Error Rate Analysis of Labeling by Crowdsourcing

Hongwei Li HWLI@STAT.BERKELEY.EDU
Department of Statistics, UC Berkeley

Bin Yu BINYU@STAT.BERKELEY.EDU
Department of Statistics, UC Berkeley

Dengyong Zhou DENGYONG.ZHOU@MICROSOFT.COM
Microsoft Research, Redmond

Abstract

Crowdsourcing label generation has been a crucial component for many real-world machine learning applications. In this paper, we provide finite-sample exponential bounds on the error rate (in probability and in expectation) of hyperplane binary labeling rules for the Dawid-Skene (and Symmetric Dawid-Skene) crowdsourcing model. The bounds can be applied to analyze many commonly used prediction methods, including the majority voting, weighted majority voting and maximum a posteriori (MAP) rules. These bound results can be used to control the error rate and design better algorithms. In particular, under the Symmetric Dawid-Skene model we use simulation to demonstrate that the data-driven EM-MAP rule is a good approximation to the oracle MAP rule which approximately optimizes our upper bound on the mean error rate for any hyperplane binary labeling rule. Meanwhile, the average error rate of the EM-MAP rule is bounded well by the upper bound on the mean error rate of the oracle MAP rule in the simulation.

1. Introduction

There are many tasks that can be easily carried out by people but tend to be hard for computers, e.g. image annotation and visual design. When these tasks require large scale data processing, outsourcing them to experts or well trained people may be too expensive. Crowdsourcing has recently emerged as a powerful alternative. It outsources tasks to a distributed group of people (usually called workers) who might be inexperienced on these tasks. However, if we can appropriately aggregate the outputs from a crowd, the aggregated results could be as good as the results by experts. (Dawid & Skene, 1979; Smyth et al., 1995; Snow et al., 2008; Whitehill et al., 2009; Raykar et al., 2010; Yan et al., 2010; Welinder et al., 2010; Liu et al., 2012; Zhou et al., 2012).

The flaws of crowdsourcing are apparent. As each worker is paid purely based on how many tasks that she has completed (for example, one cent for labeling one image) and no ground truth are available to evaluate how well she has performed in the tasks, some workers may randomly submit answers independent of the questions when the tasks assigned to them are beyond their expertise. Moreover, workers are usually not persistent. Some workers may complete many tasks, while the others may only finish very few tasks even one task.

Then we have a question. In spite of these drawbacks, is it still possible to get reliable answers in a crowdsourcing system? In fact, majority voting has been able to generate fairly reasonable results (Snow et al., 2008; Raykar et al., 2010; Liu et al., 2012; Zhou et al., 2012). However, majority voting treats each worker's result as equal in quality, and does not distinguish a spammer from a diligent worker. So we can expect that majority voting can be significantly improved upon.

The first work on crowdsourcing might be due to Dawid & Skene (1979). They assumed that each worker is associated with an unknown confusion matrix. Each off-diagonal element represents misclassification rate from one class to the other, while the diagonal elements represent the accuracy in each class. Ac-

International Conference on Machine Learning 2013 Workshop: Machine Learning Meets Crowdsourcing. Atalanta, Georgia, USA. 2013.

cording to the observed labels by the workers, the maximum likelihood principle is applied to jointly estimate unobserved true labels and worker confusion matrices. The likelihood function is non-convex. However, a local optimum can be obtained by using expectation-maximization (EM) algorithm. It can be naturally initialized by using majority voting.

The method by Dawid & Skene (1979) can be naturally extended by adding feature vectors (Raykar et al., 2010) or using prior over worker confusion matrices (Raykar et al., 2010; Liu et al., 2012). When considering binary labeling, we call the confusion probabilistic model by Dawid & Skene (1979) the *Dawid-Skene* model. If one simplifies the assumption to consider a symmetric confusion matrix for binary labeling, then it is referred to as the *Symmetric Dawid-Skene* model.

Karger et al. (2011a) provide asymptotic error bounds for majority voting and also their iterative algorithm. However, the error bound for their specific iterative algorithm cannot be generalized to other prediction rules in crowdsourcing and the asymptotic bounds may not be that practical since we always only have finite number of tasks. Additionally, their results depend on the assumption that the same number of items were assigned to each worker, and the same number of workers labeled each item. According to the analysis in (Liu et al., 2012) by Liu et al., this assumption is restrictive in practical case.

In this paper, we focus on providing bounds on the error rate under crowdsourcing models of which the effectiveness on real data have been evaluated in (Dawid & Skene, 1979; Raykar et al., 2010; Liu et al., 2012). Our main contributions are as follows: (1) We provide error rate bounds in probability and in expectation for a finite number of workers and instances under the Dawid-Skene model (with the Symmetric Dawid-Skene model as a special case). (2) We provide error bounds for an oracle Maximum A Posterior (MAP) labeler. Under the Symmetric Dawid-Skene model, we use simulation to demonstrate that the data-driven EM-MAP rule approximates well the oracle MAP rule and the MAP rule approximately optimizes the upper bound on the mean error rate for any hyperplane rule.

2. Problem setting

Assume that a set of workers are assigned to label certain items that are available on the Internet. For instance, whether an image of an animal is that of a cat or a dog, or if a face image is male or female.

In this paper, we will focus on binary labeling. Formally, suppose we have M workers, and N items. For convenience, we denote $[M] = \{1, \cdots, M\}$ and $[N] = \{1, \cdots, N\}$. The response matrix is denoted by $\widetilde{Z} \in \{\pm 1\}^{M \times N}$, in which $\widetilde{Z}_{ij}$ is the label of the j-th item given by the i-th worker. What we observe is Z, which is a realization of random matrix $\widetilde{Z}$ with missing entries and called *label matrix*. Therefore, $Z \in \{\pm 1, 0\}^{M \times N}$,where Z_{ij} is the label of the $j-$th item given by the i-th worker. It will be 0 if the corresponding label is missing.

Throughout the paper, we use y_j as the true label for j-th item, and $\hat{y}_j$ as the predicted label for the j-th item by an algorithm. At the same time, any parameter with a hat $\hat{}$ is an estimator for this parameter.

2.1. Problem formulation

Let $\pi = \mathbb{P}(y_j = 1)$ for any $j \in [N]$ denotes the prevalence of label "+1" in the true labels of the items.

We introduce the indicator matrix $T = (T_{ij})_{M \times N}$, where $T_{ij} = 1$ indicates that entry (i, j) is observed, i.e. the i-th worker has labeled the j-th item. And $T_{ij} = 0$ indicates entry (i, j) is unobserved. Note that T and Z are observed together in our crowdsourcing setting, and $T_{ij} = \mathrm{I}(Z_{ij} \neq 0) = \mathrm{I}(Z_{ij} = 1) + \mathrm{I}(Z_{ij} = -1)$.

The sampling probability matrix $Q = (q_{ij})_{M \times N}$, where $q_{ij} = \mathbb{P}(T_{ij} = 1)$. If every entry is sampled with the same probability, which is called *constant probability sampling strategy* or *with constant probability sampling distribution*, we will denote the sampling probability matrix Q as scalar $q \in (0, 1]$. Therefore, the entries in label matrix are constant probability sampled means that each worker has probability q to label any item.

We will discuss two models that we use for modeling the quality of the workers. They were first proposed by Dawid & Skene (1979):

Dawid-Skene model: we distinguish the accuracy of workers on the positive class and the negative class. Some workers might work better on labeling the items with true label "+1", and some might work better at labeling the items with true label "-1". That is, we model each worker as tossing two biased coins, and their true positive rate (sensitivity) and true negative rate (specificity) are denoted as follows respectively: for $i = 1, 2, \cdots, M$

$$p_i^+ := P(Z_{ij} = 1 | y_j = 1, T_{ij} = 1), \quad (1)$$

$$p_i^- := P(Z_{ij} = -1 | y_j = -1, T_{ij} = 1). \quad (2)$$

Then the parameter set will be $\Theta = \left\{ \{p_i^+, p_i^-\}_{i=1}^M, Q, \pi \right\}$ under this model.

Symmetric Dawid-Skene model: we assume that the i-th worker labels the item correctly with a fixed probability $w_i = \mathbb{P}(Z_{ij} = y_j | y_j, T_{ij} = 1)$. In this case, no matter whether the item is from positive class or negative class, the worker can label it correctly with the same accuracy. Therefore, the parameter set is $\Theta = \left\{ \{w_i\}_{i=1}^M, Q, \pi \right\}$.

Under both models above, the posterior probability of the label for each item to be "+1" is defined as:

$$\rho_j = \mathbb{P}(y_j = 1 | Z, T, \Theta), \quad \forall j \in [N]. \tag{3}$$

Given an estimation or a prediction rule, suppose that its predicted label for item j is $\hat{y}_j$, then our objective is to minimize the error rate, i.e. $\mathbf{e} = \frac{1}{N} \sum_{j=1}^N \mathrm{I}(\hat{y}_j \neq y_j)$.

Since the error rate is random, we are also interested in its expected value (i.e., the *mean error rate*), defined as:

$$\mathbb{E}[\mathbf{e}] = \frac{1}{N} \sum_{j=1}^N \mathbb{P}(\hat{y}_j \neq y_j). \tag{4}$$

The rest of the paper is organized as follows. In Section 3, under the Dawid-Skene model, we present finite-sample bound on the error rate of a hyperplane rule with a high probability and also a bound on the mean error rate. In Section 4, we will apply our analysis to the label inference by Maximum Likelihood method, and illustrate the bound on oracle Maximum A Posteriori classifier under the Symmetric Dawid-Skene model.

Note that we will state the main results and defer the proofs to the supplementary materials.

3. Error rate bounds

In this section, we provide finite-sample bounds on error rate of any hyperplane rule under the Dawid-Skene model and on the mean error rate as well.

3.1. Bounds on the error rate under the Dawid-Skene model

A *hyperplane rule* is a rectified linear function of the observation matrix Z in a high dimensional space: given an unnormalized weight vector $\nu = (\nu_1, \cdots, \nu_M)$ and an shift constant a, for the jth item, the rule estimates its label as

$$\hat{y}_j = \mathrm{sign}\left(\sum_{i=1}^M \nu_i Z_{ij} + a \right).$$

This method is also called *linear threshold rule*. It is a very general rule with special cases including the most common method, majority voting (MV) corresponding to all weights being 1.

Next we present two general theorems to give finite-sample error rate bounds for hyperplane (linear threshold) rule under the Dawid-Skene model. Before that, we want to define some notations as follows:

$$\Lambda_j^+ = \left(\sum_{i=1}^M q_{ij} \nu_i (2p_i^+ - 1) + a \right) \tag{5}$$

$$\Lambda_j^- = \left(\sum_{i=1}^M q_{ij} \nu_i (2p_i^- - 1) - a \right) \tag{6}$$

$$t_1 = \min_{j \in [N]} \frac{\Lambda_j^+ \wedge \Lambda_j^-}{||\nu||} \quad \text{and} \quad t_2 = \max_{j \in [N]} \frac{\Lambda_j^+ \vee \Lambda_j^-}{||\nu||}$$

$$\phi(x) = e^{-\frac{x^2}{2}} \qquad x \in \mathbb{R}$$

$$\mathrm{D}(x||y) = x \ln \frac{x}{y} + (1-x) \ln \frac{1-x}{1-y} \quad , \ x, y \in (0,1),$$

where $||\cdot||$ is L_2 norm, $x \wedge y$ is $\min\{x, y\}$ and $x \vee y$ is $\max\{x, y\}$.

Theorem 1. *For a given sampling probability matrix $Q = (q_{ij})_{M \times N}$, let $\hat{y}_j$ be the estimate obtained using hyperplane rule with weight vector ν and shift constant a. For any $\epsilon \in (0,1)$, we have the following bounds on the error rate in probability.*

(1) When $t_1 \geq \sqrt{2 \ln \frac{1}{\epsilon}}$,

$$\mathbb{P}\left(\frac{1}{N} \sum_{j=1}^N I(\hat{y}_j \neq y_j) \leq \epsilon \right) \ \geq \ 1 - e^{-N D(\epsilon || \phi(t_1))}. \tag{7}$$

(2) When $t_2 \leq -\sqrt{2 \ln \frac{1}{1-\epsilon}}$,

$$\mathbb{P}\left(\frac{1}{N} \sum_{j=1}^N I(\hat{y}_j \neq y_j) \leq \epsilon \right) \ \leq \ e^{-N D(\epsilon || 1 - \phi(t_2))}. \tag{8}$$

Remark: This theorem implies that t_1 and t_2 are two very important quantities for controlling the error rate. They depend on Λ_j^+ and Λ_j^-, which are functions of the sampling probabilities q_{ij}, weights ν_i and shift a in the prediction rule, and labeling quality measures p_i^+ and p_i^- of the i-th worker. For a fixed sampling probability matrix, if the weights are positive, then the better the worker over random guessing (or the bigger $2p_i^+ - 1$) for "+1" labels and the larger the shift a, the larger the Λ_j^+. Similarly we can interpret Λ_j^-. Note that usually we are free to choose ν and a, and in some situation we can also control Q, so the most important factors that we cannot control are p_i^+ and $p_i^-, i \in [M]$.

To control the probability of bounding the error rate to be at least $1 - \delta$, we have to solve the equation

$\exp\{-ND(\epsilon||\phi(t_1))\} = \delta$, which can not been solved analytically, so we need to consider about a method which can tell us what's the minimum t_1 for bounding the error rate with probability at least $1-\delta$. The next theorem serves this purpose.

For notation convenience, we define two constants C and G for $\epsilon, \delta \in (0,1)$:

$$C(\epsilon,\delta) = 1+\exp\left(\frac{1}{\epsilon}\left[H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right]\right), \quad (9)$$

$$G(\epsilon,\delta) = 1+\exp\left(\frac{1}{1-\epsilon}\left[H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right]\right), \quad (10)$$

where $H_e(\epsilon) = -\epsilon\ln\epsilon-(1-\epsilon)\ln(1-\epsilon)$.

Theorem 2. *Following all the setting and notation in Theorem 1, for $\forall\epsilon,\delta\in(0,1)$, we have*

(1) if $t_1 \geq \sqrt{2\ln C(\epsilon,\delta)}$, then

$$\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N} I(\hat{y}_j \neq y_j) \leq \epsilon\right) \geq 1-\delta.$$

(2) If $t_2 \leq -\sqrt{2\ln G(\epsilon,\delta)}$, then

$$\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N} I(\hat{y}_j \neq y_j) \leq \epsilon\right) < \delta.$$

To gain more insights, we now give some corollaries to cover special cases of Theorem 2.

Note that the majority voting (MV) method uses the prediction $\hat{y}_j = \text{sign}\left(\sum_{i=1}^{M} Z_{ij}\right)$, i.e. $\nu_i = 1\ \forall i\in[M]$. Bounds on error rate of MV under Dawid-Skene model thus follow easily from Theorem 2. Moreover, we are interested in the case that majority voting with the entries constant-probability sampled, or when the investigator randomly assigns tasks to random workers, and then takes the majority as the prediction.

Corollary 3. *For the majority voting rule with constant probability sampling $q\in(0,1]$, and $\forall\epsilon,\delta\in(0,1)$, (1) if*

$$\min\left\{\frac{\sum_{i=1}^{M} p_i^+}{M}, \frac{\sum_{i=1}^{M} p_i^-}{M}\right\} \geq \frac{1}{2}+\frac{1}{q}\sqrt{\frac{\ln C(\epsilon,\delta)}{2M}},$$

then $\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N} I(\hat{y}_j\neq y_j)\leq\epsilon\right)\geq 1-\delta$.

(2) If

$$\max\left\{\frac{\sum_{i=1}^{M} p_i^+}{M}, \frac{\sum_{i=1}^{M} p_i^-}{M}\right\} \leq \frac{1}{2}-\frac{1}{q}\sqrt{\frac{\ln G(\epsilon,\delta)}{2M}},$$

then $\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N} I(\hat{y}_j\neq y_j)\leq\epsilon\right)<\delta$.

Proof. By plugging in $q_{ij}=q$ and manipulating the inequality in Theorem 2, we can easily see this result. □

Remark: we can see that, under constant probability sampling, the average accuracies of the workers on both positive samples and negative samples have to be high enough, relative to random guessing, to maintain a good performance of prediction by majority voting.

3.2. Bounds on the error rate under the Symmetric Dawid-Skene model

Under the Symmetric Dawid-Skene model, we assume that the labeling accuracies on both positive samples and negative samples are the same for each worker, which is to say

$$w_i = \mathbb{P}(Z_{ij}=1|y_j=1,T_{ij}=1) = \mathbb{P}(Z_{ij}=-1|y_j=-1,T_{ij}=1). \quad (11)$$

Let $\bar{w} = \frac{1}{M}\sum_{i=1}^{M} w_i$ be the average accuracy of the workers.

There is another method as a special case of hyperplane rule called *weighted majority voting (WMV)*. For WMV, the predicted label for the j-th item is given by $\hat{y}_j = \text{sign}\left(\sum_{i=1}^{M}\alpha_i Z_{ij}\right)$ with the weight vector $\alpha=(\alpha_1,\cdots,\alpha_M)$ and $\|\alpha\|^2=\sum_{i=1}^{M}\alpha_i^2=1$.

The conditions simplify and hence are easier to interpret in the Symmetric Dawid-Skene model for the WMV method to have an error rate below a certain threshold.

Corollary 4. *Under the Symmetric Dawid-Skene model, suppose the prediction function is $\hat{y}_j = sign\left(\sum_{i=1}^{M}\alpha_i Z_{ij}\right)$. Let $t_1 = \min_{j\in[N]}\sum_{i=1}^{M} q_{ij}\alpha_i(2w_i-1)$ and $t_2=\max_{j\in[N]}\sum_{i=1}^{M} q_{ij}\alpha_i(2w_i-1)$. Then $\forall\epsilon,\delta\in(0,1)$,*

(1) if $t_1\geq\sqrt{2\ln\frac{1}{\epsilon}}$, then
$\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N} I(\hat{y}_j\neq y_j)\leq\epsilon\right) \geq 1-e^{-ND(\epsilon||\phi(t_1))}$;

(2) if $t_2\leq-\sqrt{2\ln\frac{1}{1-\epsilon}}$, then
$\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N} I(\hat{y}_j\neq y_j)\leq\epsilon\right) \leq e^{-ND(\epsilon||1-\phi(t_2))}$.

Proof. It follows from Theorem 1 by taking $p_i^+=w_i$ and $p_i^-=w_i$, $\nu_i=\alpha_i$. □

Next we consider another special case of majority voting with constant probability sampling entries, i.e. $q_{ij}=q$ for $\forall i\in[M], j\in[N]$.In this case, the weight of each worker is the same, so $\alpha_i=\frac{1}{\sqrt{M}}$.

Corollary 5. *Under the Symmetric Dawid-Skene model, for majority voting with constant probability sampling $q \in (0,1]$, if*

$$\bar{w} \geq \frac{1}{2} + \frac{1}{q}\sqrt{\frac{1}{2M}\ln\frac{1}{\epsilon}}, \tag{12}$$

then $\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N} I(\hat{y}_j \neq y_j) \leq \epsilon\right) \geq 1 - e^{-ND(\epsilon||\psi)}$, *where* $\psi = e^{-2Mq^2(\bar{w}-0.5)^2}$.

Proof. It follows from Theorem 1 by taking $q_{ij} = q$, $\nu_i = \alpha_i = \frac{1}{\sqrt{M}}$ and $a = 0$. □

3.3. Bounding the mean error rate

Oftentimes, one is also interested in bounding the mean error rate for general hyperplane rule. The mean error rate gives the expected proportion of items wrongly labeled. The results specified with majority voting under both the Symmetric Dawid-Skene and Dawid-Skene models will be obtained in this section.

Theorem 6. *Under the same setting with Theorem 1. For any $\epsilon, \delta \in (0,1)$,*
(1) if $t_1 \geq 0$, then

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \leq \exp\left(-\frac{t_1^2}{2}\right);$$

(2) if $t_2 \leq 0$, then

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \geq 1 - \exp\left(-\frac{t_2^2}{2}\right).$$

To prove this theorem, we are going to present another theorem which can provide the bounds on the mean error rate of each individual item. This can also have individual interest since it tells us the probability of labeling a specific item wrong.

Theorem 7. ***(Bounding the mean error rate of each item)*** *Under the assumptions of Theorem 6 ,and for any $j \in [N]$, let*

$$\tau_{j,min} = \frac{\Lambda_j^+ \wedge \Lambda_j^-}{||\nu||} \quad \text{and} \quad \tau_{j,max} = \frac{\Lambda_j^+ \vee \Lambda_j^-}{||\nu||}, \tag{13}$$

(1) if $\tau_{j,min} \geq 0$, then $\mathbb{P}(\hat{y}_j \neq y_j) \leq \exp\left(-\frac{\tau_{j,min}^2}{2}\right)$;

(2) if $\tau_{j,max} \leq 0$, then $\mathbb{P}(\hat{y}_j \neq y_j) \geq 1 - \exp\left(-\frac{\tau_{j,max}^2}{2}\right)$.

Due to the complicated form, Theorem 6 might not be very intuitive. Let us look at some special simple cases to gain more insights.

Corollary 8. *For majority voting with constant sampling probability $q \in (0,1]$,*

(1) if $\frac{\sum_{i=1}^{M} p_i^+}{M} \geq \frac{1}{2}$ and $\frac{\sum_{i=1}^{M} p_i^-}{M} \geq \frac{1}{2}$, then

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \leq e^{-2Mq^2t_1^2},$$

where $t_1 = \min\left\{\frac{\sum_{i=1}^{M} p_i^+}{M}, \frac{\sum_{i=1}^{M} p_i^-}{M}\right\} - \frac{1}{2}$.

(2) If $\frac{\sum_{i=1}^{M} p_i^+}{M} \leq \frac{1}{2}$ and $\frac{\sum_{i=1}^{M} p_i^-}{M} \leq \frac{1}{2}$, then

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \geq 1 - e^{-2Mq^2t_2^2},$$

where $t_2 = \min\left\{\frac{\sum_{i=1}^{M} p_i^+}{M}, \frac{\sum_{i=1}^{M} p_i^-}{M}\right\} - \frac{1}{2}$.

Proof. It follows from Theorem 6 by taking $a = 0$, $\frac{\nu_i}{||\nu||} = \frac{1}{\sqrt{M}}$ $\forall i \in [M]$, and $q_{ij} = q$. □

Remark: The corollary above implies that the mean error rate will decay exponentially w.r.t. M as long as the average accuracies of labeling on both positive samples and negative samples are better than random guessing.

One of the main results in (Karger et al., 2011b) is similar to the following corollary for majority voting under the Symmetric Dawid-Skene model. Note their result is asymptotic ($N \to \infty$) but our result is applicable to both asymptotic and finite sample scenarios.

Corollary 9. *For the majority voting under the Symmetric Dawid-Skene model and constant probability sampling $q \in (0,1]$,*

(1) if $\bar{w} > \frac{1}{2}$, then
$\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \leq e^{-2Mq^2(\bar{w}-\frac{1}{2})^2}$;

(2) if $\bar{w} < \frac{1}{2}$, then
$\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \geq 1 - e^{-2Mq^2(\bar{w}-\frac{1}{2})^2}$.

Proof. As we have mentioned in the previous sections, Symmetric Dawid-Skene model is a special case of Dawid-Skene model by adding constraints $w_i = p_i^+ = p_i^-$ $\forall i \in [M]$. Using w_i to replace both p_i^+ and p_i^- in Corollary 8, we easily obtain the results. Note that when $\bar{w} = \frac{1}{2}$, the inequalities in this corollary are trivial. □

Remark: this corollary tells us that the mean error rate will exponentially decay with M increase if the average accuracy of labeling by the workers is better than

random guessing, and the gap between the average accuracy and 0.5 plays an important role in the bound. In particular, it implies that (1) if $\lim_{M\to\infty}\bar{w} > \frac{1}{2}$, then $\lim_{M\to\infty}\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right)\right) = 0$; and (2) if $\lim_{M\to\infty}\bar{w} < \frac{1}{2}$, then $\lim_{M\to\infty}\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right)\right) = 1$.

4. Crowdsourcing via data-driven EM-MAP rule

If we know the posterior probability ρ_j (defined in (3)) of the label of each item, then the Bayesian classifier predict $\hat{y}_j = 2\mathrm{I}\left(\rho_j > 0.5\right) - 1$.

Thus if we estimate the posterior ρ_j well, we can apply the same rule to predict the true label with the estimated posterior probability. One natural way to approach it is to apply the Maximum Likelihood method to the observed label matrix in order to estimate the parameter set Θ and consequently the posterior.

4.1. Maximum A Posteriori (MAP) rule and the oracle MAP rule

Generally, we can apply the EM algorithm to obtain the maximum likelihood estimate for the parameters (Dawid & Skene, 1979) and the posterior $\hat{\rho}_j$. With $\hat{\rho}_j$, each item can be assigned with the label which has the largest posterior, that is, the prediction function with MAP rule is $\hat{y}_j = 2\mathrm{I}\left(\hat{\rho}_j > 0.5\right) - 1$, where $\hat{\rho}_j$ is the estimated posterior probability. We call this method *the EM-MAP rule*.

However, the EM algorithm can not guarantee convergence to the global maximum of the likelihood function. Thus the estimated parameters might not be close to the true parameters and similarly for the estimated posterior.

Now, we consider the oracle MAP rule, which knows the true parameters and thus uses the true posterior in MAP rule to label items. Recall that the prediction function for the oracle MAP rule is $\hat{y}_j = 2\mathrm{I}\left(\rho_j > 0.5\right) - 1$, where ρ_j is the true posterior.

4.2. Bounds on the mean error rate of the oracle MAP rule

We first provide bounds on the mean error rate of the oracle MAP rule under the Dawid-Skene model in the next theorem.

To simplify notation, we define:

$$u_i = \ln\frac{p_i^+}{1-p_i^-}, \quad \nu_i = \ln\frac{1-p_i^+}{p_i^-}, \quad a = \ln\frac{\pi}{1-\pi}, \tag{14}$$

$$\Gamma = \frac{1}{2}\sqrt{\sum_{i=1}^{M}\left(\max\left\{u_i,\nu_i,0\right\} - \min\left\{u_i,\nu_i,0\right\}\right)^2}. \tag{15}$$

Furthermore, let us define:

$$\Upsilon_j^+ = \left(\sum_{i=1}^{M} q_{ij}\left[(u_i-\nu_i)p_i^+ + \nu_i\right] + a\right), \tag{16}$$

$$\Upsilon_j^- = \left(\sum_{i=1}^{M} q_{ij}\left[(u_i-\nu_i)p_i^- - u_i\right] - a\right), \tag{17}$$

$$t_1 = \min_{j\in[N]}\frac{\Upsilon_j^+ \wedge \Upsilon_j^-}{\Gamma}, \quad \text{and} \quad t_2 = \max_{j\in[N]}\frac{\Upsilon_j^+ \vee \Upsilon_j^-}{\Gamma} \tag{18}$$

The prediction function of the oracle MAP rule is $\hat{y}_j = \mathrm{sign}\left(\sum_{i=1}^{M}\left[u_i\mathrm{I}\left(Z_{ij}=1\right) + \nu_i\mathrm{I}\left(Z_{ij}=-1\right)\right] + a\right)$. Note that Z_{ij} could be 0, thus the oracle MAP rule under the Dawid-Skene model is not a hyperplane rule.

Theorem 10. *For the oracle MAP rule knowing the true parameters $\Theta = \left\{(p_i^+, p_i^-)_{i=1}^M, Q, \pi\right\}$, its prediction function is $\hat{y}_j = I(\rho_j > 0.5)$. With t_1 and t_2 defined as in (18) respectively, we have*

(1) if $t_1 \geq 0$, then $\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \leq e^{-\frac{1}{2}t_1^2}$;

(2) if $t_2 \leq 0$,then $\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \geq 1 - e^{-\frac{1}{2}t_2^2}$.

Since the Symmetric Dawid-Skene model is a special case of the Dawid-Skene model by assuming that $p_i^+ = p_i^- = w_i$, results in the last corollary can be adapted to the Symmetric Dawid-Skene model directly. The only difference is that t_1 and t_2 simplify as follows:

$$t_1 = \min_{j\in[N]}\frac{1}{||\nu||}\left(\sum_{i=1}^{M} q_{ij}\nu_i(2w_i-1) - |a|\right), \tag{19}$$

$$t_2 = \max_{j\in[N]}\frac{1}{||\nu||}\left(\sum_{i=1}^{M} q_{ij}\nu_i(2w_i-1) + |a|\right), \tag{20}$$

where $a = \ln\frac{\pi}{1-\pi}$ and $\nu_i = \ln\frac{w_i}{1-w_i}$. Note that oracle MAP rule under the Symmetric Dawid-Skene model is a hyperplane rule with ν_i as weight and a as shift in the prediction function. With the t_1 and t_2 defined as above, the same results as Theorem 10 hold under the Symmetric Dawid-Skene model.

As the Symmetric Dawid-Skene model is intuitive and easy to visualize, it is a good example for illustration and simulation.

4.3. Exploring the oracle MAP rule under the Symmetric Dawid-Skene model

In this section, we will be exploring the relation between the oracle and the error rate bound. For simplicity, we consider the situation that the entries in the observed label matrix is sampled with a constant probability $q \in (0, 1]$ under the Symmetric Dawid-Skene model. It's to mention that though we can obtain the similar results under the Dawid-Skene model as well, we use the Symmetric Dawid-Skene model here for clarification.

Let us look closely at the mean error rate bound in Theorem 6.(1). When $t_1 \geq 0$, we have the upper bound for the mean error rate. Note that for prediction by a general hyperplane rule with the entries in the label matrix sampled with a constant probability, we have

$$t_1 = q \sum_{i=1}^{M} \frac{\nu_i}{||\nu||}(2w_i - 1) - \frac{|a|}{||\nu||}.$$

For a fixed q and $t_1 \geq 0$, we can optimize the upper bound to get

$$\begin{aligned}(\nu^\star, a^\star) &= \underset{\nu\in\mathbb{R}^M, a\in\mathbb{R}}{\operatorname{argmin}} \; e^{-\frac{1}{2}t_1^2} = \underset{\nu\in\mathbb{R}^M, a\in\mathbb{R}}{\operatorname{argmax}} \; t_1 \\ &= \underset{\nu\in\mathbb{R}^M, a\in\mathbb{R}}{\operatorname{argmax}} \left(q \sum_{i=1}^{M} \frac{\nu_i}{||\nu||}(2w_i - 1) - \frac{|a|}{||\nu||} \right) \\ &\Rightarrow \quad \nu_i^\star \propto 2w_i - 1 \quad \text{and} \quad a^\star = 0. \end{aligned} \tag{21}$$

Since the strategy to choose the hyperplane as in (21) is optimizing the upper bound of the mean error rate instead of the mean error rate itself, we call this strategy oracle bound-optimal rule.

The prediction function of oracle MAP rule is

$$\hat{y}_j = \operatorname{sign}\left(\sum_{i=1}^{M} \ln\left(\frac{w_i}{1-w_i}\right) \cdot Z_{ij} + \ln\frac{\pi}{1-\pi} \right),$$

which is a hyperplane rule with weight $\nu_i^{\text{oracMAP}} = \ln\frac{w_i}{1-w_i}$ and $a^{\text{oracMAP}} = \ln\frac{\pi}{1-\pi}$.

Note that as $|\nu_i^{\text{oracMAP}}| = |\ln\frac{w_i}{1-w_i}|$, which will be very large if w_i is close to 0 or 1. So if any $i \in [M]$, w_i is too small (close to 0) or too large (close to 1), then $||\nu^{\text{oracMAP}}||$ will be large and $\frac{|a^{\text{oracMAP}}|}{||\nu^{\text{oracMAP}}||}$ will be close to zero, especially for balanced label classes (π is close to 0.5, i.e. the chance that one item is from the positive class is close to the chance that one item is from a negative class).

Since by Taylor expansion of $\ln\frac{x}{1-x}$ around $\frac{1}{2}$,

$$\ln\frac{x}{1-x} = (4x-2) + O\left(\left(x - \frac{1}{2}\right)^2\right).$$

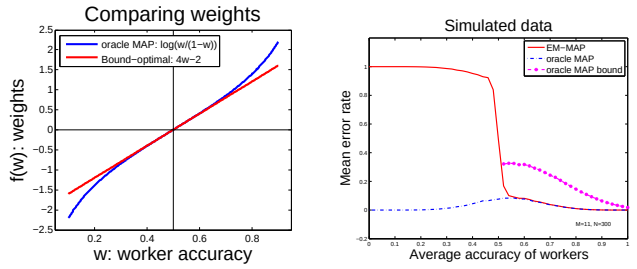
We can see that the weight of the oracle bound-optimal rule is the first order Taylor expansion of the oracle MAP rule. See Fig.1(a).

By observing that the oracle MAP rule is very close to the oracle bound-optimal rule, the oracle MAP rule approximately optimizes the upper bound of the mean error rate under Symmetric Dawid-Skene model. This fact also indicates that our bound is meaningful since the oracle MAP rule is the oracle Bayes classifier.

4.4. Comparing the EM algorithm with the oracle MAP rule by simulation

Since it is generally hard to provide the mean error rate bound for the EM-MAP rule due to its local optimality property, we compare it with the oracle MAP rule through simulation. The simulation is run under the Symmetric Dawid-Skene model with a constant sampling probability $q = 0.8$ to pick workers to label items. In the simulation, the average accuracy of the workers varies from 0 to 1 with step size 0.02. Given the average accuracy of the workers, it is inefficient to use uniform sampling to sample each accuracy of workers for matching the average accuracy. Therefore we sampled the accuracy of workers from beta distribution $Beta(a, b)$ with $b = 2$ and let the expectation $\frac{a}{a+b}$ be set to the average accuracy of workers we want to obtain. Although the average accuracy wouldn't be exact $\frac{a}{a+b}$, it will be very close to it so that we can roughly change the average accuracy of workers each time. For the labels of the items, there will be half of the items with the positive labels. We used 11 simulated workers to label 300 items. The simulation results are presented in Figure 1(b).

From Figure 1(b), we see that when the average accuracy of workers is better than random guessing, the EM-MAP rule converges to the oracle MAP rule very fast (the red curve in Figure 1(b) decreases really fast when $\bar{w}$ is around 0.5). It is interesting that when $\bar{w} < 0.5$, the gap is large (close to 1). When $\bar{w} < 0.5$, the mean error rate of the oracle MAP rule is close to 0, which is good, while the mean error rate by the EM-MAP rule is close to 1, i.e., nearly all of the items are labeled wrong. This is because when all the workers have low labeling accuracy, the EM algorithm cannot recognize since there is no ground truth. But in this case, since the oracle knows the true accuracies of workers, it can simply flip the label from the workers.



(a) (b)

Figure 1. (a) Comparing the weights between oracle MAP rule with the bound-optimal rule. (b)Comparison of the mean error rate between the EM-MAP rule with the oracle MAP rule.

We also plot the upper bound of the oracle MAP rule when $\bar{w} > 0.5$, from which we can see that this theoretical bound also bounded the mean error rate of the EM-MAP rule well.

Figure 1(b) has shown that the EM-MAP rule is a good algorithm because it approximates the oracle MAP rule. And from last section, we know that oracle MAP rule is close to suboptimal oracle rule (by minimizing the upper bound of the mean error rate). Therefore, the EM-MAP rule is effective or has good error rate properties.

5. Conclusion and Future work

In this paper, we have provided bounds on error rate of hyperplane labeling rules under the Dawid-Skene crowdsourcing model that includes the Symmetric Dawid-Skene model as a special case.

Our bounds are in probability and in expectation, and they hold for finite numbers of workers and items. Optimizing the mean error rate bound in the Symmetric Dawid-Skene model and for constant probability sampling leads to a prediction rule that is a good approximation to the Bayesian classifier . And the EM-MAP rule is shown via simulation to be close to the oracle MAP rule under the Symmetric Dawid-Skene model, and its error rate is bounded well by the bound on the mean error rate of the oracle MAP rule in the simulation.

To the best of our knowledge, this is the first work on error rate bounds for the more applicable Dawid-Skene model for crowdsourcing. The bounds we have obtained are useful for explaining the effectiveness of different prediction rules/functions. In the future, we plan to extend our results to the multiple-labeling situation and explore other types of crowdsourcing applications.

References

Dawid, A. P. and Skene, A. M. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society.*, 28(1):20–28, 1979.

Karger, David R., Oh, Sewoong, and Shah, Devavrat. Iterative learning for reliable crowdsourcing systems. In *NIPS*, 2011a.

Karger, David R., Oh, Sewoong, and Shah, Devavrat. Budget-optimal crowdsourcing using low-rank matrix approximations. In *the 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 284–291. Ieee, September 2011b. doi: 10.1109/Allerton.2011.6120180.

Liu, Qiang, Peng, Jian, and Ihler, Alexander. Variational Inference for Crowdsourcing. In *NIPS*, 2012.

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

Smyth, Padhraic, Fayyad, Usama, Burl, Michael, Perona, Pietro, and Baldi, Pierre. Inferring Ground Truth from Subjective Labelling of Venus Images. In *NIPS*, 1995.

Snow, Rion, Connor, Brendan O, Jurafsky, Daniel, Ng, Andrew Y, Labs, Dolores, and St, Capp. Cheap and Fast - But is it Good ? Evaluating Non-Expert Annotations for Natural Language Tasks. *EMNLP*, 2008.

Welinder, Peter, Branson, Steve, Belongie, Serge, and Perona, Pietro. The Multidimensional Wisdom of Crowds. In *NIPS*, 2010.

Whitehill, Jacob, Ruvolo, Paul, Wu, Tingfan, Bergsma, Jacob, and Movellan, Javier. Whose Vote Should Count More : Optimal Integration of Labels from Labelers of Unknown Expertise. In *NIPS*, 2009.

Yan, Yan, Rosales, Romer, Fung, Glenn, Schmidt, Mark, Hermosillo, Gerardo, Bogoni, Luca, Moy, Linda, and Dy, Jennifer G. Modeling annotator expertise : Learning when everybody knows a bit of something. In *ICML*, volume 9, pp. 932–939, 2010.

Zhou, Dengyong, Platt, John, Basu, Sumit, and Mao, Yi. Learning from the Wisdom of Crowds by Minimax Entropy. In *NIPS*, 2012.

Supplementary materials: Error Rate Analysis of Labeling by Crowdsourcing

1 Proofs of main results

Since the proof of Theorem 1 requires other theorems in this paper, we will not present the proofs in the same order as in the paper. The oder of our proofs will be: Theorem 7 , Theorem 6 , Theorem 1 and then Theorem 2 .

1.1 Proof of Theorem 7 (bounding the mean error rate of individual item)

Proof. For simplicity of notation, let's denote

$$\lambda_{j,min} = \left[\left(\sum_{i=1}^{M} q_{ij}\nu_i(2p_i^+ - 1) + a\right) \wedge \left(\sum_{i=1}^{M} q_{ij}\nu_i(2p_i^- - 1) - a\right)\right], \quad (1)$$

$$\lambda_{j,max} = \left[\left(\sum_{i=1}^{M} q_{ij}\nu_i(2p_i^+ - 1) + a\right) \vee \left(\sum_{i=1}^{M} q_{ij}\nu_i(2p_i^- - 1) - a\right)\right]. \quad (2)$$

So $\tau_{j,\min} = \frac{\lambda_{j,min}}{||\nu||}$ and $\tau_{j,\max} = \frac{\lambda_{j,max}}{||\nu||}$.

And let's also define

$$\begin{aligned} \mathbb{P}_+(Z_{ij}) &= \mathbb{P}(Z_{ij}|y_j = 1) \quad \text{and} \quad \mathbb{E}_+[Z_{ij}] = \mathbb{E}[Z_{ij}|y_j = 1], \\ \mathbb{P}_-(Z_{ij}) &= \mathbb{P}(Z_{ij}|y_j = -1) \quad \text{and} \quad \mathbb{E}_-[Z_{ij}] = \mathbb{E}[Z_{ij}|y_j = -1], \end{aligned}$$

and let $A = \left(H_e(\epsilon) + \frac{1}{N}\ln\frac{1}{\delta}\right)$.

Note that

$$\begin{cases} \mathbb{P}_+(Z_{ij} = 1) & = q_{ij}p_i^+ \\ \mathbb{P}_+(Z_{ij} = -1) & = q_{ij}(1 - p_i^+) \\ \mathbb{P}_+(Z_{ij} = 0) & = 1 - q_{ij} \end{cases} \implies \mathbb{E}_+[Z_{ij}] = q_{ij}(2p_i^+ - 1). \quad (3)$$

$$\begin{cases} \mathbb{P}_-(Z_{ij} = 1) & = q_{ij}(1 - p_i^-) \\ \mathbb{P}_-(Z_{ij} = -1) & = q_{ij}p_i^- \\ \mathbb{P}_-(Z_{ij} = 0) & = 1 - q_{ij} \end{cases} \implies \mathbb{E}_-[Z_{ij}] = -q_{ij}(2p_i^- - 1). \quad (4)$$

The accuracy of the weighted majority algorithm on item j is

$$\begin{aligned} \theta_j &= \mathbb{P}(\text{The WMA label the item } j \text{ correctly}) = \mathbb{P}(\hat{y}_j = y_j) \\ &= \mathbb{P}(y_j = 1)\mathbb{P}(\hat{y}_j = +1|y_j = 1) + \mathbb{P}(y_j = -1)\mathbb{P}(\hat{y}_j = -1|y_j = -1) \\ &= \pi\mathbb{P}_+\left(\sum_{i=1}^{M}\nu_i Z_{ij} + a > 0\right) + (1-\pi)\mathbb{P}_-\left(\sum_{i=1}^{M}\nu_i Z_{ij} + a < 0\right) \\ &= \pi\theta_j^+ + (1-\pi)\theta_j^-. \end{aligned}$$

where $\theta_j^+ = \mathbb{P}_+\left(\sum_{i=1}^{M}\nu_i Z_{ij} + a > 0\right)$ and $\theta_j^- = \mathbb{P}_-\left(\sum_{i=1}^{M}\nu_i Z_{ij} + a < 0\right)$

Let's define $\xi_{ij} = \nu_i Z_{ij}$, then $\xi_{ij} \in [-|\nu_i|, +|\nu_i|]$.
Let

$$r_j = \mathbb{E}_+\left[\sum_{i=1}^{M}\xi_{ij}\right] = \sum_{i=1}^{M}\mathbb{E}_+\left[\xi_{ij}\right] = \sum_{i=1}^{M} q_{ij}\nu_i(2p_i^+ - 1), \tag{5}$$

$$s_j = -\mathbb{E}_-\left[\sum_{i=1}^{M}\xi_{ij}\right] = -\sum_{i=1}^{M}\mathbb{E}_-\left[\xi_{ij}\right] = \sum_{i=1}^{M} q_{ij}\nu_i(2p_i^- - 1). \tag{6}$$

Then apparently, we have

$$\lambda_{j,min} = \{(r_j + a) \wedge (s_j - a)\} \quad \text{and} \quad \lambda_{j,max} = \{(r_j + a) \vee (s_j - a)\}.$$

Proof of Theorem 7 (1): Assume that $\tau_{j,\min} \geq 0$, which implies $\lambda_{j,min} \geq 0$, then we will have

$$r_j + a \geq \lambda_{j,min} \geq 0 \qquad \text{and} \qquad s_j - a \geq \lambda_{j,min} \geq 0. \tag{7}$$

Using Hoeffding inequality, we have

$$\begin{aligned}
\theta_j^+ = \mathbb{P}_+\left(\sum_{i=1}^{M}\xi_{ij} + a > 0\right) &= \mathbb{P}_+\left(\sum_{i=1}^{M}\xi_{ij} - \mathbb{E}_+\left[\sum_{i=1}^{M}\xi_{ij}\right] > -(r_j + a)\right) \\
&\geq 1 - \exp\left(-\frac{(r_j + a)^2}{2||\nu||^2}\right) \qquad \text{(by Hoeffding ineq.)} \\
&\geq 1 - \exp\left(\frac{-\lambda_{j,min}^2}{2||\nu||^2}\right). \qquad \text{(because (7))} \qquad (8) \\
\theta_j^- = \mathbb{P}_-\left(\sum_{i=1}^{M}\xi_{ij} + a < 0\right) &= \mathbb{P}_-\left(\sum_{i=1}^{M}\xi_{ij} - \mathbb{E}_-\left[\sum_{i=1}^{M}\xi_{ij}\right] < s_j - a\right) \\
&\geq 1 - \exp\left(-\frac{(s_j - a)^2}{2||\nu||^2}\right) \qquad \text{(by Hoeffding ineq.)} \\
&\geq 1 - \exp\left(\frac{-\lambda_{j,min}^2}{2||\nu||^2}\right). \qquad \text{(because (7))} \qquad (9)
\end{aligned}$$

then $\theta_j = \pi\theta_j^+ + (1-\pi)\theta_j^- \geq 1 - \exp\left(\frac{-\lambda_{j,min}^2}{2||\nu||^2}\right)$, which leads to,

$$\mathbb{P}(\hat{y}_j \neq y_j) = 1 - \theta_j \leq \exp\left(-\frac{\tau_{j,\min}^2}{2}\right), \forall j \in [N].$$

Thus we have proved Theorem 7 (1).

Proof of Theorem 7 (2): Assume that $\tau_{j,\max} \leq 0$,then we have $(r_j + a) \leq \lambda_{j,max} \leq 0$ and $(s_j - a) \leq \lambda_{j,max} \leq 0$. Therefore

$$\exp\left(\frac{-(r_j + a)^2}{2||\nu||^2}\right) \leq \exp\left(\frac{-\lambda_{j,max}^2}{2||\nu||^2}\right), \tag{10}$$

$$\exp\left(-\frac{(s_j - a)^2}{2||\nu||^2}\right) \leq \exp\left(\frac{-\lambda_{j,max}^2}{2||\nu||^2}\right), \tag{11}$$

then

$$\begin{aligned}
\theta_j^+ &= \mathbb{P}(\sum_{i=1}^{M} \xi_{ij} + a > 0) \\
&= \mathbb{P}\left(\sum_{i=1}^{M} \xi_{ij} - \mathbb{E}_+\left[\sum_{i=1}^{M} \xi_{ij}\right] > -(r_j + a)\right) \qquad (\text{because } -(r_j+a) > 0) \\
&\leq \exp\left(-\frac{(r_j+a)^2}{2||\nu||^2}\right) \qquad (\text{by Hoeffding Ineq. }) \\
&\leq \exp\left(\frac{-\lambda_{j,max}^2}{2||\nu||^2}\right). \quad (\text{because (10)}) \qquad (12)
\end{aligned}$$

$$\begin{aligned}
\theta_j^- &= \mathbb{P}(\sum_{i=1}^{M} \xi_{ij} - \mathbb{E}_-\left[\sum_{i=1}^{M} \xi_{ij}\right] < s_j - a) \qquad (\text{because } s_j - a > 0) \\
&\leq \exp\left(-\frac{(s_j-a)^2}{2||\nu||^2}\right) \qquad (\text{by Hoeffding Ineq. }) \\
&\leq \exp\left(\frac{-\lambda_{j,max}^2}{2||\nu||^2}\right). \qquad (\text{because (11)}) \qquad (13)
\end{aligned}$$

It follows that

$$\theta_j = \pi\theta_j^+ + (1-\pi)\theta_j^- \leq \exp\left(\frac{-\lambda_{j,max}^2}{2||\nu||^2}\right),$$

which implies,

$$\mathbb{P}(\hat{y}_j \neq y_j) = 1 - \theta_j \geq 1 - \exp\left(-\frac{\tau_{j,\max}^2}{2}\right) \qquad , \forall j \in [N].$$

Thus, we have proved Theorem 7 (2).

□

1.2 Proof of Theorem 6 (Bounding the mean error rate):

Proof. In this proof, we will follow the same notations in the proof of Theorem 7 .

For simplicity of the notation, let's define:

$$t_* = \min_{j\in[N]}\left[\left(\sum_{i=1}^{M} q_{ij}\nu_i(2p_i^+ - 1) + a\right) \wedge \left(\sum_{i=1}^{M} q_{ij}\nu_i(2p_i^- - 1) - a\right)\right] = \min_{j\in[N]} \lambda_{j,min}, \qquad (14)$$

$$t^\star = \max_{j\in[N]}\left[\left(\sum_{i=1}^{M} q_{ij}\nu_i(2p_i^+ - 1) + a\right) \vee \left(\sum_{i=1}^{M} q_{ij}\nu_i(2p_i^- - 1) - a\right)\right] = \max_{j\in[N]} \lambda_{j,max}. \qquad (15)$$

So, $t_1 = \frac{t_*}{||\nu||} \leq \frac{\lambda_{j,min}}{||\nu||}$ and $t_2 = \frac{t^\star}{||\nu||} \geq \frac{\lambda_{j,max}}{||\nu||}$, and the definition of $\lambda_{j,min}$ and $\lambda_{j,max}$ are (1) and (2) respectively.

From Theorem 7 , we know that if $t_1 \geq 0$, then $\forall j \in [N]$, we have $\lambda_{j,min} \geq 0$,

$$\begin{aligned}
\theta_j &= 1 - \mathbb{P}(\hat{y}_j \neq y_j) \\
&\geq 1 - \exp\left(\frac{-\lambda_{j,min}^2}{2||\nu||^2}\right) \\
&\geq 1 - \exp\left(-\frac{t_*^2}{2||\nu||^2}\right),
\end{aligned}$$

hence

$$\bar{\theta} = \frac{1}{N}\sum_{j=1}^{N}\theta_j \geq 1 - \exp\left(-\frac{t_*^2}{2||\nu||^2}\right) = 1 - e^{-\frac{t_1^2}{2}}. \tag{16}$$

and note that

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) = \frac{1}{N}\sum_{j=1}^{N}\left[1 - \mathbb{P}(\hat{y}_j = y_j)\right] = 1 - \bar{\theta} \leq e^{-\frac{t_1^2}{2}}.$$

Thus, we proved (1).

Similarly, we can easily get (2) proved. □

1.3 Proof of Theorem 1

So far, we have bounded the mean error rate, but we still need more tools for bounding the error rate in practical case with high probability. The following lemma is another form of Bernstein-Chernoff-Hoeffding theorem [5].

Lemma 1. ***(Bernstein-Chernoff-Hoeffding)*** *Let $\xi_i \in [0,1]$ be independent random variables where $\mathbb{E}\xi_i = p_i, i \in [n]$. Let $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ and $\bar{p} = \frac{1}{n}\sum_{i=1}^{n}p_i$. Then,*

(1) for any m such that $\bar{p} \leq \frac{m}{n} < 1$, $\mathbb{P}\left(\bar{\xi} > m\right) \leq e^{-nD(m/n||\bar{p})}$,

(2) for any m such that $0 < \frac{m}{n} \leq \bar{p}$, $\mathbb{P}\left(\bar{\xi} \leq m\right) \leq e^{-nD(m/n||\bar{p})}$.

We are going to present the proof of the Theorem 1 as following:

Proof. (Theorem 1)

Proof of Theorem 1 (1)

Let $\mu = \frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \leq e^{-t_1^2/2}$. By Theorem 6 .(1), we obtained that $\mu \leq e^{-t_1^2/2} = \phi(t_1)$. Assume that $t_1 \geq \sqrt{2\ln\frac{1}{\epsilon}}$, then we can get $\phi(t_1) \leq \epsilon$, which gives us $0 \leq \mu \leq \phi(t_1) \leq \epsilon$. Then by Bernstein-Chernoff-Hoeffding Theorem, i.e. Lemma 1, we can get $\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N}\mathrm{I}\left(\hat{y}_j \neq y_j\right) > \epsilon\right) \leq e^{-N\mathrm{D}(\epsilon||\mu)} \leq e^{-N\mathrm{D}(\epsilon||\phi(t_1)}$. Therefore, we have

$$\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N}\mathrm{I}\left(\hat{y}_j \neq y_j\right) \leq \epsilon\right) \geq 1 - e^{-N\mathrm{D}(\epsilon||\phi(t_1))},$$

.

Proof of Theorem 1 (2) With the same argument as above, assuming $t_2 \leq -\sqrt{2\ln\frac{1}{1-\epsilon}}$, then $1 \geq \mu \geq 1 - \phi(t_2) \geq \epsilon$, which gives us

$$\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N}\mathrm{I}\left(\hat{y}_j \neq y_j\right) \leq \epsilon\right) \leq e^{-N\mathrm{D}(\epsilon||1-\phi(t_2))}.$$

Thus, we have proved Theorem 1 .

□

1.4 Proof of Theorem 2

Before proving the Theorem 2 , we first prove an important lemma.

The following lemma is to bound the average of a group of independent Bernoulli random variables. Its proof will rely on Hoeffding bounds and Bernstein-Chernoff-Hoeffding theorem [5].

Lemma 2. *Suppose $\forall j \in [N]$, $\xi_j \sim$ Bernoulli(p_j) with $p_j \in (0,1)$, and ξ_j's are independent of each other. Let $\bar{\xi} = \frac{1}{N}\sum_{j=1}^{N}\xi_j$ and $\bar{p} = \mathbb{E}\bar{\xi} = \frac{1}{N}\sum_{j=1}^{N}p_j$ Given any $\epsilon, \delta \in (0,1)$:*

(1) If

$$0 < \bar{p} \leq \frac{1}{1+\exp\left(\frac{1}{\epsilon}\left[H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right]\right)}, \tag{17}$$

then

$$\mathbb{P}(\bar{\xi} \leq \epsilon) \geq 1-\delta.$$

(2) If

$$\frac{1}{1+\exp\left(-\frac{1}{1-\epsilon}\left(H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right)\right)} \leq \bar{p} < 1, \tag{18}$$

then

$$\mathbb{P}(\bar{\xi} \leq \epsilon) < \delta.$$

Proof. For simplicity, let's define $A = \left(H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right)$.

Proof of Lemma 2.(1): We will finish the proof in several steps.

We assume $0 < \bar{p} \leq \frac{1}{1+\exp\left(\frac{1}{\epsilon}\left[H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right]\right)}$.

Step 1. we want to show $\bar{p} < \epsilon$:

$$\begin{aligned}
\exp\left(\frac{A}{\epsilon}\right) &= \exp\left(\frac{\epsilon\ln\frac{1}{\epsilon}+(1-\epsilon)\ln\frac{1}{1-\epsilon}+\frac{1}{N}\ln\frac{1}{\delta}}{\epsilon}\right) \\
&= \exp\left(\ln\frac{1}{\epsilon}+\frac{1-\epsilon}{\epsilon}\ln\frac{1}{1-\epsilon}+\frac{1}{N\epsilon}\ln\frac{1}{\delta}\right) \\
&> \exp\left(\ln\frac{1}{\epsilon}\right) \qquad (\text{because } \epsilon,\delta \in (0,1), N>0) \\
&= \frac{1}{\epsilon},
\end{aligned}$$

$$\implies \quad 1+\exp\left(\tfrac{A}{\epsilon}\right) > \tfrac{1}{\epsilon} \quad \implies \quad \tfrac{1}{1+\exp\left(\frac{A}{\epsilon}\right)} < \epsilon.$$

Since $0 < \bar{p} \leq \frac{1}{1+\exp(A/\epsilon)}$, then we will have $\bar{p} < \epsilon$.

Step 2. We want to show $\mathbb{P}\left(\bar{\xi} \leq \epsilon\right) \geq 1 - e^{-N\cdot D(\epsilon||\bar{p})}$:
This is obtained by Bernstein-Chernoff-Hoeffding inequality ([4, 5]), which leads to:

$$\text{If } 0 < \bar{p} \leq \epsilon, \text{ then } \mathbb{P}(\bar{\xi} \leq \epsilon) \leq 1 - e^{-ND(\epsilon||\bar{p})}, \tag{19}$$

$$\text{If } \epsilon \leq \bar{p} < 1, \text{ then } \mathbb{P}(\bar{\xi} \leq \epsilon) \leq e^{-ND(\epsilon||\bar{p})}\ . \tag{20}$$

Since we have shown in step 1 that $\bar{p} < \epsilon$, we can prove step 2 by directly applying Bernstein-Chernoff-Hoeffding Theorem.

Step 3. We want to show $e^{-ND(\epsilon||\bar{p})} \leq \delta$.
Note

$$\begin{aligned}
& e^{-ND(\epsilon||\bar{p})} \leq \delta \\
\iff \quad & D(\epsilon||\bar{p}) \geq \frac{1}{N}\ln\frac{1}{\delta} \\
\iff \quad & \ln\frac{\epsilon^\epsilon(1-\epsilon)^{1-\epsilon}}{\bar{p}^\epsilon(1-\bar{p})^{1-\epsilon}} \geq \ln\left(\frac{1}{\delta}\right)^{\frac{1}{N}} \\
\iff \quad & \bar{p}^\epsilon(1-\bar{p})^{1-\epsilon} \leq \exp\left(-\left(H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right)\right) = e^{-A}
\end{aligned} \tag{21}$$

From the condition we have ,

$$\bar{p} \leq \frac{1}{1+\exp(A/\epsilon)} \qquad \Longrightarrow \qquad \left(\frac{\bar{p}}{1-\bar{p}}\right)^{\epsilon} \leq e^{-A}.$$

Note that

$$\bar{p}^{\epsilon}(1-\bar{p})^{1-\epsilon} = \left(\frac{\bar{p}}{1-\bar{p}}\right)^{\epsilon}(1-\bar{p}) < \left(\frac{\bar{p}}{1-\bar{p}}\right)^{\epsilon}. \qquad \text{(because } 1-\bar{p}<1\text{)}$$

$\Longrightarrow$ Inequality (21) holds: $\bar{p}^{\epsilon}(1-\bar{p})^{1-\epsilon} \leq e^{-A}$ $\Longrightarrow$ $e^{-ND(\epsilon||\bar{p})} \leq \delta$.
By step 2 and step 3, we can easily get that if $\bar{p} \leq \frac{1}{1+e^{A/\epsilon}}$, then $\mathbb{P}(\bar{\xi} \leq \epsilon) \geq 1-\delta$, which is the result we want.

Proof of Lemma 2.(2): We will also carry out the proof in several steps:
Assuming that

$$\frac{1}{1+\exp\left(-\frac{1}{1-\epsilon}\left(H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right)\right)} \leq \bar{p} < 1.$$

Step 1. We want to show $\bar{p} > \epsilon$.
The proof of step 1 is as follows :

$$\begin{aligned}
&\frac{1}{1+\exp\left(-\frac{1}{1-\epsilon}\left(H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right)\right)} > \epsilon \\
\Longleftrightarrow\quad & 1+\exp\left(-\frac{1}{1-\epsilon}\left(H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right)\right) < \frac{1}{\epsilon} \\
\Longleftrightarrow\quad & \left(H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right) > (1-\epsilon)\ln\frac{\epsilon}{1-\epsilon} \\
\Longleftrightarrow\quad & \epsilon\ln\frac{1}{\epsilon}+(1-\epsilon)\ln\frac{1}{1-\epsilon}+\frac{1}{N}\ln\frac{1}{\delta} > (1-\epsilon)\ln\frac{1}{1-\epsilon}-(1-\epsilon)\ln\frac{1}{\epsilon} \\
\Longleftrightarrow\quad & \ln\frac{1}{\epsilon}+\frac{1}{N}\ln\frac{1}{\delta} > 0
\end{aligned}$$

which is true since $\epsilon, \delta \in (0,1)$. Therefore, we have proved $\bar{p} > \epsilon$.

Step 2. We want to show $\mathbb{P}(\bar{\xi} \leq \epsilon) \leq e^{-ND(\epsilon||\bar{p})}$.
By Bernstein-Chernoff-Hoeffding Theorem, since $\epsilon < \bar{p} = \mathbb{E}\bar{\xi}$ and $\xi_j \sim$ Bernoulli(p_j) independently, we can get step 2 proved.

Step 3. we want to show $e^{-ND(\epsilon||\bar{p})} < \delta$
Note that

$$\begin{aligned}
& e^{-ND(\epsilon||\bar{p})} < \delta \\
\Longleftrightarrow\quad & D(\epsilon||\bar{p}) > \frac{1}{N}\ln\frac{1}{\delta} \\
\Longleftrightarrow\quad & \bar{p}^{\epsilon}(1-\bar{p})^{1-\epsilon} < \exp\left(-\left(H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right)\right) = e^{-A} \qquad (22)
\end{aligned}$$

From the conditions we have

$$\begin{aligned}
& \bar{p} \geq \frac{1}{1+\exp-\frac{A}{1-\epsilon}} \\
\Longrightarrow\quad & \frac{1-\bar{p}}{\bar{p}} \leq e^{-\frac{A}{1-\epsilon}} \\
\Longrightarrow\quad & \left(\frac{1-\bar{p}}{\bar{p}}\right)^{1-\epsilon} \leq \exp\left(-\left(H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right)\right) \qquad (23)
\end{aligned}$$

And note that

$$\bar{p}^{\epsilon}(1-\bar{p})^{1-\epsilon} = \left(\frac{1-\bar{p}}{\bar{p}}\right)^{1-\epsilon} \cdot \bar{p} < \left(\frac{1-\bar{p}}{\bar{p}}\right)^{1-\epsilon} \qquad \text{(because } \bar{p} < 1\text{)} \tag{24}$$

By combining inequalities (23) and (24), we can get inequality (22) proved.
Thus we obtain $e^{-ND(\epsilon||\bar{p})} < \delta$

Finally, by step 2 and 3, we can get : if $\bar{p} \geq \frac{1}{1+\exp\left(-\frac{1}{1-\epsilon}\left(H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right)\right)}$, then $\mathbb{P}\left(\bar{\xi} \leq \epsilon\right) < \delta$.

□

Now, we are going to prove Theorem 2 .

Proof. In this proof, we will follow the notations from the proof of theorem 6 . So we should keep in mind the definitions such as $\theta_j \doteq \mathbb{P}(\hat{y}_j = y_j)\ \forall j \in [N]$ and $\bar{\theta} = \frac{1}{N}\theta_j$.
Let $\zeta_j = \mathrm{I}\left(\hat{y}_j \neq y_j\right) \sim \text{Bernoulli}(\theta_j)$, then let $\bar{p} = \mathbb{E}\bar{\zeta} = \frac{1}{N}\mathbb{E}\zeta_j = 1 - \bar{\theta}$.

Proof of Theorem 2 (1):

Assume that $\frac{t_*}{||\nu||} \geq \sqrt{2\ln C}$, then $t_* \geq 0$.
By theorem 6 , we can get

$$\begin{aligned} \bar{\theta} &= 1 - \frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \\ &\geq 1 - \exp\left(-\frac{t_*^2}{2||\nu||^2}\right). \end{aligned} \tag{25}$$

Since

$$\begin{aligned} & \frac{t_*}{||\nu||} \geq \sqrt{2\ln C} = \sqrt{2\ln(1+\exp(A/\epsilon))} \\ \Longrightarrow \quad & \exp\left(-\frac{t_*^2}{2||\nu||^2}\right) \leq \frac{1}{1+\exp\left(\frac{A}{\epsilon}\right)} \\ \Longrightarrow \quad & \bar{\theta} \geq 1 - \exp\left(-\frac{t_*^2}{2||\nu||^2}\right) \geq 1 - \frac{1}{1+\exp\left(\frac{A}{\epsilon}\right)} \qquad \text{(because (25))} \\ \Longrightarrow \quad & 1 - \bar{\theta} \leq \frac{1}{1+\exp\left(\frac{A}{\epsilon}\right)}. \end{aligned} \tag{26}$$

where $A = \left(H_e(\epsilon) + \frac{1}{N}\ln\frac{1}{\delta}\right)$.
By inequality (26) and by Lemma 2, we can get

$$\mathbb{P}(\bar{\zeta} \leq \epsilon) \geq 1 - \delta,$$

which implies,

$$\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N}\mathrm{I}\left(\hat{y}_j \neq y_j\right) \leq \epsilon\right) \geq 1 - \delta.$$

Therefore, we have proved Theorem 2 .(1).

Proof of Theorem 2 (2):

Assuming that

$$\frac{t^\star}{||\nu||} \leq -\sqrt{2\ln G} \leq 0,$$

where $G = 1 + \exp\left(\frac{1}{1-\epsilon}\left(H_e(\epsilon) + \frac{1}{N}\ln\frac{1}{\delta}\right)\right)$. Then by theorem 6 , we have

$$\bar{\theta} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{P}(\hat{y}_j = y_j) \leq \exp\left(-\frac{t^{\star 2}}{2||\nu||^2}\right). \tag{27}$$

From the conditions in (2)

$$\begin{aligned}
& \frac{t^\star}{||\nu||} \leq -\sqrt{2\ln\left(1+\exp\left(\frac{1}{1-\epsilon}\left[H_e(\epsilon)+\frac{1}{N}\ln\frac{1}{\delta}\right]\right)\right)}, \\
\Longrightarrow \quad & \exp\left(-\frac{t^{\star 2}}{2||\nu||^2}\right) \leq \frac{1}{1+\exp\left(\frac{A}{1-\epsilon}\right)}, \\
\Longrightarrow \quad & 1-\bar{\theta} \geq 1-\exp\left(-\frac{t^{\star 2}}{2||\nu||^2}\right) \geq 1-\frac{1}{1+\exp\left(\frac{A}{1-\epsilon}\right)} = \frac{1}{1+\exp\left(-\frac{A}{1-\epsilon}\right)}.
\end{aligned}$$

By Lemma 2.(2), we can have

$$\mathbb{P}(\bar{\zeta} \leq \epsilon) < \delta,$$

which leads to

$$\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N}\mathrm{I}\left(\hat{y}_j \neq y_j\right) \leq \epsilon\right) < \delta.$$

□

2 Proof of Theorem 10

Remark: since the oracle MAP rule is not a hyperplane rule for prediction, we cannot directly apply the Theorem 6 to derive relative results.

Proof. For simplicity of the notation, let's define:

$$t_* = \min_{j\in[N]}\left\{\left(\sum_{i=1}^{M} q_{ij}\left[(u_i-\nu_i)p_i^+ + \nu_i\right] + a\right) \wedge \left(\sum_{i=1}^{M} q_{ij}\left[(u_i-\nu_i)p_i^- - u_i\right] - a\right)\right\}, \tag{28}$$

$$t^\star = \max_{j\in[N]}\left\{\left(\sum_{i=1}^{M} q_{ij}\left[(u_i-\nu_i)p_i^+ + \nu_i\right] + a\right) \vee \left(\sum_{i=1}^{M} q_{ij}\left[(u_i-\nu_i)p_i^- - u_i\right] - a\right)\right\}. \tag{29}$$

So, $t_1 = \frac{t_*}{\Gamma}$ and $t_2 = \frac{t^\star}{\Gamma}$.

We know that

$$\rho_j = \frac{\pi\eta_j^+}{\pi\eta_j^+ + (1-\pi)\eta_j^-},$$

where

$$\begin{aligned}
\eta_j^+ &= \mathbb{P}(Z_{\star j}|y_j = 1) = \prod_{i=1}^{M}\left(p_i^+\right)^{\mathrm{I}(Z_{ij}=1)}\left(1-p_i^+\right)^{\mathrm{I}(Z_{ij}=-1)}, \\
\eta_j^- &= \mathbb{P}(Z_{\star j}|y_j = -1) = \prod_{i=1}^{M}\left(1-p_i^-\right)^{\mathrm{I}(Z_{ij}=1)}\left(p_i^-\right)^{\mathrm{I}(Z_{ij}=-1)}.
\end{aligned}$$

For the oracle MAP rule,

$$
\begin{aligned}
\hat{y}_j^{\text{oracle}} = +1 \quad &\Longleftrightarrow \quad \rho_j > \frac{1}{2} \Longleftrightarrow \frac{\pi\eta_j^+}{(1-\pi)\eta_j^-} > 1 \\
&\Longleftrightarrow \quad \ln\frac{\pi}{1-\pi} + \sum_{i=1}^{M}\left[\mathrm{I}\left(Z_{ij}=1\right)\ln\frac{p_i^+}{1-p_i^-} + \mathrm{I}\left(Z_{ij}=-1\right)\ln\frac{1-p_i^+}{p_i^-}\right] > 0 \\
&\Longleftrightarrow \quad a + \sum_{i=1}^{M}\left[u_i\mathrm{I}\left(Z_{ij}=1\right) + \nu_i\mathrm{I}\left(Z_{ij}=-1\right)\right] > 0 \\
&\Longleftrightarrow \quad a + \sum_{i=1}^{M}\xi_{ij} > 0,
\end{aligned}
$$

where $\xi_{ij} = u_i\mathrm{I}\left(Z_{ij}=1\right) + \nu_i\mathrm{I}\left(Z_{ij}=-1\right)$.

Let $a_i = \min\left\{u_i, \nu_i, 0\right\}$ and $b_i = \max\left\{u_i, \nu_i, 0\right\}$. then apparently, $\xi_{ij} \in [a_i, b_i] \forall i \in [M]$. Then from the theorem statement, we have $\Gamma = \sqrt{\sum_{i=1}^{M}(b_i - a_i)^2}$.

For any event Ω_j with respect to the j-th item, let's define the following for simplicity of the notation:

$$
\begin{aligned}
\mathbb{P}_+\left(\Omega_j\right) &= \mathbb{P}(\Omega_j | y_j = 1), \qquad \mathbb{E}_+\left[\Omega_j\right] = \mathbb{E}[\Omega_j | y_j = 1], \\
\mathbb{P}_-\left(\Omega_j\right) &= \mathbb{P}(\Omega_j | y_j = -1), \qquad \mathbb{E}_-\left[\Omega_j\right] = \mathbb{E}[\Omega_j | y_j = -1].
\end{aligned}
$$

So, we have

$$
\begin{aligned}
\mathbb{E}_+\left[\xi_{ij}\right] &= u_i\mathbb{P}(Z_{ij}=1|y_j=1) + \nu_i\mathbb{P}(Z_{ij}=-1|y_j=-1) \\
&= q_{ij}[u_i p_i^+ + \nu_i(1-p_i^+)] \\
&= q_{ij}[(u_i-\nu_i)p_i^+ + \nu_i], \qquad (30) \\
\mathbb{E}_-\left[\xi_{ij}\right] &= q_{ij}[u_i(1-p_i^-) + \nu_i p_i^-] = -q_{ij}[(u_i-\nu_i)p_i^- - u_i]. \qquad (31)
\end{aligned}
$$

And we define

$$
\begin{aligned}
r_j &\doteq \mathbb{E}_+\left[\sum_{i=1}^{M}\xi_{ij}\right] = \sum_{i=1}^{M} q_{ij}\left[(u_i-\nu_i)p_i^+ + \nu_i\right], \qquad (32) \\
s_j &\doteq \mathbb{E}_-\left[-\sum_{i=1}^{M}\xi_{ij}\right] = \sum_{i=1}^{M} q_{ij}\left[(u_i-\nu_i)p_i^- - u_i\right]. \qquad (33)
\end{aligned}
$$

Then from the definition of t_* and $t^\star$, we have

$$
\begin{aligned}
t_* &= \min_{j\in[N]}\left\{(r_j+a)\wedge(s_j-a)\right\} \quad \text{and} \quad r_j + a \geq t_*,\ s_j - a \geq t_*,\ \forall j \in [N], \\
t^\star &= \max_{j\in[N]}\left\{(r_j+a)\vee(s_j-a)\right\} \quad \text{and} \quad r_j + a \leq t^\star,\ s_j - a \leq t^\star,\ \forall j \in [N].
\end{aligned}
$$

Let's define the probability of labeling item j correctly as

$$
\theta_j \doteq \mathbb{P}(\hat{y}_j = y_j) = \pi\mathbb{P}_+\left(\hat{y}_j = +1\right) + (1-\pi)\mathbb{P}_-\left(\hat{y}_j = -1\right) = \pi\theta_j^+ + (1-\pi)\theta_j^-, \qquad (34)
$$

where

$$
\begin{aligned}
\theta_j^+ &= \mathbb{P}_+\left(\hat{y}_j = +1\right) = \mathbb{P}_+\left(a + \sum_{i=1}^{M}\xi_{ij} > 0\right) = \mathbb{P}_+\left(\sum_{i=1}^{M}\xi_{ij} - \mathbb{E}_+\left[\sum_{i=1}^{M}\xi_{ij}\right] > -(r_j+a)\right), \qquad (35) \\
\theta_j^- &= \mathbb{P}_-\left(\hat{y}_j = -1\right) = \mathbb{P}_-\left(a + \sum_{i=1}^{M}\xi_{ij} < 0\right) = \mathbb{P}_-\left(\sum_{i=1}^{M}\xi_{ij} - \mathbb{E}_-\left[\sum_{i=1}^{M}\xi_{ij}\right] < s_j - a\right). \qquad (36)
\end{aligned}
$$

Proof of Theorem 10 (1):

Since $t_1 \geq 0$, which implies $\frac{1}{\Gamma}(t_*) \geq 0$, then $r_j + a \geq t_* \geq 0$, $s_j - a \geq t_* \geq 0$, $\frac{1}{\Gamma}(r_j + a) \geq \frac{t_*}{\Gamma} \geq 0$ and $\frac{1}{\Gamma}(s_j - a) \geq \frac{t_*}{\Gamma} \geq 0$, thus we have

$$\exp\left(-\frac{2(r_j+a)^2}{\Gamma^2}\right) \leq \exp\left(-\frac{2t_*^2}{\Gamma^2}\right) \quad \text{and} \quad \exp\left(-\frac{2(s_j-a)^2}{\Gamma^2}\right) \leq \exp\left(-\frac{2t_*^2}{\Gamma^2}\right). \tag{37}$$

Note that since $r_j + a \geq 0$ and $s_j - a \geq 0$. $\xi_{ij} \in [a_i, b_i]$ $\forall i \in [M]$, and ξ_{kj} is independent of ξ_{lj} if $k \neq l$. Then by Hoeffding inequality:

$$\begin{aligned}
\theta_j^+ &= \mathbb{P}_+\left(\sum_{i=1}^M \xi_{ij} - \mathbb{E}_+\left[\sum_{i=1}^M \xi_{ij}\right] > -(r_j+a)\right) \\
&\geq 1 - \exp\left(-\frac{2(r_j+a)^2}{\Gamma^2}\right) \geq 1 - \exp\left(-\frac{2t_*^2}{\Gamma^2}\right). \qquad (38) \\
\theta_j^- &= \mathbb{P}_-\left(\sum_{i=1}^M \xi_{ij} - \mathbb{E}_-\left[\sum_{i=1}^M \xi_{ij}\right] < s_j - a\right) \\
&\geq 1 - \exp\left(-\frac{2(s_j-a)^2}{\Gamma^2}\right) \geq 1 - \exp\left(-\frac{2t_*^2}{\Gamma^2}\right). \qquad (39) \\
\implies \theta_j &= \pi\theta_j^+ + (1-\pi)\theta_j^- \;\geq\; 1 - \exp\left(-\frac{2t_*^2}{\Gamma^2}\right) \qquad (40) \\
\implies \bar{\theta} &= \frac{1}{N}\sum_{j=1}^N \theta_j \quad \geq \quad 1 - \exp\left(-\frac{2t_*^2}{\Gamma^2}\right). \qquad (41)
\end{aligned}$$

Since $\frac{1}{N}\sum_{j=1}^N \mathbb{P}\left(\hat{y}_j \neq y_j\right) = 1 - \frac{1}{N}\sum_{j=1}^N \mathbb{P}(\hat{y}_j = y_j) = 1 - \bar{\theta}$, we have

$$\frac{1}{N}\sum_{j=1}^N \mathbb{P}\left(\hat{y}_j \neq y_j\right) \leq \exp\left(-\frac{2t_*^2}{\Gamma^2}\right) = e^{-2t_1^2}.$$

Therefore, we have proved Theorem 10 .(1).

Proof of Theorem 10 (2):

Since $t_2 \leq 0$, which implies $\frac{t^\star}{\Gamma} \leq 0$, then $r_j \leq t^\star \leq 0$, $s_j \leq t^\star \leq 0$, $\frac{1}{\Gamma}(r_j + a) \leq \frac{t^\star}{\Gamma} \leq 0$ and $\frac{1}{\Gamma}(s_j - a) \leq \frac{t^\star}{\Gamma} \leq 0$, thus we have

$$\exp\left(-\frac{2(r_j+a)^2}{\Gamma^2}\right) \leq \exp\left(-\frac{2t^{\star 2}}{\Gamma^2}\right) \quad \text{and} \quad \exp\left(-\frac{2(s_j-a)^2}{\Gamma^2}\right) \leq \exp\left(-\frac{2t^{\star 2}}{\Gamma^2}\right). \tag{42}$$

Note that since $-(r_j + a) \geq 0$ and $s_j - a \leq 0$. $\xi_{ij} \in [a_i, b_i]$ $\forall i \in [M]$, and ξ_{kj} is independent of ξ_{lj} if $k \neq l$.

Then by Hoeffding inequality:

$$\begin{aligned}
\theta_j^+ &= \mathbb{P}_+\left(\sum_{i=1}^{M}\xi_{ij} - \mathbb{E}_+\left[\sum_{i=1}^{M}\xi_{ij}\right] > -(r_j+a)\right) \\
&\leq \exp\left(-\frac{2(r_j+a)^2}{\Gamma^2}\right) \leq \exp\left(-\frac{2t^{\star 2}}{\Gamma^2}\right), \qquad (43) \\
\theta_j^- &= \mathbb{P}_-\left(\sum_{i=1}^{M}\xi_{ij} - \mathbb{E}_-\left[\sum_{i=1}^{M}\xi_{ij}\right] < s_j - a\right) \\
&\leq \exp\left(-\frac{2(s_j-a)^2}{\Gamma^2}\right) \leq \exp\left(-\frac{2t^{\star 2}}{\Gamma^2}\right), \qquad (44) \\
\implies \theta_j &= \pi\theta_j^+ + (1-\pi)\theta_j^- \leq \exp\left(-\frac{2t^{\star 2}}{\Gamma^2}\right) \qquad (45) \\
\implies \bar{\theta} &= \frac{1}{N}\sum_{j=1}^{N}\theta_j \leq \exp\left(-\frac{2t^{\star 2}}{\Gamma^2}\right). \qquad (46)
\end{aligned}$$

Since $\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) = 1-\bar{\theta}$, we have

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}\left(\hat{y}_j \neq y_j\right) \geq 1 - \exp\left(-\frac{2t^{\star 2}}{\Gamma^2}\right) = 1 - e^{-2t_2^2}.$$

Therefore, we have also proved Theorem 10 (2).

□

References

[1] Bennett, George. 1962. Probability Inequalities for the Sum of Independent Random Variables. Journal of the American Statistical Association (February 2013): 37C41.

[2] Chernoff, Herman. 1952. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations. The Annals of Mathematical Statistics 23 (4): 493C507.

[3] Hoeffding, Wassily. 1963. Probability Inequalities for Sums of Bounded Random Variables.Journal of the American Statistical Association.

[4] McDiarmid, Colin. 1998. Concentration. Technique Report. Unversity of Oxford. http://cgm.cs.mcgill.ca/ breed/conc/colin.pdf.

[5] Ngo, Hung Q. 2011. Tail and Concentration Inequalities. Lecture Notes. http://www.cse.buffalo.edu/ hungngo/classes/2011/Spring-694/lectures/l4.pdf.