Exact Exponent in Optimal Rates for Crowdsourcing

Chao Gao CHAO.GAO@YALE.EDU
Yale University, 24 Hillhouse Ave, New Haven, CT 06511 USA

Yu Lu YU.LU@YALE.EDU
Yale University, 24 Hillhouse Ave, New Haven, CT 06511 USA

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

Abstract

In many machine learning applications, crowdsourcing has become the primary means for label collection. In this paper, we study the optimal error rate for aggregating labels provided by a set of non-expert workers. Under the classic Dawid-Skene model, we establish matching upper and lower bounds with an exact exponent $mI(\pi)$ in which m is the number of workers and $I(\pi)$ the average Chernoff information that characterizes the workers' collective ability. Such an exact characterization of the error exponent allows us to state a precise sample size requirement $m > \frac{1}{I(\pi)} \log \frac{1}{\epsilon}$ in order to achieve an ϵ misclassification error. In addition, our results imply the optimality of various EM algorithms for crowdsourcing initialized by consistent estimators.

1. Introduction

In many machine learning problems such as image classification and speech recognition, we need a large amount of labeled data. Crowdsourcing provides an efficient while inexpensive way to collect labels. On a commercial crowdsourcing platform like Amazon Mechanical Turk (Amazon Mechanical Turk), in general, it takes only few hours to obtain hundreds of thousands labels from crowdsourcing workers worldwide, and each label costs only several cents.

Though massive in amount, the crowdsourced labels are usually fairly noisy. The low quality is partially due to the lack of domain expertise from the workers and presence of spammers. To overcome this issue, a common strategy is to repeatedly label each item by different workers, and then estimate truth from the redundant labels, for example, using majority voting. Since the pioneering work by Dawid and Skene (Dawid & Skene, 1979), which jointly estimates truth and workers' abilities via a simple EM algorithm, various approaches have been developed in recent years for aggregating noisy crowdsourced labels. See (Whitehill et al., 2009; Welinder et al., 2010; Raykar et al., 2010; Ghosh et al., 2011; Bachrach et al., 2012; Liu et al., 2012; Zhou et al., 2012; Dalvi et al., 2013; Zhou et al., 2014; Venanzi et al., 2014; Parisi et al., 2014; Tian & Zhu, 2015) and references therein.

Compared with the active progress in aggregation algorithms, statistical understandings of crowdsourcing do not get much attention except (Gao & Zhou, 2013; Karger et al., 2014; Zhang et al., 2014; Berend & Kontorovich, 2015). These papers not only show exponential convergence rates for several estimators, they also provide lower bounds to justify the optimality of the rates. However, the exponents found in these work are not matched in their upper and lower bounds. They are optimal only up to some constants. The main focus of this paper is to find the *exact* error exponent to better guide algorithm design and optimization.

Main Contribution. We study the minimax rate of misclassification for estimating the truth from crowdsourced labels. We provide upper and lower bounds with exact exponents that match each other. The exponent has a natural interpretation of the collective wisdom of a crowd. In the special case where each worker's ability is modeled by a real number $p_i \in [0,1]$, the exponent takes a simple form $-(1+o(1))mI(p)$ with $I(p) = -\frac{1}{m}\sum_{i=1}^{m} \log\left(2\sqrt{p_i(1-p_i)}\right)$ being the average Rényi divergence of order $1/2$. Therefore, in order to achieve an error of ϵ in the misclassification proportion, it is necessary and sufficient that the number of workers

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

m satisfies $m \geq (1+o(1))I(p)^{-1}\log(1/\epsilon)$. Note that in previous work, only $m = \Omega\left(I(p)^{-1}\log(1/\epsilon)\right)$ can be claimed. Moreover, our general theorem has implications on the convergence rates of several existing algorithms.

This paper is organized as follows. In Section 2, we present the problem setting. In Section 3, given the workers' abilities, we derive the optimal error exponent. In Section 4, we show that spectral methods can be used to achieve the optimal error exponent, followed by a discuss on other algorithms in Section 5. The main proofs are given in Section 6, and the remaining proofs are gathered in the supplementary material.

2. Problem Setting

Let us start from the classic model proposed by Dawid and Skene (Dawid & Skene, 1979). Assume there are m workers and n items to label. Denote the true label of the jth item by y_j that takes on a value in $[k] = \{1, 2, ..., k\}$. Let X_{ij} be the label given by the ith worker to the jth item. The ability of the ith worker is assumed to be fully characterized by a confusion matrix

$$\pi^{(i)}_{gh} = \mathbb{P}(X_{ij} = h | y_j = g). \tag{1}$$

which satisfies the probabilistic constraint $\sum_{h=1}^{k}\pi^{(i)}_{gh} = 1$. Given $y_j = g$, X_{ij} is generated by a multinomial distribution with parameter $\pi^{(i)}_{g*} = \left(\pi^{(i)}_{g1}, ..., \pi^{(i)}_{gk}\right)$. Our goal is to estimate the true labels $y = (y_1, \cdots, y_n)$ using the observed labels $\{X_{ij}\}$. Denote the estimate by $\hat{y} = (\hat{y}_1, ..., \hat{y}_n)$. The loss is measured by the error rate

$$L(\hat{y}, y) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\{\hat{y}_j \neq y_j\}. \tag{2}$$

We would like to remark that the true labels are considered as deterministic here. It is straightforward to generalize our results to stochastic labels generated from a distribution. Also, we assume that every worker has labeled every item. Otherwise, we can regard the missing labels as a new category and the results in this paper stay the same.

3. Main Results

In this section, we assume the confusion matrices $\{\pi^{(i)}\}$ are known. Our goal is to establish the optimal error rate with respect to the loss in Equation (2). Let $\mathbb{P}_{\pi,y}$ be the joint probability distribution of the data $\{X_{ij}\}$ given π and y specified in (1), and let $\mathbb{E}_{\pi,y}$ be the associated expectation operator. Then the optimality is characterized by

$$\mathcal{M} = \inf_{\hat{y}}\sup_{y\in[k]^n}\mathbb{E}_{\pi,y}L(\hat{y}, y), \tag{3}$$

which identifies the lowest error rate that we can achieve uniformly over all possible true labels.

Our main result of the paper is to show that under some mild condition the minimax risk (3) converges to zero exponentially fast with an exponent that characterizes the collective wisdom of a crowd. Specifically, the error exponent is $-mI(\pi)$ with

$$I(\pi) = \min_{g\neq h} C(\pi_{g*}, \pi_{h*}), \tag{4}$$

where $C(\pi_{g*}, \pi_{h*})$ is given as

$$-\min_{0\leq t\leq 1}\frac{1}{m}\sum_{i=1}^{m}\log\left(\sum_{l=1}^{k}\left(\pi^{(i)}_{gl}\right)^{1-t}\left(\pi^{(i)}_{hl}\right)^{t}\right).$$

To better present our main result, let us introduce some notations. Let $\rho_m = \min_{i,g,l}\pi^{(i)}_{gl}$. Suppose the minimum of $C(\pi_{g*}, \pi_{h*})$ is achieved at $g = a$ and $h = b$. For any $\alpha > 0$, we define a set of workers

$$\mathcal{A}_\alpha = \left\{i \in [m] : \pi^{(i)}_{aa} \geq (1+\alpha)\pi^{(i)}_{ab}, \pi^{(i)}_{bb} \geq (1+\alpha)\pi^{(i)}_{ba}\right\}.$$

These workers in $\mathcal{A}_\alpha$ have better expertise in distinguishing between categories a and b. Then, our main result can be summarized into the following theorem.

Theorem 3.1. *Assume* $\log k = o(mI(\pi))$, $|\log\rho_m| = o(\rho_m|\mathcal{A}_{0.01}|^{1/2})$ *and* $|\log\rho_m| = o(\sqrt{m}I(\pi))$, *as* $m \to \infty$. *Then, we have*

$$\inf_{\hat{y}}\sup_{y\in[k]^n}\mathbb{E}_{\pi,y}L(\hat{y}, y) = \exp\left(-(1+o(1))mI(\pi)\right),$$

where $I(\pi)$ *is defined by (4).*

In Theorem 3.1, the assumption that $|\log\rho_m| = o(\rho_m|\mathcal{A}_{0.01}|^{1/2})$ can be relaxed to that $|\log\rho_m| = o(\rho_m\alpha|\mathcal{A}_\alpha|^{1/2})$ for some $\alpha > 0$. To better present our result, we set $\alpha = 0.01$ in the theorem. To prove the upper bound, we only need the first assumption $\log k = o(mI(\pi))$. The other two assumptions on ρ_m are used for proving the lower bound. One could imagine that the larger ρ_m is, the more mistake we might make to estimate the true labels. When there is a constant c (independent of m) such that $\rho_m \geq c$, the last two assumptions reduce to $|\mathcal{A}_{0.01}| \to \infty$ and $\sqrt{m}I(\pi) \to \infty$. That means as long as $I(\pi) = \Omega(1/\sqrt{m})$ and the number of experts goes to infinity as m grows, $\exp(-(1+o(1))mI(\pi))$ serves as a valid lower bound.

Theorem 3.1 characterizes the optimal error rate for estimating the ground truth with crowdsourced labels. It implies $\exp\left(-(1+o(1))mI(\pi)\right)$ is the best error rate that can be achieved by any algorithm. Moreover, it also implies there exists an algorithm that can achieve this optimal

rate. The error exponent depends on an important quantity $I(\pi)$. When $m = 1$ and $k = 2$, this theorem reduces to the Chernoff-Stein Lemma (Cover & Thomas, 2006), in which $I(\pi)$ is the Chernoff information between probability distributions. For the general problem, $C(\pi_{g*}, \pi_{h*})$ can be understood as the average Chernoff information between $\{\pi_{g*}^{(i)}\}_{i=1}^m$ and $\{\pi_{h*}^{(i)}\}_{i=1}^m$, which measures the collective ability of the m workers to distinguish between items with label g and items with label h. Then, $I(\pi)$ is the collective ability of the m workers to distinguish between any two items of different labels. The higher the overall collective ability $mI(\pi)$, the smaller the optimal rate.

By Markov's inequality, Theorem 3.1 implies

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\{\hat{y}_j \neq y_j\} \leq \exp\left(-(1+o(1))mI(\pi)\right),$$

with probability tending to 1. This allows a precise statement for a sample size requirement to achieve a prescribed error. If it is required that the misclassification proportion is no greater than ϵ, then the number of workers should satisfy $m \geq (1+o(1))\frac{1}{I(\pi)}\log\frac{1}{\epsilon}$. A special case is $\epsilon < n^{-1}$. Since $\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\{\hat{y}_j \neq y_j\}$ only takes value in $\{0, n^{-1}, 2n^{-1}, ..., 1\}$, an error rate smaller than n^{-1} implies that every item is correctly labeled. Therefore, as long as $m > (1+o(1))\frac{1}{I(\pi)}\log n$, the misclassification rate is 0 with high probability.

When $k = 2$, a special case of the general Dawid-Skene model takes the simple form

$$\begin{bmatrix}\pi_{11}^{(i)} & \pi_{12}^{(i)}\\ \pi_{21}^{(i)} & \pi_{22}^{(i)}\end{bmatrix} = \begin{bmatrix}p_i & 1-p_i\\ 1-p_i & p_i\end{bmatrix}. \tag{5}$$

This is referred to as the one-coin model, because the ability of each worker is parametrized by a biased coin with bias p_i. In this special case, $I(\pi)$ takes the following simple form

$$I(\pi) = I(p) = -\frac{1}{m}\sum_{i=1}^{m}\log\left(2\sqrt{p_i(1-p_i)}\right). \tag{6}$$

Note that $-2\log\left(2\sqrt{p_i(1-p_i)}\right)$ is the Rényi divergence of order $1/2$ between Bernoulli(p_i) and Bernoulli$(1-p_i)$. Let us summarize the optimal convergence rate for the one-coin model in the following corollary.

Corollary 3.1. *Assume* $\max_{1\leq i\leq m}(|\log(p_i)| \vee |\log(1-p_i)|) = o(mI(p))$*, Then, we have*

$$\begin{aligned}&\inf_{\hat{y}}\sup_{y\in\{1,2\}^n}\mathbb{E}_{p,y}L(\hat{y},y)\\ =\;&\exp\left(-(1+o(1))mI(p)\right),\end{aligned}$$

where $I(p)$ is defined by (6).

Corollary 3.1 has a weaker assumption than that of Theorem 3.1. When each p_i is assumed to be in the interval $[c, 1-c]$ with some constant $c \in (0, 1/2)$, the assumption of Corollary 3.1 reduces to $mI(p) \to \infty$, which is actually the necessary and sufficient condition for consistency. The result of Corollary 3.1 is very intuitive. Note that the Rényi divergence $-2\log\left(2\sqrt{p_i(1-p_i)}\right)$ is decreasing for $p_i \in [0, 1/2]$ and increasing for $p_i \in [1/2, 1]$. When most workers have p_i's that are close to $1/2$, then the rate of convergence will be slow. On the other hand, when p_i is either close to 0 or close to 1, that worker has a high ability, which will contribute to a smaller convergence rate. It is interesting to note that the result is symmetric around $p_i = 1/2$. This means for adversarial workers with $p_i < 1/2$, an optimal algorithm can invert their labels and still get useful information.

4. Adaptive Estimation

The optimal rate in Theorem 3.1 can be achieved by the following procedure:

$$\hat{y}_j = \arg\max_{g\in[k]}\prod_{i\in[m]}\prod_{h\in[k]}\left(\pi_{gh}^{(i)}\right)^{\mathbb{I}\{X_{ij}=h\}}. \tag{7}$$

This is the maximum likelihood estimator. When $k = 2$, it reduces to the likelihood ratio test by Neyman and Pearson (Neyman & Pearson, 1933). However, (7) is not practical because it requires the knowledge of the confusion matrix $\pi^{(i)}$ for each $i \in [m]$. A natural data-driven alternative is to first get an accurate estimator $\hat{\pi}$ of π in (7) and then consider the plug-in estimator,

$$\hat{y}_j = \arg\max_{g\in[k]}\prod_{i\in[m]}\prod_{h\in[k]}\left(\hat{\pi}_{gh}^{(i)}\right)^{\mathbb{I}\{X_{ij}=h\}}. \tag{8}$$

In the next theorem, we show that as long as $\hat{\pi}$ is sufficiently accurate, (8) will also achieve the optimal rate in Theorem 3.1.

Theorem 4.1. *Assume that, as* $m \to \infty$*,*

$$\mathbb{P}\left(\max_{g\in[k]}\sum_{i\in[m]}\max_{h\in[k]}\left|\log\hat{\pi}_{gh}^{(i)} - \log\pi_{gh}^{(i)}\right| > \delta\right) \to 0 \tag{9}$$

with δ such that $\delta + \log k = o(mI(\pi))$. Then, for any $y \in [k]^n$, we have

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\{\hat{y}_j \neq y_j\} \leq \exp\left(-(1+o(1))mI(\pi)\right),$$

with probability tending to 1, where $I(\pi)$ is defined by (4).

Theorem 4.1 guarantees that as long as the confusion matrices can be consistently estimated, the plugged-in MLE

(8) achieves the optimal error rate. In what follows, we apply this result to verify the optimality of some methods proposed in the literature.

4.1. Spectral Methods

Let us first look at the spectral method proposed in (Zhang et al., 2014). They compute the second and third order empirical moments and then estimate the confusion matrices by using tensor decomposition. In particular, they randomly partition the m workers into three different groups G_1, G_2 and G_3 to formulate the moments equations. For $(a, h) \in [3] \times [k]$, let

$$\pi^{\diamond}_{ah} = \frac{1}{|G_a|} \sum_{i \in G_a} \pi^{(i)}_{h*}, \quad \omega_h = \frac{|\{j : y_j = h\}|}{n}.$$

Note that $\pi^{\diamond}_{ah}$ is a k dimensional vector and we denote its lth component as $\pi^{\diamond}_{ahl}$. They use two steps to estimate the individual confusion matrices. They first estimate the aggregated confusion matrices $\pi^{\diamond}_{a*}$ by deriving equations between the moments of the labels $\{X_{ij}\}$ and the following moments of $\pi^{\diamond}_{ah}$,

$$M_2 = \sum_{h \in [k]} \omega_h \pi^{\diamond}_{ah} \otimes \pi^{\diamond}_{ah}, \quad M_3 = \sum_{h \in [k]} \omega_h \pi^{\diamond}_{ah} \otimes \pi^{\diamond}_{ah} \otimes \pi^{\diamond}_{ah}.$$

Empirical moments are used to approximate the population moments. Due to the symmetric structure of M_2 and M_3, a robust tensor power method (Anandkumar et al., 2014) is applied to approximately solve these equations. Then they use another moment equation to get an estimator $\hat{\pi}^{(i)}$ of the confusion matrices $\pi^{(i)}$ from the estimator of $\pi^{\diamond}_{ah}$.

Let $\omega_{min} = \min_{h \in [k]} \omega_h$, $\kappa = \min_{a \in [3], l \neq h \in [k]} \{\pi^{\diamond}_{ahh} - \pi^{\diamond}_{ahl}\}$ and σ_k be the minimum kth eigenvalue of the matrices $S_{ab} = \sum_{h \in [k]} \omega_h \pi^{\diamond}_{ah} \otimes \pi^{\diamond}_{bh}$ for $a, b \in [3]$. Applying Theorem 1 in (Zhang et al., 2014) to Theorem 4.1, we have the following result.

Theorem 4.2. *Assume* $\log k = o(mI(\pi))$ *and* $\rho_m I(\pi) \leq \min\{\frac{36k\kappa \log m}{\omega_{min}\sigma_L}, 2 \log m\}$. *Let* $\hat{y}$ *be the estimated labels from (8) using the estimated confusion matrices returned by Algorithm 1 in (Zhang et al., 2014). If the number of items* n *satisfies*

$$n = \Omega\left(\frac{k^5 \log^3 m \log k}{\rho_m^2 I^2(\pi) \omega_{min}^2 \sigma_k^{13}}\right),$$

then for any $y \in [k]^n$, *we have*

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\{\hat{y}_j \neq y_j\} \leq \exp\left(-(1 + o(1))mI(\pi)\right),$$

with probability tending to 1, *where* $I(\pi)$ *is defined by (4).*

Combined with Theorem 3.1, this result shows that an one-step update (8) of the spectral method proposed in (Zhang et al., 2014) can achieve the optimal error exponent.

4.2. One-coin Model

For the one-coin model, a simpler method of moments for estimating p_i is proposed in (Gao & Zhou, 2013). Let $n_1 = |\{j : y_j = 1\}|$, $n_2 = n - n_1$, and $\gamma = n_2/n$. They observe the equation $\frac{1}{n}\sum_{j=1}^{n} \mathbb{P}\{X_{ij} = 2\} = \gamma p_i + (1-\gamma)(1-p_i)$. This leads to a natural estimator

$$\hat{p}_i = \frac{\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\{X_{ij} = 2\} - (1 - \hat{\gamma})}{2\hat{\gamma} - 1}, \tag{10}$$

where $\hat{\gamma}$ is a consistent estimator of γ proposed in (Gao & Zhou, 2013). Combining the consistency result of $\hat{p}_i$ in (Gao & Zhou, 2013) and Theorem 4.1, we have the following result.

Theorem 4.3. *Assume* $|2\gamma - 1| \geq c$ *for some constant* $c > 0$, $\rho_m \leq p_i \leq 1 - \rho_m$ *for all* $i \in [m]$ *and* $\frac{1}{m}\sum_{i \in [m]}(2p_i - 1)^2 \leq 1 - \frac{4}{m}$. *Let* $\hat{y}$ *be the estimated labels from (8) using (10). If the number of items* n *satisfies*

$$n = \Omega\left(\frac{\log^2 m \log n}{\rho_m^2 I^2(p)}\right),$$

then for any $y \in [k]^n$, *we have*

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\{\hat{y}_j \neq y_j\} \leq \exp\left(-(1 + o(1))mI(p)\right),$$

with probability tending to 1, *where* $I(p)$ *is defined by (6).*

5. Discussion

In this section, we show the implications of our results on analyzing two popular crowdsourcing algorithms, EM algorithm and majority voting.

5.1. EM Algorithm

In the probabilistic model of crowdsourcing, the true labels can be regarded at latent variables. This naturally leads to apply the celebrated EM algorithm (Dempster et al., 1977) to obtain a local optimum of maximum marginal likelihood with the following iterations (Dawid & Skene, 1979):

- (M-step) update the estimate of workers' abilities

$$\pi^{(i)}_{(t+1),gh} \propto \sum_{j} \mathbb{P}_{(t)}\{y_j = g\} \mathbb{I}\{X_{ij} = h\} \tag{11}$$

- (E-step) update the estimate of true labels

$$\mathbb{P}_{(t+1)}\{y_j = g\} \propto \prod_{i,h} \left(\pi^{(i)}_{(t+1),gh}\right)^{\mathbb{I}\{X_{ij} = h\}} \tag{12}$$

The M-step (11) is essentially the maximum likelihood estimator. Bayesian versions of (11) are considered in (Raykar et al., 2010; Liu et al., 2012). Though the E-step (12) gives a probabilistic predication of the true label, a hard label can be obtained as $\hat{y}_j = \arg\max_{g\in[k]} \mathbb{P}_{(t+1)}\{y_j = g\}$. According to Theorem 4.1, as long as the M-step gives a consistent estimate of the workers' confusion matrices, the E-step will achieve the optimal error rate. This may explain why the EM algorithm for crowdsourcing works well in practice. In particular, as we have shown, when it is initialized by moment methods (Zhang et al., 2014; Gao & Zhou, 2013), the EM algorithm is provably optimal after only one step of iteration.

5.2. Majority Voting

Majority voting is perhaps the simplest method for aggregating crowdsourced labels. In what follows, we establish the exact error exponent of the majority voting estimator and show that it is inferior compared with the optimal error exponent. For simplicity, we only discuss the one-coin model. Then, the majority voting estimator is given by

$$\hat{y}_j = \arg\max_{g\in\{1,2\}} \sum_{i=1}^{m} \mathbb{I}\{X_{ij} = g\}.$$

Its error rate is characterized by the following theorem.

Theorem 5.1. *Assume $p_i \le 1-\rho_m$ for all $i \in [m]$, $\rho_m^2 \sum_{i\in[m]} p_i(1-p_i) \to \infty$ as $m \to \infty$ and $|\log \rho_m| = o(\sqrt{m}J(p))$. Then, we have*

$$\sup_{y\in\{1,2\}^n} \mathbb{E}_{p,y} L(\hat{y}, y) = exp\left(-(1+o(1))mJ(p)\right),$$

where

$$J(p) = -\min_{t\in(0,1]} \frac{1}{m}\sum_{i=1}^{m} \log\left[p_i t + (1-p_i)t^{-1}\right].$$

The theorem says that $-mJ(p)$ is the error exponent for the majority voting estimator. Given the simple relation

$$\begin{aligned} J(p) &= -\min_{t\in(0,1]} \frac{1}{m}\sum_{i=1}^{m} \log\left[p_i t + (1-p_i)t^{-1}\right] \\ &\le -\frac{1}{m}\sum_{i=1}^{m} \min_{t>0} \log\left[p_i t + (1-p_i)t^{-1}\right] \quad (13) \\ &= -\frac{1}{m}\sum_{i=1}^{m} \log\left(2\sqrt{p_i(1-p_i)}\right) \\ &= I(p), \end{aligned}$$

we can see that the majority voting estimator has an inferior error exponent $J(p)$ to that of the optimal rate $I(p)$ in Theorem 4.3. In fact, the inequality (13) holds if and only if p_i's are all equal, in which case, the majority voting is equivalent to the MLE (7). When p_i's are varied among workers, majority voting cannot take the varied workers' abilities into account, thus being sub-optimal.

6. Proofs

Proof of Theorem 3.1. The main proof idea is as follows. Consider the maximum likelihood estimator (7), we first derive the upper bound by union bound and Markov's inequality. The proof of lower bound is quite involved and it consists of three steps. Based on a standard lower bound technique, we first lower bound the misclassification rate by testing error. Then we calculate the testing error using the Neyman-Person Lemma. Finally, we give a lower bound for the tail probability of a sum of random variables, using the technique from the proof of the Cramer-Chernoff Theorem (Van der Vaart, 2000, Proposition 14.23).

Upper Bound. Let $\hat{y} = (\hat{y}_1, ..., \hat{y}_n)$ be defined as in (7). In the following, we give a bound for $\mathbb{P}(\hat{y}_j \ne y_j)$. Let us denote by $\mathbb{P}_l$ the joint probability distribution of $\{X_{ij}, i \in [m]\}$ given π and $y_j = l$. Without loss of generality, let $y_j = 1$. Using union bound, we have

$$\mathbb{P}_1(\hat{y}_j \ne 1) \le \sum_{g=2}^{k} \mathbb{P}_1(\hat{y}_j = g).$$

For each $g \ge 2$, we have

$$\begin{aligned} &\mathbb{P}_1(\hat{y}_j = g) \\ \le\ & \mathbb{P}_1\left(\prod_{i\in[m]}\prod_{h\in[k]}\left(\frac{\pi^{(i)}_{gh}}{\pi^{(i)}_{1h}}\right)^{\mathbb{I}\{X_{ij}=h\}} > 1\right) \\ \le\ & \min_{t\ge 0}\prod_{i\in[m]} \mathbb{E}_1 \prod_{h\in[k]}\left(\frac{\pi^{(i)}_{gh}}{\pi^{(i)}_{1h}}\right)^{t\mathbb{I}\{X_{ij}=h\}} \quad (14) \\ =\ & \min_{t\ge 0}\prod_{i\in[m]}\sum_{h\in[k]}\left(\pi^{(i)}_{1h}\right)^{1-t}\left(\pi^{(i)}_{gh}\right)^{t}, \end{aligned}$$

where (14) is due to Markov's inequality for each $t \ge 0$. Therefore, we have

$$\begin{aligned} \mathbb{P}_1(\hat{y}_j \ne 1) &\le \sum_{g=2}^{k} \exp\left(-mC\left(\pi_{1*}, \pi_{g*}\right)\right) \\ &\le (k-1)\exp\left(-m\min_{g\ne 1} C(\pi_{1*}, \pi_{g*})\right), \end{aligned}$$

which leads to

$$\begin{aligned} \frac{1}{n}\sum_{j\in[n]} \mathbb{P}_{y_j}(\hat{y}_j \ne y_j) &\le (k-1)\exp\left(-mI(\pi)\right) \\ &= \exp\left(-(1+o(1))mI(\pi)\right), \end{aligned}$$

when $\log k = o(mI(\pi))$.

Lower Bound. Now we establish a matching lower bound. We first introduce some notations. Define

$$B_t(\pi^{(i)}_{g*}, \pi^{(i)}_{h*}) = \sum_{l=1}^{k} \left(\pi^{(i)}_{gl}\right)^{1-t} \left(\pi^{(i)}_{hl}\right)^{t}.$$

Without loss of generality, we let

$$C(\pi_{1*}, \pi_{2*}) = \min_{g \neq h} C(\pi_{g*}, \pi_{h*}) = I(\pi).$$

Using the fact that the supremum over $[k]^n$ is bigger than the average over $[k]^n$, the minimax rate $\mathcal{M}$ can be lower bounded as

$$\begin{aligned}
&\sup_{y\in[k]^n} \mathbb{E}_{\pi,y} L(\hat{y}, y) \\
\geq\ & \frac{1}{k^n} \sum_{y\in[k]^n} \mathbb{E}_{\pi,y} L(\hat{y}, y) \\
=\ & \frac{1}{kn} \sum_{l=1}^{k} \sum_{j=1}^{n} \mathbb{P}_l \left\{\hat{y}_j \neq l\right\} \\
\geq\ & \frac{2}{kn} \sum_{j=1}^{n} \left[\frac{1}{2}\mathbb{P}_1\{\hat{y}_j \neq 1\} + \frac{1}{2}\mathbb{P}_2\{\hat{y}_j \neq 2\}\right].
\end{aligned}$$

Taking an infimum of $\hat{y}$ on both sides leads to

$$\begin{aligned}
&\inf_{\hat{y}} \sup_{y\in[k]^n} \mathbb{E}_{\pi,y} L(\hat{y}, y) \\
\geq\ & \inf_{\hat{y}} \frac{2}{kn} \sum_{j=1}^{n} \inf_{\hat{y}_j} \left[\frac{1}{2}\mathbb{P}_1\{\hat{y}_j = 2\} + \frac{1}{2}\mathbb{P}_2\{\hat{y}_j = 1\}\right] \\
=\ & \frac{2}{kn} \sum_{j=1}^{n} \inf_{\hat{y}_j} \left[\frac{1}{2}\mathbb{P}_1\{\hat{y}_j = 2\} + \frac{1}{2}\mathbb{P}_2\{\hat{y}_j = 1\}\right].
\end{aligned}$$

By the Neyman-Pearson Lemma (Neyman & Pearson, 1933), for any fixed $j \in [n]$, the Bayes testing error

$$\frac{1}{2}\mathbb{P}_1\{\hat{y}_j = 2\} + \frac{1}{2}\mathbb{P}_2\{\hat{y}_j = 1\}$$

is minimized by the likelihood ratio test

$$\hat{y}_j = \arg\max_{g\in\{1,2\}} \prod_{i\in[m]} \prod_{h\in[k]} \left(\pi^{(i)}_{gh}\right)^{\mathbb{I}\{X_{ij}=h\}}.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}_1(\hat{y}_j = 2) \\
=\ & \mathbb{P}_1\left(\prod_{i\in[m]} \prod_{h\in[k]} \left(\frac{\pi^{(i)}_{2h}}{\pi^{(i)}_{1h}}\right)^{t\mathbb{I}\{X_{ij}=h\}} > 1\right) \\
=\ & \mathbb{P}(S_m > 0).
\end{aligned}$$

Here t is a positive constant that we will specify later. And $S_m = \sum_{i\in[m]} W_i$, with the random variable W_i defined as

$$\mathbb{P}\left(W_i = t \log\left(\frac{\pi^{(i)}_{2h}}{\pi^{(i)}_{1h}}\right)\right) = \pi^{(i)}_{1h}. \tag{15}$$

We lower bound $\mathbb{P}(S_m > 0)$ by

$$\begin{aligned}
&\sum_{0<S_m} \prod_{i\in[m]} \mathbb{P}(W_i) \\
\geq\ & \sum_{0<S_m<L} \prod_{i\in[m]} \mathbb{P}(W_i) \\
=\ & \sum_{0<S_m<L} \prod_{i\in[m]} \frac{\mathbb{P}(W_i)e^{W_i}}{B_t(\pi^{(i)}_{1*}, \pi^{(i)}_{2*})} \prod_{i\in[m]} \frac{B_t(\pi^{(i)}_{1*}, \pi^{(i)}_{2*})}{e^{W_i}} \\
\geq\ & \prod_{i\in[m]} B_t(\pi^{(i)}_{1*}, \pi^{(i)}_{2*}) e^{-L} \sum_{0<S_m<L} \mathbb{Q}_i(W_i) \\
\geq\ & \prod_{i\in[m]} B_t(\pi^{(i)}_{1*}, \pi^{(i)}_{2*}) e^{-L} \mathbb{Q}(0 < S_m < L),
\end{aligned}$$

where the distribution $\mathbb{Q}_i$ is defined as

$$\mathbb{Q}_i\left(W_i = t \log\left(\frac{\pi^{(i)}_{2h}}{\pi^{(i)}_{1h}}\right)\right) = \frac{\left(\pi^{(i)}_{1h}\right)^{1-t}\left(\pi^{(i)}_{2h}\right)^{t}}{B_t(\pi^{(i)}_{1*}, \pi^{(i)}_{2*})}, \tag{16}$$

and $\mathbb{Q}$ is defined as the joint distribution of $\mathbb{Q}_1, \cdots, \mathbb{Q}_m$.

To precede, we will need the following two lemmas.

Lemma 6.1. *If $\mathcal{A}_\alpha$ is not empty, there is an unique t_0 such that*

$$t_0 = \operatorname*{argmin}_{t\in[0,1]} \prod_{i\in[m]} B_t(\pi^{(i)}_{1*}, \pi^{(i)}_{2*}). \tag{17}$$

Moreover, we have $0 < t_0 < 1$.

Lemma 6.2. *Let $t = t_0$ defined in (17). Then under the assumption of Theorem 3.1, S_m is a zero mean random variable satisfying the central limit theorem, i.e. for any x,*

$$\mathbb{Q}\left(\frac{S_m}{\sqrt{Var(S_m)}} \leq x\right) \to \Phi(x), \text{ as } m \to \infty,$$

where Φ is the cumulative distribution function of a $N(0,1)$ random variable.

The proof of Lemma 6.1 and Lemma 6.2 are given in the supplementary material. Let $t = t_0$ and $L = 2\sqrt{\mathrm{Var}_Q(S_m)}$. Using Lemma 6.2 and Chebyshev's inequality, we have

$$\begin{aligned}
\mathbb{Q}(0 < S_m < L) &\geq 1 - \mathbb{Q}(S_m \leq 0) - \mathbb{Q}(S_m \geq L) \\
&\geq 1 - 5/8 - 1/4 = 1/8
\end{aligned}$$

for sufficiently large m. Note that

$$\begin{aligned}
&\mathbb{E}_Q W_i^2 \\
&= \sum_{h\in[k]} \left(t\log\left(\frac{\pi_{2h}^{(i)}}{\pi_{1h}^{(i)}}\right)\right)^2 \mathbb{Q}_i\left(W_i = t\log\left(\frac{\pi_{2h}^{(i)}}{\pi_{1h}^{(i)}}\right)\right) \\
&\le \max_{i,h}\left(t\log\left(\frac{\pi_{2h}^{(i)}}{\pi_{1h}^{(i)}}\right)\right)^2 \\
&\le \log^2\rho_m.
\end{aligned}$$

Consequently,

$$\mathrm{Var}_{\mathbb{Q}}(S_m) = \sum_{i\in[m]} \mathrm{Var}_{\mathbb{Q}}(W_i) \le \sum_{i\in[m]} \mathbb{E}_Q W_i^2 \le m\log^2\rho_m.$$

Under the assumption that $\log^2\rho_m = o(mI^2(\pi))$, we have $e^{-L} \ge e^{-\sqrt{m\log^2\rho_m}} \ge e^{-o(mI(\pi))}$. This leads to the lower bound

$$\begin{aligned}
\mathbb{P}_1(\hat{y}_j = 2) &\ge \prod_{i\in[m]} B_t(\pi_{1*}^{(i)}, \pi_{2*}^{(i)}) e^{-o(mI(\pi))} \\
&= \exp\left(-(1+o(1))mI(\pi)\right).
\end{aligned}$$

Note that the same bound holds for $\mathbb{P}_2(\hat{y}_j = 1)$. Hence,

$$\begin{aligned}
\inf_{\hat{y}} \sup_{y\in[k]^n} \mathbb{E}L(\hat{y}, y) &\ge \frac{2}{k}\exp\left(-(1+o(1))mI(\pi)\right) \\
&= \exp\left(-(1+o(1))mI(\pi)\right),
\end{aligned}$$

under the assumption that $\log k = o\left(mI(\pi)\right)$. This completes the proof. □

Proof of Theorem 4.1. Define

$$E = \left\{\max_{g\in[k]} \sum_{i\in[m]} \max_{h\in[k]} \left|\log\hat{\pi}_{gh}^{(i)} - \log\pi_{gh}^{(i)}\right| \le \delta\right\}.$$

Then, we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_j \mathbb{I}\{\hat{y}_j \ne y_j\} > \epsilon\right) \\
&\le \mathbb{P}\left(\frac{1}{n}\sum_j \mathbb{I}\{\hat{y}_j \ne y_j\} > \epsilon, E\right) + \mathbb{P}(E^c) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_j \mathbb{I}\{\hat{y}_j \ne y_j\} > \epsilon \middle| E\right)\mathbb{P}(E) + \mathbb{P}(E^c) \\
&\le \frac{1}{n}\sum_j \mathbb{P}\left(\hat{y}_j \ne y_j | E\right)\mathbb{P}(E)/\epsilon + \mathbb{P}(E^c) \\
&= \frac{1}{n}\sum_j \mathbb{P}\left(\hat{y}_j \ne y_j, E\right)/\epsilon + \mathbb{P}(E^c).
\end{aligned}$$

Let us give a bound for $\mathbb{P}\left(\hat{y}_j \ne y_j, E\right)$. Without loss of generality, let $y_j = 1$. Then,

$$\begin{aligned}
&\mathbb{P}\left(\hat{y}_j \ne y_j, E\right) \\
&\le \sum_{g=2}^k \mathbb{P}\left(\hat{y}_j = g, E\right) \\
&\le \sum_{g=2}^k \mathbb{P}\left(\prod_{i\in[m]}\prod_{h\in[k]}\left(\frac{\hat{\pi}_{gh}^{(i)}}{\hat{\pi}_{1h}^{(i)}}\right)^{\mathbb{I}\{X_{ij}=h\}} > 1, E\right) \\
&= \sum_{g=2}^k \mathbb{P}\left(\prod_{i\in[m]}\prod_{h\in[k]}\left(\frac{\pi_{gh}^{(i)}}{\pi_{1h}^{(i)}}\right)^{\mathbb{I}\{X_{ij}=h\}}\right. \\
&\qquad \left.\prod_{i\in[m]}\prod_{h\in[k]}\left(\frac{\hat{\pi}_{gh}^{(i)}\pi_{1h}^{(i)}}{\pi_{gh}^{(i)}\hat{\pi}_{1h}^{(i)}}\right)^{\mathbb{I}\{X_{ij}=h\}} > 1, E\right)
\end{aligned}$$

On the event E,

$$\begin{aligned}
&\log\left(\prod_{i\in[m]}\prod_{h\in[k]}\left(\frac{\hat{\pi}_{gh}^{(i)}\pi_{1h}^{(i)}}{\pi_{gh}^{(i)}\hat{\pi}_{1h}^{(i)}}\right)^{\mathbb{I}\{X_{ij}=h\}}\right) \\
&\le \sum_{i\in[m]}\sum_{h\in[k]}\left(\log\frac{\hat{\pi}_{gh}^{(i)}}{\pi_{gh}^{(i)}} - \log\frac{\hat{\pi}_{1h}^{(i)}}{\pi_{1h}^{(i)}}\right)\mathbb{I}\{X_{ij}=h\} \le 2\delta.
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbb{P}\left(\hat{y}_j \ne y_j, E\right) \\
&\le \sum_{g=2}^k \mathbb{P}\left(e^{2\delta}\prod_{i\in[m]}\prod_{h\in[k]}\left(\frac{\pi_{gh}^{(i)}}{\pi_{1h}^{(i)}}\right)^{\mathbb{I}\{X_{ij}=h\}} > 1\right) \\
&\le \sum_{g=2}^k e^{2\delta}\min_{0\le t\le 1}\prod_{i\in[m]}\sum_{h\in[k]}\left(\pi_{1h}^{(i)}\right)^{1-t}\left(\pi_{gh}^{(i)}\right)^t \\
&\le (k-1)\exp\left(-m\min_{g\ne 1} C(\pi_{1*}, \pi_{g*}) + 2\delta\right).
\end{aligned}$$

Thus,

$$\frac{1}{n}\sum_{j\in[n]}\mathbb{P}\left(\hat{y}_j \ne y_j, E\right) \le (k-1)\exp\left(-mI(\pi) + 2\delta\right).$$

Letting $\epsilon = (k-1)\exp\left(-(1-\eta)mI(\pi) + 2\delta\right)$ with $\eta = 1/\sqrt{mI(\pi)}$, we have

$$\frac{1}{n}\sum_j \mathbb{P}\left(\hat{y}_j \ne y_j, E\right)/\epsilon \le \exp\left(-\sqrt{mI(\pi)}\right).$$

Thus, the proof is complete under the assumption that $\log k + \delta = o(mI(\pi))$ and $\mathbb{P}(E^c) = o(1)$. □

Proof of Theorem 5.1. The risk is $\frac{1}{n}\sum_{j=1}^n \mathbb{P}\{\hat{y}_j \ne y_j\}$. Consider the random variable $\mathbb{I}\{\hat{y}_j \ne y_j\}$. It has the

same distribution as $\mathbb{I}\{\sum_{i=1}^m (T_i - 1/2) > 0\}$, where $T_i \sim \text{Bernoulli}(1-p_i)$. Therefore,

$$\frac{1}{n}\sum_{j=1}^n \mathbb{P}\{\hat{y}_j \neq y_j\} = \mathbb{P}\left\{\sum_{i=1}^m (T_i - 1/2) > 0\right\}.$$

We first derive the upper bound. Using Chernoff's method, we have

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^m (T_i - 1/2) > 0\right\} &\leq \prod_{i=1}^m \mathbb{E}e^{\lambda(T_i - 1/2)} \\ &= \exp\left(\sum_{i=1}^m \log\left[(1-p_i)e^{\lambda/2} + p_i e^{-\lambda/2}\right]\right).\end{aligned}$$

The desired upper bound follows by letting $t = e^{-\lambda/2}$ and optimizing over $t \in (0, 1]$.

Now we show the lower bound using the similar arguments as in the proof of Theorem 3.1. Define $W_i = \lambda(T_i - 1/2)$ and $S_m = \sum_{i=1}^m W_i$. Then, we have

$$\begin{aligned}&\mathbb{P}\left\{\sum_{i=1}^m (T_i - 1/2) > 0\right\} \\ =\ &\mathbb{P}\{S_m > 0\} \\ \geq\ &\sum_{0<S_m<L}\prod_{i=1}^m \mathbb{P}(W_i) \\ =\ &\sum_{0<S_m<L}\left(\prod_{i=1}^m \frac{\mathbb{P}(W_i)e^{W_i}}{(1-p_i)e^{\lambda/2} + p_i e^{-\lambda/2}}\right. \\ &\left.\prod_{i=1}^m \frac{(1-p_i)e^{\lambda/2} + p_i e^{-\lambda/2}}{e^{W_i}}\right) \\ \geq\ &\prod_{i=1}^m \left((1-p_i)e^{\lambda/2} + p_i e^{-\lambda/2}\right) e^{-L}\mathbb{Q}\{0 < S_m < L\}.\end{aligned}$$

Note that under $\mathbb{Q}$, W_i has distribution

$$\begin{aligned}Q_i(W_i = \lambda/2) &= \frac{(1-p_i)e^{\lambda/2}}{(1-p_i)e^{\lambda/2} + p_i e^{-\lambda/2}}, \\ Q_i(W_i = -\lambda/2) &= \frac{p_i e^{-\lambda/2}}{(1-p_i)e^{\lambda/2} + p_i e^{-\lambda/2}}.\end{aligned}$$

We choose $\lambda_0 \in [0, \infty)$ to minimize $f(\lambda) = \prod_{i=1}^m \left((1-p_i)e^{\lambda/2} + p_i e^{-\lambda/2}\right)$. This leads to the equation $\mathbb{E}_Q S_m = 0$. It is sufficient to lower bound $e^{-L}\mathbb{Q}\{0 < S_m < L\}$ to finish the proof. To do this, we need the following result.

Lemma 6.3. *Suppose $p_i \leq 1 - \rho_m$ for all $i \in [m]$ and $\rho_m^2 \sum_{i\in[m]} p_i(1-p_i) \to \infty$ as $m \to \infty$. Then we have*

i) $\lambda_0 \leq -2\log\rho_m$.

ii) $\frac{S_m}{\sqrt{Var_Q(S_m)}} \rightsquigarrow N(0,1)$ *under the distribution* $\mathbb{Q}$.

The proof of Lemma 6.3 is given in the supplementary file. Let $L = 2\sqrt{\text{Var}_Q(S_m)}$, and we have

$$e^{-L}\mathbb{Q}(0 < S_m < L) \geq 0.25e^{-2\sqrt{\text{Var}_Q(S_m)}}.$$

Finally, we need to show $\sqrt{\text{Var}_Q(S_m)} = o(mJ(p))$. This is because

$$\begin{aligned}\text{Var}(S_m) &\leq \sum_{i=1}^m \mathbb{E}_Q W_i^2 \leq m\lambda_0^2/4 \\ &\leq m\log^2\rho_m = o(m^2 J(p)^2),\end{aligned}$$

where the last equality is implied by the assumption $|\log\rho_m| = o(\sqrt{m}J(p))$. The proof is complete. □

References

Amazon Mechanical Turk. https://www.mturk.com/mturk.

Anandkumar, Animashree, Ge, Rong, Hsu, Daniel, Kakade, Sham M, and Telgarsky, Matus. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Bachrach, Yoram, Graepel, Thore, Minka, Tom, and Guiver, John. How to grade a test without knowing the answers — a Bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 1183–1190, 2012.

Berend, Daniel and Kontorovich, Aryeh. A finite sample analysis of the naive bayes classifier. *Journal of Machine Learning Research*, 16:1519–1545, 2015.

Cover, Thomas M and Thomas, Joy A. *Elements of information theory*. John Wiley & Sons, 2006.

Dalvi, N., Dasgupta, A., Kumar, R., and Rastogi, V. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1220–1229, 2013.

Dawid, A. P. and Skene, A. M. Maximum likeihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society*, 28(1):20–28, 1979.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1): 1–38, 1977.

Gao, Chao and Zhou, Dengyong. Minimax optimal convergence rates for estimating ground truth from crowdsourced labels. *arXiv preprint arXiv:1310.5764*, 2013.

Ghosh, Arpita, Kale, Satyen, and McAfee, Preston. Who moderates the moderators? Crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 167–176, 2011.

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

Liu, Q., Peng, J., and Ihler, A. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems 25*, pp. 701–709, 2012.

Neyman, J and Pearson, ES. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 231(694-706):289–337, 1933.

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

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

Tian, Tian and Zhu, Jun. Max-margin majority voting for learning from crowds. In *Advances in Neural Information Processing Systems 28*, pp. 1612–1620, 2015.

Van der Vaart, Aad W. *Asymptotic statistics*. Cambridge university press, 2000.

Venanzi, Matteo, Guiver, John, Kazai, Gabriella, Kohli, Pushmeet, and Shokouhi, Milad. Community-based Bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, pp. 155–164, 2014.

Welinder, Peter, Branson, Steve, Perona, Pietro, and Belongie, Serge J. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23*, pp. 2424–2432, 2010.

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J. Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pp. 2035–2043, 2009.

Zhang, Yuchen, Chen, Xi, Zhou, Dengyong, and Jordan, Michael I. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. In *Advances in neural information processing systems*, pp. 1260–1268, 2014.

Zhou, D., Platt, J. C., Basu, S., and Mao, Y. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems 25*, pp. 2204–2212, 2012.

Zhou, Dengyong, Liu, Qiang, Platt, John, and Meek, Christopher. Aggregating ordinal labels from crowds by minimax conditional entropy. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 262–270, 2014.