Iterative Bayesian Learning for Crowdsourced Regression

Jungseul Ok[*] **Sewoong Oh**[*] **Yunhun Jang**[†] **Jinwoo Shin**[†] **Yung Yi**[†]
[*]University of Washington, [†]Korea Advanced Institute of Science and Technology

Abstract

Crowdsourcing platforms emerged as popular venues for purchasing human intelligence at low cost for large volume of tasks. As many low-paid workers are prone to give noisy answers, a common practice is to add redundancy by assigning multiple workers to each task and then simply average out these answers. However, to fully harness the wisdom of the crowd, one needs to learn the heterogeneous quality of each worker. We resolve this fundamental challenge in crowdsourced regression tasks, i.e., the answer takes continuous labels, where identifying good or bad workers becomes much more non-trivial compared to a classification setting of discrete labels. In particular, we introduce a Bayesian iterative scheme and show that it provably achieves the optimal mean squared error. Our evaluations on synthetic and real-world datasets support our theoretical results and show the superiority of the proposed scheme.

1 INTRODUCTION

Crowdsourcing systems provide a labor market where numerous pieces of classification and regression tasks are electronically distributed to a crowd of workers, who are willing to solve such human intelligence tasks at a low cost. However, because the pay is low and the tasks are tedious, error is common even among those who are willing. This is further complicated by abundant spammers trying to make easy money with little effort. To cope with such noise in the collected data, adding redundancy is a common and powerful strategy widely used in real-world crowdsourcing. Each task is assigned to multiple workers and these responses are aggregated by inference algorithms such as averaging (for real-valued answers) or majority voting (for categorial answers). As workers' qualities are heterogeneous, such simple approaches can be significantly improved upon by re-weighting the answers from reliable workers. Here, the fundamental challenge is identifying such workers, which requires estimating ground truth answers and vice-versa. Our focus is solving this inference problem, when neither true answers nor worker reliabilities are known.

For a simpler problem of *classification tasks*, where each task asks a worker to choose one label from a discrete set, significant advances have been made in the past decade (Karger et al., 2011; Liu et al., 2012; Khetan and Oh, 2016; Shah et al., 2016; Zhou et al., 2015; Zhang et al., 2014) based on the model proposed in the seminal work of (Dawid and Skene, 1979). Deep theoretical understanding of the model under a simple but canonical case of binary classification has led to the design of powerful inference algorithms, which significantly improve upon the common practice of majority voting on real-world datasets. However, neither the model nor the algorithms generalize to *regression tasks*, where each task asks for a continuous valued assessment, and possibly in multiple dimensions. Despite of the significance of the crowdsourced regression evidenced by the empirical studies (Everingham et al., 2015; Su et al., 2012; Deng et al., 2009; De Alfaro and Shavlovsky, 2014; Piech et al., 2013), the theoretical understanding of the crowdsourced regression has remained limited.

To bridge this gap, we take a principled approach on this crowdsourced regression problem to theoretically investigate the tradeoff involved. More precisely, we ask the fundamental question of how to achieve the best accuracy given a budget constraint, or equivalently how to achieve a target accuracy with minimum budget. As in typical crowdsourcing systems, we assume we pay a fixed amount for each response, and thus the budget per task is proportional to the redundancy: how many answers we collect for each task.

Contribution. Inspired by the simplicity of the model in (Dawid and Skene, 1979) for crowdsourced classification, we propose a simple, yet effective model

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

for crowdsourced regression. We introduce a Bayesian Iterative algorithm (BI) to solve the inference problem efficiently. We provide an upper bound on the error achieved by the proposed BI (Theorem 1) that captures (i) the fundamental tradeoff between redundancy and the accuracy, and (ii) the performance loss due to the difficulty in estimating workers' reliability. Further, we prove that it is information theoretically impossible for any other algorithm to improve upon BI. This is achieved by coupling the proposed inference algorithm with a carefully constructed oracle estimator, and showing that there is no gap in the performance between those two algorithms (Theorem 2). Such strong guarantees are only known for a few other cases even under more strict assumptions (which we discuss later). Finally, in numerical evaluation, we confirm our theoretical findings on synthetic and real-world datasets.

Related work. Crowdsourcing systems are widely used in practice for a variety of real-world tasks such as protein folding (Peng et al., 2013), searching videos (Bernstein et al., 2011; Salvo et al., 2013), ranking (Lee et al., 2012), peer assessment (Piech et al., 2013; Goldin and Ashley, 2011) and natural language processing (Wu et al., 2012). However, recent theoretical advances have been focused on crowdsourced classification tasks to (a) design algorithms for aggregating answers from multiple workers on the same task; (b) analyze the performance achieved by such algorithms; and (c) identify and compare against the fundamental limit (Karger et al., 2011, 2013; Ghosh et al., 2011; Zhang et al., 2014; Ok et al., 2016; Zhou et al., 2012; Dalvi et al., 2013; Liu et al., 2012; Karger et al., 2014). In this paper, we theoretically investigate these fundamental questions for *crowdsourced regression*.

There has been several novel algorithms recently proposed for the crowdsourced regression. Raykar et al. (2010) proposed a probabilistic model and a corresponding maximum likelihood estimator, but no supporting theoretical or empirical analysis is provided (as the estimator is intractable). Zhou et al. (2015) propose a heuristic of quantizing the continuous valued answers and reducing it to discrete models, i.e. crowdsourced classification. On top of being sensitive to hyperparameter choices such as the quantization level, treating the answers as categories loses the fundamental aspect that the answers are given in a metric space where distances are well-defined.

A related work is (Liu et al., 2013), in which the authors provide a theoretical understanding in a *semi-supervised* setting. All workers are first asked golden questions with known answers, which is used to estimate all unknown parameters of the workers. Then, they are assigned to tasks with unknown answers, and their responses are aggregated using the estimated parameters. As this two phase approach completely decouples the uncertainty in worker parameters and task answers, the analysis is extremely simple and is not applicable to our *unsupervised* setting.

Finally, we remark that the proposed algorithm BI is a variant of the popular Belief propagation (BP). Although BP enjoys numerous empirical successes in various fields (Jordan, 2004), its theoretical analysis has been limited to a few instances including community detection (Mossel et al., 2014) and error correcting codes (Kudekar et al., 2013). In particular, those analyses showing the optimality of loopy BP (Mossel et al., 2014; Ok et al., 2016) are limited to cases where the corresponding factor graph has only factor degree two. Our main result (Theorem 2) extends the horizon of such cases where BI provably finds the optimal inference under an arbitrary factor degree while the regression problem is more challenging to analyze than the discrete models studied in (Mossel et al., 2014; Ok et al., 2016) as the regression error is unbounded.

2 PROBLEM FORMULATION

2.1 Crowdsourced Regression Model

The task requester has a set of n regression tasks, denoted by $V = \{1, \dots, n\}$, where task $i \in V$ is associated with the true position $\mu_i \in \mathbb{R}^d$. To estimate these unknown true positions, we assign the tasks to a set of m workers, denoted by $W = \{1, \dots, m\}$ according to a bipartite graph $G = (V, W, E)$, where edge $(i, u) \in E$ indicates that task i is assigned to worker u. We also let $N_u := \{i \in V : (i, u) \in E\}$ and $M_i := \{u \in W : (i, u) \in E\}$ denote the set of tasks assigned to worker u and the set of workers to whom task i is assigned, respectively.

When task i is assigned to worker u, she provides her estimation/guess $A_{iu} \in \mathbb{R}^d$ for the true location μ_i. Each worker u is parameterized by her noise level σ_u^2, such that the response A_{iu} suffers from an additive spherical Gaussian noise with variance σ_u^2. Precisely, conditioned on μ_i and σ_u^2, A_{iu} is independently distributed with Gaussian pdf $f_{A_{iu}}(x \,|\, \mu_i, \sigma_u^2) = \phi(x \,|\, \mu_i, \sigma_u^2) := \exp(-\|x - \mu_i\|_2^2/(2\sigma_u^2))/\sqrt{(2\pi\sigma_u^2)^d}$.

We assume that each worker u's variance σ_u^2 is independently drawn from a finite set $\mathcal{S} = \{\sigma_1^2, ..., \sigma_S^2\}$ uniformly at random. We further assume that the true position μ_i is independently drawn from a Gaussian prior distribution $\phi(x \,|\, \nu_i, \tau^2)$ for given mean $\nu_i \in \mathbb{R}^d$ and variance $\tau^2 \in (0, \infty)$, which can be interpreted as a side information on true positions. Note that we just take the Gaussian prior for the simple expression and our analysis can be generalized to other distributions, e.g., a uniform distribution on a Euclidean ball. Our

analysis is valid for arbitrarily large τ, i.e., no prior information, and our numerical experiments assume no knowledge of the prior distribution by taking $\tau \to \infty$. Theoretical understanding of such a simple but canonical model allows us to characterize the tradeoffs involved and provides guidelines for designing practical algorithms.

2.2 Optimal but Intractable Algorithm

Under the crowdsourcing model, our goal is to design an efficient estimator $\hat{\mu}(A) \in \mathbb{R}^{d\times V}$ of the unobserved true position μ from the noisy answers $A := \{A_{iu} : (i,u) \in E\}$ reported by workers. In particular, we are interested in minimizing the average of *(expected) mean squared error (MSE)*, i.e.,

$$\underset{\hat{\mu}:\text{estimator}}{\text{minimize}} \quad \frac{1}{n}\sum_{i\in V}\mathbb{E}[\text{MSE}(\hat{\mu}_i(A))] \tag{1}$$

where we define $\text{MSE}(\hat{\mu}_i(A)) := \mathbb{E}[\|\hat{\mu}_i(A)-\mu_i\|_2^2 \,|\, A]$ as the MSE conditioned on A. Using the equality $(\hat{\mu}_i(A)-\mu_i) = (\hat{\mu}_i(A)-\mathbb{E}[\mu_i \,|\, A]) + (\mathbb{E}[\mu_i \,|\, A]-\mu_i)$, it is straightforward to check that for each $i \in V$, MSE is minimized at the *Bayesian* estimator $\hat{\mu}_i^*(A) := \mathbb{E}[\mu_i \,|\, A]$, which is

$$\hat{\mu}_i^*(A) = \sum_{\sigma^2_{M_i}\in\mathcal{S}^{M_i}} \bar{\mu}_i\left(A_i, \sigma^2_{M_i}\right)\mathbb{P}[\sigma^2_{M_i} \,|\, A] \tag{2}$$

where we let $A_i := \{A_{iu} : u \in M_i\}$ and $\bar{\mu}_i(A_i, \sigma^2_{M_i}) := \mathbb{E}[\mu_i \,|\, A_i, \sigma^2_{M_i}] = \bar{\sigma}_i^2\left(\sigma^2_{M_i}\right)(\nu_i/\tau^2 + \sum_{u\in M_i} A_{iu}/\sigma_u^2)$ with $\bar{\sigma}_i^2(\sigma^2_{M_i}) := (1/\tau^2 + \sum_{u\in M_i} 1/\sigma_u^2)^{-1}$. We provide a derivation of this formula in the supplementary material. The calculation of the marginal posterior $\mathbb{P}[\sigma^2_{M_i} \,|\, A]$ is computationally intractable in general. More formally, the marginal posterior of $\sigma^2_{M_i}$ can be calculated by marginalizing out $\sigma^2_{-i} := \{\sigma_v^2 : v \in W \setminus M_i\}$ from the joint probability of σ^2, i.e.,

$$\mathbb{P}[\sigma^2_{M_i} \,|\, A] = \sum_{\sigma^2_{-i}\in\mathcal{S}^{W\setminus M_i}} \mathbb{P}[\sigma^2 \,|\, A] \tag{3}$$

which requires exponentially many summations with respect to m. Thus, the optimal estimator $\hat{\mu}^*(A)$ in (2), requiring the marginal posterior $\mathbb{P}[\sigma^2_{M_i} \,|\, A]$ in (3), is *computationally intractable* in general.

3 ITERATIVE ALGORITHM

We now introduce a computationally tractable scheme, the Bayesian iterative (BI) algorithm, and provide its theoretical guarantees under the crowdsourced regression model. For its analytic tractablity, we consider a popular assignment scheme, referred to as (ℓ, r)-regular task assignment, widely adopted in crowdsourcing (Karger et al., 2011; Ok et al., 2016). The assignment graph G is a random (ℓ, r)-regular bipartite graph drawn uniformly at random out of all (ℓ, r)-regular graphs, where each task is assigned to ℓ workers and each worker is assigned r tasks. Nevetheless, we remark that the BI algorithm is applicable to any (even, non-regular) task assignments.

3.1 Bayesian Iterative (BI) Algorithm

We first factorize the joint probability of σ^2 in (3) as

$$\mathbb{P}[\sigma^2 \,|\, A] \propto \prod_{i\in V} \mathcal{C}_i\left(A_i, \sigma^2_{M_i}\right)$$

where $\mathcal{C}_i(A_i, \sigma^2_{M_i}) := \left(\frac{\bar{\sigma}_i^2(\sigma^2_{M_i})}{\tau^2\prod_{u\in M_i} 2\pi\sigma_u^2}\right)^{\frac{d}{2}} e^{-\mathcal{D}_i(A_i,\sigma^2_{M_i})}$, and $\mathcal{D}_i(A_i, \sigma^2_{M_i}) := \frac{\bar{\sigma}_i^2(\sigma^2_{M_i})}{2}\left(\sum_{u\in M_i}\frac{\|A_{iu}-\nu_i\|_2^2}{\sigma_u^2\tau^2} + \sum_{v\in M_i\setminus\{u\}}\frac{\|A_{iu}-A_{iv}\|_2^2}{\sigma_u^2\sigma_v^2}\right)$.

We provide a derivation of this formula in the supplementary material. This factorization of the joint probability of σ^2 given A forms a factor graph (Jordan, 1998) where each worker u's variance σ_u^2 and each task i correspond to a variable and a local factor $\mathcal{C}_i(A_i, \sigma^2_{M_i})$ on the set of workers, M_i, to whom task i is assigned, respectively. This probabilistic graphical model motivates us to use the popular (sum-product) belief propagation (BP) algorithm (Pearl, 1982) on the factor graph of $\mathbb{P}[\sigma^2|A]$ to approximate the intractable computation of $\mathbb{P}[\sigma^2_{M_i} \,|\, A]$ in (3). However, BP is typically used for approximating the marginal probability of a single variable σ_u^2, while we need the marginal probability of a subset of variables $\sigma^2_{M_i}$ depending on each other. Hence, to approximate the optimal Bayesian estimator in (2), we build upon BP and propose an iterative algorithm (BI) updating belief $b_i(\sigma^2_{M_i})$ from messages $m_{i\to u}$ and $m_{u\to i}$ between task i and worker u:

$$m_{i\to u}^{t+1}(\sigma_u^2) \propto \sum_{\sigma^2_{M_i\setminus\{u\}}} \mathcal{C}_i(A_i, \sigma^2_{M_i}) \prod_{v\in M_i\setminus\{u\}} m_{v\to i}^t(\sigma_v^2) \tag{4}$$

$$m_{u\to i}^{t+1}(\sigma_u^2) \propto \prod_{j\in N_u\setminus\{i\}} m_{j\to u}^{t+1}(\sigma_u^2) \tag{5}$$

$$b_i^{t+1}(\sigma^2_{M_i}) \propto \mathcal{C}_i(A_i, \sigma^2_{M_i}) \prod_{u\in M_i} m_{u\to i}^{t+1}(\sigma_u^2) \tag{6}$$

where we initialize the messages with a trivial constant $1/|\mathcal{S}|$ and normalize the messages and beliefs so that $\sum_{\sigma_u^2} m_{i\to u}^t(\sigma_u^2) = \sum_{\sigma_u^2} m_{u\to i}^t(\sigma_u^2) = \sum_{\sigma^2_{M_i}} b_i^t(\sigma^2_{M_i}) = 1$. At the end of k iterations, as an approximation of the optimal Bayesian estimator in (2), we estimate $\hat{\mu}^{\text{BI}(k)}(A)$ using (2) with belief $b_i^k(\sigma^2_{M_i})$ as an approximation of $\mathbb{P}[\sigma^2_{M_i} \,|\, A]$. Formally,

$$\hat{\mu}_i^{\text{BI}(k)}(A) := \sum_{\sigma^2_{M_i}\in\mathcal{S}^{M_i}} \bar{\mu}_i(A_i, \sigma^2_{M_i}) b_i^k(\sigma^2_{M_i})\,. \tag{7}$$

Although the messages and their updates are the same as those of the typical BP, we use a specific form of belief in (6) for approximating the marginal probability of a *subset* of dependent variables. This allows us to provide sharp performance guarantees in Section 3, while the typical BP for *single* variable marginalization has little known provable guarantees.

We note that if the factor graph is a *tree*, i.e., having no loop, then it is not hard to check that the iterative algorithm calculates the exact value of the marginal posterior of multiple variables $\sigma^2_{M_i}$ since

$$\mathbb{P}[\sigma^2_{M_i} \,|\, A] \propto \mathcal{C}_i(A_i, \sigma^2_{M_i}) \prod_{u \in M_i} \mathbb{P}[\sigma^2_u \,|\, A_{-i}]$$

where $A_{-i} := A \setminus A_i$. More formally, if the assignment graph G is a tree from task i with depth $2k$, then we have $b^t_i(\sigma^2_{M_i}) = \mathbb{P}[\sigma^2_{M_i} \,|\, A]$ for all $t \geq k$. However, for general graphs with loops, the typical BP has no guarantee on neither the approximation error nor the convergence of BP while it has been successfully applied to many applications (Murphy et al., 1999; Yanover et al., 2006). Perhaps surprisingly, we can analytically explain such empirical success for crowdsourced regression with strong guarantees in the following section.

3.2 Quantitative Performance Guarantee

We first present a performance guarantee of BI estimator that is close to that of an *oracle* estimator. The proof is in Section 4.1.

Theorem 1. *Consider the crowdsourced regression model with $\mathcal{S} = \{\sigma^2_1, ..., \sigma^2_S\}$ and a random (ℓ, r)-regular graph G consisting of n tasks and $(\ell/r)n$ workers. For given $\varepsilon, \sigma^2_{\min}, \sigma^2_{\max} > 0$ and $\ell \geq 2$, if* (i) $|\sigma^2_s - \sigma^2_{s'}| > \varepsilon$ *and* $\sigma^2_{\min} \leq \sigma^2_s \leq \sigma^2_{\max}$ *for all* $1 \leq s \neq s' \leq S$, *and* (ii) $2 \leq r, k \leq \log\log n$, *then for sufficiently large n, BI in (7) with k iterations achieves*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i \in V} \mathrm{MSE}(\hat{\mu}_i^{BI(k)}(A))\right] \leq \frac{d}{n}\sum_{i \in V} \mathbb{E}\left[\bar{\sigma}_i^2\left(\sigma^2_{M_i}\right)\right] \quad (8a)$$

$$+ \mathcal{E}_{\ell,\mathcal{S}} \ell^{1/4}\left(4\exp\left(-\frac{\varepsilon^2 r}{8(8\varepsilon+1)\sigma^2_{\max}}\right) + 2^{-k}\right)^{1/4} \quad (8b)$$

where $\mathcal{E}_{\ell,\mathcal{S}} := 2d(\frac{1}{\tau^2} + \ell\frac{\sigma^2_{\max}}{\sigma^4_{\min}})(\frac{1}{\tau^2} + \frac{\ell}{\sigma^2_{\min}})^{-2}$ and the expectation is taken w.r.t. G and A.

We provide three interpretations of Theorem 1. First, consider an oracle estimator that knows the hidden variances σ^2_u's and makes optimal inference as $\hat{\mu}_i^{\mathsf{ora}}(A, \sigma^2) := \mathbb{E}[\mu_i \,|\, A, \sigma^2] = \bar{\mu}_i\left(A_i, \sigma^2_{M_i}\right)$. This gives the MSE of $\hat{\mu}_i^{\mathsf{ora}}(A, \sigma^2)$:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i \in V} \mathrm{MSE}(\hat{\mu}_i^{\mathsf{ora}}(A, \sigma^2))\right] = \frac{d}{n}\sum_{i \in V} \mathbb{E}\left[\bar{\sigma}_i^2\left(\sigma^2_{M_i}\right)\right] .$$

Note that the oracle estimator $\hat{\mu}^{\mathsf{ora}}$ always outperforms even the optimal estimator $\hat{\mu}^*$ in (2), providing a lower bound on the MSE of any estimator. This coincides with (8a) in our bound, implying that the gap (8b) to the oracle performance (8a) quantifies the *difficulty* in identifying reliable workers. We stress that considering a weaker oracle that captures the difficulty in estimating worker reliability, should give a tighter lower bound than (8a). This is stated precisely in the following section (see Theorem 2).

Second, for sufficiently large n, when the number r of per-worker tasks and the total iterations k grow with n, the performance of BI quickly approaches that of the oracle estimator, as (8b) vanishes exponentially. This is because under (ℓ, r)-regular task assignment, for increasing r with the total number of tasks n, the iterative algorithm accurately infers all workers' variances and thus optimally estimates the true positions μ. Note that the above performance limit holds for any $r = \omega(1)$, implying that a reasonable number of tasks per worker is enough to achieve a performance close to the oracle bound.

Third, we compare BI with simple averaging, i.e., $\hat{\mu}_i^{\mathsf{avg}}(A) := \sum_{u \in M_i} A_{iu}/|M_i|$, which achieves

$$\mathbb{E}\left[\frac{1}{n}\sum_{i \in V} \mathrm{MSE}(\hat{\mu}_i^{\mathsf{avg}}(A))\right] = \frac{d}{n}\sum_{i \in V} \mathbb{E}\left[\frac{\sum_{u \in M_i}\sigma^2_u}{|M_i|^2}\right] .$$

Note that $\mathbb{E}[\mathrm{MSE}(\hat{\mu}_i^{\mathsf{avg}}(A))]$ increases proportionally to the arithmetic mean of variances of workers assigned to each task, while $\mathbb{E}[\mathrm{MSE}(\hat{\mu}_i^{\mathrm{BI(k)}}(A))]$ is proportional to the harmonic mean of variances of workers and prior, i.e., $\mathbb{E}[\mathrm{MSE}(\hat{\mu}_i^{\mathsf{avg}}(A))] \geq \mathbb{E}[\mathrm{MSE}(\hat{\mu}_i^{\mathrm{BI(k)}}(A))]$. This gap can be made arbitrarily large by increasing the difference between the maximum and minimum variances of workers. For example, if a single worker $u \in M_i$ assigned to task i has high accuracy, i.e., $\sigma^2_u \simeq 0$, and the others' variances are x's, then $\mathbb{E}[\mathrm{MSE}(\hat{\mu}_i^{\mathsf{avg}}(A))] \simeq (d/|M_i|)x$ but $\mathbb{E}[\mathrm{MSE}(\hat{\mu}_i^{\mathrm{BI(k)}}(A))] \simeq 0$. Hence, the existence of a single worker with high precision in each task can reduce MSE significantly. Our estimator iteratively refines its belief and identifies those good workers, when r is sufficiently large.

3.3 Relative Performance Guarantee

We present the relative performance of BI by comparing to the optimal estimator, in particular, when the quantitative guarantee in Theorem 1 is not tight, i.e., r is small and thus estimating reliability is difficult.

Theorem 2. *Consider the crowdsourced regression model with $\mathcal{S} = \{\sigma^2_{\min}, \sigma^2_{\max}\}$ and a random (ℓ, r)-regular graph G consisting of n tasks and $(\ell/r)n$ workers. For given $\varepsilon > 0$ and ℓ, there exists a constant $C_{\ell,\varepsilon}$, depending on only ℓ and ε, such that if* (i) $\sigma^2_{\min} + \varepsilon \leq \sigma^2_{\max} \leq 2\sigma^2_{\min}$, *and* (ii) $C_{\ell,\varepsilon} \leq r \leq \log\log n$, *then BI in (7) with $k = \log\log n$ iterations achieves*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i \in V}\left(\mathrm{MSE}(\hat{\mu}_i^*(A)) - \mathrm{MSE}(\hat{\mu}_i^{BI(k)}(A))\right)\right] \to 0 \quad (9)$$

as $n \to \infty$. The expectation here is taken w.r.t. the distribution of G and A.

This result is not directly comparable to Theorem 1 as it applies to different regimes of the parameters.

The *oracle* optimality gap (8b) does not vanish for finite ℓ and r. This is perhaps because the oracle is too strong to compete against when ℓ and r are small. Hence, to obtain the tight result in (9), we construct a more practical lower bound on the optimal estimator in (3) that takes account of the worker reliability estimation. We use the fact that the random (ℓ, r)-regular bipartite graph has a *locally tree-like structure* with depth $k \leq \log\log n$ and our message update is exact on the local tree (Pearl, 1982). By revealing the ground truths at the boundary of this local tree of depth k, we construct a *weaker* oracle estimator that gives a tighter lower bound. We show that the gap between our estimator (without the ground truths at the boundary) and the weaker oracle vanishes as the tree depth increases. This is made clear by establishing *decaying correlation* from the information on the outside of the local tree to the root. A formal proof of Theorem 2 is presented in Section 4.2.

For the analytic tractability, we need a constant lower bound of $r \geq C_{\ell,\varepsilon}$ and $|\mathcal{S}| = 2$. Similar conditions are also required in other BP analysis (Ok et al., 2016; Mossel et al., 2014), while ours is more general in terms of ℓ, i.e., factor degree since the other analysis made on only factor degree 2 but also more challenging due to the unboundedness of the regression error. We also assume $\sigma^2_{\min} + \varepsilon \leq \sigma^2_{\max} \leq 2\sigma^2_{\min}$. However, this is the most challenging regime for any inference algorithms since it is hard to distinguish the workers' variances. Note that when this assumption is violated, i.e., $\sigma_{\min} \ll \sigma_{\max}$, Theorem 1 provides the near-optimality of BI since the MSE gap between BI and Oracle vanishes as the variance gap increases. The experimental results in Section 5 indeed suggest the BI's optimality even when such assumptions are violated.

4 PROOFS OF THEOREMS

4.1 Proof of Theorem 1

We start with an upper bound on the conditional expectation of MSE of $\hat{\mu}_i^{\mathrm{BI}(k)}(A)$ conditioned on $\sigma^2 = \tilde{\sigma}^2 \in \mathcal{S}^W$. Let $\mathbb{E}_{\tilde{\sigma}^2}$ be the conditional expectation given $\sigma^2 = \tilde{\sigma}^2$. Using Cauchy-Schwarz inequality for random variables X and Y, i.e., $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$, it is not hard to obtain that (see the supplementary material for the detailed derivation)

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\|\hat{\mu}_i^{\mathrm{BI}(k)}(A) - \mu_i\|_2^2\right] \leq d\tilde{\sigma}_i^2(\tilde{\sigma}^2_{M_i}) + \mathcal{E}_{\ell,\mathcal{S}}\left(1 - \mathbb{E}_{\tilde{\sigma}^2}\left[b_i^k(\tilde{\sigma}^2_{M_i})\right]\right)^{1/4}. \quad (10)$$

To complete the proof, we will obtain an upper bound of the last term in the RHS of (10) using the known fact that a random (ℓ, r)-regular bipartite graph G is a locally tree-like. Pick an arbitrary task $\tau \in V$. Let $G_{\tau,2k+1} = (V_{\tau,2k+1}, W_{\tau,2k+1}, E_{\tau,2k+1})$ denote the subgraph of G induced by all the nodes within (graph) distance $2k+1$ from *root* τ. From Lemma 5 in (Karger et al., 2014), we have that for sufficiently large n,

$$\mathbb{P}[G_{\tau,2k+1} \text{ is not tree}] \leq \frac{3(\ell r)^{2k+2}}{n} \leq 2^{-k} \quad (11)$$

where the last inequality follows from the choice of $r, k \leq \log\log n$ and large n. Thus, we obtain that

$$\mathbb{E}\left[1 - \mathbb{E}_{\tilde{\sigma}^2}[b_\tau^k(\tilde{\sigma}^2_{M_\tau})]\right] \leq \mathbb{E}\left[1 - \mathbb{E}_{\tilde{\sigma}^2}[b_\tau^k(\tilde{\sigma}^2_{M_\tau}) \,|\, G_{\tau,2k+1} \text{ is a tree}]\right] + 2^{-k} \quad (12)$$

where $\mathbb{E}$ is taken w.r.t. G and σ^2.

Let $A_{\tau,2k+1} := \{A_{iu} : (i,u) \in E_{\tau,2k+1}\}$. The exactness of BI on tree implies that if $G_{\tau,2k+1}$ is tree, $b_\tau^k(\sigma'^2_{M_\tau})$ is the likelihood of $\sigma^2_{M_\tau} = \sigma'^2_{M_\tau}$ given $A_{\tau,2k+1}$ and thus

$$1 - \mathbb{E}_{\tilde{\sigma}^2}[b_\tau^k(\tilde{\sigma}^2_{M_\tau})] = \mathbb{E}_{\tilde{\sigma}^2}[\mathbb{P}[\sigma^2_{M_\tau} \neq \tilde{\sigma}^2_{M_\tau} \,|\, A_{\tau,2k+1}]] \leq \sum_{u \in M_\tau} \mathbb{E}_{\tilde{\sigma}^2}[\mathbb{P}[\sigma^2_u \neq \tilde{\sigma}^2_u \,|\, A_{\tau,2k+1}]] \quad (13)$$

where the inequality is due to the union bound. Hence, it suffices to show that if $G_{\tau,2k+1}$ is tree, for any $u \in M_\tau$, the marginal probability of σ^2_u concentrated at $\tilde{\sigma}^2_u$.

Lemma 1. *For given $\rho \in W$, suppose $G_{\rho,2k} = (V_{\rho,2k}, W_{\rho,2k}, E_{\rho,2k})$ [1] is a (ℓ, r)-regular bipartite graph with $\ell \geq 2$ and $r \geq 1$ and it is a tree rooted from worker ρ with depth $2k \geq 2$. For given $\varepsilon, \sigma^2_{\min}, \sigma^2_{\max} > 0$, consider $\mathcal{S} = \{\sigma_1^2, ..., \sigma_S^2\}$ such that (i) $|\sigma_s^2 - \sigma_{s'}^2| > \varepsilon$ and $\sigma^2_{\min} \leq \sigma_s^2 \leq \sigma^2_{\max}$ for all $1 \leq s \neq s' \leq S$. Then,*

$$\mathbb{E}\left[\mathbb{E}_{\tilde{\sigma}^2}\left[\mathbb{P}\left[\sigma_\rho^2 \neq \tilde{\sigma}_\rho^2 \,|\, A_{\rho,2k}\right]\right]\right] \leq 4e^{-\frac{\varepsilon^2 r}{8(8\varepsilon+1)\sigma^2_{\max}}}$$

where the inner expectation $\mathbb{E}_{\tilde{\sigma}^2}$ is taken w.r.t. $A_{\rho,2k}$ from the crowdsourced regression model given $\sigma^2 = \tilde{\sigma}^2 \in \mathcal{S}^W$, and the outer expectation $\mathbb{E}$ is taken w.r.t. $\tilde{\sigma}^2 \in \mathcal{S}^W$ drawn uniformly at random.

The proof of Lemma 1 is in the supplementary material. Combining (12), (13) and Lemma 1 leads to

$$\mathbb{E}\left[\left(1 - \mathbb{E}_{\tilde{\sigma}^2}[b_\tau^k(\tilde{\sigma}^2_{M_\tau})]\right)^{1/4}\right] \leq \left(1 - \mathbb{E}[\mathbb{E}_{\tilde{\sigma}^2}[b_\tau^k(\tilde{\sigma}^2_{M_\tau})]]\right)^{1/4} \leq \left(4\ell e^{-\frac{\varepsilon^2 r}{8(8\varepsilon+1)\sigma^2_{\max}}} + 2^{-k}\right)^{1/4}$$

where the first inequality is from the fact that $(1-x)^{1/4}$ is concave, i.e., $\mathbb{E}[(1-X)^{1/4}] \leq (1 - \mathbb{E}[X])^{1/4}$. This completes the proof of Theorem 1 with (10) because of the arbitrary choice of root task $\tau \in V$.

4.2 Proof of Theorem 2

Pick an arbitrary task $\tau \in V$. Recalling the exactness of BI on tree, it is clear that in the case of $\ell = 1$, $\hat{\mu}_\tau^{\mathrm{BI}(k)}(A)$ is identical to the optimal estimator $\hat{\mu}_\tau^*(A)$. We so focus on $\ell \geq 2$. Recall that the Bayesian optimal estimator $\hat{\mu}_\tau^*(A)$ minimizes the MSE given A.

[1] We denote by $\tau \in V$ and $\rho \in W$ task and worker roots.

However, its analysis is very challenging due to loops in its corresponding graphical model. To overcome this issue, we use the locally tree like structure of random (ℓ, r)-regular bipartite graph, again. Intuitively, we will first construct an artificial but (analytically) tractable estimator outperforming $\hat{\mu}^*_\tau(A)$ in terms of MSE and then we will show the diminishing gap between MSE's of BI and the constructed estimator.

Let $\partial W_{\tau,2k+1}$ be the set of all workers at distance $2k+1$ from root τ in subgraph $G_{\tau,2k+1}$. Consider an oracle estimator $\hat{\mu}^{\mathsf{ora}(k)}_\tau(A)$ of μ_τ with free access to true variances of leaf-workers $\partial W_{\tau,2k+1}$, formally defined as

$$\hat{\mu}^{\mathsf{ora}(k)}_\tau(A) := \sum_{\sigma^2_{M_\tau} \in \mathcal{S}^{M_\tau}} \bar{\mu}_\tau(A_\tau, \sigma^2_{M_\tau}) \mathbb{P}[\sigma^2_{M_\tau} \,|\, A, \sigma^2_{\partial W_{\tau,2k+1}}]$$

$$= \sum_{\sigma^2_{M_\tau} \in \mathcal{S}^{M_\tau}} \bar{\mu}_i\left(A_\tau, \sigma^2_{M_\tau}\right) \mathbb{P}[\sigma^2_{M_\tau} \,|\, A_{\tau,2k+1}, \sigma^2_{\partial W_{\tau,2k+1}}]$$

where for the last equality, we use the conditional independence between $A_{\tau,2k+1}$ and $A \setminus A_{\tau,2k+1}$ given the additional information $\sigma^2_{\partial W_{\tau,2k+1}}$. Using the equality $(\hat{\mu}^*_\tau(A) - \mu_\tau) = (\hat{\mu}^*_\tau(A) - \mathbb{E}[\mu_\tau \,|\, A, \sigma^2_{\partial W_{\tau,2k+1}}]) + (\mathbb{E}[\mu_\tau \,|\, A, \sigma^2_{\partial W_{\tau,2k+1}}] - \mu_\tau)$, it is not hard to check that $\hat{\mu}^{\mathsf{ora}(k)}_\tau$ has smaller expected MSE than $\hat{\mu}^*_\tau(A)$, i.e., $\mathbb{E}[\mathrm{MSE}(\hat{\mu}^{\mathsf{ora}(k)}_\tau(A))] \le \mathbb{E}[\mathrm{MSE}(\hat{\mu}^*_\tau(A))] \le \mathbb{E}[\mathrm{MSE}(\hat{\mu}^{\mathrm{BI}(k)}_\tau(A))]$. Thus, it is enough to show that as $n \to \infty$,

$$\mathbb{E}\left[\left|\mathrm{MSE}(\hat{\mu}^{\mathsf{ora}(k)}_\tau(A)) - \mathrm{MSE}(\hat{\mu}^{\mathrm{BI}(k)}_\tau(A))\right|\right] \to 0\,. \quad (14)$$

Since the only difference between $\hat{\mu}^{\mathsf{ora}(k)}_\tau(A)$ and $\hat{\mu}^{\mathrm{BI}(k)}_\tau(A)$ is the estimation on $\sigma^2_{M_\tau}$, i.e., BI uses $b^k_\tau(\sigma^2_{M_\tau})$ instead of $\mathbb{P}[\sigma^2_{M_\tau} \,|\, A, \sigma^2_{\partial W_{\tau,2k+1}}]$. Using Cauchy-Schwarz inequality and some calculus, similarly as (10), we derive an upper bound on the expected difference between MSE's of $\mu^{\mathsf{ora(k)}}_\tau(A)$ and $\mu^{\mathrm{BI(k)}}_\tau(A)$ as follows:

$$\mathbb{E}\left[\left|\mathrm{MSE}(\hat{\mu}^{\mathsf{ora}(k)}_\tau(A)) - \mathrm{MSE}(\hat{\mu}^{\mathrm{BI}(k)}_\tau(A))\right|\right]$$

$$\le \mathcal{E}_{\ell,\mathcal{S}} \sum_{\sigma'^2_{M_\tau}, \sigma''^2_{M_\tau} \in \mathcal{S}^\ell} \sqrt{\mathbb{E}\left[\left(D_{\tau,k}(\sigma'^2_{M_\tau}, \sigma''^2_{M_\tau})\right)^2\right]} \quad (15)$$

where $D_{\tau,k}(\sigma'^2_{M_\tau}, \sigma''^2_{M_\tau}) := b^k_\tau(\sigma'^2_{M_\tau}) b^k_\tau(\sigma''^2_{M_\tau}) - \mathbb{P}[\sigma^2_{M_\tau} = \sigma'^2_{M_\tau} \,|\, A, \sigma^2_{\partial W_{\tau,2k+1}}] \mathbb{P}[\sigma^2_{M_\tau} = \sigma''^2_{M_\tau} \,|\, A, \sigma^2_{\partial W_{\tau,2k+1}}]$. We provide the detailed steps for (15) in the supplementary material. Then, from the same decomposition in (12), it follows that for each $\sigma'^2_{M_\tau}, \sigma''^2_{M_\tau} \in \mathcal{S}^\ell$ and sufficiently large n,

$$\mathbb{E}\left[\left(D_{\tau,k}(\sigma'^2_{M_\tau}, \sigma''^2_{M_\tau}\right)^2\right] \le \mathbb{E}\left[D_{\tau,k}(\sigma'^2_{M_\tau}, \sigma''^2_{M_\tau}\right]$$

$$\le \mathbb{E}\left[\left|D_{\tau,k}(\sigma'^2_{M_\tau}, \sigma''^2_{M_\tau})\right| \,|\, G_{\tau,2k+1} \text{ is tree}\right] + 2^{-k} \quad (16)$$

where the first inequality follows from that $0 \le D_{\tau,k}(\sigma'^2_{M_\tau}, \sigma''^2_{M_\tau}) \le 1$.

Suppose $G_{\tau,2k+1}$ is tree. Then, the graph subtracted from $G_{\tau,2k+1}$ to task τ and edges between task τ and workers in M_τ is partitioned into r sub-trees denoted by $\{G_{\rho,2k} = (V_{\rho,2k}, W_{\rho,2k}, E_{\rho,2k}) : \rho \in M_\tau\}$ each of which is rooted from worker $\rho \in M_\tau$ with depth $2k$ in the subtracted graph. From the exactness of BI on tree, it follows that

$$b^k_\tau(\sigma'^2_{M_\tau}) = \mathbb{P}[\sigma^2_{M_\tau} = \sigma'^2_{M_\tau} \,|\, A_{\tau,2k+1}]$$

$$\propto \mathcal{C}_\tau(A_\tau, \sigma^2_{M_\tau}) \mathbb{P}[\sigma^2_{M_\tau} = \sigma'^2_{M_\tau} \,|\, A_{\tau,2k+1} \setminus A_\tau]$$

$$= \mathcal{C}_\tau(A_\tau, \sigma^2_{M_\tau}) \prod_{\rho \in M_\tau} \mathbb{P}[\sigma^2_\rho = \sigma'^2_\rho \,|\, A_{\rho,2k}]$$

where $A_{\rho,2k} := \{A_{iu} : (i,u) \in E_{\rho,2k}\}$ and for the last equality, we use the conditional independence among $\sigma^2_{M_\tau}$ given $A_{\tau,2k+1} \setminus A_\tau$ decomposed into $A_{\rho,2k}$. Similarly, we also obtain

$$\mathbb{P}[\sigma^2_{M_\tau} = \sigma'^2_{M_\tau} \,|\, A, \sigma^2_{\partial W_{\tau,2k+1}}]$$

$$\propto \mathcal{C}_\tau(A_\tau, \sigma^2_{M_\tau}) \mathbb{P}[\sigma^2_{M_\tau} = \sigma'^2_{M_\tau} \,|\, A_{\tau,2k+1} \setminus A_\tau, \sigma^2_{\partial W_{\tau,2k+1}}]$$

$$= \mathcal{C}_\tau(A_\tau, \sigma^2_{M_\tau}) \prod_{\rho \in M_\tau} \mathbb{P}[\sigma^2_\rho = \sigma'^2_\rho \,|\, A_{\rho,2k}, \sigma^2_{\partial W_{\tau,2k+1}}]\,.$$

Hence it is enough to show the vanishing correlation of true variances of workers at leaves to inferring the root worker's variance. Formally, we provide Lemma 2 that captures a decreasing rate of the correlation.

Lemma 2. *Suppose $G_{\rho,2k} = (V_{\rho,2k}, W_{\rho,2k}, E_{\rho,2k})$ is induced from (ℓ, r)-regular bipartite graph $G = (V, W, E)$ and it is a tree with depth $2k \ge 2$. Let $\partial W_{\rho,2k}$ be the set of workers at the leaves in $G_{\rho,2k}$. For given $\varepsilon, \sigma^2_{\min}, \sigma^2_{\max} > 0$, consider $\mathcal{S} = \{\sigma^2_{\min}, \sigma^2_{\max}\}$ such that $\sigma^2_{\min} + \varepsilon \le \sigma^2_{\max} \le 2\sigma^2_{\min}$. Then, for any given $\tilde{\sigma}^2 \in \mathcal{S}^W$, there exists a constant $C_{\ell,\varepsilon}$ such that if $r \ge C_{\ell,\varepsilon}$, then*

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\left|\mathbb{P}[\sigma^2_\rho = \tilde{\sigma}^2_\rho \,|\, A_{\rho,2k}, \sigma^2_{\partial W_{\rho,2k}}] - \mathbb{P}[\sigma^2_\rho = \tilde{\sigma}^2_\rho \,|\, A_{\rho,2k}]\right|\right]$$

$$\le 2^{-k} \quad (17)$$

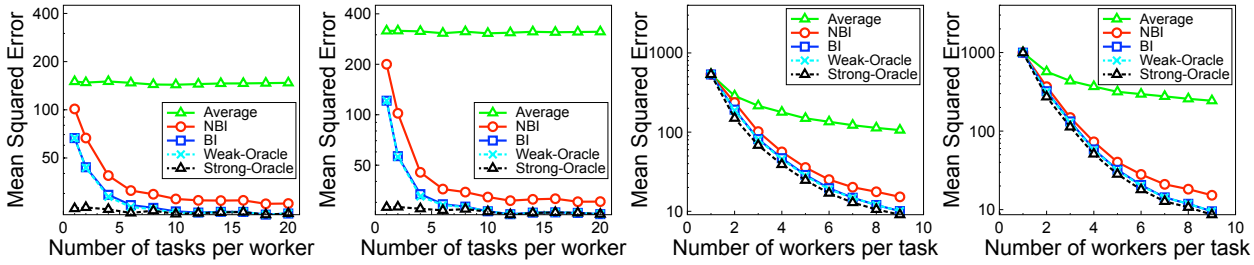
where the expectation is taken w.r.t. A from the crowdsourced regression model given $\sigma^2 = \tilde{\sigma}^2$ and G.

The proof of Lemma 2 is given in the supplementary material. This lemma completes the proof of Theorem 2 with (15) and (16).

5 EXPERIMENTAL RESULTS

We experiment the following five algorithms:

- BI is implemented without any prior information on true positions by taking the limit $\tau \to \infty$, i.e., it outputs $\mu^{\mathrm{BI}}(A)$ in (4)–(6) with $\lim_{\tau\to\infty} \mathcal{C}_i(A_i, \sigma^2_{M_i})$. Note that our theoretical guarantees on BI still hold in this regime.
- NBI is an iterative algorithm of non-Bayesian type. It recursively updating workers' variances and tasks' answers based on workers' consensus. Formally, ini-



(a) $\mathcal{S}_{\text{small}}$ with $\ell = 5$ (b) $\mathcal{S}_{\text{large}}$ with $\ell = 5$ (c) $\mathcal{S}_{\text{small}}$ with $r = 5$ (d) $\mathcal{S}_{\text{large}}$ with $r = 5$

Figure 1: The average MSE of various algorithms on the synthetic datasets consisting of 200 tasks and workers with $\mathcal{S}_{\text{small}} = \{10, 100, 1000\}$ and $\mathcal{S}_{\text{large}} = \{10, 100, 5000\}$; (a)-(b) $\ell = 5$ with varying r; (c)-(d) $r = 5$ with varying ℓ.

tialized with $\hat{\sigma}^2_{0,u} = 1$, it estimates $\mu_i^{\mathsf{NBI}(t)}(A) := \lim_{\tau\to\infty} \bar{\mu}_i(A_i, \hat{\sigma}^2_{t,M_i})$ and $\hat{\sigma}^2_{t+1,u} := \sum_{i\in N_u} \|A_{iu} - \mu_i^{\mathsf{NBI}(t)}(A)\|^2/|N_u|$.

- Average just takes the average of workers' observations without learning workers' variances, i.e., $\mu_i^{\mathsf{avg}}(A) := \frac{1}{|M_i|}\sum_{u\in M_i} A_{iu}$.
- Strong/Weak-Oracle are two artificial estimators which have free access to workers' variances σ^2. They would outperform all existing algorithms, even including the optimal estimator. For each task i, Strong-Oracle uses every worker's true variance, i.e., $\mu_i^{\mathsf{strong}}(A, \sigma^2) := \lim_{\tau\to\infty} \mathbb{E}[\mu_i \mid A, \sigma^2]$. Weak-Oracle uses just the true variances of leaf workers, denoted by ∂T_i, in Breadth-first search tree, denoted by T_i, of root i, i.e., $\mu_i^{\mathsf{weak}}(A, \sigma^2) := \lim_{\tau\to\infty} \mathbb{E}[\mu_i \mid A, \sigma^2_{\partial T_i}]$.

Recall that the true positions μ_i's are assumed to be drawn from the spherical Gaussian with mean ν_i and variance τ. As ν_i's and τ are hard to obtain in practice, in all experiments, we implement BI with no knowledge on *the prior distribution of the true positions of the tasks* by taking the limit of BI as $\tau \to \infty$. Note that our theoretical guarantees on BI still hold in this regime.

Our baseline comparisons include the ones with a simple approach of Average and the fundamental lower bounds from Strong/Weak-Oracle. Strong-Oracle and Weak-Oracle correspond to the fundamental limits that we compare with BI in Theorems 1 and 2, respectively. As we use the prior distribution in the synthetic experiments, one may ask about what is the gain of this side information. To address this, we test a non-Bayesian algorithm (NBI) that does not assume such prior information, iteratively estimating task answers and worker variances based on consensus. Note that a similar idea has been also investigated in (Raykar et al., 2010).

5.1 Synthetic Datasets

Since our theoretical results cover a large n regime, we test a more challenging regime of modest size $n = 200$. Synthetic datasets are generated by the set of random (ℓ, r)-regular bipartite graphs of 200 object detection tasks where each task i is associated with the true position μ_i chosen uniformly at random in a 100×100 image. We randomly choose each worker's variance using $\mathcal{S}_{\text{small}} = \{10, 100, 1000\}$ or $\mathcal{S}_{\text{large}} = \{10, 100, 5000\}$. The simulation results with varying r and ℓ are plotted in Figures 1(a)-(b) and 1(c)-(d), where we take the average of 50 random instances.

Optimality of BI. As discussed in Section 3, Figures 1(a)-(d) show that for *all* (ℓ, r), BI closely achieves the fundamental limit of Weak-Oracle, whereas Average and NBI have the suboptimal performance. We also observe that Weak-Oracle with the challenge of identifying reliable workers indeed provides tighter fundamental limit than Strong-Oracle, as discussed in Section 3. Overall, NBI has a small constant gap to BI, which quantifies the gain of BI using the matched prior distribution. Average shows the significant performance loss, compared to the optimal BI or even NBI. For example, in Figures 1(c)-(d), in order to make MSE less than 100 with $\mathcal{S}_{\text{small}}$, BI and NBI require only $\ell \geq 3$, but Average requires $\ell \geq 9$, implying that Average needs to hire three times more workers per task than others.

Importance of worker identification. Comparing Figures 1(c)-(d), we observe that under the minimum of workers' variances fixed, both BI and NBI, which identify reliable workers and adaptively weight their answers, sustain good performance for both small and large maximum worker variances. However, the performance of Average, which does not distinguish workers, is significantly degenerated by spammers with large variance from $\mathcal{S}_{\text{large}}$. This shows the importance of classifying workers in making the estimator robust to spammers or adversary.

Impact of (ℓ, r). As Figures 1(c)-(d) show, increasing ℓ (or budget) exponentially reduces MSE's of all algorithms, while the value of exponent varies for each algorithm. In Figures 1(a)-(b), the gap between the optimal BI (or Weak-Oracle) and Strong-Oracle quantifies the difficulty in identifying reliable workers. As studied in Theorem 1, the gap is diminishing exponentially fast with increasing r, whereas the fundamental limit of Strong-Oracle does not change with r. Hence,

Table 1: Estimation quality of Average, NBI, BI with $\mathcal{S}_{\text{est}}$, PG1 and Weak/Strong-Oracle's on crowdsourced FG-NET datasets from Amazon MTurk workers.

Estimator	Average	NBI	BI	PG1	Weak Oracle	Strong Oracle
Data noise (MSE)	34.99	32.80	28.72	33.54	28.68	28.45

for efficiency from the worker identification, the task requester needs to assign each worker so as to answer a certain number of tasks at least, while letting a worker solve too many tasks may be impractical but also unhelpful to increase accuracy.

5.2 Human Age Prediction

We also present experiment results on datasets from a *real-world* crowdsourcing system. We use FG-NET datasets which has been widely used as a benchmark for facial age estimation (Lanitis, 2008). The dataset contains $1,002$ photos of 82 individuals' faces, in which each photo is labeled with a biological age as the ground truth. Furthermore, Han et al. (2015) provide crowdsourced labels on FG-NET datasets, in which 165 workers in Amazon Mechanical Turk (MTurk) answer their own age estimation on given subset of $1,002$ photos, so that each photo has 10 answers from workers, while each worker provides a different number (from 1 to 457) of answers, and 60.73 answers on average.

Prior estimation. In processing the real-world dataset, the prior distribution on noise level is not provided in advance, while it is required to run BI. To infer the prior distribution, we first study workers' answer patterns. We often observe two extreme classes of answers for a task: a few outliers and consensus among majority. For example, in Figures 2(a) and 2(b), there exist noisy answers 5 and 7, respectively, which are far from the majority's answers, 1 and 55, respectively. Such observations suggest to choose a simple support, e.g. $\mathcal{S} = \{\sigma^2_{\text{good}}, \sigma^2_{\text{bad}}\}$. In particular, without any use of ground truth, we first run NBI and use the top 10% and bottom 10% workers' reliabilities as the binary support, which is $\mathcal{S}_{\text{est}} = \{6.687, 62.56\}$.

Validation on the estimated prior. For FG-NET, we additionally test PG1 which is Gibbs sampling algorithm[2] relying on a sophisticated worker model, called PG_1 model in (Piech et al., 2013), with biased Gaussian noises where worker variance and bias are drawn from continuous supports. In Table 1, we compare the estimation of BI to other algorithms. Observe that MSE of BI with the binary support $\mathcal{S}_{\text{est}}$ is close to those of Weak/Strong-Oracle, while the other algorithms have some gaps. Despite using a sophisticated

[2]As suggested in (Piech et al., 2013), we use 800 iterations of Gibbs sampling after discarding the initial 80 burn-in samples.

(a) 1-year-old (b) 35-year-old

Figure 2: Easy and hard samples from FG-NET in terms of average absolute error of workers' answers: (1,1,1,1,1,1,1,1,2,5) and (7,51,52,55,55,55,59,63,66,67) on photo of (a) 1-year-old, and (b) 35-year-old, resp.

model, PG_1, PG1 is observed to perform worse than BI. This is because PG1 needs non-trivial parameter optimization based on *training dataset*, which incurs overfitting. This result from the real workers supports the value of our simplified modeling on workers' noise. For interested readers, we also report the regression accuracy of deep neural networks trained using the pruned dataset produced by different algorithm in the supplementary material.

6 CONCLUSION

We study a model to address the problem of aggregating real-valued responses from a crowd of workers with heterogeneous noise level. In particular, inspired by the observation on answer pattern of Amazon MTurk workers, we use a canonical noise model. This modeling allows us to pose this crowdsourced regression problem as an inference problem over a graphical model, naturally motivating the proposed BI algorithm based on BP. Typically, the analysis of such iterative algorithms is not tractable even for estimating discrete labels. However, our theoretical framework, inspired by recent advances in BP, e.g. (Mossel et al., 2014), provides sharp guarantees on BI and shows its optimality for a broad range of parameters.

An important research direction is in generalizing the proposed noise model. First natural generalization is to allow differing task difficulties, by adding an additional independent Gaussian with variance σ_i^2 for answers on task i. Larger variance represents more difficult tasks. Second natural generalization is to allow worker biases, by adding a constant shift of value μ_u for answers given by worker u as Piech et al. (2013) considered. Our observations on the crowdsourced FG-NET datasets also suggest heterogeneous task difficulty and boundary effect on tasks. As in the examples in Figure 2, we often observe less estimation error for photos of younger individuals: MSE 15.00 for 233 photos of under-5-year olds, and MSE 85.39 for 769 photos over-5-year olds. This shows heterogeneous task difficulty and boundary effect at the same time: age prediction on older individual is more challenging due to more variations, and human age cannot be negative.

Acknowledgments

This work was supported in part by NSF grants 1553452, 1527754, and 1815535, and Google Faculty Research Award, in part by the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF-2018R1A5A1059921), and in part by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea Government (MSIT) (No.2016-0-00160, Versatile Network System Architecture for Multi-dimensional Diversity)

References

M. S. Bernstein, J. Brandt, R. C. Miller, and D. R. Karger. Crowds in two seconds: enabling realtime crowd-powered interfaces. In *Proc. of ACM UIST*, 2011.

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

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

L. De Alfaro and M. Shavlovsky. CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proc. of ACM SIGCSE*, 2014.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of IEEE CVPR*, 2009.

M. Everingham, SM. A. Eslami, L. Van Gool, C. KI Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

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

I. M. Goldin and K. D. Ashley. Peering inside peer review with bayesian models. In *Proc. of AIED*, pages 90–97. Springer, 2011.

H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1148–1161, 2015.

M. I. Jordan. *Learning in graphical models*, volume 89. Springer Science & Business Media, 1998.

M. I. Jordan. Graphical models. *Statistical Science*, 19(1):140–155, 2004.

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

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

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

A. Khetan and S. Oh. Achieving budget-optimality with adaptive schemes in crowdsourcing. In *Proc. of NIPS*, pages 4844–4852, 2016.

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

A. Lanitis. Comparative evaluation of automatic age-progression methodologies. *EURASIP Journal on Advances in Signal Processing*, 2008:101, 2008.

M. D. Lee, M. Steyvers, M. De Young, and B. Miller. Inferring expertise in knowledge and prediction ranking tasks. *Topics in cognitive science*, 4(1):151–163, 2012.

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

Q. Liu, A. T. Ihler, and M. Steyvers. Scoring workers in crowdsourcing: How many control questions are enough? In *Proc. of NIPS*, pages 1914–1922, 2013.

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

K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proc. of UAI*, 1999.

J. Ok, S. Oh, J. Shin, and Y. Yi. Optimality of belief propagtion for crowdsourced classification. In *Proc. of ICML*, 2016.

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

J. Peng, L. Qiang, A. Ihler, and B. Berger. Crowdsourcing for structured labeling with applications to protein folding. In *Proc. of ICML Machine Learning Meets Crowdsourcing Workshop*, 2013.

C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in MOOCs. In *Proc. of EDM*, 2013.

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*, 11:1297–1322, 2010.

R. Di Salvo, D. Giordano, and I. Kavasidis. A crowdsourcing approach to support video annotation. In *Proc. of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*, page 8. ACM, 2013.

N. B. Shah, S. Balakrishnan, and M. J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv preprint arXiv:1606.09632*, 2016.

H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In *Proc. of Workshops at AAAI Conference on Artificial Intelligence*, 2012.

X. Wu, W. Fan, and Y. Yu. Sembler: Ensembling crowd sequential labeling for improved quality. In *Proc. of AAAI*, 2012.

C. Yanover, T. Meltzer, and Y. Weiss. Linear programming relaxations and belief propagation–an empirical study. *Journal of Machine Learning Research*, 7 (Sep):1887–1907, 2006.

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

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

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.