rho)\log(3kn/\delta)}{\overline{D}}\quad\text{and}\quad n\ge\frac{3p_i\log(6m/\delta)}{t_i^2} \tag{65}$$

If we choose

$$t_i:=\sqrt{\frac{3\log(6m/\delta)}{n}}. \tag{66}$$

then the second part of condition (65) is guaranteed. To ensure that t_{ilc} satisfies condition (61). We need to have

$$\begin{aligned}&\sqrt{\frac{3\log(6m/\delta)}{n}}\ge\exp\Big(-m\overline{D}/4+\log(k)\Big)\quad\text{and}\\ &\sqrt{\frac{3\log(6m/\delta)}{n}}\le\alpha/2.\end{aligned}$$

The above two conditions requires that m and n satisfy

$$m \geq \frac{4\log(k\sqrt{n/(3\log(6m/\delta))})}{\overline{D}} \tag{67}$$

$$n \geq \frac{12\log(6m/\delta)}{\alpha^2}. \tag{68}$$

The five conditions (63), (64), (65), (67) and (68) are simultaneously satisfied if we have

$$m \geq \frac{33\log(1/\rho)\log(3kn/\delta)}{\overline{D}} \quad \text{and}$$
$$n \geq \frac{648\log(3m^2/\delta)}{\alpha^2\kappa_3^6}.$$

Under this setup, $\widehat{y}_j = y_j$ holds for all $j \in [n]$ with probability at least $1-\delta$. Combining equation (66) with Lemma 14, the bound

$$|\widehat{p}_i - p_i| \leq 2\sqrt{\frac{3\log(6m/\delta)}{n}}$$

holds with probability at least $1-\delta$.

D.1 Proof of Lemma 13

We claim that after initializing $\widehat{p}$ via formula (13), it satisfies

$$\min\left\{\max_{i\in[m]}\{|\widehat{p}_i - p_i|\}, \max_{i\in[m]}\{|\widehat{p}_i - (2/k - p_i)|\}\right\} \leq \frac{18t}{\kappa_3^3} \tag{69}$$

with probability at least $1-m^2\exp(-nt^2/2)$. Assuming inequality (69), it is straightforward to see that this bound is preserved by the algorithm's step (2). In addition, step (2) ensures that $\frac{1}{m}\sum_{i=1}^m \widehat{p}_i \geq \frac{1}{k}$, which implies

$$\begin{aligned}\max_{i\in[m]}\{|\widehat{p}_i - (2/k - p_i)|\} &\geq \left|\frac{1}{m}\sum_{i=1}^m \widehat{p}_i - \left(\frac{2}{k} - \frac{1}{m}\sum_{i=1}^m p_i\right)\right| \\ &\geq \frac{1}{k} - \left(\frac{1}{k} - \overline{\kappa}\right) = \overline{\kappa} > \frac{18t}{\kappa_3^3}.\end{aligned} \tag{70}$$

Combining inequalities (69) and (70) establishes the lemma.

We turn to prove claim (69). For any worker a and worker b, it is obvious that $\mathbb{I}(z_{aj} = z_{bj})$ are independent random variables for $j = 1, 2, \ldots, n$. Since

$$\mathbb{E}[\mathbb{I}(z_{aj} = z_{bj})] = p_a p_b + (k-1)\,\frac{1-p_a}{k-1}\,\frac{1-p_b}{k-1} = \frac{k}{k-1}\,(p_a - 1/k)(p_b - 1/k) + \frac{1}{k}$$

and $\frac{k-1}{k}(\mathbb{I}(z_{aj} = z_{bj}) - \frac{1}{k})$ belongs to the interval $[-1, 1]$, applying Hoeffding's inequality implies that

$$\mathbb{P}(|N_{ab} - (p_a - 1/k)(p_b - 1/k)| \leq t) \geq 1 - \exp(-nt^2/2) \quad \text{for any } t > 0.$$

By applying the union bound, the inequality

$$|N_{ab} - (p_a - 1/k)(p_b - 1/k)| \le t \tag{71}$$

holds for all $(a, b) \in [m]^2$ with probability at least $1 - m^2 \exp(-nt^2/2)$. For the rest of the proof, we assume that this high-probability event holds.

Given an arbitrary index i, we take indices (a_i, b_i) such that

$$(a_i, b_i) = \arg\max_{(a,b)} \{|N_{ab}| : a \neq b \neq i\}. \tag{72}$$

We consider another two indices (a^*, b^*) such that $|p_{a^*} - 1/k|$ and $|p_{b^*} - 1/k|$ are the two greatest elements in $\{|p_a - 1/k| : a \in [m] \backslash \{i\}\}$. Let $\beta_i := p_i - 1/k$ be a shorthand notation, then inequality (71) and equation (72) yields that

$$|\beta_{a_i}\beta_{b_i}| \ge |N_{a_ib_i}| - t \ge |N_{a^*b^*}| - t \ge |\beta_{a^*}\beta_{b^*}| - 2t \ge |\beta_{a^*}\beta_{b^*}|/2, \tag{73}$$

where the last step follows since $2t \le \kappa_3^2/2 \le |\beta_{a^*}\beta_{b^*}|/2$. Note that $|\beta_{b_i}| \le |\beta_{a^*}|$ (since $|\beta_{a^*}|$ is the largest entry by its definition), inequality (73) implies that $|\beta_{a_i}| \ge \frac{|\beta_{b^*}\beta_{a^*}|}{2|\beta_{b_i}|} \ge \frac{|\beta_{b^*}|}{2} \ge \kappa_3/2$. By the same argument, we obtain $|\beta_{b_i}| \ge |\beta_{b^*}|/2 \ge \kappa_3/2$. To upper bound the estimation error, we write $|N_{ia_i}|$, $|N_{ib_i}|$, $|N_{a_ib_i}|$ in the form of

$$\begin{aligned} |N_{ia_i}| &= |\beta_i\beta_{a_i}| + \delta_1 \\ |N_{ib_i}| &= |\beta_i\beta_{b_i}| + \delta_2 \\ |N_{a_ib_i}| &= |\beta_{a_i}\beta_{b_i}| + \delta_3, \end{aligned}$$

where $|\delta_1|, |\delta_2|, |\delta_3| \le t$. Firstly, notice that $N_{ia_i}, N_{ib_i} \in [-1, 1]$, thus,

$$\left|\sqrt{\frac{|N_{ia_i}N_{ib_i}|}{|N_{a_ib_i}|}} - \sqrt{\frac{|N_{ia_i}N_{ib_i}|}{|\beta_{a_i}\beta_{b_i}|}}\right| \le \left|\frac{1}{\sqrt{|N_{a_ib_i}|}} - \frac{1}{\sqrt{|\beta_{a_i}\beta_{b_i}|}}\right| \le \frac{t}{2(|\beta_{a_i}\beta_{b_i}| - t)^{3/2}} \le \frac{t}{(\kappa_3^2/4)^{3/2}}, \tag{74}$$

where the last step relies on the inequality $|\beta_{a_i}\beta_{b_i}| - t \ge \kappa_3^2/4$ obtained by inequality (73). Secondly, we upper bound the difference between $\sqrt{|N_{ia_i}N_{ib_i}|}$ and $\sqrt{|\beta_i^2\beta_{a_i}\beta_{b_i}|}$. If $|\beta_i| \le t$, using the fact that $|\beta_{a_i}|, |\beta_{b_i}| \le 1$, we have

$$\left|\sqrt{|N_{ia_i}N_{ib_i}|} - \sqrt{|\beta_i^2\beta_{a_i}\beta_{b_i}|}\right| \le \sqrt{|N_{ia_i}N_{ib_i}|} + \sqrt{|\beta_i^2\beta_{a_i}\beta_{b_i}|} \le \sqrt{4t^2} + \sqrt{t^2} \le 3t.$$

If $|\beta_i| > t$, using the fact that $|\beta_{a_i}|, |\beta_{b_i}| \in [\kappa_3/2, 1]$ and $|\beta_{a_i}\beta_{b_i}| \ge \kappa_3^2/2$, we have

$$\begin{aligned} \left|\sqrt{|N_{ia_i}N_{ib_i}|} - \sqrt{|\beta_i^2\beta_{a_i}\beta_{b_i}|}\right| &\le \frac{|\beta_i\beta_{b_i}\delta_1| + |\beta_i\beta_{a_i}\delta_2| + |\delta_1\delta_2|}{\sqrt{|\beta_i^2\beta_{a_i}\beta_{b_i}|}} \\ &\le \frac{|\delta_1|}{\sqrt{|\beta_{a_i}/\beta_{b_i}|}} + \frac{|\delta_2|}{\sqrt{|\beta_{b_i}/\beta_{a_i}|}} + \frac{|\delta_1\delta_2|}{t\sqrt{|\beta_{a_i}\beta_{b_i}|}} \\ &\le 3\sqrt{2}t/\kappa_3. \end{aligned}$$

Combining the above two upper bounds implies

$$\left|\sqrt{\frac{|N_{ia_i}N_{ib_i}|}{|\beta_{a_i}\beta_{b_i}|}}-\sqrt{\frac{|\beta_i^2\beta_{a_i}\beta_{b_i}|}{|\beta_{a_i}\beta_{b_i}|}}\right|=\frac{\left|\sqrt{|N_{ia_i}N_{ib_i}|}-\sqrt{|\beta_i^2\beta_{a_i}\beta_{b_i}|}\right|}{\sqrt{|\beta_{a_i}\beta_{b_i}|}}\leq\frac{6t}{\kappa_3^2}. \tag{75}$$

Combining inequalities (74) and (75), we obtain

$$\left|\sqrt{\frac{|N_{ia_i}N_{ib_i}|}{|N_{a_ib_i}|}}-|\beta_i|\right|\leq\frac{14t}{\kappa_3^3}. \tag{76}$$

Finally, we turn to analyzing the sign of N_{ia_1}. According to inequality (71), we have

$$N_{ia_1}=\beta_i\beta_{a_1}+\delta_4,$$

where $|\delta_4|\leq t$. Following the same argument for β_{a_i} and β_{b_i}, it was shown that $|\beta_{a_1}|\geq\kappa_3/2$. We combine inequality (76) with a case study of $\mathrm{sign}(N_{ia_1})$ to complete the proof. Let

$$\widehat{p}_i:=\frac{1}{k}+\mathrm{sign}(N_{ia_1})\sqrt{\frac{|N_{ia_i}N_{ib_i}|}{|N_{a_ib_i}|}}.$$

If $\mathrm{sign}(N_{ia_1})\neq\mathrm{sign}(\beta_i\beta_{a_1})$, then $|\beta_i\beta_{a_1}|\leq|\delta_4|\leq t$. Thus, $|\beta_i|\leq t/|\beta_{a_i}|\leq 2t/\kappa_3$, and consequently,

$$\begin{aligned}\max\{|\widehat{p}_i-p_i|,|\widehat{p}_i-(2/k-p_i)|\}&\leq\left|\sqrt{\frac{|N_{ia_i}N_{ib_i}|}{|N_{a_ib_i}|}}\right|+|p_i-1/k|\\&\leq\left|\sqrt{\frac{|N_{ia_i}N_{ib_i}|}{|N_{a_ib_i}|}}-|p_i-1/k|\right|+2|p_i-1/k|\leq\frac{18t}{\kappa_3^3}\end{aligned} \tag{77}$$

Otherwise, we have $\mathrm{sign}(N_{ia_1})=\mathrm{sign}(\beta_i\beta_{a_1})$ and consequently $\mathrm{sign}(\beta_i)=\mathrm{sign}(N_{ia_1})\mathrm{sign}(\beta_{a_1})$. If $\mathrm{sign}(\beta_{a_1})=1$, then $\mathrm{sign}(\beta_i)=\mathrm{sign}(N_{ia_1})$, which yields that

$$|\widehat{p}_i-p_i|=\left|\mathrm{sign}(N_{ia_1})\sqrt{\frac{|N_{ia_i}N_{ib_i}|}{|N_{a_ib_i}|}}-\mathrm{sign}(\beta_i)|\beta_i|\right|\leq\frac{14t}{\kappa_3^3}. \tag{78}$$

If $\mathrm{sign}(\beta_{a_1})=-1$, then $\mathrm{sign}(\beta_i)=-\mathrm{sign}(N_{ia_1})$, which yields that

$$|\widehat{p}_i-(2/k-p_i)|=\left|\mathrm{sign}(N_{ia_1})\sqrt{\frac{|N_{ia_i}N_{ib_i}|}{|N_{a_ib_i}|}}+\mathrm{sign}(\beta_i)|\beta_i|\right|\leq\frac{14t}{\kappa_3^3}. \tag{79}$$

Combining inequalities (77), (78) and (79), we find that

$$\min\left\{\max_{i\in[m]}\{|\widehat{p}_i-p_i|\},\max_{i\in[m]}\{|\widehat{p}_i-(2/k-p_i)|\}\right\}\leq\frac{18t}{\kappa_3^3}.$$

which establishes claim (69).

D.2 Proof of Lemma 14

The proof follows the argument in the proof of Lemma 9. We present two lemmas upper bounding the error of update (14) and update (15), assuming proper initialization.

Lemma 15 *Assume that event $\mathcal{E}_1$ holds. If p and its estimate $\widehat{p}$ satisfies*

$$\rho \le p_i \le 1-\rho \quad \text{and} \quad |\widehat{p}_i - p_i| \le \delta_1 \qquad \text{for all } i \in [m], \tag{80}$$

and $\widehat{q}$ is updated by formula (14), then $\widehat{q}$ is bounded as:

$$\max_{l\in[k]}\{|\widehat{q}_{jl} - \mathbb{I}(y_j = l)|\} \le \exp\left(-m\left(\frac{\overline{D}}{2} - \frac{2\delta_1}{\rho-\delta_1}\right) + \log(k)\right) \qquad \text{for all } j \in [n]. \tag{81}$$

Proof Following the proof of Lemma 11, the lemma is established since both $|\log(\widehat{p}_i/p_i)|$ and $|\log((1-\widehat{p}_i)/(1-p_i))|$ are bounded by $\log(\rho/(\rho-\delta_1))$. ■

Lemma 16 *Assume that event $\mathcal{E}_2$ holds. If $\widehat{q}$ satisfies*

$$\max_{l\in[k]}\{|\widehat{q}_{jl} - \mathbb{I}(y_j = l)|\} \le \delta_2 \qquad \text{for all } j \in [n], \tag{82}$$

and $\widehat{p}$ is updated by formula (15), then $\widehat{p}$ is bounded as:

$$|\widehat{p}_i - p_i| \le t_i + \delta_2. \qquad \text{for all } i \in [m]. \tag{83}$$

Proof By formula (15), we have

$$\widehat{p}_i - p_i = \frac{1}{n}\left(\sum_{j=1}^{n}\mathbb{I}(z_{ij} = e_{y_i}) - np_i\right) + \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k}(\widehat{q}_{il} - \mathbb{I}(y_j = l))\mathbb{I}(z_{ij} = e_l).$$

Combining inequality (82) with the inequality implied by event $\mathcal{E}_2$ completes the proof. ■

Following the steps in the proof of Lemma 9, we assign specific values to δ_1 and δ_2. Let

$$\delta_1 := \min\left\{\frac{\rho}{2}, \frac{\rho\overline{D}}{16}\right\} \quad \text{and} \quad \delta_2 := \min_{i\in[m]}\{t_i\}.$$

By the same inductive argument for proving Lemma 9, we can show that the upper bounds (81) and (83) always hold after the first iteration. Plugging the assignments of δ_1 and δ_2 into upper bounds (81) and (83) completes the proof.

Appendix E. Basic Lemmas

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

Lemma 17 (Matrix Inversion) *Let $A, E \in \mathbb{R}^{k\times k}$ be given, where A is invertible and E satisfies that $\|E\|_{\mathrm{op}} \le \sigma_k(A)/2$. Then*

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

Proof A little bit of algebra reveals that

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

Thus, we have

$$\|(A+E)^{-1} - A^{-1}\|_{\mathrm{op}} \le \frac{\|E\|_{\mathrm{op}}}{\sigma_k(A)\sigma_k(A+E)}.$$

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

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

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

Lemma 18 (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 $\|A_i\|_{\mathrm{op}} \le K_i$, $\|E_i\|_{\mathrm{op}} \le K_i$. Then*

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

Proof By the triangle inequality, we have

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

which completes the proof. ■

Lemma 19 (Matrix and Tensor Concentration) *Let $\{X_j\}_{j=1}^n$, $\{Y_j\}_{j=1}^n$ and $\{Z_j\}_{j=1}^n$ be i.i.d. samples from some distribution over $\mathbb{R}^k$ with bounded support ($\|X\|_2 \leq 1$, $\|Y\|_2 \leq 1$ and $\|Z\|_2 \leq 1$ with probability 1). Then with probability at least $1-\delta$,*

$$\left\| \frac{1}{n} \sum_{j=1}^{n} X_j \otimes Y_j - \mathbb{E}[X_1 \otimes Y_1] \right\|_F \leq \frac{1+\sqrt{\log(1/\delta)}}{\sqrt{n}}. \tag{84}$$

$$\left\| \frac{1}{n} \sum_{j=1}^{n} X_j \otimes Y_j \otimes Z_j - \mathbb{E}[X_1 \otimes Y_1 \otimes Z_1] \right\|_F \leq \frac{1+\sqrt{\log(k/\delta)}}{\sqrt{n/k}}. \tag{85}$$

Proof Inequality (84) is proved in Lemma D.1 of Anandkumar et al. (2015). To prove inequality (85), we note that for any tensor $T \in \mathbb{R}^{k\times k\times k}$, we can define k-by-k matrices $T_1, \ldots, T_k$ 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 (85), 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 (84), we find that with probability at least $1-k\delta'$, we have

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

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

References

Anima Anandkumar, Daniel Hsu, and Sham M. Kakade. A method of moments for mixture models and hidden Markov models. In *Proceedings of the Annual Conference on Learning Theory*, 2012.

Anima Anandkumar, Rong Ge, Daniel Hsu, and Sham M. Kakade. A tensor spectral approach to learning mixed membership community models. In *Proceedings of the Annual Conference on Learning Theory*, 2013.

Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

Animashree Anandkumar, Dean P. Foster, Daniel Hsu, Sham M. Kakade, and Yi-Kai Liu. A spectral algorithm for latent Dirichlet allocation. *Algorithmica*, 72(1):193–214, 2015.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statisics (to appear)*, 2016.

Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. In *Proceedings of the International Conference on Machine Learning*, 2013.

Xi Chen, Qihang Lin, and Dengyong Zhou. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *Proceedings of the International Conference on Machine Learning*, 2013.

Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of the World Wide Web Conference*, 2013.

Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society, Series C*, 28 (1):20–28, 1979.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Chao Gao and Dengyong Zhou. Minimax optimal convergence rates for estimating ground truth from crowdsourced labels. *arXiv preprint arXiv:1310.5764*, 2014.

Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators? Crowdsourcing abuse detection in user-generated content. In *Proceedings of the ACM Conference on Electronic Commerce*, 2011.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Prateek Jain and Sewoong Oh. Learning mixtures of discrete product distributions using spectral decompositions. In *Proceedings of the Conference on Learning Theory*, 2014.

David R. Karger, Sewoong Oh, and Devavrat Shah. Efficient crowdsourcing for multi-class labeling. In *Proceedings of the ACM SIGMETRICS*, 2013.

David R. Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

Matthew Lease and Gabriella Kazai. Overview of the TREC 2011 crowdsourcing track. In *Proceedings of TREC 2011*, 2011.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2nd edition, 2003.

Percy Liang. Partial information from spectral methods. NIPS Spectral Learning Workshop, 2013.

Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems*, 2012.

Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Klugera. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258, 2014.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11: 1297–1322, 2010.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.

Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, 2010.

Jacob Whitehill, Ting F. Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, 2009.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Dengyong Zhou, John C. Platt, Sumit Basu, and Yi Mao. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems*, 2012.

Dengyong Zhou, Qiang Liu, John C. Platt, and Christopher Meek. Aggregating ordinal labels from crowds by minimax conditional entropy. In *Proceedings of the International Conference on Machine Learning*, 2014.

James Zou, Daniel Hsu, David Parkes, and Ryan Adams. Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems*, 2013.