Optimality of Belief Propagation for Crowdsourced Classification

Jungseul Ok[*] OCKJS@KAIST.AC.KR
Sewoong Oh[†] SWOH@ILLINOIS.EDU
Jinwoo Shin[*] JINWOOS@KAIST.AC.KR
Yung Yi[*] YIYUNG@KAIST.AC.KR
[*]EE Department, Korea Advanced Institute of Science and Technology, Daejeon 34141 South Korea
[†]IESE Department, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA

Abstract

Crowdsourcing systems are popular for solving large-scale labelling tasks with low-paid (or even non-paid) workers. We study the problem of recovering the true labels from noisy 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 under a simple but canonical scenario where each worker is assigned at most two tasks. In particular, we introduce a tighter lower bound on the fundamental limit and prove 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. In the general setting, when more than two tasks are assigned to each worker, we establish the dominance result on BP that it outperforms other existing algorithms with known provable guarantees. Experimental results suggest that BP is close to optimal for all regimes considered, while existing state-of-the-art algorithms exhibit suboptimal performances.

1. 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 (Completely Automated Public Turing test to tell Computers and Humans Apart, 2000) 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 labelling quality. 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, and there are adversaries, who may deliberately give 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 outperforms majority voting, have been recently proposed, e.g. (Smyth et al., 1995; Jin & Ghahramani, 2003; Whitehill et al., 2009; Welinder et al., 2010; Raykar et al., 2010). The key idea is to identify the good workers and give more weights to the answers from those workers. Although the ground truth 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 of (Dawid & Skene, 1979). 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 experi-

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

ments. These include algorithms based on spectral methods (Ghosh et al., 2011; Dalvi et al., 2013; Karger et al., 2011; 2013; 2014), Belief Propagation (BP) (Liu et al., 2012), Expectation Maximization (EM) (Liu et al., 2012; Zhang et al., 2014), maximum entropy (Zhou et al., 2012; 2015), weighted majority voting (Littlestone & Warmuth, 1989; Li et al., 2013; Li & Yu, 2014), and combinatorial approaches (Gao & Zhou, 2013).

Despite the algorithmic advances, theoretical advances have been relatively slow. Some upper bounds on the performances are known (Karger et al., 2011; Zhang et al., 2014; Gao & Zhou, 2013), 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, under some simple but canonical scenarios. Under some assumptions on the parameters of the problem, we show the following:

Belief propagation is exactly optimal.

To the best of our knowledge, this is the first result proving an algorithm is not only computationally efficient but also provably maximizing the fraction of correctly labeled tasks, i.e., achieving exact optimality.

Contribution. We identify simple but canonical regimes where the standard BP achieves the performance of the optimal MAP estimator. Namely, when tasks are binary-classifiable, each worker is assigned at most two tasks, and each task is assigned to a sufficient 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 parameters.

Although our analysis techniques do not generalize to the case when each worker is assigned more than two tasks, our experimental results suggest that the optimality of BP generally holds for all regimes considered, while all other algorithms are sub-optimal in certain regimes for synthetic and real datasets we considered (See Section 5). Under this general scenario, we prove the following dominance result of our approach: BP correctly labels more tasks than two provable methods, the KOS algorithm from (Karger et al., 2011) and the majority voting.

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 (Mossel et al., 2014), error correcting codes (Kudekar et al., 2013) and combinatorial optimization (Park & Shin, 2015). Technically, our proof strategy for the optimality of BP is similar to that in (Mossel et al., 2014) 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, for example, in analyzing censored block models (Saade et al., 2015; Hajek et al., 2015) with observations that depend on the direction of the edges.

Related work. The crowdsourced classification problem has been first studied in the *dense regime*, where all tasks are assigned all the workers (Ghosh et al., 2011; Zhang et al., 2014). In this paper, we focus on the *sparse regime*, where each task is assigned to a few workers. Suppose ℓ workers are assigned each task. In practical crowdsourcing systems, a typical choice of ℓ is three or five. For a fixed ℓ, the probability of error now does not decay with increasing dimension of the problem. The theoretical interest is 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 (Karger et al., 2011), 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. A similar bound was also proved for another spectral approach, but under more restricted conditions in (Dalvi et al., 2013). Our main results provide an algorithm that (when $\ell > C$ for some constant C and each worker is assigned two tasks) correctly labels the optimal fraction of tasks, in the sense that it is information-theoretically impossible to correctly label a larger fraction for any other algorithms.

These spectral approaches are popular due to simplicity, but empirically do not perform as well as BP. In fact, the authors in (Liu et al., 2012) showed that the state-of-the-art spectral approach proposed in (Karger et al., 2011) is a special case of BP with a specific choice of the prior on the worker qualities. Since the algorithmic prior might be in mismatch with the true prior, the spectral approach is in general suboptimal.

Organization. In Section 2, we provide necessary backgrounds including the Dawid-Skene model for crowdsourced classification and the BP algorithm. Section 3 pro-

vides the main results of this paper, and their proofs are presented in Section 4. Our experimental results on the performance of BP are reported in Section 5 and we conclude in Section 6.

2. Preliminaries

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

2.1. Crowdsourced Classification Problem

We consider a set of n binary tasks, denoted by V. Each task $i \in V$ is associated with an arbitrary but latent ground truth $s_i \in \{-1, +1\}$. We assume that s_i's are independently chosen uniformly at random. 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 (Dawid & Skene, 1979). 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 constants ℓ, r using, for example, the configuration model (Bollobás, 1998).[1] 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 (Karger et al., 2011), and empirically shown to have good performances. This is due to the fact that the random graphs have large spectral gaps.

[1] We assume constants ℓ, r for simplicity, but our results hold as long as $\ell r = O(\log n)$.

2.2. 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.,

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

where we define

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

The probability is taken with respect to s and A for given G and π. From standard Bayesian arguments, the maximum a posteriori (MAP) estimator is an optimal solution of (1):

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

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

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

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,

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

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

$$\Pr[s_i \mid A] = \sum_{s_{-i} \in \{\pm 1\}^{V \setminus i}} \Pr[s \mid A] \propto \sum_{s_{-i}} \prod_{u \in W} f_u\left(s_{N_u}\right). \quad (5)$$

We note that the summation in (5) is taken over exponentially many $s_{-i} \in \{-1, +1\}^{n-1}$ 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*.

2.3. 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:

$$m_{i \to u}^{t+1}(s_i) \propto \prod_{v \in M_i \setminus \{u\}} m_{v \to i}^{t}(s_i)\,, \quad (6)$$

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

$$b_i^{t+1}(s_i) \propto \prod_{u \in M_i} m_{u \to i}^{t+1}(s_i)\,, \quad (8)$$

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:

$$\hat{s}_i^{\mathsf{BP}} = \arg\max_{s_i} b_i^k(s_i). \quad (9)$$

We note that if the factor graph is a tree, then it is known that the belief converges, and computes the exact marginal probability (Pearl, 1982).

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

$$b_i^t(s_i) = \Pr[s_i \mid A] \quad \textit{for all } t \geq n$$

where $b_i^t(s_i)$ 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 $b_i(s_i) \neq \Pr[s_i \mid A]$. Further the convergence of BP is not guaranteed, i.e., the value of $\lim_{t \to \infty} b_i^t(s_i)$ may not exist.

3. 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 $k = \log\log n$. Then, one can check that the overall complexity of BP is bounded by $O(n\ell r \log r \cdot \log\log n)$ because each iteration of BP requires $O(n\ell r \log r)$ operations (Liu et al., 2012).

3.1. Exact Optimality of BP for $r \leq 2$

We first state the following theorem on exact optimality of BP when the number of tasks assigned to each worker is at most two.

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. For $\mu := \mathbb{E}[2p_u - 1] > 0$ and $r \in \{1, 2\}$, there exists a constant $C_{\mu,r}$ that only depends on μ and r such that if $\ell \geq C_{\mu,r}$, then*

$$\lim_{n \to \infty} \mathbb{E}\left[\min_{\hat{s}:\textit{estimator}} P_{\mathsf{err}}(\hat{s}) - P_{\mathsf{err}}(\hat{s}^{\mathsf{BP}})\right] = 0\,,$$

where $\hat{s}^{\mathsf{BP}}$ is the output of BP in (9) *with $k = \log\log n$ and the expectation is taken with respect to the graph G.*

A proof is provided in Section 4.1. Our analysis compares BP to an *oracle* estimator. This 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* (Bollobás, 1998) and BP is exact on the local tree (Pearl, 1982), 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. This establishes that the gap between BP and the best estimator vanishes, in the large system limit.

The assumption on μ is mild, 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. We require $k = \Theta(\log\log n)$ to ensure we have a tree within the neighborhood of depth k, and at the same time include enough labels in the process to ensure convergence to optimality.

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 $C_{\mu,1} = 1$. The interesting non-trivial case is when $r = 2$. The sufficient condition is for ℓ to be larger than some $C_{\mu,2}$. Although experimental

results in Section 5 suggest that BP is optimal in all regimes considered, proving optimality for $r > 2$ or $\ell < C_{\mu,2}$ requires new analysis techniques, beyond those we develop in this paper. Both the problem of analyzing BP for $\ell < C_{\mu,r}$ (sample sparse regime) and for $r > 2$ (with higher-order factor nodes) are challenging problems. Similar challenges have not been resolved even in a simpler models of stochastic block models, where BP and other efficient inference algorithms have been analyzed extensively (Mossel et al., 2014; Bordenave et al., 2015).

3.2. Relative Dominance of BP for $r \geq 3$

For general ℓ and 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 (Karger et al., 2011). In the sparse regime, where $\ell r = O(\log n)$, these are the only algorithms with 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. Then, for any $\ell, r \geq 1$ such that $\ell r = O(\log n)$,*

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

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

A proof of the above theorem is presented in Section 4.2. Using Theorem 2 and the known error rates of MV and KOS algorithms in (Karger et al., 2011), one can derive the following upper bound on the error rate of BP:

$$\lim_{n\to\infty} \mathbb{E}\left[P_{\mathsf{err}}(\hat{s}^{\mathsf{BP}})\right] \leq \min\left\{e^{-\left(\frac{\ell\mu^2}{2}\right)}, e^{-\left(\frac{\ell q}{2}\cdot\frac{q^2(\ell-1)(r-1)-1}{3q^2(\ell-1)(r-1)+q(\ell-1)}\right)}\right\} \quad (10)$$

where $q := \mathbb{E}\left[(2p_u - 1)^2\right]$.

This is particularly interesting, since it has been observed empirically and conjectured with some non-rigorous analysis in (Karger et al., 2014) 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 1). 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 1.

4. Proofs of Theorems

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

4.1. 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 $\Pr[s_\rho|A] = \Pr[s_\rho|A_{\rho,1}]$, 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

$$\hat{s}_\rho^{\mathsf{BP}} = \arg\max_{s_\rho} \Pr[s_\rho \mid A_{\rho,1}] = \hat{s}_\rho^*(A_{\rho,1}).$$

Therefore, for any $\ell \geq 1$, the optimal MAP estimator $\hat{s}_\rho^*(A)$ in (2) is identical to the output $\hat{s}_\rho^{\mathsf{BP}}$ with any $k \geq 1$.

From now on, we focus on the case $r = 2$, and we condition on a fixed task assignment graph G. All the arguments holds for general r, but the key technical lemma requires r to be exactly two (Lemma 3). We will identify a condition under which the desired claim holds, and show that under the random (ℓ, r)-regular model this condition holds with sufficiently large probability.

Define $\rho \in V$ as a random node chosen uniformly at random and let $\Delta(\hat{s}_\rho)$ denote the gain of estimator $\hat{s}_\rho$ compared to random guessing, i.e.,

$$\Delta(\hat{s}_\rho) := \frac{1}{2} - \Pr[s_\rho \neq \hat{s}_\rho] \text{ and } P_{\mathsf{err}}(\hat{s}) = \frac{1}{2} - \Delta(\hat{s}_\rho)$$

where the expectation is taken with respect to the distribution of G. Then it is enough to show that $\Delta(\hat{s}_\rho^*(A))$ and $\Delta(\hat{s}_\rho^{\mathsf{BP}})$ converge to the same value, i.e., the limit value of $\lim_{n\to\infty} \mathbb{E}[\Delta(\hat{s}_\rho^*(A))]$ exists and as $n \to \infty$,

$$\mathbb{E}\left[\Delta(\hat{s}_\rho^*(A)) - \Delta(\hat{s}_\rho^{\mathsf{BP}})\right] \to 0 \quad (11)$$

where the expectation is taken with respect to the distribution of G.

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. 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[2] whose distance from ρ is exactly $2k$. We

[2]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.

now define the following *oracle* estimator:

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

where

$$A_{\rho,2k} := \{A_{iu} : (i,u) \in E_{\rho,2k}\} \tag{12}$$

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 the supplementary material.

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

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

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

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

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

On estimating task ρ, BP at k-th iteration on G is identical to BP on $G_{\rho,2k}$. If $G_{\rho,2k}$ is a tree, then from Property 1, BP calculates the exact marginal probability of s_ρ given $A_{\rho,2k}$, i.e., if $G_{\rho,2k}$ is a tree

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

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

$$\begin{aligned}\Delta(\hat{z}^*_\rho(A_{\rho,2k})) &\geq \Delta(\hat{s}^*_\rho(A)) \\ &\geq \Delta(\hat{s}^{\mathsf{BP}}_\rho) = \Delta(\hat{s}^*_\rho(A_{\rho,2k}))\end{aligned} \tag{13}$$

where we define $\hat{s}^*(A_{\rho,2k}) := \arg\max \Pr[s_\rho \mid A_{\rho,2k}]$.

Consider now a random (ℓ, r)-regular bipartite graph G, which is a locally tree-like. More formally, from Lemma 5 in (Karger et al., 2014), if follows that

$$\Pr[G_{\rho,2k} \text{ is not a tree}] \leq \frac{3\ell r}{n}\left((\ell-1)(r-1)\right)^{2k}. \tag{14}$$

Hence, by taking the expectation with respect to G and applying (14) to (13), we get

$$\begin{aligned}0 &\leq \mathbb{E}\left[\Delta(\hat{s}^*_\rho(A)) - \Delta(\hat{s}^{\mathsf{BP}}_\rho)\right] \\ &\leq \mathbb{E}\left[\Delta(\hat{z}^*_\rho(A_{\rho,2k})) - \Delta(\hat{s}^*_\rho(A_{\rho,2k}))\right] + \frac{3}{n}(\ell r)^{2k+1}\end{aligned} \tag{15}$$

where the last term in the RHS goes 0 as $n \to \infty$ if $\ell r = O(\log n)$ and $k = O(\log\log n)$. In addition, from the following lemma, the first term in the RHS also converges to 0 since we set $k = \Theta(\log\log n)$. Hence, this implies (11) and the existence of the limit of $\lim_{n\to\infty} \mathbb{E}[\Delta(\hat{s}^*_\rho(A))]$ due to the bounded and non-increasing sequence of $\Delta(\hat{z}^*_\rho(A_{\rho,2k}))$ in Lemma 1. We complete the proof of Theorem 1.

Lemma 3. *Suppose $G_{\rho,2k} = (V_{\rho,2k}, W_{\rho,2k}, E_{\rho,2k})$ is 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 two tasks. For a given $\mu := \mathbb{E}[2p_u - 1] > 0$, there exists a constant $C_{\mu,2}$ such that if $\ell \geq C_{\mu,2}$, then as $k \to \infty$*

$$\left|\Delta(\hat{z}^*_\rho(A_{\rho,2k})) - \Delta(\hat{s}^*_\rho(A_{\rho,2k}))\right| \to 0 \tag{16}$$

A rigorous proof of Lemma 3 is given in the supplementary material. 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.

4.2. Proof of Theorem 2

We note that that KOS is an iterative algorithm where for each $\rho \in V$ and $k \geq 1$, $\hat{s}^{\mathsf{KOS},k}_\rho$ depends on only $A_{\rho,2k}$ defined in (12). 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.,

$$\Delta(\hat{s}^*_\rho(A_{\rho,2k})) \geq \max\left\{\Delta(\hat{s}^{\mathsf{MV}}_\rho), \Delta(\hat{s}^{\mathsf{KOS},k}_\rho)\right\}. \tag{17}$$

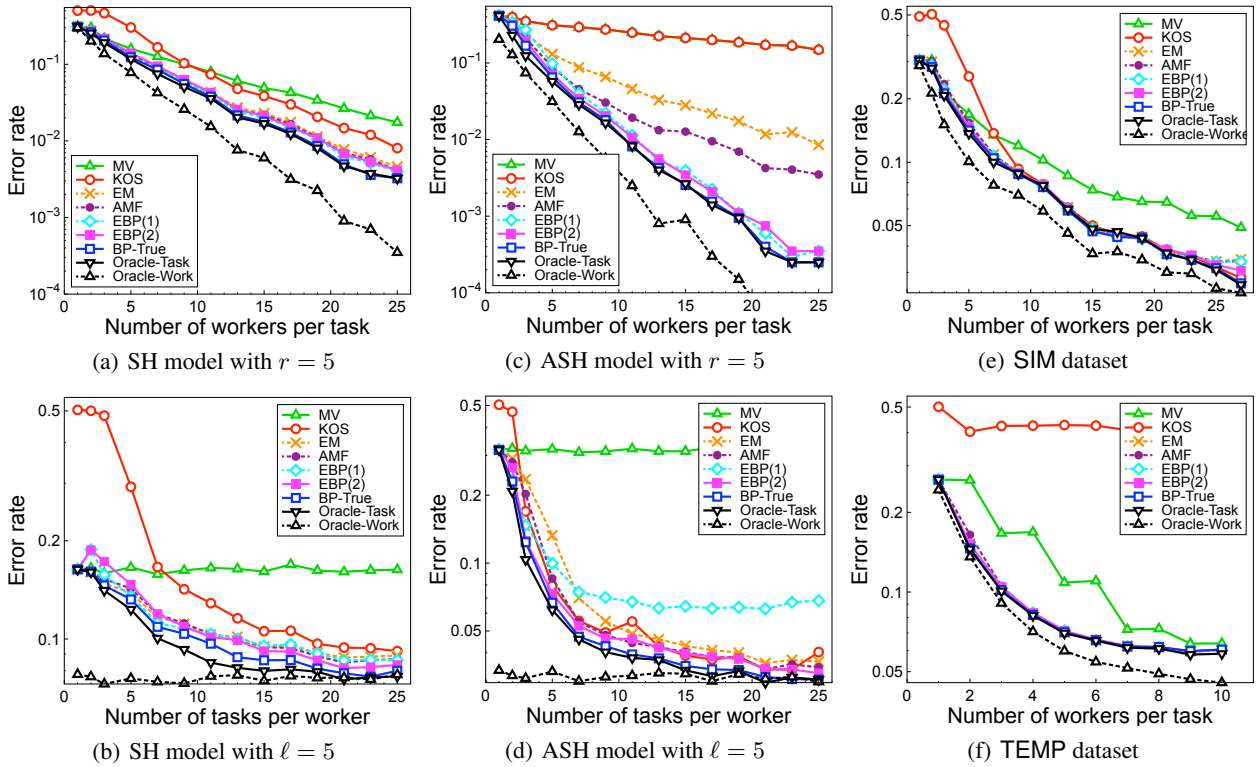
Recall that if $G_{\rho,2k}$ is a tree, we have $\hat{s}^{\mathsf{BP},k}_\rho = \hat{s}^*_\rho(A_{\rho,2k})$. Similarly to (15), by taking the expectation with respect to G, it follows that

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

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

5. 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) 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

Figure 1. 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 (Karger et al., 2014); (b) Temporal ordering (TEMP) dataset with 462 tasks and 76 workers obtained in (Snow et al., 2008).

5.1. Tested Algorithms

We compare BP and a variant of BP to two oracle algorithms and several state-of-the-art algorithms in (Dawid & Skene, 1979; Karger et al., 2011; Liu et al., 2012), 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, i.e., reliability distribution π. However, in practice, the distribution π 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 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, called Oracle-Task, which has an omniscient access to a subset of the true labels of tasks to label each task. Oracle-Task estimates 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. Due to the exactness of BP on a tree in Property 1 and Lemma 1, we can obtain the lower bound in time $O(n^2 lr \log r)$.

Existing algorithms. For comparison to the state-of-the-art algorithms, we test the majority voting (MV), an iterative algorithm (KOS) (Karger et al., 2011)), the expectation maximization (EM) (Dawid & Skene, 1979)) and an approach based on approximate mean field (AMF) (Liu et al., 2012)). Specifically, as the authors in (Liu et al., 2012) suggested, we run EM and AMF with 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.

5.2. 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 (Bollobás, 1998), 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 $\pi(0.1) = \pi(0.5) = 1/2$ and $\pi(0.9) = 1/2$, whose results are plotted in Figures 1(a)-1(b) and Figures 1(c)-1(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 1(c)). As discussed in (Karger et al., 2011), we observe a threshold behavior at $(\ell - 1)(r - 1) = 1/q^2$ where 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 two rounds (EBP(2)), achieves almost the same performance as BP-True, as shown in Figure 1(d). Note that MV performs poorly since the number of workers per task is small and the quality of workers, $\mu = \mathbb{E}[2p_u - 1]$, is small. Figure 1(d) shows that EBP with a single round leads to moderate performance improvement, but one additional round in EBP(2) provides 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 (Karger et al., 2011) 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 1(b) and 1(d)).

5.3. Performance on Real Datasets

We use two real-world Amazon Mechanical Turk datasets from (Karger et al., 2011) and (Snow et al., 2008): 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 1(e) and 1(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.

6. Conclusion and Discussion

In this paper, we investigate the question of optimality and computational gap for a canonical scenario for the crowdsourced classification where the tasks are binary. Here we list some interesting theoretical questions left open for future research.

First, it would be interesting to tighten the constants in the error exponent of the upper bound in (10) since the actual performance of BP is better than predicted by this upper bound. We provide an oracle estimator that is significantly tighter than the naive oracle estimators presented in (Karger et al., 2011). This strong oracle can be numerically evaluated, as we showed 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 (10) would complete the investigation of optimality of BP. It has been observed in (Karger et al., 2011; 2013) 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, but a theoretical understanding is lacking.

It is also interesting to generalize our analysis in models with more practical aspects. First, one can consider multi-alphabet tasks having more than two classes. In this case, BP is naturally extended while KOS requires some modification (Karger et al., 2013). It is not hard to show the superiority of BP over MV and KOS with the same analysis in this paper. However, on the optimality of BP over a larger alphabet, we need new proof techniques to handle multi-alphabet tasks. Another generalization can be considering tasks with different difficulty levels. To capture such heterogeneity, several generalized models have been proposed (Raykar et al., 2010; Whitehill et al., 2009; Welinder et al., 2010; Snow et al., 2008; Sheng et al., 2008; Zhou et al., 2012; 2015). 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.

Acknowledgments

This work is supported in part by NSF SaTC award CNS-1527754, and NSF CISE award CCF-1553452. This work is supported in part by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government(MSIP) (No.R0132-15-1005), Content visual browsing technology in the online and offline environments. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.B0717-16-0034,Versatile Network System Architecture for Multi-dimensional Diversity)

References

Bollobás, B. *Random graphs*. Springer, 1998.

Bordenave, C., Lelarge, M., and Massoulié, L. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Proceedings of IEEE FOCS*, 2015.

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

Dalvi, N., Dasgupta, A., Kumar, R., and Rastogi, V. Aggregating crowdsourced binary ratings. In *Proceedings of WWW*, 2013.

Dawid, A. P. and Skene, A. M. 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.

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

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

Hajek, B., Wu, Y., and Xu, J. Exact recovery threshold in the binary censored block model. In *Proceedings of IEEE Information Theory Workshop*, 2015.

Jin, R. and Ghahramani, Z. Learning with multiple labels. In *Proceedings of NIPS*, 2003.

Karger, D. R., Oh, S., and Shah, D. Iterative learning for reliable crowdsourcing systems. In *Proceedings of NIPS*, 2011.

Karger, D. R., Oh, S., and Shah, D. Efficient crowdsourcing for multi-class labeling. In *Proceedings of ACM SIGMETRICS*, 2013.

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

Kudekar, Shrinivas, Richardson, Tom, and Urbanke, Rudiger L. Spatially coupled ensembles universally achieve capacity under belief propagation. *IEEE Transactions on Information Theory*, 59(12):7761–7813, 2013.

Li, H. and Yu, B. Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv preprint arXiv:1411.4086*, 2014.

Li, H., Yu, B., and Zhou, D. Error rate analysis of labeling by crowdsourcing. In *Proceedings of ICML*, 2013.

Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. In *Proceedings of IEEE FOCS*, 1989.

Liu, Q., Peng, J., and Ihler, A. T. Variational inference for crowdsourcing. In *Proceedings of NIPS*, 2012.

Mossel, E., Neeman, J., and Sly, A. Belief propagation, robust reconstruction and optimal recovery of block models. In *Proceedings of COLT*, 2014.

Park, S. and Shin, J. Max-product belief propagation for linear programming: applications to combinatorial optimization. In *Proceedings of UAI*, 2015.

Pearl, J. Reverend bayes on inference engines: A distributed hierarchical approach. In *Proceedings of AAAI*, 1982.

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

Saade, A., Lelarge, M., Krzakala, F., and Zdeborová, L. Spectral detection in the censored block model. In *Proceedings of IEEE ISIT*, pp. 1184–1188. IEEE, 2015.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of ACM SIGKDD*, 2008.

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

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. Cheap and fast-but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*. Association for Computational Linguistics, 2008.

Welinder, P., Branson, S., Belongie, S., and Perona, P. The multidimensional wisdom of crowds. In *Proceedings of NIPS*, 2010.

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of NIPS*, 2009.

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

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

Zhou, D., Liu, Q., Platt, J. C., Meek, C., and Shah, N. B. Regularized minimax conditional entropy for crowdsourcing. *arXiv preprint arXiv:1503.07240*, 2015.