Permuation Models meet Dawid-Skene: A Generalised Model for Crowdsourcing

Ankur Mallick
Electrical and Computer Engineering
Carnegie Mellon University
amallic1@andrew.cmu.edu

Abstract

The advent of machine learning and deep learning applications has increased the need for obtaining large amounts of labelled data required to train machine learning algorithms. Due to the high cost associated with labelling of data by experts, crowdsourcing has emerged as a popular option for labelling data. In this paper we propose a permutation-based model for estimating labels from crowdsourced data that is a significant generalization of previous works. While most existing literature on this topic is based on the classical Dawid-Skene model we argue that the assumptions made by this model are highly restrictive. We propose a new error metric to compare different estimators and show that despite greatly relaxing the assumtions of the Dawid-Skene model, we only incur a minor statistical penalty in estimation. While previous work on the permutation-based model focused on binary classification, our work extends this model to the M-ary case. We work in a high-dimensional, non-asymptotic setting and allow the number of workers, tasks, and classes to scale. In this setting we derive a sharp lower bound for all estimators under our proposed model. We also propose a computationally efficient estimator for a subclass of our model and derive a sharp upper bound on the rate of convergence of this estimator.

1 Introduction

Crowdsourcing is defined as the practice of obtaining information (such as labels for items in a large dataset) from a large number of people (the crowd). The motivation for crowdsourcing in data-labelling applications stems from the fact that most supervised learning algorithms need to be trained on huge amounts of labelled data to produce an acceptable level of performance. The lack of availability of such large labelled datasets and the infeasibility of getting a small group of experts to label all the data has made researchers turn to the crowd.

In a typical crowdsourcing setting, each item in the dataset is assumed to belong to one out of M classes. Each crowd worker is asked to label (classify) a subset of the data and each data item is labelled by more than one person. Thus each item has multiple labels assigned to it by the crowd. Since the individuals in the crowd are typically non-experts it is expected that there will be a large amount of error in the labels collected. The goal of any effective crowdsourcing framework is to design a suitable model for this uncertainty and to develop inference algorithms that accept the multiple noisy labels for the items in the data set and return a single inferred label for each item with minimum error. The varying skill levels of the crowd workers and the varying difficulty levels of labelling each item are the major reasons for heterogenity in the data collected and this heterogenity is the major challenge in designing models and inference algorithms for crowdsourcing.

Final project report for course CMU 10709 "Fundamentals of Learning from the Crowd", Fall 2017. Instructor: Nihar B. Shah, TA: Ritesh Noothigattu. The formatting style file is borrowed from the NIPS conference.

Existing works like [1],[2] rely on the classical Dawid and Skene model [3] for estimating the true labels from crowdsourced data. The Dawid and Skene model assumes that for all items belonging to a particular class, the probability distribution of a worker's responses, to the task of classifying these items, is the same. This assumption of homogenity across different items belonging to the same class is often not satisfied in practical settings where one items might be harder to classify than another even if both belong to the same class. Recent work [4] suggeests that the assumptions of the Dawid and Skene model can be relaxed to allow different items from the same class to have a different distribution of worker responses while still yielding computationally efficient estimators. However the authors only propose their permutation-based model for the case of binary classificaton. In this work we generalize the permutation-based approach to the case where items belong to one of out M classes. We also generalize the difficulty-weighted error metric proposed in [4] that accounts for the differences in difficulty of classifying items, as opoosed to the classical Hamming error metric used by Dawid-Skene models, thus ensuring that estimators are not penalised too heavily for misclassifying items which are difficult to classify for most crowdowrkers. We derive sharp statistical lower bounds on the estimation error for our model. We show that the Dawid-Skene model is a subclass of our model and thus the aforementioned lower bound applies to all previous estimators derived for this model. We propose a computationally efficient estimator for our model when the approximate ordering of worker abilities is known and derive a statistical upper bound on this estimator that is identical to our lower bound upto a logarithmic factor. The rest of the paper is organized as follows. In Section 2 we discuss our observation model and survey some existing literature in this area. Section 3 contains our main results viz. the lower bound on our model, the estimation algorithm, and the upper bound on the proposed estimator. We present proofs of our results in Section 4 . We conclude the paper with a summary of our work and a discussion of future directions in Section 6.

2 Observation Model and Prior Work

In this section we dicuss our observation model for crowd-labelled data and the loss function we use for evaluating estimators. We conclude the section with a survey of prior work and comparison of earlier models with our model.

2.1 Observation Model

We consider a crowdsourcing system that consists of n workers and d questions or items to be classified. Each item belongs to 1 out M classes (an item is not allowed to belong to more than one class under our model). Thus the quantity representing the true class of an item can take 1 out of M possible values. Integers in the set $\mathbb{F}_M = \{1, \ldots, M\}$ are used to represent class labels and the true classs labels of the d items are stored in a vector $x \in \mathbb{F}_M^d$. A tensor $Q^* \in [0,1]^{n \times d \times M}$ is used to model the probability of selecting different classes for an item. $Q(k,l,m)$ is the probability that worker k selects class m for item l.

We assume that a worker is asked to label an item with probability p_0 and the worker responses are stored in a $n \times d$ matrix Y. Specifically, Y_{kl} is the label assigned by worker k to item l if the worker is asked to label the item and is 0 otherwise. Assuming that worker responses are independent given x^*, the vector of true class labels, and Q^*, the tensor of probabilities, we have the following distribution on the entries of Y:

$$Y_{kl} = \begin{cases} m, & \text{with probability } p_0 Q^*(k,l,m) \\ 0, & \text{with probability } 1 - p_0 \end{cases}$$

Given this random matrix Y our goal is to estimate the vector $x^* \in \mathbb{F}_M^d$ of true class labels.

To ensure identifiability of the proposed model and obtain non-trivial guarantess, we need to impose some structure on the probability tensor Q^*. These assumptions are in the spirit of the structural assumptions imposed by [4] on the permutation-based model for the binary classification case and are considerably weaker than those introduced by the Dawid-Skene model and works [1],[2],[3], based on this model. Before we introduce our structural assumptions, we define a matrix $R^* \in [0,1]^{n \times d}$ whose $(k,l)^{\text{th}}$ entry is given by:

$$R^*_{kl} = Q^*(k,l,m^*_l) - \underset{m \neq x^*_l}{\operatorname{argmax}} Q(k,l,m)$$

i where m^*_l is the true class label of item l.

We impose the following structural assumptions on the tensor Q^*:

1. Law of total probabilities: As each worker must pick one of the M classes when asked to classify an item, $\sum_{m=1}^{m=M} Q^*(k,l,m) = 1, \forall k,l$
2. Non-adverserial workers: For each item, the probabiity of choosing an incorrect class is smaller than the probability of choosing the correct class for all workers i.e. $Q^*(k,l,m^*_l) \geq Q^*(k,l,m) \forall m \neq m^*_l$
3. Worker monotonicity: There exists a permutation $\pi^* : [n] \to [n]$ such that for every pair of workers k, k' for which $\pi^*(k) < \pi^*(k')$ and every item l, we have $R^*_{kl} \geq R^*_{k'l}$
4. Item monotonicity: There exists a permutation $\sigma^* : [d] \to [d]$ such that for every pair of items l, l' for which $\sigma^*(l) < \sigma^*(l')$ and every worker k, we have $R^*_{kl} \geq R^*_{kl'}$

The first assumption is a simple consequence of the law of total probabilities for our model. The second assures that workers are not adversarial in nature and classify items uniformly at random in the worst case since it guarantees that $Q^*(k,l,m^*_l) \geq \frac{1}{M}$. It also assures that the entries of the matrix R^* are always non-negative. This is a generalization of a similar assumption in [3] that the probability of a worker picking the correct option for an item is always greater than 0.5 for the binary classification case. To understand the third and fourth assumptions we need to take a closer look at the matrix R^* which is a measure of the "certainty" of a worker about an item. It can be seen that the maximum value $R_{kl} = 1$ occurs when worker $Q^*(k,l,m^*_l) = 1$ and $Q^*(k,l,m) = 0$ for all other m i.e. worker k is absolutely certain about the true class of item l, while the minimum value $R_{kl} = 0$ occurs when $\exists m \neq m^*_l$ for which $Q^*(k,l,m) = Q^*(k,l,m^*_l)$ i.e the worker k is uncertain about item l and can pick an incorrect class with as much probability as the correct class. In between these two extremes, a larger value of R_{kl} implies a higher difference between the probability of choosing the correct class and that of choosing an incorrect class i.e. higher certainty about the true class of the item. With this in mind, we can see that our third and fourth assumptions regarding worker and item monotonicity imply the existence of an ordering of workers based on their abilities and on questions based on their difficulties such that a worker with a higher ability is more certain about any item than a worker with lower ability, while an item with higher difficulty causes greater uncertainty among all workers than an item with lower difficulty.

In light of the above description of the structural assumptions of our model, it is easy to see that the popular Dawid-Skene model satisfies all of the above assumptions and thus corresponds to a particular type of probability tensors Q^* which, in addition to satisfying the aforementioned assumptions also includes the highly restrictive constraint that for any pair of items l, l', if $m^*_l = m^*_{l'}$ then $Q^*(k,l,m) = Q^*(k,l',m) \forall k,m$ i.e. the distribution of worker responses is identical for items belonging to the same class.

In summary, let $\mathbb{C}_{\text{perm}}(x^*)$ denote the class of probability tensors Q^* that satisfy our model assumptions, and let $\mathbb{C}_{DS}$ denote the class of probability tensors satisfying the Dawid-Skene assumptions. Thus:

$$\mathbb{C}_{\text{perm}}(x^*) := \{Q \in [0,1]^{n \times d \times M} \mid Q \text{ satisfies (i) - (iv)}\}$$
$$\mathbb{C}_{\text{DS}}(x^*) := \{Q \in \mathbb{C}_{\text{perm}} \mid m^*_l = m^*_{l'} \text{ implies } Q^*(k,l,m) = Q^*(k,l',m) \forall k,m\}$$

2.2 Loss Function for Estimators

i We define an estimator as a function that maps the matrix of worker responses Y to a vector $\hat{x} \in \mathbb{F}_{\mathbb{M}}{}^{d}$. A popular loss function for evaluating such estimators is the Hamming error which can be generalised to our case as:

$$d_H(\hat{x}, x^*) = \frac{1}{d}\sum_{l=1}^{d} \mathbb{1}\{\hat{x}_l \neq x_l^*\} \tag{1}$$

Since we assume non-uniform difficulty of questions, we would like to weight them differently while evaluating estimators in order to factor in the difference in difficulty levels. Consider the following, weighted version of the Hamming error:

$$L_{Q^*}(\hat{x}, x^*) = \frac{1}{nd}\sum_{l=1}^{d}\sum_{k=1}^{n}(Q^*(k,l,m_l^*) - Q^*(k,l,\hat{m}_l))^2 \tag{2}$$

Here $\hat{m}_l$ is the estimated class label corresponding to item l for the estimator $\hat{x}$. Thus using the above loss function (henceforth referred to as the Q^*-Error) the error is 0 if $\hat{x}_l = x_l$, and otherwise it is $\frac{1}{n}\sum_{k=1}^{n}(Q^*(k,l,m_l^*) - Q^*(k,l,\hat{m}_l))^2$ which is the mean squared difference between the probability of picking the estimated and true classes for item l by each worker. In general if the probability of picking the estimated class is much smaller than the probability of picking the true class, then such an estimator is penalised more heavily under this criterionas opposed to one where the two have similar probabilities of being picked since we expect that in such a scenario it would be harder to distinguish between the two classes. For the case of binary classification this error metric reduces to the Q^*-Loss defined in [4].

2.3 Prior Work

The prior work in this area can be broadly divided into two classes. The first consists of a vast body of work of which [1],[2] are some representative samples. These works are based on the classical Dawid-Skene model [3] where the underlying assumption is that the questions (or items to be labelled) are of the same difficulty. Thus these models assume that each worker is associated with a confusion matrix π which defines the probability of the worker making an error in labelling an item such that π_{kl} is the probability of the worker labelling an item as class k when it belongs to class l. The confusion matrix is distinct for each worker but is the same for all questions for a particular worker. Various algorithms have been designed and guarantees have been proved for estimating the true labels of the data items in this setting. However the assumption of homogenity in the difficulty levels of each question is often not realised in practice.

The second class consists of a relatively new work [4] where a new "permutation-based" model has been proposed to account for the difficulty level of each question. The authors of [4] consider only binary choice questions i.e. each item to be labelled lies in one of two classes, and define a matrix Q that models the probability of correctness of labelling each item by each person such that Q_{ij} is the probability of person i labelling item j correctly. In this setting they establish bounds on the minimax error and provide computationally efficient estimators that achieve the minimax limits when certain additional constraints are satisfied. However the model has so far only been explored for the binary choice case.

3 Main Results

In this section we present our main results which include a sharp lower bound on the minimax risk for our permutation based model, an algorithm for estimating the true class labels when the ordering of the worker abilities is known exactly or approximately, and an upper bound on the estimation error under the proposed algorithm. In our analysis we focus on a non-asymptotic regime and evaluate the error in terms of n, d, M and p_0. We assume that $d \geq n$. We also assume that $p_0 \geq \frac{1}{n}$ which ensures that on average, at least one worker answers any question. We use symbols c, c', c_E, c'_E to represent universal constants.

3.1 Lower Bound on Minimax Risk

The minimax risk of an estimator is defined as:

$$F(\hat{x}) = \sup_{x* \in \mathbb{F}_M^d} \sup_{Q^* \in \mathbb{C}_{\text{perm}}(x^*)} \mathbb{E}[L_{Q^*}(\hat{x}, x^*)]$$

Thus, it represents the worst case expected error over the set of true class labels and probability tensors (the expectation is taken over the random matrix Y of worker responses).

The following theorem gives a lower bound on the minimax risk:

Theorem 1 *For a universal constant c, any estimator $\hat{x}$ has minimax risk at least:*

$$F(\hat{x}) \geq \frac{c}{np_0} \tag{3}$$

An important point to note is that the above lower bound holds even when the set of allowed probability tensors Q^* is restricted to the Dawid-Skene model i.e. $\mathbb{C}_{DS} \subset \mathbb{C}_{\text{perm}}$. Indeed it even holds for the case when Q^* is known at the estimator. In what follows, we propose an estimation algorithm with an upper bound that matches the above lower bound upto a logarithmic factor for a sub-class of our observation model. This shows that at least for this subclass we incur atmost a logarithmic penalty in estimation as compared to the Dawid-Skene model which is far more restrictive than our model.

3.2 Estimation Algorithm

This estimation algorithms is applicable to the case where workers are calibrated, in the sense that the ordering of worker abilities is known. This is indeed the case in several real-life crowdsourcing scenarios where workers are employed only after being thoroughly tested and calibrated. In this setting we expect to know the permuation π^* of worker abilities, on the matrix R^* defined earlier, which represents the certainty of workers regarding questions. However we do not know the permuation on the question difficulties or the true values of either R^* or Q.

For a given permuation of workers π, we first define the notion of bias in responses of the top r workers for an item l as:

$$b_l^r = \min_{m \neq m_l^r} \sum_{k \in [r]} (\mathbb{1}\{(Y_{\pi^{-1}(k)l} = m_l^r\} - \mathbb{1}\{Y_{\pi^{-1}(k)l} = m\}) \tag{4}$$

where

$$m_l^r = \max_m \sum_{k \in [r]} \mathbb{1}\{Y_{\pi^{-1}(k)l} = m\} \tag{5}$$

Essentially our definition of bias corresponds to the difference between the number of votes received by the most popular response among the top r workers and the number of votes received by the second most popular response among the top r workers.

We now use the following 2-step procedure is used to estimate the true class labels:

1. Identifying smart workers: Compute the integer r_0

$$r_0 \in \operatorname*{argmax}_{r \in p_0^{-1} \log^{1.5}(dn), \ldots, n} \sum_{l \in [d]} \mathbb{1}\{b_l^r \geq \frac{1}{2}\sqrt{kp_0 \log^{1.5}(dn)}\} \tag{6}$$

2. Majority voting by smart workers: Set $\hat{m}_l^\pi, \hat{x}(\pi)$ as a majority vote of the best r_0 workers

$$\hat{m}_l^\pi = \operatorname*{argmax}_m \sum_{k \in [r_0]} \mathbb{1}\{Y_{\pi^{-1}(k)l} = m\} \tag{7}$$

Thus our algorithm first finds a set of "good" workers such that their responses for any item are significantly biased towards one of the classes. In the next step we simply take the majority vote of the responses of these workers. Choosing a resonably good value of r_0 is crucial in our algorithm since choosing a large value would include too many workers with low abilities in the final majority vote, while a small value would result in too little data for estimating the final answers. We now give an upper bound on the estimation error under our algorithm.

3.3 Upper Bound on Estimation Error

The following theorem gives an upper bound on the estimation error under our proposed estimation algorithm:

Theorem 2 *For any x^* in $\mathbb{F}_M^d$ and $Q^* \in \mathbb{C}_{perm}(x^*)$ with R^* corresponding to Q^* if our estimation algorithm is provided with a permutation π of workers, then for every item $l \in [d]$ such that*

$$||R_l^*||_2^2 \geq \frac{5\log^{2.5}(dn)}{p_0} \text{ and } ||R_l^\pi - R_l^*||_2^2 \leq \frac{||R_l^*||_2^2}{\sqrt{9log(dn))}} \tag{8}$$

we have

$$P(\hat{x}(\pi)_l = x*_l) \geq 1 - M\exp^{-c_E log^{1.5}(dn)} \tag{9}$$

and if π is the correct permutation of workers:

$$L_{Q*}(\hat{x}(\pi), x^*) \leq c' \frac{1}{np_0} \log^{2.5} d \tag{10}$$

with probability at least $1 - M\exp^{-c'_E log^{1.5}(dn)}$

4 Proofs

In the section we present the proofs of the aforementioned theorems. These proofs closely follow the proofs for the corresponding results in [4].

4.1 Proof of Theorem 1

The Gilbert-Varshamov bound [5] guarantees that for a universal constant $c_1 > 0$ there exists a collection of $\beta = \exp(c_1 d)$ vectors, $x_1, \ldots, x_\beta$ all lying in $\mathbb{F}_M^d$, such that the generalised Hamming error defined in (1) is lower bounded as:

$$d_H(x^l, x^{l'}) \geq \frac{1}{10} \; \forall \, l, l' \in [\beta] \tag{11}$$

We define the following probability tensor $Q^* \in \mathbb{C}_{\text{perm}}(x^*)$ for a given x^* such that for each item l:

$$Q^*_{klm} = \begin{cases} \frac{1}{M} + \delta, & \text{if } k \leq \frac{1}{p_0}, m = m_l^* \\ \frac{1}{M} - \delta, & \text{if } k \leq \frac{1}{p_0}, m = m_l^* +_M 1 \\ \frac{1}{M}, & \text{otherwise} \end{cases}$$

where $+_M$ denotes modulo-M addition and δ is a constant in the interval $(0, \frac{1}{M^2})$. For each $l \in [\beta]$ let $\mathbb{P}^l$ denote the probability distribution of observed worker responses Y, induced by setting $x^* = x^l$ and choosing $Q^* \in \mathbb{C}_{\text{perm}}(x^l)$ as described above. Some algebra gives the following upper bound on the KL-divergence between any pair of distributions from this collection:

$$D_{KL}(\mathbb{P}^l || \mathbb{P}^{l'}) \leq \frac{c_2 d\delta^2}{M} \tag{12}$$

for another universal constant $c_2 \geq 0$. Combining the above with Fano's inequality [6] yields that any estimator $\hat{x}$ has an expected Hamming error lower bounded as:

$$\mathbb{E}[d_H(\hat{(x)}, x^*)] \geq \frac{1}{20}(1 - \frac{c_2 d\delta^2 + M]log2}{M \log \beta}) \tag{13}$$

Consequently, for our choice of Q^*, the Q^*-Error is lower bounded as:

$$\mathbb{E}[L_{Q^*}(\hat{(x)}, x^*)] \geq \frac{\delta^2}{np_0}(1 - \frac{c_2\delta^2}{c_1 M} - \frac{\log 2}{c_1 d}) \geq \frac{c}{np_0} \tag{14}$$

where the choice of δ depends on c_1, c_2 and c.

4.2 Proof of Theorem 2

The proof of this theorem makes use of the following lemma from [4]:

Lemma 1 *For any vector $v \in [0,1]^n$ such that $v_1 \geq \ldots \geq v_n$, there must be some $\alpha \geq \lceil \frac{1}{2}||v||_2^2 \rceil$ such that*

$$\sum_{i=1}^{\alpha} v_i \geq \sqrt{\frac{\alpha ||v||_2^2}{2\log n}} \tag{15}$$

Using this result, we can follow the corresponding proof in [4] to show that that for any item l which satisfies the bounds:

$$||R_l^*||_2^2 \geq \frac{5\log^{2.5}(dn)}{p_0} \text{ and } ||R_l^\pi - R_l^*||_2^2 \leq \frac{||R_l^*||_2^2}{\sqrt{9log(dn))}} \tag{16}$$

there exist $r_l \geq \frac{1}{p_0}\log^{1.5}(dn)$ such that:

$$\sum_{l=1}^{r_l} R^*_{\pi^{-1}(k)l} \geq \frac{3}{4}\sqrt{\frac{k}{p_0}\log^{1.5}(dn)} \tag{17}$$

This essentially tells us that, for the "easy" questions, i.e. those satisfying the bounds of (16) we can find a set of top r_l workers based on our permutation π who are significantly "certain" of the correct class label.

For any item l, class m, and integer $r \in \{\frac{1}{p_0}\log^{1.5}(dn), \ldots, dn\}$, we generalise our definition of the bias in responses b_l^r in (4) as:

$$\gamma(r, l, m) = \min_{\hat{m} \neq m} \sum_{k \in [r]} (\mathbb{1}\{(Y_{\pi^{-1}(k)l} = m\} - \mathbb{1}\{Y_{\pi^{-1}(k)l} = \hat{m}\}) \tag{18}$$

$$b_l^r = \gamma(r, l, m_l^r) \tag{19}$$

We can observe that the following properties hold with probability at least $1 - M\exp^{-c_E log^{1.5}(dn)}$, for $\gamma(r, l, m)$ as a consequence of the Bernstein inequality [7]:

1. If $\sum_{k=1}^{r} R^*_{\pi^{-1}(k)l} \geq \frac{3}{4}\sqrt{\frac{k}{p_0}\log^{1.5}(dn)}$ then $\gamma(r, l, m_l^*) \geq \frac{1}{2}\sqrt{kp_0\log^{1.5}(dn)}$
2. $\gamma(r, l, m) \geq \frac{1}{2}\sqrt{kp_0\log^{1.5}(dn)}$ only if $m = m_l^*$ and $\sum_{k=1}^{r} R^*_{\pi^{-1}(k)l} \geq \frac{1}{2}\sqrt{\frac{k}{p_0}\log^{1.5}(dn)}$
3. If $\sum_{k=1}^{r} R^*_{\pi^{-1}(k)l} \geq \frac{1}{2}\sqrt{\frac{k}{p_0}\log^{1.5}(dn)}$ then $\gamma(r, l, m_l^*) > 0$

For any item l_0 that satisfies the bounds of (16), we define a set $L_0 = \{l \in [d] : \sigma^*(l) \geq \sigma^*(l_0)\}$ i.e. the set of questions at least as easy as L_0. Thus, each question in L_0 also satisfies the bounds in (16) by our assumption on question monotonicity for R^*. Thus the function $\gamma(r, l, m_l^*) \forall\, l \in L_0$ satisfies property (1) above.

We also define the set $L(r) = \{l \in [d] : \gamma(r,l,m) \geq \frac{1}{2}\sqrt{kp_0 \log^{1.5}(dn)}$ for some $m\}$. Thus $L(r)$ is the set of questions for which the responses of the top r workers are biased towards some answer. As per our definition of r_0 in (6), it is the integer for which the maximum number of questions have responses by the top r_0 workers that are biased towards some answer. Thus $|L(r_0)| \geq |L(r_{l_0})| \geq |L_0|$ where r_{l_0} corresponds to the upper limit of the summation in (17) for item l_0. This means either $L(r_0) = L_0$ or there is some question outside the set L_0 in $L(r_0)$.

If $L(r_0) = L_0$, by property (2) we can conclude that the response of the top r_0 workers is correct for all items in L_0 i.e. $[\hat{x}(\pi)]_{l_0} = [x^*_{l_0}]$.

If $|L(r_0| > |L_0|$ then there exists some item l' outside the set L_0 for which $\gamma(r_0, l', m_{l'}) \geq \frac{1}{2}\sqrt{kp_0 \log^{1.5}(dn)}$. From property (2) we know that $\sum_{k=1}^{r_0} R^*_{\pi^{-1}(k)l'} \geq \frac{1}{2}\sqrt{\frac{k}{p_0}\log^{1.5}(dn)}$. Since this item must be harder to classify than l_0 for it to lie outside L_0, by our assumption on question monotonicity, $\sum_{k=1}^{r_0} R^*_{\pi^{-1}(k)l} \geq \frac{1}{2}\sqrt{\frac{k}{p_0}\log^{1.5}(dn)}$ for all items l in L_0 and thus by properties (2) and (3) we conclude that $[\hat{x}(\pi)]_{l_0} = [x^*_{l_0}]$ even in this case.

Thus for every item l_0 that satisfies the bounds of (16), $\mathbb{P}([\hat{x}(\pi)]_{l_0} = [x^*_{l_0}]) \geq 1 - M\exp^{-c_E log^{1.5}(dn)}$. This along with the definition of the Q^*-Error yields the claimed result.

5 Conclusion

We propose a flexible permutation based model for the problem of multi-class labelling using crowdsourced data. Our model can handle non-binary classification and varying difficulty of tasks, and is thus a significant generalisation over previous work. We propose a modified error metric weighted according to the difficulty of each task for our model. We derive statistical lower bounds on the error, propose an estimation algorithm, and give an upper bound on the estimation error for this algorithm. Extending the proposed algorithm to the setting where ordering of the workers is not known, and performing real world experiments to verify our approach is left for future work.

6 References

[1] Zhang, Yuchen, et al. "Spectral methods meet EM: A provably optimal algorithm for crowdsourcing." Advances in neural information processing systems. 2014.

[2] Karger, David R., Sewoong Oh, and Devavrat Shah. "Efficient crowdsourcing for multi-class labeling." ACM SIGMETRICS Performance Evaluation Review 41.1 (2013): 81-92

[3] Dawid, Alexander Philip, and Allan M. Skene. "Maximum likelihood estimation of observer error-rates using the EM algorithm." Applied statistics (1979): 20-28.

[4] Shah, Nihar B., Sivaraman Balakrishnan, and Martin J. Wainwright. "A permutation-based model for crowd labeling: Optimal estimation and robustness." arxiv preprint arxiv:1606.09632 (2016).

[5] E. N. Gilbert. A comparison of signalling alphabets. Bell System Technical Journal, 31(3):504–522, 1952.

[6] T. M. Cover and J. A. Thomas. Elements of information theory. John Wiley Sons, 2012.

[7] S. Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. Ann. Sci. Inst. Sav. Ukraine, Sect. Math, 1(4):38–49, 1924.