Optimal Inference in Crowdsourced Classification via Belief Propagation

Jungseul Ok, *Member, IEEE,* Sewoong Oh *Member, IEEE,* Jinwoo Shin, *Member, IEEE,* and Yung Yi, *Member, IEEE*

***Abstract*—Crowdsourcing systems are popular for solving large-scale labelling tasks with low-paid workers. We study the problem of recovering the true labels from the possibly erroneous crowdsourced labels under the popular Dawid-Skene model. To address this inference problem, several algorithms have recently been proposed, but the best known guarantee is still significantly larger than the fundamental limit. We close this gap by introducing a tighter lower bound on the fundamental limit and proving that Belief Propagation (BP) exactly matches this lower bound. The guaranteed optimality of BP is the strongest in the sense that it is information-theoretically impossible for any other algorithm to correctly label a larger fraction of the tasks. Experimental results suggest that BP is close to optimal for all regimes considered and improves upon competing state-of-the-art algorithms.**

***Index Terms*—Crowdsourcing, Belief Propagation, Optimal Inference**

I. Introduction

CROWDSOURCING platforms provide scalable human-powered solutions to labelling large-scale datasets at minimal cost. They are particularly popular in domains where the task is easy for humans but hard for machines, e.g., computer vision and natural language processing. For example, the CAPTCHA system [2] uses a pair of scanned images of English words, one for authenticating the user and the other for the purpose of getting high-quality character recognitions to be used in digitizing books. However, because the tasks are tedious and the pay is low, one of the major issues is the quality of the labels. Errors are common even among those who put in efforts. In real-world systems, spammers are abundant, who submit random answers rather than good-faith attempts to label. There are adversaries deliberately giving wrong answers.

A common and powerful strategy to improve reliability is to add redundancy: assigning each task to multiple workers and aggregating their answers by some algorithm such as majority voting. Although majority voting is widely used in practice, several novel approaches, which outperform majority voting, have been recently proposed, e.g. [3]–[7]. The key idea is to identify the good workers and give more weights to the answers from those workers. Although the ground truths may never be exactly known, one can compare one worker's answers to those from other workers on the same tasks, and infer how reliable or trustworthy each worker is.

The standard probabilistic model for representing the noisy answers in labelling tasks is the model introduced by Dawid and Skene in [8]. Under this model, the core problem of interest is how to aggregate the answers to maximize the accuracy of the estimated labels. This is naturally posed as a statistical inference problem that we call the *crowdsourced classification* problem. Due to the combinatorial nature of the problem, the Maximum A Posteriori (MAP) estimate is optimal but computationally intractable. Several algorithms have recently been proposed as approximations, and their performances are demonstrated only by numerical experiments. These include algorithms based on spectral methods [9]–[13], Belief Propagation (BP) [14], expectation maximization [14], [15], maximum entropy [16], [17], weighted majority voting [18]–[20], and combinatorial approaches [21].

Despite the algorithmic advances, theoretical advances have been relatively slow. Some upper bounds on the performances are known [11], [15], [21], but fall short of answering which algorithm should be used in practice. In this paper, we ask the fundamental question of whether it is possible to achieve the performance of the optimal MAP estimator with a computationally efficient inference algorithm. In other words, we investigate the computational gap between what is information-theoretically possible and what is achievable with a polynomial time algorithm.

Our main result is that there is no computational gap in the crowdsourced classification problem for a broad range of problem parameters. Under some mild assumptions on the parameters of the problem, we show the following:

Belief propagation is exactly optimal.

To the best of our knowledge, our algorithm is the only computationally efficient approach that provably maximizes the fraction of correctly labeled tasks, achieving exact optimality.

Contribution. We consider binary classification tasks and identify regimes where the standard BP achieves the performance of the optimal MAP estimator. When each task is assigned enough number of workers, we prove that it is impossible for any other algorithm to correctly label a larger fraction of tasks than BP. This is the only known algorithm to achieve such a strong notion of optimality and settles the question of whether there is a computational gap in the crowdsourced classification problem for a broad range of

Jungseul Ok is with the Department of Automatic Control at Royal Institute of Technology KTH, Stockholm, Sweden (e-mail: jungseul@kth.se). Jinwoo Shin and Yung Yi are with the Department of Electrical Engineering at Korea Advanced Institute of Science and Technology, Daejeon, South Korea (e-mail: {jinwoos, yiyung}@kaist.ac.kr). Sewoong Oh is with the Department of Industrial and Enterprise Systems Engineering at University of Illinois at Urbana-Champaign, Illinois, USA (e-mail: swoh@illinois.edu).

This article is partially based on preliminary results [1] presented in the 33rd International Conference on Machine Learning (ICML 2016).

parameters. We provide experimental results confirming the optimality of BP for both synthetic and real datasets.

The provable optimality of BP-based algorithms in graphical models with loops (such as those in our model) is known only in a few instances including community detection [22], error correcting codes [23] and combinatorial optimization [24]. Technically, our proof strategy for the optimality of BP is similar to that in [22] where another variant of BP algorithm is proved to be optimal to recover the latent community structure among users. However, our proof technique overcomes several unique challenges, arising from the complicated correlation among tasks that can only be represented by weighted and directed hyper-edges, as opposed to simpler unweighted undirected edges in the case of stochastic block models. This might be of independence interest in analyzing censored block models with some directed observations [25], and clustering from hyper-edge information [26].

The crowdsourced classification problem has been first studied in the *dense regime*, where all tasks are assigned all the workers [9], [15]. In such dense regimes, as the problem size increases, each task receives increasing number of answers. Thus, previous work has studied the probability of labelling all tasks correctly [9], [15]. However, in practice, the number of workers per task (or tasks per worker), denoted by ℓ (or r, resp.), is relatively small comparing to the total number of tasks since the budget per task and the capacity of a worker are constrained. For example, a typical choice of ℓ is three or five. For such a fixed ℓ, i.e., *sparse regime*, the probability of error does not decay with increasing dimension of the problem. The theoretical interest has been focused on identifying how the error scales with ℓ, that represents how much redundancy should be introduced in the system. An upper bound that scales as $e^{-\Omega(\ell)}$ (when $\ell > \ell^*$ for some ℓ^* that depends on the problem parameters) was proved by [11], analyzing a spectral algorithm that is modified to use the spectral properties of the non-backtracking operators instead of the usual adjacency matrices. This scaling order is also shown to be optimal by comparing it to the error rate of an oracle estimator. Tighter bounds are also provided for other spectral methods, but under more restricted conditions in [10], [27].

In this paper, we focus on the *sparse regime* with small ℓ, r. More precisely, we show that for any given $r \geq 1$[1] and some constant C_r depending on r, if $\ell > C_r$, i.e., ℓ is possibly constant, BP is *information-theoretically* optimal. This coincides with the *empirical* study in [14], where BP outperforms other algorithms including the state-of-the-art spectral approach proposed in [11]. In fact, the authors of [14] showed that the spectral approach in [11] is a special case of BP with a specific choice of the prior distribution on the worker quality. This implies that the spectral approach is suboptimal since the algorithmic prior might be in mismatch with the true prior. Since the true prior is often unavailable in practice, we propose a practical version of BP, which we call EBP (Estimation and Belief Propagation) that has an additional procedure estimating the prior from the observed data. Our experimental result suggests that both EBP and BP with the true prior closely achieve the optimal performance for all ℓ, r, although we show the optimality of BP in the certain regime of ℓ, r.

Organization. In Section II, we provide necessary backgrounds including the Dawid-Skene model for crowdsourced classification and the BP algorithm. Section III provides the main results of this paper, and their proofs are presented in Section IV. Our experimental results on the performance of BP are reported in Section VI and we conclude in Section VII.

II. Preliminaries

We describe the mathematical model and present the standard MAP and the BP approaches.

A. Crowdsourced Classification Model

We consider a set of n binary tasks, denoted by V. Each task $i \in V$ is associated with a ground truth $s_i \in \{-1, +1\}$. Without loss of generality, we assume s_i's are independently chosen with equal probability. We let W denote the set of workers who are assigned tasks to answer. Hence, this task assignment is represented by as a bipartite graph $G = (V, W, E)$, where edge $(i, u) \in E$ indicates that task i is assigned to worker u. For notational simplicity, let $N_u := \{i \in V : (i, u) \in E\}$ denote the set of tasks assigned to worker u and conversely let $M_i := \{u \in W : (i, u) \in E\}$ denote the set of workers to whom task i is assigned.

When task i is assigned to worker u, worker u provides a binary answer $A_{iu} \in \{-1, +1\}$, which is a noisy assessment of the true label s_i. Each worker u is parameterized by a *reliability* $p_u \in [0, 1]$, such that each of her answers is correct with probability p_u. Namely, for given $p := \{p_u : u \in W\}$, the answers $A := \{A_{iu} : (i, u) \in E\}$ are independent random variables such that

$$A_{iu} = \begin{cases} s_i & \text{with probability} \quad p_u \\ -s_i & \text{with probability } 1 - p_u \end{cases}.$$

We assume that the average reliability is greater than $1/2$, i.e., $\mu := \mathbb{E}[2p_u - 1] > 0$.

This Dawid-Skene model is the most popular one in crowdsourcing dating back to [8]. The underlying assumption is that all the tasks share a homogeneous difficulty; the error probability of a worker is consistent across all tasks. We assume that the reliability p_u's are i.i.d. according to a *reliability distribution* on $[0, 1]$, described by a probability density function π.

For the theoretical analysis, we assume that the bipartite graph is drawn uniformly over all (ℓ, r)-regular graphs for some small ℓ, r using, for example, the configuration model [28], where each task is assigned to ℓ random workers and each worker is assigned r random tasks. In real-world crowdsourcing systems, the designer gets to choose which graph to use for task assignments. Random regular graphs have been proven to achieve minimax optimal performance in [11], and empirically shown to have good performances. This is due to the fact that the random graphs have large spectral gaps.

[1]In our previous work [1], we showed the optimality of BP for $r = 1$ or 2. In this paper, we extend it for $r \geq 1$, where a worker's answers correspond to a *hyper*-edge information on multiple tasks.

B. MAP Estimator

Under this crowdsourcing model with given assignment graph $G=(V,W,E)$ and reliability distribution π, our goal is to design an efficient estimator $\hat{s}(A)\in\{-1,+1\}^V$ of the unobserved true answers $s:=\{s_i : i\in V\}$ from the noisy answers A reported by workers. In particular, we are interested in the optimal estimator minimizing the (expected) average bit-wise *error rate*, i.e.,

$$\underset{\hat{s}:\text{estimator}}{\text{minimize}} \quad P_{\mathsf{err}}(\hat{s}) \tag{1}$$

where we define

$$P_{\mathsf{err}}(\hat{s};G) := \frac{1}{n}\sum_{i\in V}\mathbb{P}\left[s_i \neq \hat{s}_i\left(A\right) \mid G\right] \ .$$

The probability here is taken with respect to the conditional distribution of s, $\hat{s}$ and A given G. For simplicity, we often omit G in the argument of P_{err} if it is clear from context. From standard Bayesian arguments, the maximum a posteriori (MAP) estimator is an optimal solution of (1):

$$\hat{s}^*_i(A) := \arg\max_{s_i}\ \mathbb{P}[s_i \mid A] \ . \tag{2}$$

However, this MAP estimator is challenging to compute, as we show below. Note that

$$\begin{aligned}\mathbb{P}[s,p\mid A] &\propto \mathbb{P}[p]\cdot\mathbb{P}[A\mid s,p]\\ &= \prod_{u\in W}\mathbb{P}[p_u]\prod_{i\in N_u}\mathbb{P}[A_{iu}\mid s_i,p_u]\\ &= \prod_{u\in W}\pi(p_u)\cdot p_u^{c_u}(1-p_u)^{r_u-c_u}\ ,\end{aligned} \tag{3}$$

where $r_u := |N_u|$ is the number of the tasks assigned to worker u and $c_u := |\{i\in N_u : A_{iu}=s_i\}|$ is the number of the correct answers from worker u. Then,

$$\begin{aligned}\mathbb{P}[s\mid A] &= \int_{[0,1]^W}\mathbb{P}[s,p\mid A]dp\\ &\propto \prod_{u\in W}\underbrace{\int_0^1 \pi(p_u)\cdot p_u^{c_u}(1-p_u)^{r_u-c_u}dp_u}_{:=f_u(s_{N_u})}\ ,\end{aligned} \tag{4}$$

where we let $f_u\left(s_{N_u}\right) := \mathbb{E}[p_u^{c_u}(1-p_u)^{r_u-c_u}]$ denote the local factor associated with worker u for given G. We note that the factorized form of the joint probability of s in (4) corresponds to a standard graphical model with a *factor graph* $G=(V,W,E)$ that represents the joint probability of s given A, where each task $i\in V$ and each worker $u\in W$ correspond to the random variable s_i and the local factor f_u, respectively, and the edges in E indicate couplings among the variables and the factors.

The marginal probability $\mathbb{P}[s_i\mid A]$ in the optimal estimator $\hat{s}^*_i(A)$ is calculated by marginalizing out $s_{\text{-}i} := \{s_j : i\neq j\in V\}$ from (4), i.e.,

$$\mathbb{P}[s_i\mid A] \ = \sum_{s_{\text{-}i}\in\{\pm1\}^{V\setminus i}}\mathbb{P}[s\mid A] \ \propto\ \sum_{s_{\text{-}i}}\prod_{u\in W}f_u\left(s_{N_u}\right) \ . \tag{5}$$

We note that the summation in (5) is taken over exponentially many $s_{\text{-}i}\in\{-1,+1\}^{n-1}$ with respect to n. Thus in general, the optimal estimator $\hat{s}^*$, which requires to obtain the marginal probability of s_i given A in (2), is *computationally intractable* due to the exponential complexity in (5).

C. Belief Propagation

Recalling the factor graph described by (4), the computational intractability in (5) motivates us to use a standard sum-product belief propagation (BP) algorithm on the factor graph as a heuristic method for approximating the marginalization. The BP algorithm is described by the following iterative update of messages $m_{i\to u}$ and $m_{u\to i}$ between task i and worker u and belief b_i on each task i:

$$m^{t+1}_{i\to u}(s_i) \propto \prod_{v\in M_i\setminus\{u\}} m^t_{v\to i}(s_i)\ , \tag{6}$$

$$m^{t+1}_{u\to i}(s_i) \propto \sum_{s_{N_u\setminus\{i\}}} f_u(s_{N_u}) \prod_{j\in N_u\setminus\{i\}} m^{t+1}_{j\to u}(s_j)\ , \tag{7}$$

$$b^{t+1}_i(s_i) \propto \prod_{u\in M_i} m^{t+1}_{u\to i}(s_i)\ , \tag{8}$$

where the belief $b_i(s_i)$ is the estimated marginal probability of s_i given A. We here initialize messages with a trivial constant $\frac{1}{2}$ and normalize messages and beliefs, i.e., $\sum_{s_i} m_{i\to u}(s_i) = \sum_{s_i} m_{u\to i}(s_i) = \sum_{s_i} b_i(s_i) = 1$. Then at the end of k iterations, we estimates the label of task i as follows:

$$\hat{s}^{\mathsf{BP}(k)}_i = \arg\max_{s_i}\ b^k_i(s_i)\ . \tag{9}$$

We note that if the factor graph is a tree, then it is known that the belief converges, and computes the exact marginal probability [29].

Property 1. *If assignment graph G is a tree so that the corresponding factor graph is a tree as well, then*

$$b^t_i(s_i) = \mathbb{P}[s_i\mid A] \quad \textit{for all } t\geq n\ ,$$

where $b^t_i(s_i)$ is iteratively updated by BP in (6)–(8).

However, for general graphs which may have loops, e.g., random (ℓ,r)-regular graphs, BP has no performance guarantee, i.e., BP may output $b_i(s_i)\neq\mathbb{P}[s_i\mid A]$. Further the convergence of BP is not guaranteed, i.e., the value of $\lim_{t\to\infty} b^t_i(s_i)$ may not exist.

III. Performance Guarantees of BP

In this section, we provide the theoretical guarantees on the performance of BP. To this end, we consider the output of BP in (9) with a choice of $k=O(\log\log n)$. It follows that the overall complexity of BP is bounded by $O(n\ell r\log r\cdot\log\log n)$ as each iteration of BP requires $O(n\ell r\log r)$ operations [14].

A. Exact Optimality of BP for large ℓ

We show in the following that BP is asymptotically optimal when each task is assigned to sufficiently large (but possibly constant with respect to the number of tasks) number of workers, i.e., $\ell > C_{r,\pi}$. This follows from a tighter bound in the non-asymptotic regime, where we upper bound the optimality gap, exponentially vanishing in the number of iterations k.

Theorem 1. *Consider the Dawid-Skene model under the task assignment generated by a random bipartite (ℓ, r)-regular graph G consisting of n tasks and $(\ell/r)n$ workers. Let $\hat{s}^{\mathsf{BP}(k)}$ denote the output of BP in* (9) *after k iterations. For $\mu := \mathbb{E}[2p_u - 1] > 0$, $\mathbb{E}[(2p_u-1)^2] < 1$, and constant $r \geq 1$, there exists a constant $C_{r,\pi}$ that only depends on r and π such that if $\ell \geq C_{r,\pi}$, then for sufficiently large n:*

$$\mathbb{E}\left[P_{\mathsf{err}}(\hat{s}^{\mathsf{BP}(k)}) - \min_{\hat{s}:\textit{estimator}} P_{\mathsf{err}}(\hat{s})\right] \;\leq\; 2^{-k} + \frac{3(\ell r)^{2k+1}}{n} \;, \tag{10}$$

where the expectation is taken with respect to the graph G. Hence, when we set $k = \log\log n$ with constant r and $\ell = O(\log n)$, for sufficiently large n,

$$\mathbb{E}\left[P_{\mathsf{err}}(\hat{s}^{\mathsf{BP}(k)}) - \min_{\hat{s}:\textit{estimator}} P_{\mathsf{err}}(\hat{s})\right] \;\leq\; 2^{-k+1} \;, \tag{11}$$

which converges to 0 as $n \to \infty$.

A proof is provided in Section IV-A. Our analysis compares BP to an *oracle* estimator. This oracle estimator not only has access to the observed crowdsourced labels, but also the ground truths of a subset of tasks. Given this extra information, it performs the optimal estimation, outperforming any algorithm that operates only on the observations. Using the fact that the random (ℓ, r)-regular bipartite graph has a *locally tree-like structure* [28] and BP is exact on the local tree [29], we prove that the performance gap between BP and the oracle estimator vanishes due to *decaying correlation* from the information on the outside of the local tree to the root. To be specific, the decaying correlation and the probability to have local tree within the neighborhood of depth k are captured in the first and second terms in the RHS of (10), respectively. This establishes that the gap between BP and the best estimator vanishes in the large system limit under some condition on ℓ, r, k.

Although empirically, BP works well in all regimes of parameters as suggested in Section VI, for the theoretical analysis, we limit the parameters to verify (i) $(\ell r)^k = o(n)$ for the locally tree-like graph structure, and (ii) $\ell > C_{r,\pi}$ with fixed r for the decaying correlation. The locally tree-like structure is essential also in other applications such as community detection [22], since it provides the conditional independence between two consecutive generations of the computation tree. When $r = 1$, there is nothing to learn about the workers and simple majority voting is also the optimal estimator. BP also reduces to majority voting in this case, achieving the same optimality, and in fact $C_{1,\pi} = 1$. The interesting non-trivial case is when $r \geq 2$. The sufficient condition is for ℓ to be larger than some $C_{r,\pi}$. The problem of analyzing BP for $\ell < C_{r,\pi}$ is challenging. Similar challenges have not been resolved even in a simpler models[2] of stochastic block models, where BP and other efficient inference algorithms have been analyzed extensively [22], [30].

[2]The stochastic block model is simpler than our model in the sense that it has only pair-wise factors which is the special case of our model with $r = 2$.

The assumption on μ is canonical, since it only requires that the crowd as a whole can distinguish what the true label is. In the case $\mu < 0$, one can flip the sign of the final estimate to achieve the same guarantee. It is more intuitive to understand this assumption as formally defining a ground truths, as what the majority crowd would agree on (on average) if we asked the same question to all the workers in the crowd. Hence, this assumption is without loss of generality.

The assumption on $\mathbb{E}[(2p_u-1)^2] < 1$ is mild, as the only case when $\mathbb{E}[(2p_u-1)^2] = 1$ is if p_u is a a binary random variable taking values only in $\{0, 1\}$. In such cases, every worker is either telling the exact truths consistently or exact the opposite of the truths. It follows from Perron-Frobenius theorem [31] that a naive spectral method would work (and so does several other simple techniques). However, BP messages are not smooth in this case, which is required for our analysis. We believe optimality of BP still holds but requires a different analysis technique.

B. Relative Dominance of BP for all ℓ

For general ℓ, r, we establish the dominance of BP over two existing algorithms with known guarantees: the majority voting (MV) and the state-of-the-art iterative algorithm (KOS) in [11]. In the sparse regime, these are the only existing algorithms with tight provable guarantees.

Theorem 2. *Consider the Dawid-Skene model under the task assignment generated by a random bipartite (ℓ, r)-regular graph G consisting of n tasks and $(\ell/r)n$ workers. Let $\hat{s}^{\mathsf{MV}}$ and $\hat{s}^{\mathsf{KOS}}$ denote the outputs of MV and KOS algorithms, respectively. If $(\ell r)^k = o(n)$,*

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{E}\left[P_{\mathsf{err}}(\hat{s}^{\mathsf{BP}})\right] \\ &\leq \min\left\{\lim_{n\to\infty} \mathbb{E}\left[P_{\mathsf{err}}(\hat{s}^{\mathsf{MV}})\right], \lim_{n\to\infty} \mathbb{E}\left[P_{\mathsf{err}}(\hat{s}^{\mathsf{KOS}})\right]\right\}\end{aligned}$$

where $\hat{s}^{\mathsf{BP}}$ is the output of BP in (9) *after k iterations and the expectations are taken with respect to the graph G.*

A proof of the above theorem is presented in Section IV-B, where we also use the locally tree-like structure so that we need to assume $(\ell r)^k = o(n)$. Using Theorem 2 and the known error rates of MV and KOS algorithms in [11], one can derive the following upper bound on the error rate of BP:

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{E}\left[P_{\mathsf{err}}(\hat{s}^{\mathsf{BP}})\right] \qquad (12)\\ &\leq \min\left\{\lim_{n\to\infty} e^{-\left(\frac{\ell\mu^2}{2}\right)}, \lim_{n\to\infty} e^{-\left(\frac{\ell q}{2}\cdot\frac{q^2(\ell-1)(r-1)-1}{3q^2(\ell-1)(r-1)+q(\ell-1)}\right)}\right\}\end{aligned}$$

where $q := \mathbb{E}\left[(2p_u-1)^2\right]$ and all the parameters ℓ, r, μ, and q can depend on n.

This is particularly interesting, since it has been observed empirically and conjectured with some non-rigorous analysis in [13] that there exists a threshold $(\ell-1)(r-1) = 1/q^2$, above which KOS dominates over MV, and below which MV dominates over KOS (see Figure 2). This is due to the fact that KOS is inherently a spectral algorithm relying on the singular vectors of a particular matrix derived from A. Below the threshold, the sample noise overwhelms the signal in the spectrum of the matrix, which is known as the spectral barrier, and spectral methods fail. However, in practice, it is not clear which of the two algorithms should be used, since the threshold depends on latent parameters of the problem. Our dominance result shows that one can safely use BP, since it outperforms both algorithms in both regimes governed by the threshold. This is further confirmed by numerical experiments in Figure 2.

IV. Proofs of Theorems

In this section, we provide the proofs of Theorems 1 and 2.

A. *Proof of Theorem 1*

We first consider the case $r = 1$. Then, G is the set of disjoint *one-level* trees, i.e., star graphs, where the root of each tree corresponds to task $\rho \in V$ and the leaves are the set M_ρ of workers assigned to the task ρ. Since the graphs are disjoint, we have $\mathbb{P}[s_\rho|A] = \mathbb{P}[s_\rho|A_{\rho,1}]$, where $A = \{A_{iu} : (i,u) \in E\}$ and $A_{\rho,1} = \{A_{\rho u} : u \in M_\rho\}$. From Property 1, it follows that $\hat{s}^{\mathsf{BP}}_\rho(A) = \arg\max_{s_\rho} \mathbb{P}[s_\rho \mid A_{\rho,1}] = \hat{s}^*_\rho(A_{\rho,1})$. Therefore, for any $\ell \geq 1$, the optimal MAP estimator $\hat{s}^*_\rho(A)$ in (2) is identical to the output $\hat{s}^{\mathsf{BP}}_\rho$ with any $k \geq 1$.

We now focus on the case $r \geq 2$. Recall that random (ℓ, r)-regular bipartite graph G is locally tree-like. More formally, from Lemma 5 in [13], it follows that for $\rho \in V$,

$$\mathbb{P}[G_{\rho,2k} \text{ is not a tree}] \leq \frac{3\ell r}{n}\left((\ell-1)(r-1)\right)^{2k} , \qquad (13)$$

where we let $G_{\rho,2k} = (V_{\rho,2k}, W_{\rho,2k}, E_{\rho,2k})$ denote the subgraph of G induced by all the nodes within (graph) distance $2k$ from *root* ρ and $\partial V_{\rho,2k}$ denote the set of (task) nodes[3] whose distance from ρ is exactly $2k$. Let

$$\Delta(\hat{s}_\rho(A); G) := \frac{1}{2} - \mathbb{P}[s_\rho \neq \hat{s}_\rho(A) \mid G] ,$$

where the probability is taken with respect to the conditional distribution of s, $\hat{s}$, and A given G. We here note that $\Delta(\hat{s}_\rho(A); G)$ is a function of the distribution of $\hat{s}_\rho(A)$ given G. For simplicity, we often omit G in the notation if it is clear from context. Then, $\Delta(\hat{s}_\rho(A); G)$ is the expected gain of estimator $\hat{s}_\rho(A)$ compared to random guessing, i.e., $P_{\mathsf{err}}(\hat{s}(A)) = \frac{1}{2} - \frac{1}{n}\sum_{\rho \in V} \Delta(\hat{s}_\rho(A))$. Using (13), we obtain

$$\begin{aligned} &\mathbb{E}\left[P_{\mathsf{err}}(\hat{s}^{\mathsf{BP}(k)}) - \min_{\hat{s}:\text{estimator}} P_{\mathsf{err}}(\hat{s})\right] \\ &\leq \frac{1}{n}\sum_{\rho \in V} \mathbb{E}\left[\Delta(\hat{s}^*_\rho(A); G) - \Delta(\hat{s}^{\mathsf{BP}}_\rho; G) \mid G_{\rho,2k} \text{ is a tree}\right] \\ &\quad + \frac{3(\ell r)^{2k+1}}{n} , \end{aligned} \qquad (14)$$

where the expectation is taken with respect to the graph G.

Fix an arbitrary task $\rho \in V$ and G, and assume $G_{\rho,2k}$ is a tree. Then, it is enough to show that $\Delta(\hat{s}^*_\rho(A))$ and $\Delta(\hat{s}^{\mathsf{BP}(k)}_\rho)$ converge to the same value at exponential rate with respect to k, i.e.,

$$\left|\Delta(\hat{s}^*_\rho(A)) - \Delta(\hat{s}^{\mathsf{BP}}_\rho)\right| \leq 2^{-k} . \qquad (15)$$

[3]Since G is a bipartite graph, the distance from task ρ to every task is even and the distance from task ρ to every worker is odd.

To this end, we introduce two estimators, $\hat{z}^*_\rho(A_{\rho,2k})$ and $\hat{s}^*_\rho(A_{\rho,2k})$, which have accesses to different amounts and types of information. We now define the following *oracle* estimator:

$$\hat{z}^*_\rho(A_{\rho,2k}) := \arg\max_{s_\rho} \mathbb{P}[s_i \mid A_{\rho,2k}, s_{\partial V_{\rho,2k}}] ,$$

where we denote

$$A_{\rho,2k} := \{A_{iu} : (i,u) \in E_{\rho,2k}\} . \qquad (16)$$

We note that $\hat{z}^*_\rho(A_{\rho,2k})$ uses the exact label information of $\partial V_{\rho,2k}$ separating the inside and the outside of $G_{\rho,2k}$. Hence one can show that $\hat{z}^*_\rho(A_{\rho,2k})$ outperforms the optimal estimator $\hat{s}^*_\rho(A)$. We formally provide the following lemma whose proof is given in Section V-A.

Lemma 1. *Consider the Dawid-Skene model with a given task assignment graph $G = (V, W, E)$ and let A denote the set of workers' labels. For $\rho \in V$ and $k \geq 1$,*

$$\Delta(\hat{z}^*_\rho(A_{\rho,2k})) \;\geq\; \Delta(\hat{z}^*_\rho(A_{\rho,2k+2})) \;\dots\; \geq \; \Delta(\hat{s}^*_\rho(A)) .$$

Conversely, if an estimator uses less information than another, it performs worse. Formally, we provide the following lemma whose proof is given in Section V-B.

Lemma 2. *Consider the Dawid-Skene model with a given task assignment graph $G = (V, W, E)$ and let A denote the set of workers' labels. For any $\rho \in V$ and subset $A' \subset A$,*

$$\Delta(\hat{s}^*_\rho(A)) \geq \Delta(\hat{s}^*_\rho(A')) .$$

We note that the limit of $\lim_{n\to\infty} \mathbb{E}[\Delta(\hat{s}^*_\rho(A))]$ exists due to the non-increasing sequence of $\Delta(\hat{z}^*_\rho(A_{\rho,2k})) \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ in Lemma 1. Recalling the assumption that $G_{\rho,2k}$ is a tree and Property 1, it follows that

$$\hat{s}^{\mathsf{BP}(k)}_\rho := \arg\max_{s_\rho} b^k_\rho(s_\rho) = \arg\max_{s_\rho} \mathbb{P}[s_\rho \mid A_{\rho,2k}] .$$

Thus, using Lemmas 1 and 2 with $A_{\rho,2k} \subset A$, we have that

$$\begin{aligned} \Delta(\hat{z}^*_\rho(A_{\rho,2k})) &\geq \Delta(\hat{s}^*_\rho(A)) \\ &\geq \Delta(\hat{s}^{\mathsf{BP}(k)}_\rho) = \Delta(\hat{s}^*_\rho(A_{\rho,2k})) , \end{aligned} \qquad (17)$$

where we define $\hat{s}^*(A_{\rho,2k})_\rho := \arg\max_{s_\rho} \mathbb{P}[s_\rho \mid A_{\rho,2k}]$. Hence, the following lemma concludes (15) and completes the proof of Theorem 1.

Lemma 3. *Suppose $G_{\rho,2k} = (V_{\rho,2k}, W_{\rho,2k}, E_{\rho,2k})$ is given as a tree of which root is task ρ and depth is $2k$, where every task except the leaves $\partial V_{\rho,2k}$ is assigned to ℓ workers and every worker labels r tasks. For a given $\mu := \mathbb{E}[2p_u - 1] > 0$, $\mathbb{E}[(2p_u - 1)^2] < 1$, and constant $r \geq 1$, there exists a constant $C_{r,\pi}$ such that if $\ell \geq C_{r,\pi}$, then for sufficiently large k,*

$$\left|\Delta(\hat{z}^*_\rho(A_{\rho,2k})) - \Delta(\hat{s}^*_\rho(A_{\rho,2k}))\right| \leq 2^{-k}. \qquad (18)$$

A rigorous proof of Lemma 3 is given in Section V-C. Here, we briefly provide the underlying intuition on the proof. As long as μ is strictly greater than 0 and ℓ is sufficiently large, the majority voting of the one-hop information $\{A_{\rho u} : u \in M_\rho\}$ can achieve high accuracy. On the other hand, intuitively the information in two or more hops is less useful. In the proof of Lemma 3, we also provide a quantification of the *decaying rate of the correlation* from the information on $\partial V_{\rho,2k}$ to ρ as the distance $2k$ increases.

B. Proof of Theorem 2

We note that that KOS is an iterative algorithm where for each $\rho \in V$ and $k \geq 1$, $\hat{s}_\rho^{\mathsf{KOS},k}$ depends on only $A_{\rho,2k}$ defined in (16). In addition, it is clear that MV uses only one-hop information $A_{\rho,1} \subset A_{\rho,2k}$. Hence for given $A_{\rho,2k}$, the MAP estimator $\hat{s}^*_\rho(A_{\rho,2k})$ outperforms MV and KOS, i.e.,

$$\Delta(\hat{s}^*_\rho(A_{\rho,2k})) \geq \max\left\{\Delta(\hat{s}_\rho^{\mathsf{MV}}), \Delta(\hat{s}_\rho^{\mathsf{KOS},k})\right\}. \qquad (19)$$

Recall that if $G_{\rho,2k}$ is a tree, we have $\hat{s}_\rho^{\mathsf{BP},k} = \hat{s}^*_\rho(A_{\rho,2k})$. Similarly to (14), by separating the expectation with respect to G into the conditional expectations given $G_{\rho,2k}$ is tree or not, it is not hard to show that

$$\begin{aligned}&\mathbb{E}\left[\Delta(\hat{s}_\rho^{\mathsf{BP},k})\right]\\ &\geq \mathbb{E}\left[\max\left\{\Delta(\hat{s}_\rho^{\mathsf{MV}}), \Delta(\hat{s}_\rho^{\mathsf{KOS},k})\right\}\right] - \frac{3\ell r}{n}\left((\ell-1)(r-1)\right)^{2k},\end{aligned}$$

where the last term goes 0 as $n \to \infty$ if $(\ell r)^k = o(n)$. This completes the proof of Theorem 2.

V. Proofs of Lemmas

A. Proof of Lemma 1

We start with the conditional probability of error given A in the following:

$$\mathbb{P}[s_\rho \neq \hat{s}^*_\rho(A) \mid A] = \min\left\{\mathbb{P}[s_\rho = +1 \mid A], \mathbb{P}[s_\rho = -1 \mid A]\right\}.$$

This directly implies that

$$\begin{aligned}\Delta(\hat{s}^*_\rho(A)) &= \mathbb{E}\left[\frac{1}{2} - \mathbb{P}[s_\rho \neq \hat{s}^*_\rho(A) \mid A]\right]\\ &= \frac{1}{2}\,\mathbb{E}\left[\big|\mathbb{P}[s_\rho = +1 \mid A] - \mathbb{P}[s_\rho = -1 \mid A]\big|\right]. \qquad (20)\end{aligned}$$

Then, by simple algebra, it follows that

$$\begin{aligned}\Delta(\hat{s}^*_\rho(A)) &= \frac{1}{2}\sum_A \mathbb{P}[A]\cdot\big|\mathbb{P}[s_\rho = +1 \mid A] - \mathbb{P}[s_\rho = -1 \mid A]\big|\\ &= \frac{1}{2}\sum_A \left|\mathbb{P}[A, s_\rho = +1] - \mathbb{P}[A, s_\rho = -1]\right|\\ &= \frac{1}{2}\sum_A \frac{1}{2}\left|\mathbb{P}[A \mid s_\rho = +1] - \mathbb{P}[A \mid s_\rho = -1]\right|,\end{aligned}$$

where for the last equality we use $\mathbb{P}[s_\rho = +1] = \mathbb{P}[s_\rho = -1] = 1/2$.

Let ϕ_ρ^+ denote the distribution of A given $s_\rho = +1$, and let ϕ_ρ^- be the distribution of A given $s_\rho = -1$, i.e.,

$$\phi_i^+(A) = \mathbb{P}[A \mid s_i = +1] \text{ and } \phi_i^-(A) = \mathbb{P}[A \mid s_i = -1].$$

Then we have a simple expression of $\Delta(\hat{s}^*_\rho(A))$ as follows:

$$\Delta(\hat{s}^*_\rho(A)) = d_{\mathtt{TV}}(\phi_\rho^+, \phi_\rho^-), \qquad (21)$$

where we let $d_{\mathtt{TV}}$ denotes the total variation distance, i.e., for distributions ϕ and ψ on the same space Ω, we define

$$d_{\mathtt{TV}}(\phi, \psi) := \frac{1}{2}\sum_{\sigma\in\Omega}|\phi(\sigma) - \psi(\sigma)|.$$

Next we note that since $\partial V_{\rho,2k}$ blocks every path from the outside of $G_{\rho,2k}$ to ρ, the information on the outside of $G_{\rho,2k}$, $A \setminus A_{\rho,2k}$, is independent of s_ρ given $s_{\partial V_{\rho,2k}}$, i.e.,

$$\mathbb{P}[s_\rho \mid A_{\rho,2k}, s_{\partial V_{\rho,2k}}] = \mathbb{P}[s_\rho \mid A, s_{\partial V_{\rho,2k}}]. \qquad (22)$$

Hence if we set $\psi^+_{\rho,2k}$ to be the distribution of A and $s_{\partial V_{\rho,2k}}$ given $s_\rho = +1$ and similarly for $\psi^-_{\rho,2k}$, we have

$$\Delta(\hat{z}^*_\rho(A_{\rho,2k})) = d_{\mathtt{TV}}(\psi^+_{\rho,2k}, \psi^-_{\rho,2k}).$$

Noting that ϕ^+_ρ and ϕ^-_ρ can be obtained by marginalizing out $s_{\partial V_{\rho,2k}}$ in $\psi^+_{\rho,2k}$ and $\psi^-_{\rho,2k}$, it follows that

$$\begin{aligned}&d_{\mathtt{TV}}(\phi^+_\rho, \phi^-_\rho)\\ &= \frac{1}{2}\sum_A |\phi^+_\rho(A) - \phi^-_\rho(A)|\\ &= \frac{1}{2}\sum_A \Bigg|\sum_{s_{\partial V_{\rho,2k}}}\left(\psi_i^+(A, s_{\partial V_{\rho,2k}}) - \psi_i^-(A, s_{\partial V_{\rho,2k}})\right)\Bigg|\\ &\leq \frac{1}{2}\sum_A\sum_{s_{\partial V_{\rho,2k}}}\left|\psi_i^+(A, s_{\partial V_{\rho,2k}}) - \psi_i^-(A, s_{\partial V_{\rho,2k}})\right|\\ &= d_{\mathtt{TV}}(\psi^+_{\rho,2k}, \psi^-_{\rho,2k}), \qquad (23)\end{aligned}$$

which implies $\Delta(\hat{z}^*(A_{\rho,2k})) \geq \Delta(\hat{s}^*(A))$.

We now study $\Delta(\hat{z}^*(A_{\rho,2k}))$ with different k. Observe that $\partial V_{\rho,2k}$ blocks every path from $\partial V_{\rho,2k+2}$ to ρ, i.e., $s_{\partial V_{\rho,2k+2}}$ is independent of s_ρ given $s_{\partial V_{\rho,2k}}$. Thus, from (22), it follows that

$$\mathbb{P}[s_\rho \mid A, s_{\partial V_{\rho,2k}}] = \mathbb{P}[s_\rho \mid A, s_{\partial V_{\rho,2k}}, s_{\partial V_{\rho,2k+2}}].$$

Therefore, $\psi^+_{\rho,2k+2}$ and $\psi^-_{\rho,2k+2}$ can be obtained from $\psi^+_{\rho,2k}$ and $\psi^-_{\rho,2k}$ by marginalizing out $s_{\partial V_{\rho,2k+2}}$. Similarly to (23), we have

$$d_{\mathtt{TV}}(\psi^+_{\rho,2k+2}, \psi^-_{\rho,2k+2}) \leq d_{\mathtt{TV}}(\psi^+_{\rho,2k}, \psi^-_{\rho,2k}),$$

which completes the proof of Lemma 1.

B. Proof of Lemma 2

The proof of Lemma 2 is analog to that of Lemma 1. Let φ^+_ρ be the distribution of A' given $s_\rho = +1$ and φ^-_ρ be the distribution of A' given $s_\rho = -1$, i.e.,

$$\Delta(\hat{s}^*_\rho(A')) = d_{\mathsf{TV}}(\varphi^+_\rho, \varphi^-_\rho).$$

Since φ^+_ρ and φ^-_ρ can be obtained by marginalizing out $A \setminus A'$ from ϕ^+_ρ and ϕ^-_ρ in (21), using the same logic for (23), we have

$$d_{\mathtt{TV}}(\varphi^+_\rho, \varphi^-_\rho) \leq d_{\mathtt{TV}}(\phi^+_\rho, \phi^-_\rho),$$

which completes the proof of Lemma 2.

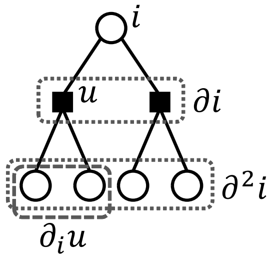


Fig. 1. A graphical representation of notations: $\partial i, \partial_i u$, and $\partial^2 i$.

C. Proof of Lemma 3

We start with several notations which we use in the proof. For $i \in V_{\rho,2k}$, let $T_i = (V_i, W_i, E_i)$ be the subtree rooted from i including all the offsprings of i in tree $G_{\rho,2k}$. We let ∂V_i denote the leaves in T_i and $A_i := \{A_{ju} : (j,u) \in E_i\}$. Define

$$X_i := \mathbb{P}[s_i = +1 \mid A_i] - \mathbb{P}[s_i = -1 \mid A_i] \,.$$

Here X_i is often called the *magnetization* of s_i given A_i. Similarly, given A_i and $s_{\partial V_i}$, we define the *biased* magnetization Y_i:

$$Y_i := \mathbb{P}[s_i = +1 \mid A_i, s_{\partial V_i}] - \mathbb{P}[s_i = -1 \mid A_i, s_{\partial V_i}] \,.$$

We note that X_i and Y_i correspond to the messages of BP from task i to i parent worker given different initialization at leaf tasks, where X_i and Y_i eventually calculate the marginal probability of s_i given only A_i, or both A_i and s_j for all $j \in \partial V_i$, respectively. Thus, using the alternative expression of Δ in (20), one can check that

$$\begin{aligned} 0 \le \Delta(\hat{z}_i^*(A_i)) - \Delta(\hat{s}_i^*(A_i)) &= \frac{1}{2}\,\mathbb{E}\left[|Y_i| - |X_i|\right] \\ &\le \mathbb{E}[|Y_i - X_i|] \,, \end{aligned} \tag{24}$$

where the expectation is taken with respect to A_i and $s_{\partial V_i}$.

In what follows, we fix $0 < t \le k$ and $i \in \partial V_{\rho,2k-2t}$, where we let $\partial V_{\rho,0} = \{\rho\}$. Let ∂i be the set of all the offspring of i and $\partial_i u$ be the set of all the offspring of u in tree T_i, i.e., $\partial i := \{u \in W_i : (i,u) \in E_i\}$ and $\partial_i u := \{j \in V_i : (j,u) \in E_i\}$. We further let $\partial_i^2 := \{j \in \partial_i u : u \in \partial i\}$ denote the set of all the second offspring of i. (See Figure 1 for a graphical explanation of the notations.) We note that if $i \in \partial V_{\rho,2k}$ is leaf node in $G_{\rho,2k}$, then $X_i = 0$ and $|X_i - Y_i| \le 1$. Hence, we will show that

$$\mathbb{E}\left[\left|X_i - Y_i\right|\right] \le \frac{1}{2|\partial^2 i|} \sum_{j \in \partial^2 i} \mathbb{E}\left[\left|X_j - Y_j\right|\right] \,, \tag{25}$$

since this implies $\mathbb{E}\left[\left|Y_\rho - X_\rho\right|\right] \le 2^{-k}$ and completes the proof of Lemma 3 with (24). We note that this implies the convergence of the oracle estimator $\hat{z}^*$ and the BP estimator $\hat{s}^{\mathsf{BP}}$ in a strong sense that X_ρ and Y_ρ converges to the same random variable in L_1-norm as k increases, i.e., $\mathrm{Var}[X_\rho - Y_\rho] = O(2^{-k})$, where the signs of X_ρ and Y_ρ are the estimated labels of task ρ from the oracle and BP estimators, respectively.

To show (25), we study certain recursions describing relations among X and Y. Define $A_u := \{A_{iu} : (i,u) \in E\}$ and $\mu_u := (2p_u - 1) \in [-1, 1]$ such that $\mu = \mathbb{E}[2p_u - 1] = \mathbb{E}[\mu_u] > 0$. Then f_u in (4) can be expressed as follows:

$$f_u(s_{N_u}) = \mathbb{E}\left[\prod_{j \in N_u} \frac{1 + A_{ju} s_j \mu_u}{2} \mid A_u, s_{N_u}\right] ,$$

where the expectation is taken with respect to μ_u given A_u, s_{N_u}. Using the above expression of f_u and the fact that $\mathbb{P}[s_j \mid A_j] = \frac{1+s_j X_j}{2}$, we can write the marginal probability of s_i given A_u and $X_{\partial_i u}$:

$$\begin{aligned} &\mathbb{P}\left[s_i \mid A_u, X_{\partial_i u}\right] \\ &= \sum_{s_{\partial_i u}} f_u(s_i, s_{\partial_i u}) \prod_{j \in \partial_i u} \frac{1 + s_j X_j}{2} \\ &= \sum_{s_{\partial_i u}} \mathbb{E}\left[\tfrac{1 + A_{iu} s_i \mu_u}{2} \prod_{j \in \partial_i u} \tfrac{(1 + A_{ju} s_j \mu_u)(1 + s_j X_j)}{4} \mid A_u, X_{\partial_i u}, s_{\partial_i u}\right] \\ &= \mathbb{E}\left[\frac{1 + A_{iu} s_i \mu_u}{2} \prod_{j \in \partial_i u} \frac{1 + A_{ju} \mu_u X_j}{2} \mid A_u, X_{\partial_i u}\right] . \end{aligned}$$

For notational convenience, we define g_{iu}^+ and g_{iu}^- as follows:

$$g_{iu}^+(X_{\partial_i u}; A_u) := \mathbb{E}\left[\frac{1 + A_{iu} \mu_u}{2} \prod_{j \in \partial_i u} \frac{1 + A_{ju} \mu_u X_j}{2} \mid A_u, X_{\partial_i u}\right],$$

$$g_{iu}^-(X_{\partial_i u}; A_u) := \mathbb{E}\left[\frac{1 - A_{iu} \mu_u}{2} \prod_{j \in \partial_i u} \frac{1 + A_{ju} \mu_u X_j}{2} \mid A_u, X_{\partial_i u}\right] ,$$

where we may omit A_u in the argument of g_{iu}^+ and g_{iu}^- if A_u is clear from the context. We note that $g_{iu}^+(X_{\partial_i u}; A_u) = \mathbb{P}\left[s_i = +1 \mid A_u, X_{\partial_i u}\right]$ and $g_{iu}^-(X_{\partial_i u}; A_u) = \mathbb{P}\left[s_i = -1 \mid A_u, X_{\partial_i u}\right]$. Hence, using Bayes' rule with g_{iu}^+ and g_{iu}^-, we obtain the following *recurrence* for X:

$$X_i = h_i(X_{\partial^2 i}) := \frac{\prod_{u \in \partial i} g_{iu}^+(X_{\partial_i u}) - \prod_{u \in \partial i} g_{iu}^-(X_{\partial_i u})}{\prod_{u \in \partial i} g_{iu}^+(X_{\partial_i u}) + \prod_{u \in \partial i} g_{iu}^-(X_{\partial_i u})} \,. \tag{26}$$

Similarly, we also have $Y_i = h_i(Y_{\partial_i^2})$.

For simplicity, we assume that i is not leaf or root node so that $\left|\partial^2 i\right| = (\ell - 1) \cdot (r - 1)$. Also, without loss of generality, we focus on the case where $s_j = +1$ for all $j \in V_{\rho,2k}$ since the true label s_j is uniformly distributed and the choice of $i(t)$ in (25) is uniform. Let $\mathbb{E}^+$ denote the conditional expectation given $s_j = +1$ for all $j \in V_{\rho,2k}$. Then, to complete the proof of (25), using the mean value theorem we will show

$$\mathbb{E}^+\left[|X_i - Y_i|\right] \le \frac{1}{2(\ell - 1)(r - 1)} \sum_{j \in \partial^2 i} \mathbb{E}^+\left[|X_j - Y_j|\right] \,. \tag{27}$$

We first obtain a bound on gradient of $h_i(x)$ for $x \in [-1, 1]^{\partial^2 i}$. Define $g_i^+(x) := \prod_{u \in \partial i} g_{iu}^+(x_{\partial_i u})$ and $g_i^-(x) := \prod_{u \in \partial i} g_{iu}^-(x_{\partial_i u})$. Then, using basic calculus, we obtain that for $j \in \partial_i u$,

$$\begin{aligned} \frac{\partial h_i}{\partial x_j} &= \frac{\partial}{\partial x_j} \frac{g_i^+ - g_i^-}{g_i^+ + g_i^-} \\ &= \frac{2}{(g_i^+ + g_i^-)^2} \left(g_i^- \cdot \frac{\partial g_i^+}{\partial x_j} - g_i^+ \cdot \frac{\partial g_i^-}{\partial x_j}\right) \end{aligned}$$

$$= \frac{2g_i^+ g_i^-}{(g_i^+ + g_i^-)^2} \left(\frac{1}{g_{iu}^+} \frac{\partial g_{iu}^+}{\partial x_j} - \frac{1}{g_{iu}^-} \frac{\partial g_{iu}^-}{\partial x_j} \right).$$

Using the fact that for $x \in [-1,1]^{\partial^2 i}$, both g_i^+ and g_i^- are positive, it is not hard to show that[4]

$$\frac{g_i^+ g_i^-}{(g_i^+ + g_i^-)^2} \le \sqrt{\frac{g_i^-}{g_i^+}} \,. \tag{28}$$

We note here that one can replace g_i^-/g_i^+ with g_i^+/g_i^- in the upper bound. However, in our analysis, we use (28) since we focus on the case of $s_i = +1$ where plugging $X_{\partial^2 i}$ or $Y_{\partial^2 i}$ into x in (28), $h_i(x)$, which is the magnetization X_i or Y_i, will be large thus g_i^-/g_i^+ will be a tighter upper bound than g_i^+/g_i^-. Our analysis covers all the general cases because the same analysis with g_i^+/g_i^- will work with $s_i = -1$ conversely.

From (28), it follows that for $x \in [-1,1]^{\partial^2 i}$,

$$\left| \frac{\partial h_i}{\partial x_j}(x) \right| \le \left| g'_{ij}(x_{\partial_i u}) \right| \cdot \prod_{u' \in \partial i \,:\, u' \neq u} \sqrt{\frac{g_{iu'}^-(x_{\partial_i u'})}{g_{iu'}^+(x_{\partial_i u'})}} \,,$$

where we define

$$g'_{ij}(x_{\partial_i u}) := 2\sqrt{\frac{g_{iu}^-(x_{\partial_i u})}{g_{iu}^+(x_{\partial_i u})}} \left(\frac{1}{g_{iu}^+(x_{\partial_i u})} \frac{\partial g_{iu}^+(x_{\partial_i u})}{\partial x_j} - \frac{1}{g_{iu}^-(x_{\partial_i u})} \frac{\partial g_{iu}^-(x_{\partial_i u})}{\partial x_j} \right).$$

From the assumption on μ_u (or p_u), i.e., $\mathbb{E}[\mu_u] > 0$ and $\mathbb{E}[\mu_u^2] < 1$, it follows that for all $x_{\partial_i u} \in [-1,1]^{\partial_i u}$, $g_{iu}^+(x_{\partial_i u}) > 0$ and $g_{iu}^-(x_{\partial_i u}) > 0$. Thus, for given r, we can find finite η, which is a constant with respect to ℓ, such that

$$\max_{x_{\partial_i u} \in [-1,1]^{\partial_i u}} |g'_{ij}(x_{\partial_i u})| \;\le\; \eta \;<\; \infty.$$

Hence, we have

$$\left| \frac{\partial h_i}{\partial x_j}(x) \right| \le \eta \cdot \prod_{u' \in \partial i \,:\, u' \neq u} \sqrt{\frac{g_{iu'}^-(x_{\partial_i u'})}{g_{iu'}^+(x_{\partial_i u'})}} \,. \tag{29}$$

Let $\mathbb{E}_{x,y}^+$ denote the conditional expectation given $X_{\partial^2 i} = x_{\partial^2 i}$, $Y_{\partial^2 i} = y_{\partial^2 i}$, and $s_j = +1$ for all j. Then, using the mean value theorem with (29), it follows that for given $X_{\partial^2 i}$ and $Y_{\partial^2 i}$, there exists $\lambda' \in [0,1]$ such that

$$\begin{aligned}
&\mathbb{E}_{x,y}^+ \left| h_i(X_{\partial^2 i}) - h_i(Y_{\partial^2 i}) \right| \\
&\le \sum_{u \in \partial i} \sum_{j \in \partial_i u} |X_j - Y_j| \;\; \mathbb{E}_{x,y}^+ \left[\left| \frac{\partial h_i}{\partial x_j} \left(\lambda' X_{\partial^2 i} + (1-\lambda') Y_{\partial^2 i} \right) \right| \right] \\
&\le \sum_{u \in \partial i} \sum_{j \in \partial_i u} |X_j - Y_j| \\
&\quad \times \eta \prod_{u' \in \partial i \setminus \{u\}} \max_{\lambda \in [0,1]} \left\{ \mathbb{E}_{x,y}^+ \left[\sqrt{\frac{g_{iu'}^-(\lambda X_{\partial_i u'} + (1-\lambda) Y_{\partial_i u'})}{g_{iu'}^+(\lambda X_{\partial_i u'} + (1-\lambda) Y_{\partial_i u'})}} \right] \right\}.
\end{aligned} \tag{30}$$

We note that each term in an element of the summation in the RHS is independent to each other. Thus, from the symmetry among $\{X_{\partial_i u}\}_{u \in \partial i}$, it follows that

$$\begin{aligned}
&\mathbb{E}^+ \left[|X_i - Y_i| \right] \\
&\le \sum_{u \in \partial i} \sum_{j \in \partial_i u} \mathbb{E}^+ \left[|X_j - Y_j| \right] \\
&\quad \times \eta \left(\mathbb{E}^+ \left[\max_{\lambda \in [0,1]} \Gamma(\lambda X_{\partial_i u} + (1-\lambda) Y_{\partial_i u}) \right] \right)^{\ell - 1},
\end{aligned} \tag{31}$$

where we define function $\Gamma(x_{\partial_i u})$ for given $x_{\partial_i u} \in [-1,1]^{\partial_i u}$ as

$$\Gamma(x_{\partial_i u}) := \mathbb{E}_{x,y}^+ \left[\sqrt{\frac{g_{iu}^-(x_{\partial_i u})}{g_{iu}^+(x_{\partial_i u})}} \right] .$$

We may calculate $\Gamma(x_{\partial_i u})$ as follows:

$$\begin{aligned}
\Gamma(x_{\partial_i u}) &= \sum_{A_u \in \{-1,+1\}^{N_u}} \mathbb{P}^+[A_u] \cdot \sqrt{\frac{g_{iu}^-(x_{\partial_i u}; A_u)}{g_{iu}^+(x_{\partial_i u}; A_u)}} \\
&= \sum_{A_u \in \{-1,+1\}^{N_u}} \mathbb{E} \left[\prod_{j \in N_u} \frac{1 + A_{ju}\mu_u}{2} \mid A_u \right] \\
&\quad \times \sqrt{\frac{\mathbb{E} \left[\frac{1 - A_{iu}\mu_u}{2} \prod_{j \in \partial_i u} \frac{1 + A_{ju}\mu_u x_j}{2} \mid A_u, X_{\partial_i u} = x_{\partial_i u} \right]}{\mathbb{E} \left[\frac{1 + A_{iu}\mu_u}{2} \prod_{j \in \partial_i u} \frac{1 + A_{ju}\mu_u x_j}{2} \mid A_u, X_{\partial_i u} = x_{\partial_i u} \right]}} \,,
\end{aligned}$$

where we let $\mathbb{P}^+$ denote the conditional probability measure given that s_j for all j. We bound the last term in (31) as:

Lemma 4. *For given π such that $\mu := \mathbb{E}[\mu_u] > 0$ and $\mathbb{E}[\mu_u^2] < 1$, there exists constant $C'_{r,\pi}$ such that for any $\ell \ge C'_{r,\pi}$,*

$$\mathbb{E}^+ \left[\max_{\lambda \in [0,1]} \Gamma(\lambda X_{\partial_i u} + (1-\lambda) Y_{\partial_i u}) \right] \;\le\; \sqrt{1 - \frac{\mu^2}{4}} \;<\; 1 \,.$$

Using the above lemma, we can find a sufficiently large constant $C_{r,\pi} \ge C'_{r,\pi}$ such that if $\ell - 1 \ge C_{r,\pi}$,

$$\eta \left(\sqrt{1 - \frac{\mu^2}{4}} \right)^{C_{r,\pi}} \;\le\; \frac{1}{2C_{r,\pi}(r-1)} \;\le\; \frac{1}{2(\ell-1)(r-1)} \,.$$

This implies (27) with (31) and completes the proof of Lemma 3.

We now focus on the proof of Lemma 4. We first obtain a bound on X_j and Y_j for $j \in \partial_i u$. The MAP estimator $\hat{s}_j^*(A_j)$ of s_j given A_j is identical to estimating $s_j = +1$ if X_j is positive and $s_j = -1$ otherwise. From the definition of the MAP estimator, it is straightforward to check

$$\mathbb{P}[s_j \neq \hat{s}_j^*(A_j)] = \frac{1 - \mathbb{E}^+[X_j]}{2} \,.$$

In addition, as Lemma 2 states, the MAP estimator $\hat{s}_j^*(A_j)$ outperforms MV with $\{A_{ju(jj')} : j' \in \partial j\}$. Using Hoeffding's bound, the error probability of MV is bounded as follows:

$$\frac{1 - \mathbb{E}^+[X_j]}{2} \;\le\; \mathbb{P}^+[s_j \neq \hat{s}_j^{\mathsf{MV}}] \;\le\; \exp\left(-\frac{(|\partial j| - 1)\mu^2}{2} \right) \,,$$

where Lemma 2 implies the first inequality. Similarly, $\hat{z}_j^*(A_j)$ of s_j given A_j and ∂V_i is identical to estimating $s_j = +1$ if Y_j is positive and $s_j = -1$ otherwise. Using Lemma 1 and the Markov inequality, it follows that for small $\varepsilon > 0$,

$$\mathbb{P}^+[Y_j < 1 - \varepsilon] \;\le\; \mathbb{P}^+[X_j < 1 - \varepsilon] \;\le\; \frac{2\exp\left(-\frac{(|\partial j| - 1)\mu^2}{2} \right)}{\varepsilon} \,, \tag{32}$$

[4]We can further obtain $\frac{g_i^+ g_i^-}{(g_i^+ + g_i^-)^2} \le \frac{3\sqrt{3}}{16}\sqrt{g_i^-/g_i^+}$, but we use (28) for simplicity.

where we use Lemma 1 for the first inequality and the Markov inequality for the second one.

Since $0 < \mathbb{E}[\mu_u]$ and $\mathbb{E}[\mu_u^2] < 1$, we can find finite constants η' and η'' such that for all $x \in [0,1]^{\partial_i u}$,

$$|\Gamma(x)| \le \eta' \quad \text{and} \quad \left|\frac{\partial \Gamma(x)}{\partial x_j}\right| \le \eta'' \ .$$

Let $\varepsilon(\ell) := \exp\left(-\frac{(\ell-1)\mu^2}{4}\right) \le \exp\left(-\frac{(|\partial j|-1)\mu^2}{4}\right)$. Then,

$$\begin{aligned}
&\mathbb{E}^+\left[\max_{\lambda\in[0,1]}\{\Gamma(\lambda X_{\partial_i u} + (1-\lambda)Y_{\partial_i u})\}\right] \\
&\le \left(1 - \mathbb{P}^+[X_j > 1-\varepsilon \text{ and } Y_j > 1-\varepsilon \ \forall j \in \partial_i u]\right) \times \max_{x\in[-1,1]^{\partial_i u}} \Gamma(x) \\
&\quad + \mathbb{P}^+[X_j > 1-\varepsilon \text{ and } Y_j > 1-\varepsilon \ \forall j \in \partial_i u] \times \max_{x\in[1-\varepsilon,1]^{\partial_i u}} \Gamma(x) \\
&\underset{(a)}{\le} \left(\sum_{j\in\partial_i u} \mathbb{P}^+[X_j \le 1-\varepsilon] + \mathbb{P}^+[Y_j \le 1-\varepsilon]\right) \times \max_{x\in[-1,1]^{\partial_i u}} \Gamma(x) \\
&\quad + 1 \times \max_{x\in[1-\varepsilon,1]^{\partial_i u}} \Gamma(x) \\
&\underset{(b)}{\le} 4r\eta'\varepsilon(\ell) + \max_{x\in[1-\varepsilon,1]^{\partial_i u}} \Gamma(x) \\
&\underset{(c)}{\le} 4r\eta'\varepsilon(\ell) + \Gamma(1) + \varepsilon(\ell)\eta'' \ ,
\end{aligned}$$

where we use the union bound, (32), and the mean value theorem for (a), (b), and (c), respectively. Since $\varepsilon(\ell)$ decreases as ℓ increases, it is enough to show $\Gamma(1) \le \sqrt{1-\mu^2}$. Using the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}
\Gamma(1_{\partial_i} u) = &\sum_{A_{\partial_i u}} \sqrt{\mathbb{E}\left[\frac{1+\mu_u}{2}\prod_{j\in\partial_i u}\frac{1+A_{ju}\mu_u}{2} \mid A_u\right]} \\
&\times \sqrt{\mathbb{E}\left[\frac{1-\mu_u}{2}\prod_{j\in\partial_i u}\frac{1+A_{ju}\mu_u}{2} \mid A_u\right]} \\
&+ \sum_{A_{\partial_i u}} \sqrt{\mathbb{E}\left[\frac{1-\mu_u}{2}\prod_{j\in\partial_i u}\frac{1+A_{ju}\mu_u}{2} \mid A_u\right]} \\
&\times \sqrt{\mathbb{E}\left[\frac{1+\mu_u}{2}\prod_{j\in\partial_i u}\frac{1+A_{ju}\mu_u}{2} \mid A_u\right]} \\
&\le \sqrt{\frac{1+\mu}{2}}\sqrt{\frac{1-\mu}{2}} + \sqrt{\frac{1-\mu}{2}}\sqrt{\frac{1+\mu}{2}} = \sqrt{1-\mu^2}
\end{aligned}$$

which completes the proof.

VI. Experimental Result

In this section, we evaluate the performance of BP using both synthetic datasets and real-world Amazon Mechanical Turk datasets to study how our theoretical findings are demonstrated in practice.

A. Tested Algorithms

We compare BP and a variant of BP to two oracle algorithms and several state-of-the-art algorithms in [8], [11], [14], each of which are briefly summarized next.

A practical version of BP. We note that BP, named BP-True in our plots, requires the knowledge of the prior on p_u's. However, in practice, the prior is typically unknown. Thus, we design a practical version of BP, which we call EBP (Estimation and Belief Propagation) that has an additional procedure that extracts the required statistics on the prior of p_u's from the observed data. In EBP, starting with a certain initialization[5] of labels, it first estimates the statistics of each worker's reliability assuming the labels are true, and updates the labels via BP using the estimated statistics as the reliability distribution, over multiple rounds in an iterative manner. We will focus on two versions of EBP with one and two rounds, respectively, marked as EBP(1) and EBP(2), which is motivated by our empirical observation that two rounds are enough to achieve good performance, and the gain from more rounds is marginal.

Oracle algorithm. Since computing the MAP estimate is computationally intractable, we instead compute the lower bound on the error rate, using the following estimator with access to an oracle. We consider an oracle MAP estimator which has an omniscient access to a subset of the true labels of tasks to label each task. We consider the Oracle-Task that, to estimate task ρ, uses the true labels of the only tasks separating the inside and the outside of the breadth-first searching tree rooted from task ρ in G. Then due to the exactness of BP on a tree in Property 1 and Lemma 1, we can obtain the lower bound in a polynomial time.

Tested algorithms for comparisons. For comparison to the state-of-the-art algorithms, we test the majority voting (MV), an iterative algorithm (KOS) [11]), the expectation maximization (EM) [8]) and an approach based on approximate mean field (AMF) [14]). Specifically, as the authors in [14] suggested, we run EM and AMF with $\text{Beta}(2,1)$ as the input distribution on workers' reliability.

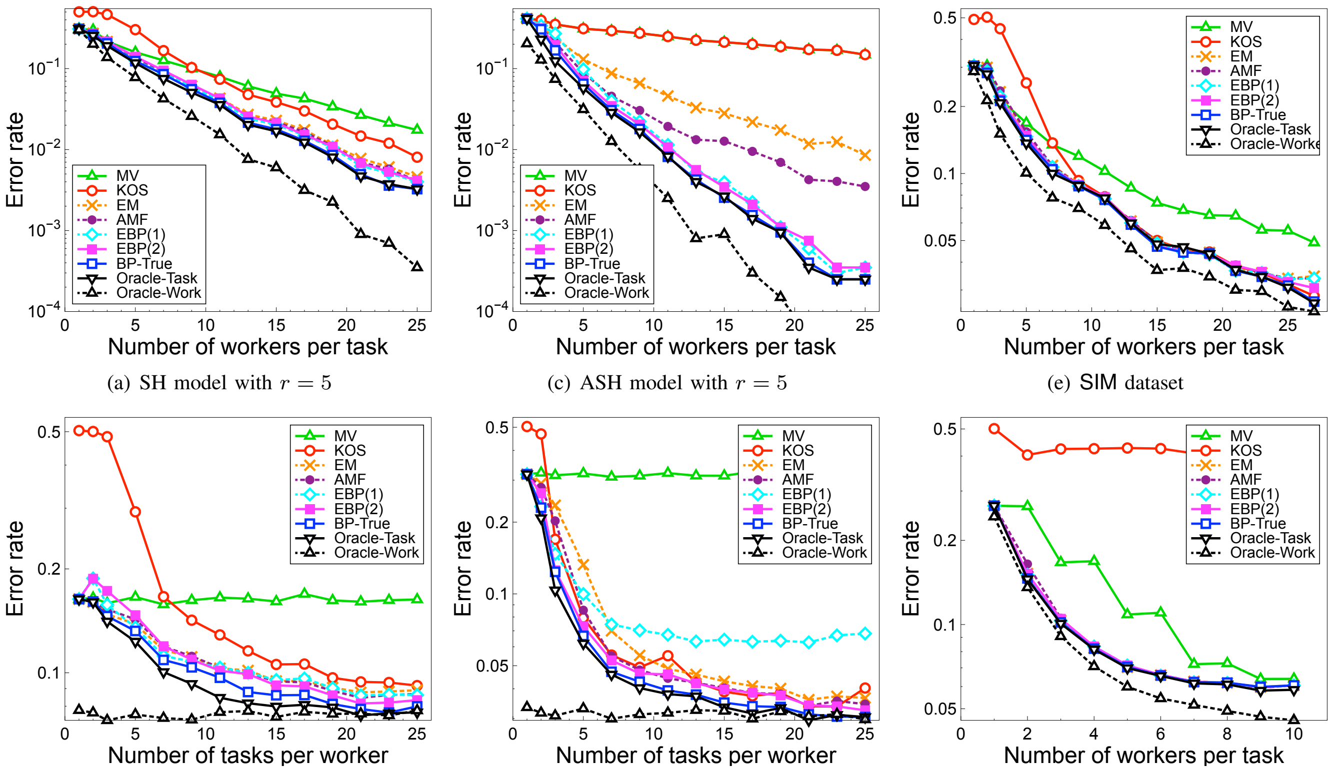
We terminate all algorithms that run in an iterative manner (i.e., all the algorithms except for MV) at the maximum of 100 iterations or with 10^{-5} message convergence tolerance, all results are averaged on 100 random samples.

B. Performance on Synthetic Datasets

We first compare all the algorithms with synthetic datasets generated by the set of random (ℓ, r)-regular bipartite graphs having 200 tasks from the configuration model [28], where we vary either ℓ or r. We randomly choose worker's reliability p_u from the *spammer-hammer* model with $\pi(0.5) = \pi(0.9) = 1/2$ and the *adversary-spammer-hammer* model with $\pi(0.1) = \pi(0.5) = 1/4$ and $\pi(0.9) = 1/2$, whose results are plotted in Figures 2(a)-2(b) and Figures 2(c)-2(d), respectively.

Optimality of BP. We observe that BP-True with the knowledge of the true reliability distribution has the negligible performance gap from the lower bound of Oracle-Task, whereas other algorithms have the suboptimal performance and their suboptimality gap depends on ℓ, r and the reliability distribution π (see Figures 2(c).). As discussed in [11], we observe a threshold behavior at $(\ell-1)(r-1) = 1/q^2$ where

[5] In this paper, we initialize EBP with the labels from MV to deliver more interpretations. However, better initialization such as EM and AMF with $\text{Beta}(2,1)$ can be also considered in practical use of EBP.



(a) SH model with $r = 5$

(c) ASH model with $r = 5$

(e) SIM dataset

(b) SH model with $\ell = 5$

(d) ASH model with $\ell = 5$

(f) TEMP dataset

Fig. 2. The average fraction of incorrectly labeled tasks on the synthetic datasets and the real-world Amazon Mechanical Turk datasets; (a)-(b) the synthetic datasets consisting of 200 tasks with the spammer-hammer (SH) model with $\pi(0.5) = \pi(0.9) = 1/2$; (c)-(d) the synthetic datasets consisting of 200 tasks with the adversary-spammer-hammer (ASH) model with $\pi(0.1) = \pi(0.5) = 1/4$ and $\pi(0.9) = 1/2$; (a) Color-similarity comparison (SIM) dataset with 50 tasks and 28 workers obtained in [13]; (b) Temporal ordering (TEMP) dataset with 462 tasks and 76 workers obtained in [32].

for small ℓ and r MV outperforms KOS but for large ℓ and r KOS is better. However, BP-true consistently outperforms all other algorithms irrespective of the values of ℓ and r.

Near-optimality of EBP. Even without knowing the true reliability distribution, EBP with just two rounds (EBP(2)) of updating prior, achieves almost the same performance as BP-True, as shown in Figure 2(d). However, MV, KOS, EM and AMF use fixed priors[6] which are different from the true prior. Hence, such mismatches between the algorithmic and true priors cause performance degeneration, which is particularly significant when the true prior is the adversary-spammer-hammer (see Figure 2(c).). Even though EBP is initialized with the labels from MV, of which performance is poor, EBP improves the accuracy by recursively updating the estimations of prior and labels. Indeed, Figure 2(d) shows the recursive improvement of EBP from MV to EBP(1) to EBP(2), where two rounds of updates (EBP(2)) provide us the performance close to optimality.

Tighter lower bound. We recall that a lower bound in Lemma 1 (i.e., Oracle-Task) was tight enough to show the exact optimality of BP, and this tightness is demonstrated in all Figures. Note that a different lower bound is studied by [11] to show just an order-wise optimality of KOS, which is obtained by the Bayesian estimator with full information on *true workers' reliabilities*, marked as Oracle-Work in our plots. Both Oracle-Work and Oracle-Task scale well with respect to ℓ but only Oracle-Work does with r as well, thus being a tighter lower bound (see Figures 2(b) and 2(d)).

C. *Performance on Real Datasets*

We use two real-world Amazon Mechanical Turk datasets from [11] and [32]: SIM dataset and TEMP dataset. SIM dataset is a set of collected labels where 50 tasks on color-similarity comparison are assigned to 28 users in Amazon Mechanical Turk. TEMP dataset consists of 76 workers' labels on 462 questions about temporal ordering of two events in a collection of sentences of a natural language. In both datasets, we use the reliability measured from the dataset as a true workers' reliability, and we vary ℓ by subsampling the datasets. Figures 2(e) and 2(f) shows the evaluation results, where we obtain similar implications to those with the synthetic datasets, where EBP(2) is close to Oracle-Task and outperforms all other the state-of-the-art algorithms. In particular, KOS performs poorly for the TEMP dataset, because it is under the regime for small ℓ, i.e., before the threshold.

VII. Conclusion and Discussion

In this paper, we settle the question of optimality and computational gap for a canonical scenario for the crowd-sourced classification where the task assignment is random (ℓ, r)-regular bipartite graph. Here we discuss some interesting

[6] MV and KOS can be interpreted as special cases of BP with deterministic prior and Haldane prior, respectively, [14]. EM and AMF use $\mathrm{Beta}(2, 1)$ as the prior.

potential extensions of our result. First the BP optimality can be proved when the task assignment graph is *irregular*. Our proof of the BP optimality uses the locally tree-like structure in (13) and the decaying correlation in Lemma 3. These properties hold as long as the numbers of workers per task are finite. One can potentially generalize Theorem 1 to irregular bipartite graphs, where each task is assigned to sufficiently large but *different* number of workers and each worker is assigned to large but different number of tasks. This extension is important in practical setting where the workers decide how many tasks to work on.

Second it would be interesting to tighten the constants in the error exponent in (12) since the actual performance of BP is better than predicted by this upper bound. The analysis could be significantly tightened, if one can provide tighter analysis of both the majority voting and the KOS algorithm. Next, a tighter analysis of the oracle error rate is needed. We provide an oracle estimator that is significantly tighter than the naive oracle estimators presented in [11]. This strong oracle can be numerically evaluated, as we do in our experiments. However, it is not known how the error achieved by this oracle estimator scales with problem parameters. A tight analysis of this lower bound in a form similar to (12) would complete the investigation of optimality of BP. Finally, it has been observed in [11], [12] that there exists a spectral barrier at $(\ell - 1)(r - 1) = 1/q^2$, where $q = \mathbb{E}[(2p_u - 1)^2]$. Below the spectral barrier, we observe that the gap between the simple majority voting and BP becomes narrower as we step away from this threshold. It is of interest to identify where MV is optimal, in order to provide guidelines on how to design crowdsourcing experiments and which algorithms to use.

When we have more than two classes, our algorithm naturally generalizes. However, the computational complexity increases and the analysis techniques do not generalize. We need to investigate other inference algorithms, perhaps those based on semidefinite programming or expectation maximization, and provide an analysis that naturally generalizes to multiple classes. When there are k classes, characterizing the error rate when k scales as n^α for some parameter α is of interest. We expect BP to be no longer optimal for some regimes of α.

One of the major drawback of the Dawid-Skene model is that it does not account for tasks that have different difficulty levels. In real-world crowdsourcing data, it is common to see some tasks that are more difficult than the others. To capture such heterogeneity, several generalized models have been proposed [5]–[7], [16], [17], [32]–[34]. For these general models, the questions of the error rate achieved by efficient inference algorithms is widely open. Finally, in real crowdsourcing systems, adaptive design is common. One can decide to collect more data on those tasks that are more difficult. Tighter analysis of the error rate can provide guidelines on how to design such adaptive crowdsourcing experiments. Understanding such adaptive task assignments is an important topic, as they are widely used in practice. Under the standard Dawid-Skene model studied in this paper, it is known that there is not much gain in using adaptive schemes [13]. The main reason is that all tasks are inherently assumed to be equally easy (or difficult) and there is not much gain in identifying tasks with less confidence and assigning more workers on those tasks. However, recent advances work in [34] proves that under a more general variation of the Dawid-Skene model, it is possible to significantly outperform non-adaptive schemes (such as those studied in this paper), by using adaptive task assignment schemes. Understanding the optimality of BP under this more generalized Dawid-Skene model is an interesting open problem. It is not even clear how to run BP in this case, as both tasks and workers are parametrized by continuous variables.

ACKNOWLEDGMENT

A preliminary version of this work has been published as [1], where the authors showed the BP optimality when $r = 2$. In this work, we provide a generalized proof of the BP optimality with all $r \geq 1$.

This work was supported by NSF awards (CNS-1527754, CCF-1553452, and CCF-1705007), Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.R0132-15-1005, Content visual browsing technology in the online and offline environments, and No.2016-0-00160,Versatile Network System Architecture for Multi-dimensional Diversity), and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No. 2016R1A2A2A05921755).

REFERENCES

[1] J. Ok, S. Oh, J. Shin, and Y. Yi, "Optimality of belief propagtion for crowdsourced classification," in *Proceedings of ICML*, 2016.

[2] Completely Automated Public Turing test to tell Computers and Humans Apart, "Captcha," http://www.captcha.net/, 2000.

[3] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Proceedings of NIPS*, 1995.

[4] R. Jin and Z. Ghahramani, "Learning with multiple labels," in *Proceedings of NIPS*, 2003.

[5] 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 *Proceedings of NIPS*, 2009.

[6] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *Proceedings of NIPS*, 2010.

[7] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, L. Moy, and D. Blei, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[8] 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)*, vol. 28, no. 1, pp. 20–28, 1979.

[9] A. Ghosh, S. Kale, and P. McAfee, "Who moderates the moderators?: Crowdsourcing abuse detection in user-generated content," in *Proceedings of ACM EC*, 2011.

[10] N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi, "Aggregating crowdsourced binary ratings," in *Proceedings of WWW*, 2013.

[11] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Proceedings of NIPS*, 2011.

[12] ——, "Efficient crowdsourcing for multi-class labeling," in *Proceedings of ACM SIGMETRICS*, 2013.

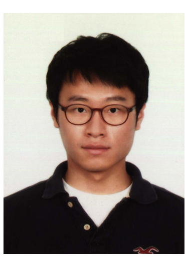
[13] ——, "Budget-optimal task allocation for reliable crowdsourcing systems," *Operations Research*, vol. 62, no. 1, pp. 1–24, 2014.

[14] Q. Liu, J. Peng, and A. T. Ihler, "Variational inference for crowdsourcing," in *Proceedings of NIPS*, 2012.

[15] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet em: A provably optimal algorithm for crowdsourcing," in *Proceedings of NIPS*, 2014.

[16] D. Zhou, J. Platt, S. Basu, and Y. Mao, "Learning from the wisdom of crowds by minimax entropy," in *Proceedings of NIPS*, 2012.

[17] D. Zhou, Q. Liu, J. C. Platt, C. Meek, and N. B. Shah, "Regularized minimax conditional entropy for crowdsourcing," *arXiv preprint arXiv:1503.07240*, 2015.
[18] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," in *Proceedings of IEEE FOCS*, 1989.
[19] H. Li, B. Yu, and D. Zhou, "Error rate analysis of labeling by crowdsourcing," in *Proceedings of ICML*, 2013.
[20] H. Li and B. Yu, "Error rate bounds and iterative weighted majority voting for crowdsourcing," *arXiv preprint arXiv:1411.4086*, 2014.
[21] C. Gao and D. Zhou, "Minimax optimal convergence rates for estimating ground truth from crowdsourced labels," *arXiv preprint arXiv:1310.5764*, 2013.
[22] E. Mossel, J. Neeman, and A. Sly, "Belief propagation, robust reconstruction and optimal recovery of block models," in *Proceedings of COLT*, 2014.
[23] S. Kudekar, T. Richardson, and R. L. Urbanke, "Spatially coupled ensembles universally achieve capacity under belief propagation," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 7761–7813, 2013.
[24] S. Park and J. Shin, "Max-product belief propagation for linear programming: applications to combinatorial optimization," in *Proceedings of UAI*, 2015.
[25] B. Hajek, Y. Wu, and J. Xu, "Exact recovery threshold in the binary censored block model," in *Proceedings of IEEE Information Theory Workshop*, 2015.
[26] M. C. Angelini, F. Caltagirone, F. Krzakala, and L. Zdeborová, "Spectral detection on sparse hypergraphs," in *Proceedings of Allerton*, 2015, pp. 66–73.
[27] C. Gao, Y. Lu, and D. Zhou, "Exact exponent in optimal rates for crowdsourcing," in *Proceedings of ICML*, 2016, pp. 603–611.
[28] B. Bollobás, *Random graphs*. Springer, 1998.
[29] J. Pearl, "Reverend bayes on inference engines: A distributed hierarchical approach," in *Proceedings of AAAI*, 1982.
[30] C. Bordenave, M. Lelarge, and L. Massoulié, "Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs," in *Proceedings of IEEE FOCS*, 2015.
[31] R. H. Keshavan, S. Oh, and A. Montanari, "Matrix completion from a few entries," in *2009 IEEE International Symposium on Information Theory*. IEEE, 2009, pp. 324–328.
[32] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast-but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of EMNLP*. Association for Computational Linguistics, 2008.
[33] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of ACM SIGKDD*, 2008.
[34] A. Khetan and S. Oh, "Achieving budget-optimality with adaptive schemes in crowdsourcing," in *Advances in Neural Information Processing Systems*, 2016, pp. 4844–4852.



Jungseul Ok Jungseul Ok currently working as postdoctoral researcher in Department of Automatic Control at KTH Royal Institute of Technology. He received Ph.D degree under the supervision of Prof. Yung Yi and Prof. Jinwoo Shin in 2016 and B.S degree in 2011 in School of Electrical Engineering at Korea Advanced Institute of Science and Technology (KAIST). He is interested in understanding and improving practical systems based on theoretical model and analysis. His research interest lies at the intersection of applied mathematics and network systems including wireless network, social network, and probabilistic graphical model, where his major focus is social models with certain engineering purposes such as crowdsourcing and information spread.

Sewoong Oh Sewoong Oh is an Assistant Professor of Industrial and Enterprise Systems Engineering at UIUC. He received his PhD from the department of Electrical Engineering at Stanford University. Following his PhD, he worked as a postdoctoral researcher at Laboratory for Information and Decision Systems (LIDS) at MIT. His research interest is in theoretical machine learning, including spectral methods, ranking, crowdsourcing, estimation of information measures, differential privacy, and generative adversarial networks. He was co-awarded the best paper award at the ACM SIGMETRICS in 2015, awarded the SIGMETRICS Rising Star Research Award in 2017, NSF CAREER award in 2016, and GOOGLE Faculty Research Award in 2017.

Jinwoo Shin Jinwoo Shin is currently an associate professor at the School of Electrical Engineering at KAIST, Korea. His current major research interest is on algorithmic questions for machine learning and related fields. He obtained the Ph.D. degree from Massachusetts Institute of Technology in 2010 with George M. Sprowls (Best MIT CS PhD Thesis) Award and B.S. degrees (in Math and CS) from Seoul National University in 2001. After spending two years (2010-2012) at Algorithms & Randomness Center, Georgia Institute of Technology, one year (2012-2013) at Business Analytics and Mathematical Sciences Department, IBM T. J. Watson Research, he joined KAIST EE in Fall 2013. Dr. Shin received Kenneth C. Sevcik Award at ACM SIGMETRICS/Performance 2009, Best Publication Award from INFORMS Applied Probability Society 2013, Best Paper Award at ACM MOBIHOC 2013, Bloomberg Scientific Research Award 2015 and ACM SIGMETRICS Rising Star Award 2015. He has served TPCs (or reviewers) at AAAI, ACML, ICC, ICML, INFOCOM, INFORMS, MOBIHOC, NIPS, SIGMETRICS, UAI, WIOPT and AEs at IEEE/ACM Transactions on Networking, ACM Modeling and Performance Evaluation of Computing Systems.

Yung Yi Yung Yi received the B.S. and M.S. degrees from the School of Computer Science and Engineering, Seoul National University, South Korea, in 1997 and 1999, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Texas at Austin, in 2006. From 2006 to 2008, he was a Post-Doctoral Research Associate with the Department of Electrical Engineering, Princeton University. He is currently a professor with the School of Electrical Engineering, KAIST, South Korea. His current research interests include applied machine learning, computer networking, and performance evaluation of distributed systems. He received the best paper awards at the IEEE SECON 2013, the ACM MOBIHOC 2013, and the IEEE William R. Bennett Award 2016.