Efficient Crowdsourcing for Multi-class Labeling

David R. Karger
Massachusetts Institute of Technology
karger@mit.edu

Sewoong Oh
University of Illinois at Urbana-Champaign
swoh@illinois.edu

Devavrat Shah
Massachusetts Institute of Technology
devavrat@mit.edu

ABSTRACT

Crowdsourcing systems like Amazon's Mechanical Turk have emerged as an effective large-scale human-powered platform for performing tasks in domains such as image classification, data entry, recommendation, and proofreading. Since workers are low-paid (a few cents per task) and tasks performed are monotonous, the answers obtained are noisy and hence unreliable. To obtain reliable estimates, it is essential to utilize appropriate inference algorithms (e.g. Majority voting) coupled with structured redundancy through task assignment. Our goal is to obtain the best possible trade-off between reliability and redundancy.

In this paper, we consider a general probabilistic model for noisy observations for crowd-sourcing systems and pose the problem of minimizing the total price (i.e. redundancy) that must be paid to achieve a target overall reliability. Concretely, we show that it is possible to obtain an answer to each task correctly with probability $1-\varepsilon$ as long as the redundancy per task is $O\big((K/q)\log(K/\varepsilon)\big)$, where each task can have any of the K distinct answers equally likely, q is the *crowd-quality* parameter that is defined through a probabilistic model. Further, effectively this is the best possible redundancy-accuracy trade-off *any* system design can achieve. Such a single-parameter crisp characterization of the (order-)optimal trade-off between redundancy and reliability has various useful operational consequences. Further, we analyze the robustness of our approach in the presence of adversarial workers and provide a bound on their influence on the redundancy-accuracy trade-off.

Unlike recent prior work [13, 17, 19], our result applies to non-binary (i.e. $K > 2$) tasks. In effect, we utilize algorithms for binary tasks (with inhomogeneous error model unlike that in [13, 17, 19]) as key subroutine to obtain answers for K-ary tasks. Technically, the algorithm is based on low-rank approximation of weighted adjacency matrix for a random regular bipartite graph, weighted according to the answers provided by the workers.

SIGMETRICS'13, June 17-21, 2013, Pittsburgh, PA, USA.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Statistical computing; F.2 [**Analysis of Algorithms and Problem of Complexity**]: Miscellaneous

Keywords

Crowd-sourcing, Low-rank Matrices, Random Graphs

1. INTRODUCTION

In this paper, we are interested in designing crowd-sourcing systems that are efficient in the sense of achieving *reliability* at the minimal cost of *redundancy*. We will provide appropriate definitions of redundancy and reliability later in this section. More generally, this work is aimed at addressing the following scenarios.

Scenario One. Using MTurk[1] platform for obtaining count of cancerous tumor cells in each microscope image for a very large collection of images leads to answers that are noisy – a good fraction of answers are either nearly correct or arbitrary (cf. see [14]) as the workers either make honest mistakes or they are not making any effort.

Scenario Two. Clinicians collect and record medical history of patients by asking them various questions and classifying the patients' symptoms for type, severity, and duration. Such medical opinions are subject to observer errors and different clinicians may give different values due to variety of reasons (cf. see [9]) such as different wording of questions, different interpretation of the scales, etc.

Scenario Three. Scores are collected from reviewers in the reviewing process of conferences such as Sigmetrics 2013. Each paper, though may have an *innate* score, receives varying scores from different reviewers for reasons such as different reviews have different subjective interpretation of score-scale, or value the contribution of papers differently.

In all of the above scenarios, we have numerous 'multiple choice' tasks at hand and means to collect noisy answers on those tasks by either assigning the tasks using MTurk, getting medical opinions from clinicians, or asking reviewers to review papers. If we are parsimonious and collect only one opinion per task, then we have no other way than to trust that opinion which could be erroneous. To increase reliability, a common practice is to utilize redundancy – each task is assigned to multiple MTurk workers, clinicians or reviewers. Naturally, the more redundancy we introduce, the better accuracy we can hope to achieve. The goal of this paper

[1]http://www.mturk.com

is to get the most accurate estimates from given amount of redundancy. To this end, we develop an algorithm for deciding which tasks to assign to which workers, and estimating the answers to the tasks from noisy answers collected from those assigned workers.

Model and problem formulation. Our interest is in finding answers to the tasks, each of which has one true answer from a set of K possible choices denoted by $\mathcal{K} \equiv \{1, \ldots, K\}$. Each worker, when given a task with true answer k, provides an answer $\ell \in \mathcal{K}$ with probability $\pi_{k\ell}$; by definition $\sum_{\ell \in \mathcal{K}} \pi_{k\ell} = 1$ for all $k \in \mathcal{K}$. We call $\pi = [\pi_{k\ell}] \in [0,1]^{K \times K}$ to be the *confusion* (probability) matrix of that worker. Without loss of generality[2], let each task have correct answer equal to k with probability θ_k independently and let worker have confusion matrix π drawn from a distribution $\mathcal{D}$ on space of confusion matrices. As one example, we can define a generalization of the *spammer-hammer* model from [18], where each worker is either a 'hammer' with probability q or is a 'spammer' with probability $1-q$. A hammer, who always gives the correct answer, has the identity confusion matrix $\pi = \mathbf{I}_{K \times K}$, where $\mathbf{I}$ is the identity matrix. A spammer, who gives answers that are independent of the true answers, has a uniform confusion matrix $\pi = (1/K)\mathbf{1}_K \cdot p^T$, where $\mathbf{1}$ is the vector of all ones, and p^T denoted the transpose of a probability vector p. For example, a spammer might always answer 'one' for any tasks, in which case $p = [1, 0, \ldots, 0]$, or give uniformly random answers, in which case $p = (1/K)[1, 1, \ldots, 1]$. We use t_i to denote the groundtruth answer to the i-th task (which we assume is drawn randomly from a distribution θ), and $\pi^{(j)}$ for the confusion matrix of the j-th worker (which we assume is drawn randomly from a distribution $\mathcal{D}$).

Given this setting, we wish to find answers to a given set of n tasks using m workers so that we are confident that answer to any particular task is correctly with probability at least $1-\varepsilon$ for some small positive ε, and hence *reliable*. Indeed, if a given task is assigned to only one worker, the probability of making an error is given by

$$\sum_{1 \le \ell \le K} \theta_\ell (1 - \mathbb{E}[\pi_{\ell\ell}]),$$

where expectation in $\mathbb{E}[\pi_{\ell\ell}]$ is with respect to $\mathcal{D}$. To further reduce error down to $1-\varepsilon$ for any ε, one might choose to assign the same task to multiple workers and then take majority of the received answers. Such an approach can lead to reduced error at the cost of increase in the *redundancy*, i.e. the average number of answers received per task. In practice, increase in redundancy typically leads to increase in the cost, e.g., payment to MTurk workers or time to finish reviews.

In general, consider the case when we have n tasks to complete and m workers available. Assigning tasks can be viewed as constructing a bipartite graph $G = (T, W, E)$ with $T = \{t_1, \ldots, t_n\}$ representing tasks, $W = \{w_1, \ldots, w_m\}$ representing workers and $E \subset T \times W$ representing task assignment: $(t_i, w_j) \in E$ if task t_i is assigned to worker w_j. In this case, the per task redundancy is $|E|/n$, that is, the average degree of task vertices in graph G. Once tasks are assigned according to a graph G, the workers provide answers $A = [A_{ij}] \in \{\mathcal{K} \cup \mathsf{null}\}^{n \times m}$ where $A_{ij} = \mathsf{null}$ if $(t_i, w_j) \notin E$, i.e. worker w_j is not assigned to task t_i, and it is equal to the answer provided by the worker w_j to the task t_i if $(t_i, w_j) \in E$. Once all the answers $\{A_{ij}\}_{(i,j) \in E}$ are collected, we want to estimate the true answers to the tasks. With abuse of notation, we shall use t_i to represent both node in bipartite graph G and the true answer (in $\mathcal{K}$) to the i-th task. Let $\hat{t}_i \in \mathcal{K}$ be the estimation produced. Then, the probability of error is defined as

$$P_{\mathsf{err}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(t_i \neq \hat{t}_i), \tag{1}$$

where the probability is taken over all realizations of $\{t_i\}$, $\{\pi^{(j)}\}$, $\{A_{ij}\}$, and any randomness in the task assignment and inference algorithm.

The goal in designing a reliable and cost-efficient crowdsourcing system is to obtain P_{err} smaller than given target $\varepsilon \in (0, 1/2)$ with minimal redundancy by appropriately choosing task assignment graph G and the inference algorithm to estimate $\{\hat{t}_i\}$.

Next, we define a few quantities that will be useful to describe the result precisely in the subsequent text (readers may skip these definitions till "summary of results"). For any task i and worker j, define the following probabilities

$$\begin{aligned} p_k^+ &\equiv \mathbb{P}(A_{ij} > k | t_i > k) = \sum_{k < \ell \le K} \sum_{k < \ell' \le K} \frac{\theta_\ell \, \pi^{(j)}_{\ell\ell'}}{\theta_{>k}}, \\ p_k^- &\equiv \mathbb{P}(A_{ij} \le k | t_i \le k) = \sum_{1 \le \ell \le k} \sum_{1 \le \ell' \le k} \frac{\theta_\ell \, \pi^{(j)}_{\ell\ell'}}{1 - \theta_{>k}}, \end{aligned} \tag{2}$$

where $\theta_{>k} = \sum_{k < \ell \le K} \theta_\ell$. Also define $q_k \equiv \mathbb{E}[(p_k^+ + p_k^- - 1)^2]$ for $1 \le k < K$ where the expectation is with respect to the distribution $\mathcal{D}$ of the confusion matrix. Define crowd-quality parameter $q = \min_{1 \le k < K} q_k$. For example, under the *spammer-hammer* model, a hammer has $p_k^+ = p_k^- = 1$ and a spammer has $p_k^+ + p_k^- = 1$ for all k. If a randomly drawn worker is a hammer with probability $\tilde{q}$ and a spammer otherwise, we have $q_k = \tilde{q}$ for all k and $q_k = \tilde{q}$.

Define the maximum *bias* of the true answers as $|\bar{s}| = \max_{k=1}^{K-1} |\bar{s}^k|$, where $\bar{s}^k = 2\theta_{>k} - 1$ is the bias in the a priori distribution of the true answers in binary classification task "is t_i larger than k?". For uniform prior $\theta_\ell = 1/K$ and hence the maximum bias is $1 - 2/K$. We will see that in order to achieve average probability of error less than ε, we need to have redundancy that scales as $(1/(q(1-|\bar{s}|)))\log(K/\varepsilon)$ which in the case of uniform prior scales as $(K/q)\log(K/\varepsilon)$.

Prior work. Though crowd-sourcing is a recent phenomenon, similar questions were considered by Dawid and Skene [9] in the context of *Scenario Two* described earlier. They introduced an iterative algorithm for inferring the solutions and reliability of workers, based on the expectation maximization (EM) [10]. EM is a heuristic inference algorithm that iteratively does the following: given workers' answers to the tasks, the algorithm attempts to estimate the reliability of the workers and given estimation of reliability (error probabilities) of workers, it estimates the solution of the tasks; and repeat. Due to particular simplicity of the EM algorithm, it has been widely applied in classification problems where the training data is annotated by low-cost noisy 'labelers' [16, 23]. Sheng et al. [26] have extensively studied the al-

[2]This is without loss of generality, as the results stated in this paper hold even if we use the empirical distribution in place of the distribution assumed for prior on tasks as well as worker confusion matrices.

gorithm's performance empirically. Now EM algorithm has various shortcomings: (i) it is a heuristic and there are no rigorous guarantees known about its correctness or overall performance; (ii) a priori it is not clear that for this particular problem EM is convergent; and (iii) the role of the task allocation is not at all understood with the EM algorithm.

More rigorous approaches towards designing task assignment graphs and inference algorithms were recently proposed starting [13, 17]. In these work, task assignment was done through random graphs (Erdos-Renyi in [13], random regular in [17]) and inference was done through low-rank approximations. They, however, assumed binary tasks (i.e. $K = 2$) and homogeneous error model (i.e. $\pi_{12} = \pi_{21}$ with $K = 2$); and resulted in sub-optimal trade-off between redundancy and error. This was further improved upon to reach order-optimal error-redundancy trade-off by means of belief propagation based iterative estimation algorithm in [19]. This algorithm uses weighted majority voting where the weights are computed by an approximate belief propagation. Our approach is similar but the weights are computed by singular value decomposition (SVD). The major difference is that SVD based approach generalizes to more general probabilistic models we study in this paper, whereas the belief propagation based approach only works for a simpler model where the underlying structure is a rank one matrix. More recently, it was shown that the resulting design and inference algorithm are optimal even with respect to adaptive system design [18]. The key limitation of all of the above, definitely very impressive, results is applicability to binary tasks with homogeneous error model.

Given graphical models such as the one studied in these prior work, one can solve the inference problem using a standard belief propagation. The main challenge in such an approach is that the inference algorithm requires the priors from which the distribution of the quality of the workers are drawn. In this paper, we do not assume any knowledge of the prior. However, it was shown through experiments on real and simulated datasets in [21] that when the prior is known, improved performance can be achieved.

It should be noted that crowdsourcing is currently extremely active research area in terms of designing actual platforms like [2, 3, 1, 4, 5], empirical results based on experiments like [16, 7, 23, 6, 28, 27] and deciding on issues like pricing such as results in [22, 15]. The main focus of this paper is rigorous treatment of crowdsourcing system design and hence we only provide a limited coverage of prior work related to general crowdsourcing. In particular, we do not address some practical questions such as embedding golden questions which you know the answers to, screening workers with accuracy thresholds, and paying only on accurate responses.

Summary of results. As the main result of this paper, we provide a crowdsourcing system design that is asymptotically order-optimal for the general noise model considered here for K-ary tasks for any $K \geq 2$. This is the first rigorous result for K-ary (even for $K = 2$) tasks with non-homogeneous error model. In a sense, it resolves the question raised by Dawid and Skene [9] in the context of medical record collection or more generally noisy computation. Formally, we show that it is possible to achieve $P_{\mathsf{err}} \leq \varepsilon$ for any $\varepsilon \in (0, 1)$ with per task redundancy $O\big(\frac{1}{q(1-|\bar{s}|)} \log \frac{K}{\varepsilon}\big)$. The minimum bias $|\bar{s}|$ depends on the prior distribution $(\theta_1, \ldots, \theta_K)$; for uniform prior, it is such that $1 - |\bar{s}| = 2/K$. That is, effectively, for uniform prior, our result states that redundancy requirement scales as $O\big(\frac{K}{q} \log \frac{K}{\varepsilon}\big)$. And, (using result of [17, 19]) for any system to achieve $P_{\mathsf{err}} \leq \varepsilon$, redundancy of $\Omega\big(\frac{1}{q} \log \frac{1}{\varepsilon}\big)$ is needed. Thus, for any fixed K (i.e. treating K as a constant), with respect to $q, \varepsilon \to 0$ asymptotic, our system design is order optimal; non-asymptotically off by $(K/q) \log K$.

2. MAIN RESULT

In this section, we describe our task allocation and inference algorithm accompanied by theorems describing it's performance.

2.1 Task allocation

Given n tasks, to utilize redundancy of $\ell \times R$ per task, we shall utilize $n \times R$ workers. Specifically, we shall choose R distinct (ℓ, ℓ) random regular graph $G_1, \ldots, G_R$ for task allocation – in each of these R graphs, we, of course, use the same n tasks but use distinct n workers; thus utilizing $n \times R$ total workers. Each graph G_r, $1 \leq r \leq R$ is generated as per the scheme known as the *configuration model*, cf. [8, 24]. Intuitively, the random regular graphs are good choice because they are known to be good 'expanders' and therefore allows us to efficiently extract the true answers from noisy data matrix using low-rank approximation. We will make independent estimates of the tasks (using low-rank matrix approximations) based on each of these R datasets collected independently. For each task, we will combine these R estimates (using majority voting) to further refine our estimate and guarantee order optimal performance.

2.2 Inference algorithm

Let $A(r) = [A_{ij}(r)] \in \big(\{\mathsf{null}\} \cup \mathcal{K}\big)^{n \times n}$ be the noisy answers obtained using the r-th, random (ℓ, ℓ)-regular task allocation graph G_r, for $1 \leq r \leq R$. From these datasets on the answers $\{A(r)\}_{1 \leq r \leq R}$, we wish to obtain estimates $\{\hat{t}_i\} \in \mathcal{K}^n$ on what the true answers are for all n tasks. We shall utilize combination of low-rank matrix approximation and majority voting to obtain estimates as described below.

We first reduce the K-ary classification tasks into a series of $K - 1$ simple binary classification tasks. Using each dataset $A(r)$ for $1 \leq r \leq R$, we first produce binary estimates $\hat{t}^k(r) = [\hat{t}^k_i(r)] \in \{-1, 1\}^n$ for $1 \leq k < K$ where

$$\hat{t}^k_i(r) \;=\; \begin{cases} -1 & \text{if we believe that } t_i \leq k \text{ based on } A(r)\,, \\ +1 & \text{if we believe that } t_i > k \text{ based on } A(r)\,. \end{cases}$$

The low-rank matrix approximation algorithm for estimating $\hat{t}^k_i(r)$ based on $A(r)$ is explained later in this section in detail. Based on these binary estimates on each independent datasets $A(r)$, we further refine our estimates by combining our estimates using majority aggregation over the whole data $\{A(r)\}_{1 \leq r \leq R}$, to get $\hat{t}^k = [\hat{t}^k_i]$. The estimate $\hat{t}^k_i$ for the i-th task is our estimated answer to the question "is t_i larger than k?", determined through majority voting as

$$\hat{t}^k_i \;=\; \mathrm{sign}\Big(\sum_{r=1}^{R} \hat{t}^k_i(r)\Big) \tag{3}$$

where $\mathrm{sign}(x) = 1$ if $x \geq 0$ and -1 if $x < 0$. As we will show in Section 3.1, the main reason we use R independent datasets is to use concentration inequalities to get a tighter bound on the probability of error.

Focusing on a particular task i, if our estimates $\hat{t}_i^k$ are accurate for all $k \in \mathcal{K}$, then we expect them to have a single switch at the true answer $k = t_i$:

$$(\hat{t}_i^1, \ldots, \hat{t}_i^{t_i-1}, \hat{t}_i^{t_i}, \ldots, \hat{t}_i^K) = (\underbrace{+1, \ldots, +1}_{t_i-1}, \underbrace{-1 \ldots, -1}_{K-t_i+1}) \ ,$$

where we define $\hat{t}_i^K = -1$ for all i. This naturally defines the following rule for producing our final estimates $\hat{t} = [\hat{t}_i]$:

$$\hat{t}_i = \min \big\{ \, k \, : \, \hat{t}_i^k = -1 \, \big\}. \tag{4}$$

Other methods for aggregating the binary estimates to get a full K-ary estimates are $\hat{t}_i = (1/2)(2+K+\sum_{k=1}^K \hat{t}_i^k)$ or finding the index that minimize the inconsistencies: $\min_k |\{a < k : \hat{t}_i^a = -1\} \cup \{a \geq k : \hat{t}_i^a = +1\}|$. However, the simple aggregation rule above is powerful enough to ensure that we achieve order-optimal performance.

Now we describe how $A(r)$ is used to produce $\hat{t}^k(r)$, for $1 \leq k < K$. Define matrices $A^k(r) = [A_{ij}^k(r)]$ for $1 \leq k < K$ where

$$A_{ij}^k(r) = \begin{cases} 0 & \text{if } A_{ij}(r) = \mathsf{null} \\ 1 & \text{if } k < A_{ij}(r) \leq K \\ -1 & \text{if } 1 \leq A_{ij}(r) \leq k. \end{cases} \tag{5}$$

That is, entries of matrix $A^k(r)$ converts (quantizes) the answers $A(r)$ into greater than k $(+1)$, less or equal to k (-1), and null (0). Define an $n \times n$ projection matrix L as

$$L \equiv \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T \ , \tag{6}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is the all-ones vector. Consider the projected matrices $B^k(r) = LA^k(r)$. Let $u^k(r)$, $v^k(r)$ be the pair of normalized (unit norm) left and right singular vectors respectively of $B^k(r)$ corresponding to the largest singular value of $B^k(r)$. Produce quantized estimates of tasks $\hat{t}^k(r) = [\hat{t}_i^k(r)]$ as

$$\hat{t}^k(r) = \begin{cases} \text{sign}(u^k(r)) & \text{if } \sum_{j:v_j^k(r)\geq 0}(v_j^k(r))^2 \geq 1/2 \ , \\ \text{sign}(-u^k(r)) & \text{if } \sum_{j:v_j^k(r)\geq 0}(v_j^k(r))^2 < 1/2 \ , \end{cases} \tag{7}$$

where $\text{sign}(\cdot)$ is a function that outputs entry-wise sign of a vector, such that $\text{sign}(x) = [\text{sign}(x_i)]$. Even when the largest singular value is unique, the left singular vector $u^k(r)$ is only determined up to a sign. To resolve this ambiguity we use the right singular vector $v^k(r)$ to determine the sign of our final estimate. We can also use other means of resolving this ambiguity up to a sign, such as asking *golden* questions with known answers, if we have them available.

We can also interpret (7) as a weighted majority voting with the right singular vector as the weights. Since $u^k(r) = A^k(r)v^k(r)$, our estimate for the i-th task is

$$\begin{aligned} \hat{t}_i^k(r) &= \text{sign}(u_i^k(r)) \\ &= \text{sign}\big(\sum_j A_{ij}^k(r)\, v_j^k(r) \big) \ , \end{aligned}$$

assuming we have resolved the ambiguity in sign. Effectively, we are weighting each response, $A_{ij}^k(r)$, by how reliable each worker is, $v_j^k(r)$. In proving the main results, we will show in (14) that $v_j^k(r)$ is an estimate for $(p_k^+ + p_k^- - 1)$ for the j-th worker. Intuitively, the larger $v_j^k(r)$ is the more reliable the worker j is.

2.3 Performance

Here we describe the performance of the algorithm introduced above. For this, define the maximum *bias* of the true answers as $|\bar{s}| = \max_{k=1}^{K-1} |\bar{s}^k|$, where $\bar{s}^k = 2\theta_{>k} - 1$ is the bias in the a priori distribution of the true answers in binary classification task "is t_i larger than k?". For results below to hold, we shall assume that the random variables p_k^+ and p_k^- defined in (2) satisfy $p_k^+ + p_k^- \geq 1$ for all $1 \leq k < K$ with probability one according to the the distribution $\mathcal{D}$ of the confusion matrix. However, this assumption is only necessary to ensure that we can resolve the ambiguity of the sign in deciding whether to use $u^k(r)$ or $-u^k(r)$ for our inference in (7). If we have alternative way of resolving this ambiguity, for instance embedding golden questions with known answers, then the following theorem holds for any $\mathcal{D}$.

THEOREM 2.1. *For any $\varepsilon \in (0, 1/2)$ and a choice of $\ell = \Theta\big(\frac{1}{q(1-|\bar{s}|)^3}\big)$ and $R = \Theta\big(\log(K/\varepsilon)\big)$, there exists a $N(\varepsilon, \ell, \bar{s}, q)$ that depends on ε, ℓ, $\bar{s}$, and q such that for all $n \geq N(\varepsilon, \ell, \bar{s}, q)$, we have*

$$P_{\mathsf{err}} \ = \ \frac{1}{n}\sum_{i=1}^n \mathbb{P}(\hat{t}_i \neq t_i) \ \leq \ \varepsilon,$$

where the probability is over the randomness in the choice of task allocation graph, true answers to the tasks, worker confusion matrices, and the the realization of the answers submitted by the workers.

In terms of the *redundancy* and *reliability* trade-off, the above theorem states that we need to collect $\ell R = \Theta\big(\frac{1}{q(1-|\bar{s}|)^3}\log(K/\varepsilon)\big)$ answers per task to ensure that we achieve error rate less than ε.

Dealing with $\bar{s}$. Let us discuss the dependence of the required redundancy on $\bar{s}$. When we have uniformly distributed true answers, $\theta_\ell = 1/K$ for $1 \leq \ell \leq K$, then $|1 - \bar{s}| = 2/K$ leading to the redundancy dependence scale as $O\big((K^3/q)\log(K/\varepsilon)\big)$ in Theorem 2.1. While K is treated as a constant, for moderate size of K, this is terrible dependence. It is, indeed, possible to improve this dependence on $\bar{s}$ by modifying the estimation step (7) as follows: let $\hat{u}^k(r) = u^k(r) + \frac{\bar{s}^k}{\sqrt{(1-(\bar{s}^k)^2)n}}\mathbf{1}$, then

$$\hat{t}^k(r) = \begin{cases} \text{sign}(\hat{u}^k(r)) & \text{if } \sum_{j:v_j^k(r)\geq 0}(v_j^k(r))^2 \geq 1/2 \\ \text{sign}(-\hat{u}^k(r)) & \text{if } \sum_{j:v_j^k(r)\geq 0}(v_j^k(r))^2 < 1/2 \end{cases} \tag{8}$$

The above estimation step, however, requires knowledge of $\bar{s}^k$ which is quite feasible as it's population level aggregation (i.e. knowledge of θ_ℓ, $1 \leq \ell \leq K$). With the above estimation, we get the following improved bound with change of $\ell = \Theta(1/q(1-|\bar{s}|))$ in place of $\ell = \Theta\big(1/q(1-|\bar{s}|)^3\big)$.

THEOREM 2.2. *Under the hypotheses of Theorem 2.1, for any $\varepsilon \in (0, 1/2)$ and a choice of $\ell = \Theta\big(\frac{1}{q(1-|\bar{s}|)}\big)$ and $R = \Theta\big(\log(K/\varepsilon)\big)$, there exists a $N(\varepsilon, \ell, \bar{s}, q)$ such that for all $n \geq N(\varepsilon, \ell, \bar{s}, q)$, the estimates in (8) achieve $P_{\mathsf{err}} \leq \varepsilon$.*

When designing a task assignment, we choose how much *redundancy* we want to add per task, which is the average number of answers we are collecting per task. Let $\gamma = \ell R$ denote the redundancy per task. According to the above theorem, to achieve an average error probability less than ε, we need the redundancy per task that scales as $\gamma =$

$O((1/q(1-|\bar{s}|))\log(K/\varepsilon))$. Then, this implies that the probability of error achieved by our approach is upper bounded by $P_{\text{err}} \leq Ke^{-C\gamma q(1-|\bar{s}|)}$ for a positive constant C. Figure 1 illustrates this exponential dependency of P_{err} on the redundancy γ for fixed K, q and $|\bar{s}|$. Compared to an algorithm-independent analytical lower bound, this shows that the constant C in the error exponent is very close to the optimal one, since the slop of the error probability is very close to that of the lower bound.

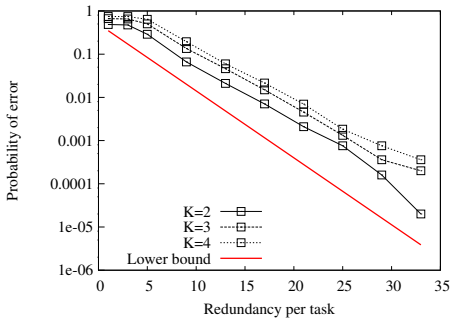

Figure 1: Average probability of error decreases exponentially as redundancy increases and is weakly dependent on the alphabet size K.

For this example, we used uniform prior of $\theta_k = 1/K$ and the spammer-hammer model described in Section 1 to generate the data with $q = q_k = 0.3$. We plot the average number of errors over 1000 tasks averaged over 50 random instances of this problem. As we increase the alphabet size, the slop of the (log) probability of error does not change. If our upper bound on P_{err} was tight, we expect the slop to scale as $(1 - |\bar{s}|)$, which in this numerical example is $1/K$.

Optimality. In [17, 19], it was shown that for binary model ($K = 2$) with homogeneous noise model (i.e. $\pi_{12} = \pi_{21}$ for all π), to obtain $P_{\text{err}} \leq \varepsilon$, the per task redundancy must scale as $\Omega\big(\frac{1}{q}\log(\frac{1}{\varepsilon})\big)$. This lower-bound on redundancy requirement is independent of the choice of any task-allocation and inference algorithm. Clearly, this is a special case of our general model and hence applies to our setting. From Theorem 2.2, it follows that our algorithm is within a factor of K of optimal redundancy requirement for $K = O(1/\epsilon)$. Equivalently, in the asymptotic of $\varepsilon, q \to 0$, our algorithm is order-optimal, since the dependencies on ϵ and q are the same as the optimal budget requirement.

Running time of algorithm. The key step in our inference algorithm is obtaining rank-1 approximation of the $n \times n$ matrices $LA^k(r)$ for $1 \leq k < K$ and $1 \leq r \leq R$. In practice, n is the number of papers submitted to a conference, number of patients, or the number of images we want to label, and it is likely to be very large. Standard iterative methods, such as the power iteration or the Lanczos method can be used to compute the leading singular vector of such large matrices. These iterative methods only rely on the matrix-vector product, which can be done quite efficiently by exploiting the structure of $LA^k(r)$.

The standard power-iteration algorithm leads to identification of rank-1 approximation (i.e. left, right singular vectors) within error of δ with number of iterations $O\big(\frac{\log(n/\delta)}{\log(\sigma_1/\sigma_2)}\big)$, where σ_1, σ_2 are the largest and second largest singular values of matrix $LA^k(r)$. In the process of establishing Theorem 2.1, we shall show that $\sigma_2/\sigma_1 = O(1/\sqrt{\ell q_k})$, and with $\ell = \Theta(1/q_k)$ this can be made as small as we want.

At each iteration of the power iteration algorithm, we compute matrix-vector multiplication of

$$x^{(t+1)} \;=\; L\,A^k(r)\,(A^k(r))^T\,L\,x^{(t)}\;,$$

and it is known that $x^{(t)}$ eventually converges to the left singular vector of matrix $LA^k(r)$ up to a normalization. Each computation of this multiplication can be done efficiently in $O(n\ell)$ time. Since $A^k(r)$ is a sparse matrix with $n\ell$ non-zero entries, we can compute $(A^k(r))^T y$ with $O(n\ell)$ operations. Since $L = \mathbf{I} - (1/n)\mathbf{1}\mathbf{1}^T$, we can compute $Ly = y - (1/n)\mathbf{1}\mathbf{1}^T y$ in $O(n)$ operations.

Finally, we only need to compute an approximate singular vector up to a certain error. Let u be the left singular vector of $LA^k(r)$ and define $t^k = [t_i^k]$ to be the true answer for a binary classification problem:

$$t_i^k \;=\; \begin{cases} -1 & \text{if } t_i \leq k\;, \\ +1 & \text{if } t_i > k\;. \end{cases}$$

In the process of establishing Theorem 2.1, we will utilize the fact that the singular vector u is at most distance $C/\sqrt{\ell q_k}$ from the true answers that we want: $\|u - (1/\sqrt{n})t^k\| \leq C/\sqrt{\ell q_k}$. Hence, we only need an approximate singular vector up to error $\delta = C/\sqrt{\ell q_k}$ and the same result holds with this approximate singular vector. Therefore, total computation cost of computing the top left singular vector of $LA^k(r)$ scales as $O\big(n\ell\frac{\log(n\ell q_k)}{\log(\ell q_k)}\big)$ (this is assuming $\ell q_k > 1$).

Operational implications. Here we discuss a few concrete and highly attractive operational implications of crisp result we obtain in Theorem 2.1. Suppose there are M classes of worker: workers of class $m, 1 \leq m \leq M$, have confusion matrix distribution $\mathcal{D}_m$ such that the corresponding quality parameter is q_m and each of them requires payment of c_k to perform a task. Theorem 2.1 immediately suggests that we should hire the worker class m^* that maximizes q_m/c_m over $1 \leq m \leq M$.

The next variation is on the assumed knowledge of q. When designing the regular bipartite graph for task assignment, it requires selecting the degree $\ell = \Theta(1/q(1 - |\bar{s}|))$. This assumes that we know a priori the value of q. One way to overcome this limitation is to do binary search for appropriate value of ℓ. This results in a cost of additional constant factor in the budget, i.e. scaling of cost per task still remains $\Theta(\,(1/q(1-|\bar{s}|))\log(K/\varepsilon)\,)$. Use following iterative procedure to test the system with $q = 2^{-a}$ at iteration a, and we stop if the resulting estimates are consistent in the following sense. At iteration a, design two replicas of the system for $q = 2^{-a}$, and compare the estimates obtained by these two replicas for all n tasks. If they agree amongst $n(1 - 2\varepsilon)$ tasks, then we stop and declare that as the final answer. Or else, we increase a to $a + 1$ and repeat. Note that by our main result, it follows that if 2^{-a} is less than the actual q then the iteration must stop with high probability.

Robustness against adversarial attacks. We consider two scenarios: first case is where the malicious workers are able to choose their own confusion matrix but still give answers according to our probabilistic model, and second case is where malicious workers are able to give any answers they

want. We want to see how robust our approach is when α proportion of the workers are adversarial, that is when αnR workers are adversarial among total nR workers.

Under the first scenario, it follows from Theorem 2.1 that the effect of such adversarial workers is fully captured in q'_k, where now each worker with probability $1-\alpha$ has $\pi^{(j)}$ coming from $\mathcal{D}$ and with probability α has $\pi^{(j)}$ chosen by the adversary. Then, even in the worst case, $q'_k \geq (1-\alpha)\mathbb{E}_{\mathcal{D}}[(p_k^+ + p_k^- - 1)^2]]$. The new 'crowd quality' is now degraded to $q' \geq (1-\alpha)q$. In terms of the redundancy necessary to achieve error rate of ε, we now need a factor of $1/(1-\alpha)$ more redundancy to achieve the same error rate with the presence of α proportion of adversarial workers. This suggests that our approach is robust, since this is the best dependency on α one can hope for. Let M be the number of workers necessary under non-adversarial setting. If we have adversaries among the workers, and let us even assume that we can detect any adversary, even then we need $M/(1-\alpha)$ total workers to get M non-adversarial workers. Our approach requires the same number of 'good' workers as the one that can detect all adversaries. A similar analysis, in the case of binary tasks (i.e., K=2) and homogeneous error model (i.e., $\pi_{12} = \pi_{21}$ with $K = 2$) was provided in [13].

Under the second scenario, we assume that αnR workers are adversarial as before, but those adversarial workers can submit any answers they want. In particular, this model includes the adversaries who are colluding to manipulate our crowdsourcing system. We want to prove a bound on how much the performance of our algorithm degrades as the number of such adversaries increases. The following theorem proves that our algorithm is robust, in the sense that the same guarantee is achieved with redundancy that scales in the same way as when there are no adversaries, as long as the proportion of adversaries is bounded by $\alpha = cq(1-|\bar{s}^k|)$ for some positive constant c.

THEOREM 2.3. *Under the hypotheses of Theorem 2.1, there exists a constant c such that when the proportion of the adversarial workers is $\alpha \leq cq(1-|\bar{s}^k|)^3$, our estimates in (7) aggregated as in (3) achieve $P_{\mathsf{err}} \leq \varepsilon$ with a choice of $\ell = \Theta\big(\frac{1}{q(1-|\bar{s}|)^3}\big)$ and $R = \Theta\big(\log(K/\varepsilon)\big)$ for $n \geq N(\varepsilon, \ell, \bar{s}, q)$. Further, if we use estimates in (8), then the same guarantee holds with $\alpha \leq cq(1-|\bar{s}^k|)$ and $\ell = \Theta\big(\frac{1}{q(1-|\bar{s}|)}\big)$.*

3. PROOF OF MAIN RESULTS

In this section, we provide the proofs of the main results and technical lemmas.

3.1 Proof of Theorem 2.1

First we consider a single binary estimation problem on a single dataset $A(r)$ and a classification threshold $k \in \mathcal{K}$. We will show that, with choice of $\ell = \Theta(1/(q(1-|\bar{s}^k|)^3))$, we can get good estimates from each dataset $A(r)$ on each binary classification task such that the probability of making an error on each task is less than $1/4$:

$$\begin{aligned} p_e^+ \equiv \mathbb{P}(\hat{t}_i^k(r) = -1 | t_i^k = +1) &\leq 1/4\ , \text{ and} \\ p_e^- \equiv \mathbb{P}(\hat{t}_i^k(r) = +1 | t_i^k = -1) &\leq 1/4\ . \end{aligned}$$

By symmetry p_e^+ and p_e^- do not depend on r or i, but it does depend on k. However, the upper bound holds for any k and we omit the dependence on k to lighten the notations. We can achieve a significantly improved accuracy by repeating the data collection and estimation process R times on independently chosen task assignment graph and completely different set of workers. These R estimates then can be aggregated using (3): $\hat{t}_i^k = \text{sign}\Big(\sum_{r=1}^R \hat{t}_i^k(r)\Big)$. We claim that each $\hat{t}_i^k(r)$ are independent estimates with error probability less than $1/4$. Applying Hoeffding's inequality, we have

$$\begin{aligned} \mathbb{P}\big(\hat{t}_i^k \neq t_i^k\big) &\leq \theta_{>k}\,\mathbb{P}\Big(\sum_{r=1}^R \hat{t}_i^k(r) \leq 0 \,\Big|\, t_i^k = +1\Big) \\ &\quad + \theta_{\leq k}\,\mathbb{P}\Big(\sum_{r=1}^R \hat{t}_i^k(r) \geq 0 \,\Big|\, t_i^k = -1\Big) \\ &\leq \theta_{>k}\,\exp\Big\{-\frac{2(2p_e^+ - 1)^2R^2}{4R}\Big\} \\ &\quad + \theta_{\leq k}\,\exp\Big\{-\frac{2(2p_e^- - 1)^2R^2}{4R}\Big\} \\ &\leq \exp\{-R/8\}\ , \end{aligned}$$

where we used the fact that $p_e^+ \leq 1/4$ and $p_e^- \leq 1/4$.

For each task i, if we did not make any errors in the $K-1$ binary estimations, than we correctly recover the true answer to this task as per rule (4). This happens with probability at least $1 - Ke^{-R/8}$, which follows from the union bound over $k \in \mathcal{K}$. It follows that the average error probability over all n tasks is also bounded by

$$P_{\mathsf{err}} \leq Ke^{-R/8}\ .$$

Setting $R = 8\log(K/\varepsilon)$, the average error probability is guaranteed to be less than ε for any $\varepsilon \in (0, 1/2)$. This finishes the proof of Theorem 2.1.

Now we are left to prove that error probabilities on a single dataset are bounded by $1/4$. Recall that $t_i^k = -1$ if $t_i \leq k$, and $+1$ if $t_i > k$. Given a single dataset $A(r)$, we 'quantize' this matrix to get $A^k(r)$ as defined in (5). Then we multiply this matrix on the left by a projection L defined in (6), and let $B^k(r) = LA^k(r)$ be the resulting matrix. We use the top left singular vector of this matrix $B^k(r)$, to get an estimate of t_i^k as defined in (7). This can be formulated as a general binary estimation problem with heterogeneous error model as follows: when $t_i^k = +1$, the 'quantized' answer of a worker is $+1$ if actual answer is greater than k. This happens with probability

$$\begin{aligned} &\mathbb{P}\big(\text{'quantized' answer} = +1 | t_i^k = +1\big) \\ &= \sum_{\ell > k} \mathbb{P}\big(\text{'actual' answer} = \ell | t_i^k = +1\big) \\ &= \sum_{\ell, \ell' > k} \mathbb{P}\big(\text{'actual' answer} = \ell,\ t_i = \ell' | t_i^k = +1\big) \\ &= \sum_{\ell, \ell' > k} \pi_{\ell'\ell} \frac{\theta_{\ell'}}{\theta_{>k}} \ \equiv\ p_k^+. \end{aligned} \tag{9}$$

Similarly, when $t_i^k = -1$,

$$\begin{aligned} &\mathbb{P}\big(\text{'quantized' answer} = -1 | t_i^k = -1\big) \\ &= \sum_{\ell, \ell' \leq k} \pi_{\ell'\ell} \frac{\theta_{\ell'}}{1 - \theta_{>k}} \ \equiv\ p_k^-. \end{aligned} \tag{10}$$

Thus, the probability of receiving correct answer for such binary tasks (i.e. $> k$ or $\leq k$) depends on whether the true answer is $+1$ (i.e. $> k$) or -1 (i.e. $\leq k$) and they are p_k^+ and

p_k^- respectively. In the prior works [13, 17, 19], the binary task model considers a setting where $p_k^+ = p_k^-$. In that sense, in this paper, we shall extend the results for binary tasks when p_k^+ need not be equal to p_k^-.

For such problem of tasks with binary answers with heterogeneous probability of correct answers, the following lemma provides an upper bound on the probability of error (technically, this is the key contribution of this work).

LEMMA 3.1. *There exists positive numerical constants C and C' such that*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\big(\hat{t}_i^k(r) \neq t_i^k\big) \quad \leq \quad \frac{C}{\ell q_k(1-|\bar{s}^k|)^2} \tag{11}$$

with probability at least $1-2e^{-n(1-|\bar{s}^k|)^2/8}-e^{-q_k^2 n/2}-n^{-C'\sqrt{\ell}}$ where $\bar{s}^k$ and q_k are parameters defined earlier. The probability is over all the realization of the random graphs, the answers submitted by the workers, worker confusion matrices, and the true answers to the tasks.

Since $q \equiv \min_k q_k$ and $\bar{s} = \max_k |\bar{s}^k|$, with our choice of $\ell = \Omega(1/(q(1-\bar{s})^3))$ and for n large enough (dependent on $q, |\bar{s}|$ and ε), we can guarantee that the probability of error is upper bounded by:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}\big(\hat{t}_i^k(r) \neq t_i^k\big) \quad \leq \quad \frac{1-|\bar{s}^k|}{8}\,. \tag{12}$$

By the symmetry of the problem, the probability of error for all the positive tasks are the same and the error probability for all the negative tasks are also the same. Let p_e^+ and p_e^- denote these error probability for positive and negative tasks respectively. Then p_e^+ cannot be larger than $1/4$, since even if we make no mistake on the negative tasks, there are $(1/2)(1+\bar{s}^k)n$ positive tasks with equal probability of error. From the upper bound on average error probability in (12), we get that $p_e^+(1/2)(1+\bar{s}^k) \leq (1-|\bar{s}^k|)/8$. Since $1-|\bar{s}^k| \leq 1+\bar{s}^k$, this implies that $p_e^+ \leq 1/4$. Similarly, we can also show that $p_e^- \leq 1/4$.

3.2 Proof of lemma 3.1

A rank-1 approximation of our data matrix $B^k(r) = LA^k(r)$ can be easily computed using singular value decomposition (SVD). Let the singular value decomposition of $B^k(r)$ be

$$B^k(r) \quad = \quad \sum_{i=1}^{n} u^{(i)}\sigma_i(v^{(i)})^T\,,$$

where $u^{(i)} \in \mathbb{R}^n$ and $v^{(i)} \in \mathbb{R}^n$ are the i-th left and right singular vectors, and $\sigma_i \in \mathbb{R}$ is the i-th singular value. Here and after, $(\cdot)^T$ denotes the transpose of a matrix or a vector. For simplicity, we use $u = u^{(1)}$ for the first left singular vector and $v = v^{(1)}$ for the first right singular vector. Singular values are typically assumed to be sorted in a non-increasing order satisfying $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. Then, the optimal rank-1 approximation is given by a rank-1 projector $\mathcal{P}_1(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ such that

$$\mathcal{P}_1(B^k(r)) \quad = \quad \sigma_1 uv^T\,, \tag{13}$$

It is a well known fact that $\mathcal{P}_1(B^k(r))$ minimizes the mean squared error. In formula,

$$\mathcal{P}_1(B^k(r)) \quad = \quad \arg\min_{X:rank(X)\leq 1}\sum_{i,j}(B^k(r)_{ij}-X_{ij})^2$$

Let $\pi^{(j)}$ be a $K\times K$ confusion matrix for worker $j \in \{1,\ldots,n\}$. In this section, we use p^+ to denote the n-dimensional vector such that p_j^+ is the probability that worker j makes an error on a positive task: $p_j^+ = \sum_{a>k,b\leq k}\pi_{ab}^{(j)}$. Similarly, we let $p_j^- = \sum_{a\leq k,b>k}\pi_{ab}^{(j)}$. We use $t^k = [t_i^k] \in \{-1,+1\}^n$ to denote the n-dimensional vector of true answers.

Recall that conditioned on a given vectors p^+, p^-, and a true answer vector t^k, the conditional expectation of the responses results in a matrix

$$\mathbb{E}[A^k(r)|t^k,p^+,p^-] \quad = \quad \frac{\ell}{n}t^k(p^+ + p^- - \mathbf{1}_n)^T + \frac{\ell}{n}\mathbf{1}_n(p^+ + p^-)^T\,.$$

Since this is a sum of two rank-1 matrices, the rank of the conditional expectation is at most two. One way to recover vector t^k from this expectation is to apply a projection that eliminates the contributions from the second term, which gives

$$\mathbb{E}[LA^k(r)|t^k,p^+,p^-] = \frac{\ell}{n}(t^k - \bar{t}^k\mathbf{1}_n)(p^+ + p^- - \mathbf{1}_n)^T\,, \tag{14}$$

where $L = \mathbf{I} - (1/n)\mathbf{1}_n\mathbf{1}_n^T$, $\bar{t}^k = (1/n)\sum_i t_i^k$, and we used the fact that $L\mathbf{1}_n = 0$. In the following, we will prove that when $A^k(r)$ is close to its expectation $\mathbb{E}[A^k(r)]$ in an appropriate spectral distance, then the top left singular vector of $LA^k(r)$ provides us a good estimate for t^k.

Let u be the left singular vector of $LA^k(r)$ corresponding to the leading singular value. Ideally, we want to track each entry u_i for most realizations of the random matrix $A^k(r)$, which is difficult. Instead, our strategy is to upper bound the spectral radius of $L(A^k(r) - \mathbb{E}[A^k(r)|t^k,p^+,p^-])$, and use it to upper bound the Euclidean distance between the left top singular vectors of those two matrices: u and $(1/\|t^k - \bar{t}^k\mathbf{1}\|)(t^k - \bar{t}^k\mathbf{1})$. Once we have this, we can related the average number of errors to the Euclidean distance between two singular vectors using the following series of inequalities:

$$\begin{aligned}\frac{1}{n}\sum_i \mathbb{I}(t_i^k \neq \mathrm{sign}(u_i)) \;&\leq\; \frac{1}{n}\sum_i \mathbb{I}(t_i^k u_i \leq 0)\\ &\leq\; \frac{1}{n}\sum_i\Big(\frac{1+|\bar{t}^k|}{1-|\bar{t}^k|}\Big)\Big(\sqrt{n}u_i - \frac{t_i^k - \bar{t}^k}{\sqrt{1-(\bar{t}^k)^2}}\Big)^2\\ &=\; \Big(\frac{1+|\bar{t}^k|}{1-|\bar{t}^k|}\Big)\Big\|u - \frac{t^k - \bar{t}^k\mathbf{1}}{\|t^k - \bar{t}^k\mathbf{1}\|}\Big\|^2\,,\end{aligned} \tag{15}$$

where we used the fact that $\|t^k - \bar{t}^k\mathbf{1}\| = \sqrt{n(1-(\bar{t}^k)^2)}$ which follows from the definition $\bar{t}^k = (1/n)\sum_i t_i^k$.

To upper bound the Euclidean distance in (15), we apply the next lemma to two rank-1 matrices: $\mathcal{P}_1(LA^k(r))$ and $\mathbb{E}[LA^k(r)|t^k,p^+p^-]$ where $\mathcal{P}_1(LA^k(r))$ is the best rank-1 approximation of the matrix $LA^k(r)$. This lemma states that if two rank-1 matrices are close in Frobenius norm, then the top singular vectors are also close in the Euclidean distance. For the proof of this lemma, we refer Section 3.3.

LEMMA 3.2. *For two rank-1 matrices with singular value decomposition $M = x\sigma y^T$ and $M' = x'\sigma'(y')^T$, we have*

$$\min\big\{\,\|x+x'\|\,,\,\|x-x'\|\,\big\} \quad \leq \quad \frac{\sqrt{2}\,\|M-M'\|_F}{\max\{\sigma,\sigma'\}}\,,$$

where $\|x\| = \sqrt{\sum_i x_i^2}$ *denotes the Euclidean norm and* $\|X\|_F = \sqrt{\sum_{i,j}(X_{ij})^2}$ *denotes the Frobenius norm.*

Define a random variable $\mathbf{q} = (1/n)\sum_{j=1}^n (p_j^+ + p_j^- - 1)^2$ such that $\mathbb{E}[\mathbf{q}] = q_k$. Then, the conditional expectation matrix $\mathbb{E}[LA^k(r)|t^k, p^+, p^-]$ has top singular value of

$$\begin{aligned}\frac{\ell}{n}\|t^k - \bar{t}^k\mathbf{1}\|\|p^+ + p^- - \mathbf{1}\| &= \frac{\ell}{n}\sqrt{n(1-(\bar{t}^k)^2)}\sqrt{n\mathbf{q}} \\ &= \sqrt{\ell^2\mathbf{q}(1-(\bar{t}^k)^2)}\end{aligned}$$

and the corresponding left and right singular vectors are $(1/\|t^k - \bar{t}^k\mathbf{1}\|)(t^k - \bar{t}^k\mathbf{1})$ and $(1/\|p^+ + p^- - \mathbf{1}\|)(p^+ + p^- - \mathbf{1})$. Before we apply the above lemma to this matrix together with $\mathcal{P}_1(LA^k(r))$, notice that we have two choices for the left singular vector. Both u and $-u$ are valid singular vectors of $\mathcal{P}_1(LA^k(r))$ and we do not know a priori which one is closer to $(t^k - \bar{t}^k\mathbf{1})$. For now, let us assume that u is the one closer to the correct solution, such that $\|u - (1/\|t^k - \bar{t}^k\mathbf{1}\|)(t^k - \bar{t}^k\mathbf{1})\| \le \|u + (1/\|t^k - \bar{t}^k\mathbf{1}\|)(t^k - \bar{t}^k\mathbf{1})\|$. Later in this section, we will explain how we can identify u with high probability of success. Then, from Lemma 3.2, we get

$$\begin{aligned}&\Big\|\frac{1}{\|t^k - \bar{t}^k\mathbf{1}\|}(t^k - \bar{t}^k\mathbf{1}) - u\Big\| \le \\ &\sqrt{\frac{2}{\ell^2\mathbf{q}(1-(\bar{t}^k)^2)}}\Big\|\mathbb{E}[LA^k(r)|t^k, p^+, p^-] - \mathcal{P}_1(LA^k(r))\Big\|_F .\end{aligned} \tag{16}$$

In the following we will prove that the Frobenius norm of the difference $\|\mathbb{E}[LA^k(r)|t^k, p^+, p^-] - \mathcal{P}_1(LA^k(r))\|_F$ is upper bounded by $C\sqrt{\ell}$ with probability at least $1 - n^{-C'\sqrt{\ell}}$ for some positive constants C and C'. Together with (16) and (15), this implies

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}\big(t^k)i \neq \text{sign}(u_i)\big) \le \frac{C}{\ell\mathbf{q}(1-|\bar{t}^k|)^2} .$$

Next, we use standard concentration inequalities to relate random quantities $\mathbf{q}$ and $\bar{t}^k$ to q_k and $\bar{s}^k$. By standard concentration results, we know that

$$\mathbb{P}\big(\mathbf{q} - q_k < -q_k/2\big) \le e^{-q_k^2 n/2} .$$

Hence, with probability at least $1 - e^{-q_k^2 n/2}$, we have $\mathbf{q} \ge q/2$. Similarly, for $\bar{t}^k = (1/n)\sum_i t_i^k$, and assuming without loss of generality that $\bar{s}^k = 2\theta_{>k} - 1$ is positive,

$$\begin{aligned}\mathbb{P}\big(1 - |\bar{t}^k| < (1/2)(1-\bar{s}^k)\big) &= \mathbb{P}\Big(|\frac{1}{n}\sum_i t_i^k| > (1/2)(1+|\bar{s}^k)\Big) \\ &\le 2e^{-n(1-\bar{s}^k)^2/8} .\end{aligned}$$

Hence, it follows that with probability at least $1 - 2e^{-n(1-\bar{s}^k)^2/8} - e^{-q_k^2 n/2} - n^{-C'\sqrt{\ell}}$,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}\big(t^k)i \neq \text{sign}(u_i)\big) \le \frac{C}{\ell q_k(1-|\bar{s}^k|)^2} .$$

This proves Lemma 3.1.

Now, we are left to prove an upper bound on the Frobenius norm in (16). Notice that for any matrix X of rank-2, $\|X\|_F \le \sqrt{2}\|X\|_2$, where $\|X\|_2 \equiv \max_{\|x\|,\|y\|\le 1} x^T X y$ denotes the operator norm. Therefore, by triangular inequity,

$$\begin{aligned}&\big\|\mathbb{E}[LA^k(r)|t^k, p^+, p^-] - \mathcal{P}_1(LA^k(r))\big\|_F \\ &\le \sqrt{2}\big\|\mathbb{E}[LA^k(r)|t^k, p^+ p^-] - \mathcal{P}_1(LA^k(r))\big\|_2 \\ &\le \sqrt{2}\big\|\mathbb{E}[LA^k(r)|t^k, p^+, p^-] - LA^k(r)\big\|_2 \\ &\quad + \sqrt{2}\big\|LA^k(r) - \mathcal{P}_1(LA^k(r))\big\|_2 \\ &\le 2\sqrt{2}\big\|\mathbb{E}[LA^k(r)|t^k, p^+, p^-] - LA^k(r)\big\|_2 \\ &\le 2\sqrt{2}\big\|\mathbb{E}[A^k(r)|t^k, p^+, p^-] - A^k(r)\big\|_2 ,\end{aligned} \tag{17}$$

where in the last inequity we used the fact that $\mathcal{P}_1(LA^k(r))$ is the minimizer of $\|LA^k(r) - X\|_2$ among all matrices X of rank one, whence $\|LA^k(r) - \mathcal{P}_1(LA^k(r))\|_2 \le \|LA^k(r) - \mathbb{E}[LA^k(r)|t^k, p^+, p^-]\|_2$.

The following key technical lemma provides a bound on the operator norm of the difference between random matrix $A^k(r)$ and its (conditional) expectation. This lemma generalizes a celebrated bound on the second largest eigenvalue of d-regular random graphs by Friedman-Kahn-Szemerédi [12, 11, 20]. The proof of this lemma is provided in Section 3.4.

LEMMA 3.3. *Assume that an* (ℓ, ℓ)*-regular random bipartite graph* G *with* n *left and right nodes is generated according to the configuration model.* $A^k(r)$ *is the weighted adjacency matrix of* G *with random weight* $A^k(r)_{ij}$ *assigned to each edge* $(i,j) \in E$*. With probability at least* $1 - n^{-\Omega(\sqrt{\ell})}$*,*

$$\|A^k(r) - \mathbb{E}[A^k(r)|t^k, p^+, p^-]\|_2 \le C' A_{\max}\sqrt{\ell} , \tag{18}$$

for all realizations of t^k*,* p^+*, and* p^-*, where* $|A^k(r)_{ij}| \le A_{\max}$ *almost surely and* C' *is a universal constant.*

Under our model, $A_{\max} = 1$ since $A^k(r)_{ij} \in \{\pm 1\}$. We then apply this lemma to each realization of p^+ and p^- and substitute this bound in (17). Together with (16) and (15), this finishes the proof Lemma 3.1.

Now, we are left to prove that between u and $-u$, we can determine which one is closer to $(1/\|t^k - \bar{t}^k\mathbf{1}\|)(t^k - \bar{t}^k\mathbf{1})$. Given a rank-1 matrix $\mathcal{P}_1(LA^k(r))$, there are two possible pairs of left and right 'normalized' singular vectors: (u, v) and $(-u, -v)$. Let $\mathcal{P}_+(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}^n$ denote the projection onto the positive orthant such that $\mathcal{P}_+(v)_i = \mathbb{I}(v_i \ge 0)v_i$. Our strategy is to choose u to be our estimate if $\|\mathcal{P}_+(v)\|^2 \ge 1/2$ (and $-u$ otherwise). We claim that with high probability the pair (u, v) chosen according to our strategy satisfies

$$\Big\|\frac{1}{\|t^k - \bar{t}^k\mathbf{1}\|}(t^k - \bar{t}^k\mathbf{1}) - u\Big\| \le \Big\|\frac{1}{\|t^k - \bar{t}^k\mathbf{1}\|}(t^k - \bar{t}^k\mathbf{1}) + u\Big\| . \tag{19}$$

Assume that the pair (u, v) is the one satisfying the above inequality. Denote the singular vectors of $\mathbb{E}[A^k(r)|t^k, p^+, p^-]$ by $x = (1/\|t^k - \bar{t}^k\mathbf{1}\|)(t^k - \bar{t}^k\mathbf{1})$ and $y = (1/\|p^+ + p^- - \mathbf{1}_n\|)(p^+ + p^- - \mathbf{1}_n)$, and singular value $\sigma' = \|\mathbb{E}[A^k(r)|t^k, p^+, p^-]\|_2$. Let $\sigma = \|\mathcal{P}_1(A^k(r))\|_2$. Then, by triangular inequality,

$$\begin{aligned}\|y - v\| &= \Big\|\frac{1}{\sigma'}\mathbb{E}[A^k(r)|t^k, p^+, p^-]^T x - \frac{1}{\sigma}\mathcal{P}_1(A^k(r))^T u\Big\| \\ &\le \Big\|\frac{1}{\sigma'}\mathbb{E}[A^k(r)|t^k, p^+, p^-]^T(x - u)\Big\| \\ &\quad + \Big\|\frac{1}{\sigma'}(\mathbb{E}[A^k(r)|t^k, p^+, p^-] - \mathcal{P}_1(A^k(r)))^T u\Big\| \\ &\quad + \Big\|\Big(\frac{1}{\sigma'} - \frac{1}{\sigma}\Big)\mathcal{P}_1(A^k(r))^T u\Big\| \\ &\le \frac{C_1}{\sqrt{\ell\mathbf{q}(1-(\bar{t}^k)^2)}} .\end{aligned}$$

The first term in the second line is upper bounded by

$$\|(1/\sigma')\mathbb{E}[A^k(r)|t^k,p^+,p^-]^T(x-u)\| \le \|x-u\| \ ,$$

which is again upper bounded by $C_2/(\ell q_k(1-(\bar{t}^k)^2))^{1/2}$ using (16). The second term is upper bounded by

$$\begin{aligned}&\|(1/\sigma')(\mathbb{E}[A^k(r)|t^k,p^+,p^-]-\mathcal{P}_1(A^k(r)))^T u\| \\ &\le (1/\sigma')\|\mathbb{E}[A^k(r)|t^k,p^+,p^-]-\mathcal{P}_1(A^k(r))\|_2 \ ,\end{aligned}$$

which is again upper bounded by $C_3/(\ell q_k(1-(\bar{t}^k)^2))^{1/2}$ using (18) and $\sigma' \ge \sqrt{(1/2)\ell^2 q_k(1-(\bar{t}^k)^2)}$. The third term is upper bounded by $\|\left(\frac{1}{\sigma'}-\frac{1}{\sigma}\right)\mathcal{P}_1(A^k(r))^T u\| \le |\sigma-\sigma'|/\sigma'$, which is again upper bounded by $C_4/(\ell q_k(1-(\bar{t}^k)^2))^{1/2}$ using the following triangular inequality:

$$\begin{aligned}&\frac{1}{\sigma'}\big|\|\mathbb{E}[A^k(r)|t^k,p^+,p^-]\|_2-\|\mathcal{P}_1(A^k(r))\|_2\big| \\ &\le \ \frac{1}{\sigma'}\|\mathbb{E}[A^k(r)|t^k,p^+,p^-]-\mathcal{P}_1(A^k(r))\|_2 \ .\end{aligned}$$

Since we assume that $p_j^+ + p_j^- \ge 1$, we have $y_j = p_j^+ + p_j^- - 1 \ge 0$ for all j. It follows that $\|y-\mathcal{P}_+(v)\| \le \|y-v\|$ for any vector v. This implies that

$$\begin{aligned}\|\mathcal{P}_+(v)\| \ &\ge \ \|y\| - \|y-\mathcal{P}_+(v)\| \\ &\ge \ 1-\|y-v\| \\ &\ge \ 1-\frac{C_1}{(\ell q(1-(\bar{t}^k)^2))^{1/2}} \ .\end{aligned}$$

Notice that we can increase the constant C in the bound (11) of the main theorem such that we only need to restrict our attention to $(\ell q(1-(\bar{t}^k)^2))^{1/2} > 4C_1$. This proves that the pair (u,v) chosen according to our strategy satisfy (19), which is all we need in order to prove Lemma 3.1.

3.3 Proof of Lemma 3.2

A more general statement for general low-rank matrices is proved in [20, Remark 6.3]. Here we provide a proof of a special case when both matrices have rank one. For two rank-1 matrices with singular value decomposition $M = x\sigma y^T$ and $M' = x'\sigma'(y')^T$, we want to upper bound $\min\{\|x+x'\|, \|x-x'\|\}$. Define the angle between the two vectors to be $\theta = \arccos(|x^T x'|)$ such that $\min\{\|x+x'\|, \|x-x'\|\} = 2\sin(\theta/2)$ and $\min_a \|x-ax'\| = \sin\theta$. It follows from $2\sin(\theta/2) = (1/\cos(\theta/2))\sin\theta \le \sqrt{2}\sin\theta$ for all $\theta \in [0,(1/2)\pi]$ that

$$\min\{\|x+x'\|, \|x-x'\|\} \ \le \ \sqrt{2}\min_a \|x-ax'\| \ .$$

Define the inner product of two vectors or matrices as $\langle A,B\rangle = \mathrm{Trace}(A^T B)$. We take $a^* = (\sigma'/\sigma)y^T y'$. Then,

$$\begin{aligned}\min\{\|x+x'\|, \|x-x'\|\} \ &\le \ \sqrt{2}\|x-a^*x'\| \\ &\le \ \max_{u\in\mathbb{R}^n,\|u\|\le 1}\sqrt{2}\langle u, x-a^*x'\rangle \\ &\le \ \max_{u\in\mathbb{R}^n,\|u\|\le 1}\sqrt{2}\langle u, x-(\sigma'/\sigma)y^T y' x'\rangle \\ &\le \ \max_{u\in\mathbb{R}^n,\|u\|\le 1}\sqrt{2}\langle u, (1/\sigma)\big(\sigma xy^T-\sigma' x'(y')^T\big)y\rangle \\ &\le \ \max_{u\in\mathbb{R}^n,\|u\|\le 1}(\sqrt{2}/\sigma)\langle uy^T, \sigma xy^T-\sigma' x'(y')^T\rangle \\ &\le \ \max_{u\in\mathbb{R}^n,\|u\|\le 1}(\sqrt{2}/\sigma)\|uy^T\|_F\,\|M-M'\|_F \\ &\le \ \frac{\sqrt{2}\,\|M-M'\|_F}{\sigma} \ .\end{aligned}$$

By symmetry, the same inequality holds with σ' in the denominator. This proves the desired claim.

3.4 Proof of Lemma 3.3

Since the proof does not depend on the specific realizations of t^k, p^+, and p^-, we will drop the conditions on these variables in this section and write $\mathbb{E}[A^k(r)]$ for $\mathbb{E}[A^k(r)|t^k,p^+,p^-]$. Define an ℓ_2-ball $\mathcal{B}_n \equiv \{x\in\mathbb{R}^n : \|x\|\le 1\}$ in n-dimensions. We want to show that, with high probability,

$$|x^T(A^k(r)-\mathbb{E}[A^k(r)])y| \le C' A_{\max}\sqrt{\ell} \ , \qquad (20)$$

for all $x\in\mathcal{B}_n$ and $y\in\mathcal{B}_n$. The technical challenge is that the left-hand side of (20) is a random variable indexed by x and y each belonging to a set with infinite number of elements. Our strategy, which is inspired by [12] and is similar to the techniques used in [11, 20], is as follows:
(i) Reduce x, y belonging to a finite discrete set $\mathcal{T}_n$;
(ii) Bound the contribution of *light couples* using concentration of measure result on a random variable

$$Z \equiv \sum_{(i,j)\in\mathcal{L}} x_i A^k(r)_{ij} y_j - x^T\mathbb{E}[A^k(r)]y$$

and applying union bound over (exponentially many but finite) choices of x and y;
(iii) Bound the contribution of *heavy couples* using *discrepancy property* of the random graph G.
The definitions of *light* and *heavy couples* and *discrepancy property* is provided later in this section.
Discretization. Fix some $\Delta\in(0,1)$ and define a discretization of $\mathcal{B}_n$ as

$$\mathcal{T}_n \equiv \left\{x\in\left\{\frac{\Delta}{\sqrt{n}}\mathbb{Z}\right\}^n \ : \ \|x\|\le 1\right\} \ .$$

Next proposition allows us to restrict our attention to discretized x and y and is proved in [12, 11].

PROPOSITION 3.4. *Let $M\in\mathbb{R}^{n\times n}$ be a matrix. If $|x^T My|\le B$ for all $x\in\mathcal{T}_n$ and $y\in\mathcal{T}_n$, then $|x'^T My'|\le(1-\Delta)^{-2}B$ for all $x'\in\mathcal{B}_n$ and $y'\in\mathcal{B}_n$.*

It is thus enough to show that the bound (20) holds with high probability for all $x,y\in\mathcal{T}_n$. A naive approach would be to apply tail bound on the random variable $\sum_{i,j} x_i(A^k(r)_{ij}-\mathbb{E}[A^k(r)_{ij}])y_j$. However, this approach fails when x or y have entries of value much larger than the typical size $O(n^{-1/2})$. Hence, we need to separate the contribution into two parts.

Define a set of *light couples* $\mathcal{L} \subseteq [n] \times [n]$ as

$$\mathcal{L} \equiv \left\{ (i,j) \,:\, |x_i y_j| \leq \frac{\sqrt{\ell}}{n} \right\} ,$$

and the set of heavy couples $\overline{\mathcal{L}}$ as its complement. Using this definition, we can separate the contribution from light and heavy couples.

$$\left| x^T (A^k(r) - \mathbb{E}[A^k(r)]) y \right| \;\leq\; \left| \sum_{(i,j)\in\mathcal{L}} x_i A^k(r)_{ij} y_j - x^T \mathbb{E}[A^k(r)] y \right| + \left| \sum_{(i,a)\in\overline{\mathcal{L}}} x_i A^k(r)_{ij} y_j \right| .$$

In the following, we prove that both of these contributions are upper bounded by $(C'/2) A_{\max} \sqrt{\ell}$ for all $x, y \in \mathcal{T}_n$. By Proposition 3.4, this finishes the proof of Lemma 3.3.

Bounding the contribution of light couples. Let $\mathbf{Z} = \sum_{(i,j)\in\mathcal{L}} x_i A^k(r)_{ij} y_j - x^T \mathbb{E}[A^k(r)] y$. Using the fact that $\mathbb{E}[\mathbf{Z}] = -\sum_{(i,j)\in\overline{\mathcal{L}}} x_i \mathbb{E}[A^k(r)_{ij}] y_j$, we get

$$\begin{aligned} |\mathbb{E}[\mathbf{Z}]| \;&\leq\; \sum_{(i,j)\in\overline{\mathcal{L}}} \frac{|\mathbb{E}[A^k(r)_{ij}]| (x_i y_j)^2}{|x_i y_j|} \\ &\leq\; \frac{n}{\sqrt{\ell}} \max_{i,j} |\mathbb{E}[A^k(r)_{ij}]| \;\leq\; A_{\max} \sqrt{\ell} \,, \end{aligned} \tag{21}$$

where, in the second inequality, we used $|x_i y_j| \geq \sqrt{\ell}/n$ for any $(i,j) \in \overline{\mathcal{L}}$, and the last inequality follows from the fact that $|\mathbb{E}[A^k(r)_{ij}]|$ is at most $(\ell/n) A_{\max}$. Together with the next lemma, this implies that when restricted to the discretized sets $\mathcal{T}_n$, the contribution of light couples is bounded by $C_5 A_{\max} \sqrt{\ell}$ with high probability.

LEMMA 3.5. *There exists numerical constants C_6 and C_7 such that, for any $x \in \mathcal{T}_n$ and $y \in \mathcal{T}_n$,*

$$\mathbb{P}\Big(\big| \mathbf{Z} - \mathbb{E}[\mathbf{Z}] \big| > C_6 A_{\max} \sqrt{\ell} \Big) \;\leq\; e^{-C_7 n} .$$

If the edges were selected independently, this lemma can be proved using routine tail bounds. In the case of (l,r)-regular graphs, we can use a martingale construction known as Doob's martingale process [25]. The proof follows closely the technique used in the proof of [12, Lemma 2.4], where an analogous statement is proved for unweighted non-bipartite random d-regular graphs. For the proof of this lemma, we refer to a journal version of this paper.

Cardinality of the discrete set $\mathcal{T}_n$ can be bounded using a simple volume argument: $|\mathcal{T}_n| \leq (10/\Delta)^n$ [11]. In Lemma 3.5, we can choose a large enough C_6 such that $C_7 > 2\log(10/\Delta)$. Applying union bound over all $x, y \in \mathcal{T}_n$, this proves that the contribution of light couples is bounded by $C_5 A_{\max} \sqrt{\ell}$ uniformly for all $x, y \in \mathcal{T}_n$ with probability $1 - e^{-\Omega(n)}$.

Bounding the contribution of heavy couples. Let $Q \in \{0,1\}^{m \times n}$ denote the standard (unweighted) adjacency matrix corresponding to the bipartite graph G. Then,

$$\left| \sum_{(i,j)\in\overline{\mathcal{L}}} x_i A_{ij} y_j \right| \;\leq\; A_{\max} \Big(\sum_{(i,j)\in\overline{\mathcal{L}}} Q_{ij} |x_i y_j| \Big) . \tag{22}$$

We can upper bound the right-hand side using discrepancy property of random graphs. It is a well-known result in graph theory that a random graph does not contain an unexpectedly dense subgraph with high probability. This discrepancy property plays an important role in the proof of structural properties such as expansion and spectrum of random graphs.

- *Bounded discrepancy.* We say that G (equivalently, the adjacency matrix Q) has bounded discrepancy property if, for any pair $L \subseteq [n]$ and $R \subseteq [n]$, (at least) one of the following is true. Here, $e(L,R) = |\{(i,j) \in E : i \in L, j \in R\}|$ denotes the number of edges between a subset L of left nodes and a subset R of right nodes, and $\mu(L,R) \equiv |L||R||E|/n^2$ denotes the average number of edges between L and R.

  (i) $e(L,R) \leq C_1 \mu(L,R)$,

  (ii) $|L|$, $|R|$, and $e(L,R)$ are all at most $C_2\sqrt{\ell}$,

  (iii) $e(L,R) \log\Big(\frac{e(L,R)}{\mu(L,R)}\Big) \leq C_3 \max\{|L|,|R|\} \log\Big(\frac{n}{\max\{|L|,|R|\}}\Big)$,

  for some constants C_1, C_2, and C_3 which only depend on m/n.

- *Bounded degree.* The graph G has degree of the left nodes bounded by ℓ and the right nodes also by ℓ.

For a random (ℓ,ℓ)-regular graph G, the bounded degree property is always satisfied. Next lemma shows that G also satisfies the discrepancy property [12, Lemma 2.5].

LEMMA 3.6. *For an (ℓ,ℓ)-regular random graph G, with probability $1 - n^{\Omega(\sqrt{\ell})}$, every pair $L \subseteq [n]$ and $R \subseteq [n]$ satisfies the bounded discrepancy property.*

Together with (22) and Lemma 3.6, the next lemma implies that the contribution of heavy couples is upper bounded by $C_4 A_{\max} \sqrt{\ell}$ with probability $1 - n^{\Omega(\sqrt{\ell})}$.

LEMMA 3.7. *If a bipartite graph G satisfies the bounded degree and bounded discrepancy properties, then there exists a positive constant C_4 such that for any $x, y \in \mathcal{T}_n$ the adjacency matrix Q satisfy*

$$\sum_{(i,j)\in\overline{\mathcal{L}}} |x_i Q_{ij} y_j| \leq C_4 \sqrt{\ell} .$$

A similar statement was proved in [20, Remark 4.5] for Erd os-Renyi graph. Due to a space constraint, the proof is omitted here.

3.5 Proof of Theorem 2.2

The proof follows closely the techniques we used in proving Theorem 2.1. In fact, the only difference is in the following key technical lemma. This lemma improves the upper bound in Lemma 3.1 by a factor of $(1 - |\bar{s}^k|)^2$. Once we have this, we can use the same arguments as in Section 3.1 to get the improved bound on error probability as in Theorem 2.2.

LEMMA 3.8. *There exists positive numerical constants C and C' such that the estimates of* (8) *achieve*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\big(\hat{t}_i^k(r) \neq t_i^k \big) \;\leq\; \frac{C}{\ell q_k} \tag{23}$$

with probability at least $1 - 2e^{-n(1-|\bar{s}^k|)^2/8} - e^{-q_k^2 n/2} - n^{-C'\sqrt{\ell}}$ where $\bar{s}^k$ and q_k are parameters defined earlier.

PROOF. In proving Theorem 2.1, we are using the left singular vectors of $A^k(r)$, which we denote by $u^k(r)$. Assuming for simplicity that $\sum_{j:v_j^k(r)\geq 0}(v_j^k(r))^2 \geq 1/2$, we make a decision on whether t_i^k is more likely to be a positive task or a negative task based on whether $u^k(r)_i$ is positive or negative: $\hat{t}^k(r)_i = \text{sign}(u^k(r)_i)$. Effectively, we are using a *threshold* of zero to classify the tasks using $u^k(r)_i$ as a noisy observation of that task.

If we have a good estimate of $\bar{s}^k$, we can use it to get a better threshold of $-\bar{s}^k/\sqrt{(1-(\bar{s}^k)^2)n}$. This gives a new decision rule of (8): $\hat{t}^k(r)_i = \text{sign}\big(\, u^k(r)_i + \frac{\bar{s}^k}{\sqrt{(1-(\bar{s}^k)^2)n}}\,\big)$, when $\sum_{j:v_j^k(r)\geq 0}(v_j^k(r))^2 \geq 1/2$. We can prove that with this new threshold, the bound on the number of errors improve by a factor of $(1-|\bar{s}^k|)^2$. The analysis of follows closely the proof of Lemma 3.1.

Let u be the left singular vector of $LA^k(r)$ corresponding to the leading singular value. From Section 3.2, we know that the Euclidean distance between u and $(1/\|t^k - \bar{t}^k\mathbf{1}\|)(t^k - \bar{t}^k\mathbf{1})$ is upper bounded by $\sqrt{C/(\ell q_k(1-(\bar{t}^k)^2))}$. We can use the following series of inequalities to related this bound to the average number of errors. From $\|t^k - \bar{t}^k\mathbf{1}\| = \sqrt{n(1-(\bar{t}^k)^2)}$ which follows from the definition $\bar{t}^k = (1/n)\sum_i t_i^k$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_i \mathbb{I}\Big(t_i^k \neq \text{sign}\big(\, u_i + \frac{\bar{t}^k}{\sqrt{(1-(\bar{t}^k)^2)n)}}\,\big)\Big) \\
&\leq\ \frac{1}{n}\sum_i \mathbb{I}\Big(t_i^k\big(u_i + \frac{\bar{t}^k}{\sqrt{(1-(\bar{t}^k)^2)n)}}\big) \leq 0\Big) \\
&\leq\ \frac{1}{n}\sum_i \big(1-(\bar{t}^k)^2\big)\Big(\sqrt{n}u_i - \frac{t_i^k - \bar{t}^k}{\sqrt{1-(\bar{t}^k)^2}}\Big)^2 \\
&=\ \big(1-(\bar{t}^k)^2\big)\Big\|u - \frac{t^k - \bar{t}^k\mathbf{1}}{\|t^k - \bar{t}^k\mathbf{1}\|}\Big\|^2 \quad\leq\ \frac{C}{\ell q_k}\,.
\end{aligned}$$

□

3.6 Proof of Theorem 2.3

Let us focus on a dataset $A(r)$ that is collected from n workers, αn of which are are adversarial. Since we assign tasks according to a random graph, the performance does not depend on the indices assigned to particular workers, and hence we let the first αn workers be the adversarial ones. Let $\tilde{A}^k(r) \in \{\mathsf{null}, +1, -1\}^{n\times n}$ be the $n \times n$ matrix of 'quantized' answers when adversaries are present. Define a random matrix $A^k(r)$ to be the answers we would get on the same graph and same set of non-adversarial workers for the last $(1-\alpha)n$ workers but this time replacing all of the adversarial workers with randomly chosen non-adversarial workers with confusion matrix drawn from $\mathcal{D}$. The dataset $A^k(r)$ represent the answers we would get if there were no adversaries. We want to bound the effect of adversaries on our estimate, by bounding difference, in spectral norm, between the random matrix $\tilde{A}^k(r)$ and conditional expectation of non-adversarial answers $\mathbb{E}[A^k(r)|t^k, p_k^+, p_k^-]$. We claim that

$$\big\|\, \tilde{A}^k(r) - \mathbb{E}[A^k(r)|t^k, p_k^+, p_k^-]\,\big\|_2 \quad\leq\quad C(\sqrt{\ell} + \ell\sqrt{\alpha}) \quad (24)$$

Once we have this bound, we can finish our analysis following closely the proof of Theorem 2.1. Let u be the top left singular vector of the matrix $\tilde{A}^k(r)$. From Section 3.2, we know that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}(t_i^k \neq \text{sign}(u_i)) \quad\leq\quad \frac{16\,\big\|\, \tilde{A}^k(r) - \mathbb{E}[A^k(r)|t^k, p_k^+, p_k^-]\,\big\|_2^2}{\ell^2 q_k(1-|\bar{s}^k|)^2}$$

Substituting (24) into the above, and using the fact that $(a+b)^2 \leq 2a^2 + 2b^2$, we get

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}(t_i^k \neq \text{sign}(u_i)) \quad\leq\quad \frac{C\,(1+\alpha\ell)}{\ell q_k(1-|\bar{s}^k|)^2}\,.$$

Further, in a similar way as we proved Theorem 2.2, if we have a good estimate of $\bar{s}^k$, then we can find a better threshold as in (8) and improve the upper bound as

$$\frac{1}{n}\sum_{i=1}^n \mathbb{I}\Big(t_i^k \neq \text{sign}\big(u_i + \frac{\bar{t}^k}{\sqrt{(1-(\bar{t}^k)^2)n)}}\big)\Big) \quad\leq\quad \frac{C\,(1+\alpha\ell)}{\ell q_k}\,.$$

Substituting this bound in Lemma 3.1 from Section 3.1 and following the same argument, this finishes the proof of Theorem 2.3.

Now, we are left to prove the upper bound in (24). Let $\tilde{A}_a^k(r)$ denote the answers from adversarial workers and $\tilde{A}_g^k(r)$ denote the answer from the non-adversarial workers such that $\tilde{A}^k(r) = [\,\tilde{A}_a^k(r)\ \tilde{A}_g^k(r)\,]$. Let $B = \mathbb{E}[A^k(r)|t^k, p_k^+, p_k^-]$ be the conditional expectation of non-adversarial data. We use B_0 to denote the first αn columns of B and B_1 for the last $(1-\alpha)n$ columns of B, such that $B = [\,B_0\ B_1\,]$. By the triangular inequality, we get

$$\begin{aligned}
\big\|\, \tilde{A}^k(r) - B\,\big\|_2 &= \big\|\, [\,\tilde{A}_a^k(r)\ \tilde{A}_g^k(r)\,] - [\,B_0\ B_1\,]\,\big\|_2 \\
&\leq \big\|\, \tilde{A}_g^k(r) - B_1\,\big\|_2 + \big\|\, B_0\,\big\|_2 + \big\|\, \tilde{A}_a^k(r)\,\big\|_2\,.
\end{aligned}$$

To upper bounded the first term, notice that it is a projection of a $n \times n$ matrix where the projection P sets the first αn columns to zeros: $\tilde{A}_g^k(r) - B_1 = P(\tilde{A}^k(r) - B)$. Since a projection can only decrease the spectral radius, we have $\big\|\tilde{A}_g^k(r) - B_1\big\|_2 \leq \big\|\tilde{A}^k(r) - B\big\|_2$. From Lemma 3.3 we know that this is upper bounded by $C\sqrt{\ell}$. For the second term, recall that $B_0 = (\ell/n)\mathbf{1}_n\mathbf{1}_{\alpha n}^T$. This gives $\|B_0\|_2 = \ell\sqrt{\alpha}$.

To upper bound the last term, let M_α be an $n \times \alpha n$ matrix that have the same pattern as $\tilde{A}_a^k(r)$, but all the nonzero entries are set to one. Statistically, this is the first αn columns of the adjacency matrix of a random (ℓ,ℓ)-regular graph. Since M_α is a result of taking the absolute value of the matrix $\tilde{A}_a^k(r)$, we have $\big\|\, \tilde{A}_a^k(r)\,\big\|_2 \leq \big\|M_\alpha\big\|_2$. By triangular inequality, $\big\|M_\alpha\big\|_2 \leq \big\|\, E[M_\alpha]\,\big\|_2 + \big\|\, M_\alpha - E[M_\alpha]\,\big\|_2$. The first term is bounded by $\big\|\, E[M_\alpha]\,\big\|_2 = \|B_0\|_2 = \ell\sqrt{\alpha}$. To bound the second term, we use the same technique of projection. Let M be an $n \times n$ adjacency matrix of a random (ℓ,ℓ)-regular graph such that the first αn columns are equal to M_α. Then, $\big\|\, A_{M_\alpha} - E[M_\alpha]\,\big\|_2 \leq \|M - \mathbb{E}[M]\|_2$. Friedman, Kahn, and Szemeredie [12] proved that this is upper bounded by $C\sqrt{\ell}$ with probability at least $1 - n^{-C'\ell}$. Collecting all the terms, this proves the upper bound in (24).

4. CONCLUSIONS

In this paper, we considered the question of designing crowd-sourcing platform with the aim of obtaining best trade-off between reliability and redundancy (equivalently, budget). Operationally, this boiled down to developing appropriate task allocation (worker-task assignment) and estimation of task answers from noisy responses of workers. We

presented task allocation based on random regular bipartite graph and estimation algorithm based on low-rank approximation of appropriate matrices. We established that the design we have presented achieves (order-)optimal performance for the generic model of crowd-sourcing (cf. model considered by Dawid and Skene [9]).

Ours is the first rigorous result for crowd-sourcing system design for generic K-ary tasks with general noise model. The algorithms presented are entirely data-driven and hence useful for the setting even when the precise probabilistic model is not obeyed.

One limitation of the current model is that the tasks are assumed to be equally difficult. It is of great practical interest to accommodate differences in task difficulties. Our approach exploits low-rank structure inherent in the probabilistic model studied in this paper. However, more general models proposed in crowdsourcing literature lack such low-rank structures, and it is an interesting future research direction to understand the accuracy-redundancy trade-offs for more general class of models.

Acknowledgements

DK is supported in parts by a grant from the National Science Foundation. DS is supported in parts by Army Research Office under MURI Award 58153-MA-MUR.

5. REFERENCES

[1] Casting Words. http://castingwords.com.

[2] Crowd Flower. http://crowdflower.com.

[3] Crowd Spring. http://www.crowdspring.com.

[4] ESP game. http://www.espgame.org.

[5] Soylent. http://projects.csail.mit.edu/soylent/.

[6] M. S. Bernstein, J. Brandt, R. C. Miller, and D. R. Karger. Crowds in two seconds: enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 33–42, 2011.

[7] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, ACM UIST, pages 313–322, New York, NY, USA, 2010.

[8] B. Bollobás. *Random Graphs*. Cambridge University Press, Jan. 2001.

[9] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

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

[11] U. Feige and E. Ofek. Spectral techniques applied to sparse random graphs. *Random Struct. Algorithms*, 27(2):251–275, 2005.

[12] J. Friedman, J. Kahn, and E. Szemerédi. On the second eigenvalue in random regular graphs. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, pages 587–598, Seattle, Washington, USA, may 1989. ACM.

[13] A. Ghosh, S. Kale, and P. McAfee. Who moderates the moderators?: crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 167–176. ACM, 2011.

[14] S. Holmes. Crowd counting a crowd. March 2011, Statistics Seminar, Stanford University.

[15] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67. ACM, 2010.

[16] R. Jin and Z. Ghahramani. Learning with multiple labels. In *Advances in neural information processing systems*, pages 921–928, 2003.

[17] D. R. Karger, S. Oh, and D. Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Proc. of the Allerton Conf. on Commun., Control and Computing*, 2011.

[18] D. R. Karger, S. Oh, and D. Shah. Budget-optimal task allocation for reliable crowd sourcing systems. 2011. http://arxiv.org/abs/1110.3564.

[19] D. R. Karger, S. Oh, and D. Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961, 2011.

[20] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, June 2010.

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

[22] V. C. Raykar and S. Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 13:491–518, mar 2012.

[23] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *J. Mach. Learn. Res.*, 99:1297–1322, August 2010.

[24] T. Richardson and R. Urbanke. *Modern Coding Theory*. Cambridge University Press, march 2008.

[25] E. Shamir and J. Spencer. Sharp concentration of the chromatic number on random graphs $G_{n,p}$. *Combinatorica*, 7:121–129, 1987.

[26] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 614–622. ACM, 2008.

[27] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, pages 2424–2432, 2010.

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